

**Exploring Molecular Adaptations of Extremophilic Proteins:  
A Platform for Protein Engineering**

**A Thesis**

**Submitted in Partial Fulfillment of the Requirements  
for the Degree of**

**DOCTOR OF PHILOSOPHY**

**BY**

**MOHD FAHEEM KHAN**

**Under the Supervision of  
Prof. Sanjukta Patra**



**Department of Biosciences and Bioengineering  
Indian Institute of Technology Guwahati**

**MAY 2018**



**INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI**  
**DEPARTMENT OF BIOSCIENCES AND BIOENGINEERING**

**STATEMENT**

I do hereby declare that the matter embodied in this thesis entitled: ***“Exploring Molecular Adaptations of Extremophilic Proteins: A Platform for Protein Engineering”***, is the result of investigations carried out by me in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India, under the guidance of Prof. Sanjukta Patra.

In keeping with the general practice of reporting scientific observations, due acknowledgements have been made wherever the work described is based on the findings of other investigator(s).

**May 2018**

**Mr. Mohd Faheem Khan**

Roll No. 126106037



**INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI**  
**DEPARTMENT OF BIOSCIENCES AND BIOENGINEERING**

**CERTIFICATE**

It is certified that the work entitled in this thesis entitled: ***“Exploring Molecular Adaptations of Extremophilic Proteins: A Platform for Protein Engineering”***, done by Mr. Mohd Faheem Khan (Roll No. 126106037), for the award of degree of Doctor of Philosophy in an authentic record of the results obtained from the research work carried out under my supervision in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India, and this work has been submitted elsewhere for a degree.

**May 2018**

**Dr. Sanjukta Patra**

Professor

Department of Biosciences and Bioengineering

**(Thesis Supervisor)**

## **Acknowledgement**

*This page is designed to note my appreciation towards all who directly or, indirectly are instrumental in contributing towards this thesis. I take this as an opportunity to express my sincere thanks to everyone who has been with me in this entire journey.*

*Firstly, I would like to express my heartfelt gratitude to my research supervisor Prof. Sanjukta Patra for her continuous and unconditional support during my Ph. D. tenure. Her originality has nurtured my intellectual maturity and I am really indebted to her for her continuous guidance, support, motivation and enthusiasm throughout these years.*

*I would like to give my sincere thanks to my doctoral committee members; Dr. Anil Mukund Limaye, Dr. Shankar Prasad Kanaujia and Dr. Gautam Kumar Das, Dr. V. V. Dasu, Prof. Bishwajit Kundu (External Examiner, IIT Delhi), Dr. Manish Kumar and Prof. Mohammad Qureshi for time to time evaluation of my progress of work with critical comments, valuable guidance and inspirational remarks. I gratefully acknowledge teaching and non-teaching staffs of Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India for providing me with all the supports and necessary facilities.*

*I am also thankful to Indian Institute of Technology Guwahati, India for providing me financial assistance through Institute fellowship for smooth functioning of my Ph. D. I would also like to give special thanks to Central Instrumentation Facility, Indian Institute of Technology Guwahati, India for providing me instrumentation facilities which were crucial and critical for my research work as well as to Bidya Mandir for their generous support throughout.*

*I would be failing in my duty if I do not convey my warm and sincere thanks to my lab seniors namely Dr. P. Saravanan, Dr. Debamitra Chakravorty, Dr. Bhaskar Das, Dr. Nivedita Singh, who have provided me with constant support, inspiration and enjoyable working environment.*

*I owe my deepest gratitude for the help rendered on professional and personal front by Dr. Debasree Kundu, Mayur Mahindra Kedare, Swagata Patra, Poulomi Saha, Gaurang Gurjar, Ghadir Nofal, Arnesh Das and Dr. Chinmay Hazra for being with me*



*in all my ups and downs of life and for maintaining an enjoyable as well as supportive research environment during the tenure. I extend my warmest thanks to my research colleagues Prithwi Chayan Chatterjee, Bhaskar Kalita, Satakshi Hazra and Mukesh Aakula who have supported me. Thanks are also due to my friends Abshar, Suraj, Zia, Tabish, Haris, Aquib, Imran, Ponkhi, Payal Didi, Arun Bhaiya, Manali, Jinti and Shalini. I am also thankful to little Paridhi for her patient presence during the tenure.*

*Back at home, I am blessed with a loving family having my grandfather (Dada, Dr. Mohd Naimullah Khan), mother (Ammi, Mrs. Shamim Rabbani), father (Papa, Mr. Mohd Waseem Khan), brothers and sisters (Azeem Bhai, Kaleem Bhai, Mustaqeem Bhai, Aleem Bhai, Nida Appi and Gulnaz Bajia) whose unconditional support, love and patience motivated me towards the completion of my work. I express my special emotions to my youngest brother and companion Tasleem who had made my stay at Guwahati most comfortable.*

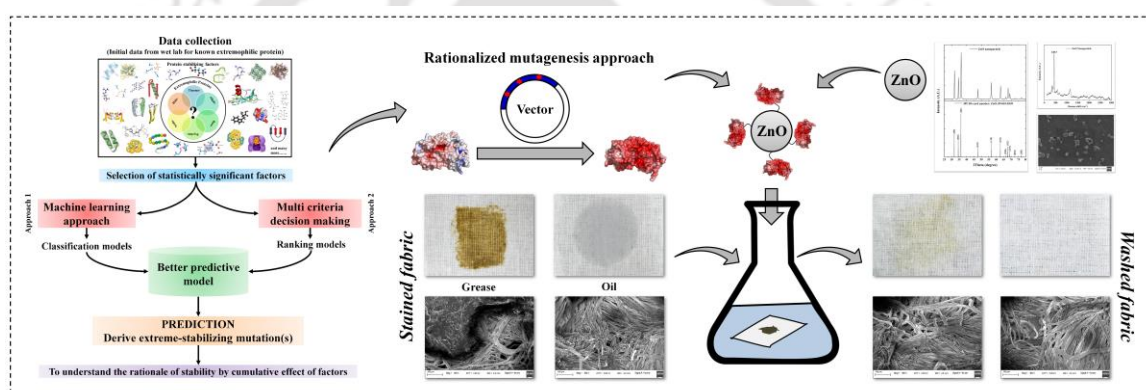
*At last, I would like to thank everybody who was involved as well as expressing my apology if I have missed anyone.*

*Last but not the least, I would like to devote my wholehearted gratitude to the Almighty Allah for his blessings to achieve the goals.*

**Mohd Faheem Khan**

**May, 2018**

# Thesis abstract



Research on extremophilic proteins has seen immense growth in the past decade. Modern mechanistic approaches to engineer such proteins *in vitro* will have impact in industrial biotechnology economy. Developing *a priori* knowledge about the mechanism behind protein extreme-stability by understanding genomic and proteomic attributes is the foundation for protein engineering. The present thesis entitled “Exploring molecular adaptations of extremophilic proteins: A platform for protein engineering” is an effort in this direction.

The **first chapter** provides an insight into till date research pertaining to the multifactorial level of extreme-stability of proteins. Protein stability can be affected at three hierarchies of protein organization – DNA level, protein sequence and structure level (including primary sequence features, various intra-protein interactions, secondary and tertiary structural features). Can these factors be exploited for protein engineering to fill the caveat of attaining pre-determined mutations? The further objectives of the thesis were proposed to fill the existing lacuna.

**Chapter two** is on creating a curated database on Extreme-stable Protein Database (ExProtDB). The database Architecture is on a platform based on the Linux-Apache-MySQL-PHP (LAMP). It is a knowledgebase about all the information available in the World Wide Web about extremophiles and their extreme-stable proteins. The novelty of the database lies in the fact that it gives consolidated and comprehensive information of all types of extremophiles and extremophilic proteins.

This is followed by **Chapter three** which intrudes into “Understanding the specific codon usage patterns in extremophiles”. This chapter throws light on primary level i.e. genomic adaptations of extremophilic proteins. It deciphers the rationale behind the specific codon usage patterns in extremophiles. Relative abundance analysis, 1-9 scale ranking, nucleotide compositions, attribute weighting and machine learning algorithms were employed to arrive at findings. Codon usage patterns exhibited harmony among different extremophiles and has been detailed. However, the codon attribute preferences and their selectivity of extremophiles varied in comparison to non-extremophiles.

Although codon study is being partially deciphered in thermophiles, this is probably the first report of codon preferences pertaining to all extremophile classes.

**Chapter four** analyses the attributes contributing to extreme-stability of proteins followed by their application in classifying extremostable proteins from their normal counter parts. An attempt was also made to generate ranking model for prepredicting desirable mutants. The positively contributing attributes have been further exploited to design mutations in a mesophilic (non-extremophilic) protein in such a way to increase the content of these attributes to endorse the prediction of generated models. This chapter reports novel contributing factors for extreme-protein stabilization, which were not reported earlier.

**Chapter five** gives a pertinent account of *in silico* prediction of plausible mutations and their validation in *Bacillus subtilis* lipase (a chosen mesophilic enzyme) by substituting the preferred protein attributes to attain extreme-stability of proteins. Six different types of extreme-stable mutants were generated and validated further. *In silico* validation of mutant was done by AHP generated ranking models, various mutation prediction servers (HotSpot Wizard, I-Mutant2, Cupsat, iPTREE-STAB, WET-STAB and ERIS webserver), Substrate Docking (Autodock 4.2 version), Ramachandran analysis (PROCHECK) and Contact Map analysis (CMView 1.1.1).

This was followed by *in vitro* validation of predicted mutations in **chapter six**. Experimentation was employed for *in vitro* cloning, mutagenesis, expression, purification and characterization of lipases. The content of secondary structure enhanced extreme-stability (with increased  $T_m$  value) which was confirmed through FT-IR (ATR mode) and Differential Scanning Fluorimetry (Thermal shift assay).

The mutants obtained were further employed for applications by work carried out in **chapter seven**. Detergent formulation was the chosen application here. To further increase the enzyme stability engineered lipases were immobilized on synthesized ZnO nanoparticles as it is economical, biocompatible and environmental friendly. The synthesized ZnO nanoparticles and lipase immobilization was confirmed using UV-Vis,

FT-IR, Raman spectroscopy, DLS (for particle size), powder XRD, TEM and FESEM analysis. For selection of better washing conditions, the washing process was done at different temperatures and time intervals using different concentrations of detergent and lipase activities. Results also showed that the maximum removal of oil and grease stains were up to 87% and 82%, respectively.

The thesis concludes with summarizing the findings in **chapter eight**. It also states the future scopes and upcoming challenges. It can be concluded that thesis work has been successful to develop a single design strategy to mutate amino acid residues which can render proteins stable at different extremophilic range.

Conclusively, this work reveals a promising perspective of using low-cost nanoparticle conjugated biocatalysts which was produced through rationalized protein engineering approach and was instrumented for detergent formulation to overcome the drawbacks of the chemical counterparts.

## TABLE OF CONTENTS

| <b>CONTENTS</b>                                                                                                        | <b>Page</b>      |
|------------------------------------------------------------------------------------------------------------------------|------------------|
| <b>Thesis Abstract</b>                                                                                                 | <b>i-iv</b>      |
| <b>Table of Contents</b>                                                                                               | <b>v-xi</b>      |
| <b>List of Tables</b>                                                                                                  | <b>xii-xiii</b>  |
| <b>List of Figures</b>                                                                                                 | <b>xiv-xviii</b> |
| <b>List of Abbreviations</b>                                                                                           | <b>xix-xx</b>    |
| <b>List of Software, Database and Web-tools</b>                                                                        | <b>xxi-xxii</b>  |
| <b>List of Protein Attributes</b>                                                                                      | <b>xxiii</b>     |
| <br><b>CHAPTER 1: Introduction and literature review</b>                                                               | <br><b>1-48</b>  |
| Abstract                                                                                                               | 1                |
| 1.1 Introduction                                                                                                       | 2                |
| 1.2 Extremophilic adaptations                                                                                          | 3                |
| 1.3 Multiple factors involved in extreme-stability of proteins                                                         | 5                |
| 1.4 Genomic adaptations of extremophiles: codon adaptability and GC-content                                            | 6                |
| 1.5 Proteomic adaptations of extremophiles: role in extreme-stabilization of proteins                                  | 10               |
| 1.5.1 Role of amino acid in rendering proteins extreme-stabilization                                                   | 10               |
| 1.5.2 The biophysical attributes responsible for extreme-stability of proteins                                         | 14               |
| 1.6 Approaches to attain extreme-stable proteins by protein engineering                                                | 24               |
| 1.7 Available theoretical prediction models for predicting stabilizing mutations                                       | 25               |
| 1.8 Salient features deduced from engineered extreme-stable mutants to better understand extreme-stability of proteins | 28               |
| 1.9 Stabilizing proteins by non-protein engineering methods                                                            | 29               |
| 1.10 Vivid role of chaperones and extremolytes in stabilizing proteins in extreme <i>milieu</i>                        | 31               |
| 1.11 Conclusion                                                                                                        | 34               |
| 1.12 Origin of the research work                                                                                       | 36               |
| 1.13 Thesis organization                                                                                               | 36               |

|                                                                                                      |              |
|------------------------------------------------------------------------------------------------------|--------------|
| 1.14 References                                                                                      | 39           |
| <b>CHAPTER 2: Extremophile Protein Database (ExProtDB)</b>                                           | <b>49-66</b> |
| Abstract                                                                                             | 49           |
| 2.1 Introduction                                                                                     | 50           |
| 2.2. Methodology                                                                                     | 51           |
| 2.2.1 Servers and software used                                                                      | 51           |
| 2.2.2 Organism and genome information                                                                | 51           |
| 2.2.3 Collection of protein data and their refinement                                                | 51           |
| 2.2.4 Amino acid composition analysis of all thermostable proteins                                   | 52           |
| 2.2.5 Structural analysis of all extremophilic proteins and feature generation                       | 52           |
| 2.2.6 Tools and downloads                                                                            | 53           |
| 2.2.7 Literature and patent                                                                          | 53           |
| 2.2.8 Data integration                                                                               | 53           |
| 2.3 Features of ExProtDB                                                                             | 53           |
| 2.3.1 Information about extremophiles, their proteins and contributing genomic and proteomic factors | 55           |
| 2.3.2 Genomic information                                                                            | 55           |
| 2.3.3 Protein information                                                                            | 55           |
| 2.3.4 Mutant information                                                                             | 55           |
| 2.3.5 Physicochemical protein factors                                                                | 56           |
| 2.3.6 Literature and patents                                                                         | 56           |
| 2.3.7 Tools                                                                                          | 56           |
| 2.4 Database architecture and integration                                                            | 57           |
| 2.5 Data collection and analysis: statistics                                                         | 59           |
| 2.6 Unique ID generation for collected extremophiles and their proteins                              | 65           |
| 2.7 Conclusion                                                                                       | 65           |
| 2.8 References                                                                                       | 66           |
| <b>CHAPTER 3: Understanding the specific codon usage patterns in extremophiles</b>                   | <b>67-98</b> |



|                                                                                                                   |               |
|-------------------------------------------------------------------------------------------------------------------|---------------|
| Abstract                                                                                                          | 67            |
| 3.1 Introduction                                                                                                  | 70            |
| 3.2 Methodology                                                                                                   | 70            |
| 3.2.1 Creation of comparative datasets and enumeration of statistically significant codons                        | 70            |
| 3.2.2 Statistical analysis of significant codons in different extremophiles                                       | 70            |
| 3.2.3 Relative abundance analysis of codons                                                                       | 71            |
| 3.2.4 Prioritizing the codons to understand their preference in extremophiles                                     | 72            |
| 3.2.5 Analysis of AT- or GC-rich and A/T- or G/C-ending codons                                                    | 72            |
| 3.2.6 Analysing data-points of highest and lowest ranked codon                                                    | 73            |
| 3.2.7 Finding codon harmony among different classes of extremophiles                                              | 73            |
| 3.2.8 Generation of machine learning models to classify and predict extremophilic codons                          | 73            |
| 3.3 Results and Discussion                                                                                        | 75            |
| 3.3.1 Dataset creation and enumeration of statistically significant codons for extremophiles                      | 75            |
| 3.3.2 Analyzing relative abundance of codons among extremophiles and non-extremophiles                            | 78            |
| 3.3.3 Understanding codon preferences in extremophiles by ranking them in 1-9 scale                               | 80            |
| 3.3.4 Analysis of AT- or GC-rich and A/T- or G/C-ending codons                                                    | 83            |
| 3.3.5 Analysis of variation in the normalized data-point of highest and lowest ranked codons w.r.t. extremophiles | 85            |
| 3.3.6 Exploration of codon harmony among various extremophiles                                                    | 86            |
| 3.3.7 Generation of machine learning models to classify and predict extremophilic codons                          | 88            |
| 3.4 Conclusion                                                                                                    | 95            |
| 3.5 References                                                                                                    | 97            |
| <b>CHAPTER 4: Analysis of attributes contributing to extreme-stability of proteins</b>                            | <b>99-134</b> |
| Abstract                                                                                                          | 99            |



|                   |                                                                                                                            |                |
|-------------------|----------------------------------------------------------------------------------------------------------------------------|----------------|
| 4.1               | Introduction                                                                                                               | 100            |
| 4.2               | Materials and methods                                                                                                      | 101            |
| 4.2.1             | Data collection and enumeration of statistically significant features                                                      | 102            |
| 4.2.2             | Finding relative abundance of selected significant features                                                                | 103            |
| 4.2.3             | Classification of extremophilic proteins by employing machine learning analysis using statistically significant attributes | 104            |
| 4.2.4             | Ranking of attributes contributing to extreme-stability through multi-criteria decision making approach                    | 107            |
| 4.2.5             | Testing the performance of AHP generated ranking model by blind test                                                       | 109            |
| 4.3               | Results and discussion                                                                                                     | 110            |
| 4.3.1             | Data collection and statistically significant feature generation                                                           | 110            |
| 4.3.2             | Relative abundance of protein feature in extremophiles                                                                     | 113            |
| 4.3.3             | Machine learning model generation for extremophilic proteins                                                               | 116            |
| 4.3.4             | Generation of ranking models for protein extreme attributes by AHP                                                         | 127            |
| 4.3.5             | Blind test analysis: validation of generated ranking model                                                                 | 129            |
| 4.3.6             | Salient findings                                                                                                           | 130            |
| 4.4               | Conclusion                                                                                                                 | 131            |
| 4.5               | References                                                                                                                 | 132            |
| <b>CHAPTER 5:</b> | <b><i>In silico</i> prediction of mutations to attain extreme-stability of proteins and their validation</b>               | <b>135-154</b> |
|                   | Abstract                                                                                                                   | 135            |
| 5.1               | Introduction                                                                                                               | 136            |
| 5.2               | Methodology                                                                                                                | 138            |
| 5.2.1             | Selection of mesophilic candidate enzyme for experimentation                                                               | 139            |
| 5.2.2             | <i>In silico</i> prediction and validation of mutations                                                                    | 139            |
| 5.3               | Results and discussion                                                                                                     | 140            |
| 5.3.1             | Predicted mutations in <i>Bacillus subtilis</i> lipase                                                                     | 141            |
| 5.3.2             | <i>In silico</i> validation of designed mutations through AHP ranking models                                               | 142            |

|                                                                                           |                |
|-------------------------------------------------------------------------------------------|----------------|
| 5.3.3 Validation of mutants through various mutation prediction tools/servers             | 144            |
| 5.3.4 Validation of mutants by Ramachandran plot                                          | 145            |
| 5.3.5 Structure superimposition validation of generated mutations                         | 147            |
| 5.3.6 Enzyme-substrate docking analysis for functional efficiency                         | 148            |
| 5.3.7 Contact map analysis for enumeration of interactions in lipases                     | 149            |
| 5.3.8 <i>In silico</i> prediction of secondary structure of mutants                       | 150            |
| 5.4 Conclusion                                                                            | 151            |
| 5.5 Reference                                                                             | 153            |
| <b>CHAPTER 6: <i>In vitro</i> validation and characterization of predicted mutations</b>  | <b>155-188</b> |
| Abstract                                                                                  | 152            |
| 6.1 Introduction                                                                          | 156            |
| 6.2 Methodology                                                                           | 157            |
| 6.2.1 Strains, plasmids, enzymes, reagents and instruments                                | 157            |
| 6.2.2 Cloning and mutagenesis of wild-type and mutant enzymes                             | 157            |
| 6.2.3 Overexpression and purification of wild-type and mutant enzymes                     | 160            |
| 6.2.4 Kinetic characterization of wild-type and mutant enzymes                            | 160            |
| 6.2.5 Effect of temperature, pH and NaCl concentration on lipase activity and stability   | 161            |
| 6.2.6 Differential scanning fluorimetry analysis for determining unfolding ( $T_m$ value) | 161            |
| 6.2.7 ATR-FTIR for determination of secondary changes after mutagenesis in lipase         | 162            |
| 6.3 Results and Discussion                                                                | 162            |
| 6.3.1 Cloning and <i>in vitro</i> mutagenesis of <i>Bacillus subtilis</i> lipase          | 163            |
| 6.3.2 Enzyme expression and purification                                                  | 163            |
| 6.3.3 Enzyme kinetics studies                                                             | 165            |
| 6.3.4 Temperature stability analysis of wild type and mutated lipases                     | 169            |
| 6.3.5 pH stability analysis of wild type and mutated lipases                              | 173            |
| 6.3.6 Halostability analysis of wild type and mutated lipases                             | 176            |

|                                                                                                                                 |                |
|---------------------------------------------------------------------------------------------------------------------------------|----------------|
| 6.3.7 Differential Scanning Fluorimetry of wild type and mutated lipases                                                        | 179            |
| 6.3.8 Secondary structure prediction of wild-type and mutant lipase by FTIR (ATR)                                               | 183            |
| 6.4 Conclusion                                                                                                                  | 186            |
| 6.5 Reference                                                                                                                   | 187            |
| <b>CHAPTER 7: Employing predicted extreme-stable mutants for applications</b>                                                   | <b>189-230</b> |
| Abstract                                                                                                                        | 189            |
| 7.1 Introduction                                                                                                                | 190            |
| 7.2 Methodology                                                                                                                 | 191            |
| 7.2.1 Chemicals and instrumentation                                                                                             | 191            |
| 7.2.2 Testing lipase compatibility with surfactants, oxidizing agents and commercial detergents before and after immobilization | 192            |
| 7.2.3 Synthesis of functionalized ZnO nanoparticles                                                                             | 193            |
| 7.2.4 Covalent immobilization of lipase on synthesized ZnO nanoparticles                                                        | 193            |
| 7.2.5 Studying morphology and size of naïve and bioconjugated ZnO nanoparticle                                                  | 194            |
| 7.2.6 Characterization of naïve and bioconjugated ZnO nanoparticle using spectroscopic analysis                                 | 195            |
| 7.2.7 Analysis of naïve and bioconjugated ZnO nanoparticles using powder X-ray diffractometers                                  | 196            |
| 7.2.8 Enzyme activity assay and biochemical characterization of immobilized lipases                                             | 197            |
| 7.2.9 Reuse and storage stability of ZnO immobilized lipases                                                                    | 198            |
| 7.2.10 Application of ZnO immobilized lipases for removal of oil and grease stains from cotton fabrics                          | 198            |
| 7.2.11 Statistical analysis                                                                                                     | 199            |
| 7.3 Results and Discussion                                                                                                      | 200            |
| 7.3.1 Compatibility of lipase with surfactants, oxidizing Agents and commercial detergents                                      | 200            |
| 7.3.2 Morphological characterization of naïve and bioconjugated ZnO nanoparticles using FESEM, TEM and DLS analysis             | 202            |
| 7.3.3 UV-Vis spectroscopic analysis for characterization of naïve and                                                           | 207            |

|                                                                                                      |                |
|------------------------------------------------------------------------------------------------------|----------------|
| bioconjugated ZnO nanoparticles                                                                      |                |
| 7.3.4 FTIR spectroscopic analysis for characterization of naïve and bioconjugated ZnO nanoparticles  | 208            |
| 7.3.5 Raman spectroscopic analysis for characterization of naïve and bioconjugated ZnO nanoparticles | 209            |
| 7.3.6 XRD analysis for characterization of naïve and bioconjugated ZnO nanoparticles                 | 212            |
| 7.3.7 Biochemical characterization of immobilized lipase: Immobilization efficiency and yield        | 213            |
| 7.3.8 Reusability and storage stability of immobilized lipases                                       | 214            |
| 7.3.9 Lipase compatibility with detergents and oxidizing agents after immobilization                 | 217            |
| 7.3.10 Destaining of oil and grease stains from cotton fabrics                                       | 218            |
| 7.4 Conclusion                                                                                       | 228            |
| 7.5 References                                                                                       | 228            |
| <b>CHAPTER 8: Conclusion and future perspectives</b>                                                 | <b>231-236</b> |
| <b>Research Output</b>                                                                               | <b>237-240</b> |
| <b>Appendices</b>                                                                                    | <b>I-LXXII</b> |

## LIST OF TABLES

| Table | Description                                                                                                                                                                                | Page No. |
|-------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| 1.1   | Classification of extremophiles and their condition for growth and habitats                                                                                                                | 2        |
| 1.2   | Examples and applications of some extremophilic proteins.                                                                                                                                  | 5        |
| 1.3   | Contribution of amino acid residues towards protein extreme-stability.                                                                                                                     | 11       |
| 1.4   | Popular available tools to predict intra-protein interactions.                                                                                                                             | 14       |
| 1.5   | Popular software that predict stability changes after mutations.                                                                                                                           | 25       |
| 1.6   | Engineered mutants of extreme-stable proteins.                                                                                                                                             | 29       |
| 2.1   | Summary of collected data from each type of extremophiles.                                                                                                                                 | 59       |
| 3.1   | List of collected gene CDS from homologous extremophilic and non-extremophilic proteins and enumerated statistically significant codon features obtained after KS test (with $p < 0.05$ ). | 77       |
| 3.2   | Summary of results obtained by using 11 different algorithms of attribute weighting employed on different extremophilic datasets.                                                          | 90       |
| 3.3   | Results of unsupervised clustering for classification of extremophiles on the basis of codon usage.                                                                                        | 91       |
| 3.4   | Predicted accuracy of supervised learning for classification and model generation for different types of extremophiles on the basis of codon usage.                                        | 93       |
| 3.5   | Summary of decision tree prediction on extremophile datasets with their criteria chosen and best discriminatory rule for classification of codons.                                         | 96       |
| 4.1   | Attributes used for predication of protein extreme-stability (29 AA and 20 ST).                                                                                                            | 111      |
| 4.2   | Software, tools and algorithms used for prediction studies.                                                                                                                                | 112      |
| 4.3   | Enumerating statistically significant protein features by KS test.                                                                                                                         | 113      |
| 4.4   | Attribute weighting analysis of protein features.                                                                                                                                          | 119      |
| 4.5   | Results of unsupervised clustering for classification of extremophiles on the basis of codon usage.                                                                                        | 120      |
| 4.6   | Classification models by machine learning approaches and their prediction accuracies.                                                                                                      | 125      |
| 4.7   | Decision tree prediction and chosen best possible discriminatory rule for extreme-stability.                                                                                               | 126      |
| 4.8   | Validation of ranking models by blind test analysis.                                                                                                                                       | 130      |
| 4.9   | Salient findings for attributes contributing different types of extreme-stability.                                                                                                         | 131      |
| 5.1   | Selected mutations in <i>B. subtilis</i> lipase for generating various extreme-stabilizing mutants.                                                                                        | 141      |
| 5.2   | Validation of predicted mutations by online available tools/servers.                                                                                                                       | 144      |
| 5.3   | Ramachandran plot enumeration of generated mutations.                                                                                                                                      | 147      |
| 5.4   | Obtained RMSD values of mutant lipases in superimposition analysis.                                                                                                                        | 147      |
| 5.5   | Molecular docking of lipase mutants with C8 ( <i>p</i> -NPO) substrate.                                                                                                                    | 148      |
| 5.6   | Predicted percentage of secondary structure of wild-type and mutant lipases.                                                                                                               | 151      |
| 6.1   | Primers designed for predicted point mutations (underlined) in <i>Bacillus subtilis</i> 168 lipase                                                                                         | 159      |

|     |                                                                                                            |     |
|-----|------------------------------------------------------------------------------------------------------------|-----|
| 6.2 | Concentration of lipases obtained after purification ( $\text{g l}^{-1}$ ).                                | 164 |
| 6.3 | Enumerated enzyme kinetics of the wild type and mutant lipases.                                            | 169 |
| 6.4 | Analysis comparing the optimum activity and stability of wild-type and mutant lipases.                     | 179 |
| 6.5 | Comparing $T_m$ values of wild-type and mutant lipases at different experimental conditions.               | 183 |
| 6.6 | Determination of secondary structures by FTIR-ATR.                                                         | 186 |
| 7.1 | Compositions of washing solution for making 100 ml solution.                                               | 199 |
| 7.2 | Effect of surfactants, oxidizing agents and commercial detergents on wild-type and mutant lipase activity. | 201 |
| 7.3 | Enumeration of immobilization parameters and formula for immobilization efficiency and yield.              | 214 |
| 7.4 | Biochemical characterization of lipase after and before immobilization.                                    | 214 |
| 7.5 | Summary of washing process optimization of stain removal from the cotton fabrics.                          | 227 |



## LIST OF FIGURES

| Figure | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Page No. |
|--------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| 1.1    | Positive contribution of features related to genome and proteome of extremophilic organisms towards extreme-stability. The type of extremophile has been color coded. The figure has been deduced from the available data in literature. It is interesting to note that thermophiles are the most robust in adapting to elevated temperature conditions by employing the maximum number of versatile.                                                                                                            | 4        |
| 1.2    | Types of $\gamma$ -turns and the variation in their interval dihedral angle at $(i + 1)^{th}$ .                                                                                                                                                                                                                                                                                                                                                                                                                  | 18       |
| 2.1    | Representative features of database.                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 54       |
| 2.2    | Graphical unit interface of ExProtDB.                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 57       |
| 2.3    | Schematic illustration of detailed integration of components associated with ExProtDB. It depicts the major aspects of ExProtDB are extremophilic organisms, extremophilic proteins, mutants and physiochemical features.                                                                                                                                                                                                                                                                                        | 58       |
| 2.4    | Distribution of extremophilic organism into – (A) Domains – archaea, bacteria and acidophile; (B) Extremophile types – alkaliphile, barophile, halophile, psychrophile and thermophiles.                                                                                                                                                                                                                                                                                                                         | 60       |
| 2.5    | Distribution of extremophilic proteins (A) thermophilic proteins, (B) psychrophilic proteins, (C) acidophilic proteins, (D) alkaliphilic proteins, (E) halophilic protein and (F) barophilic proteins into enzyme classes – oxidoreductase, transferase, hydrolase, lyase, isomerase, ligase and other proteins.                                                                                                                                                                                                 | 61       |
| 2.6    | Distribution of extremophilic protein into extreme-stability – (A) thermostability (30 to 120 °C), (B) psychrostability (-5 to 25 °C), (C) acidostability (pH 2 to pH 6), (D) alkalistability (pH 8 to pH 11), (E) halostability (0 to 30 % (w/w) NaCl) and (F) barostability (20 to 99 MPa).                                                                                                                                                                                                                    | 62       |
| 2.7    | Amino acid composition analysis in different extremophilic proteins – (A) Thermophile vs Mesophile vs Psychrophile; (B) Acidophile vs Alkaliphile; (C) Halophile vs Non-halophile; and (D) Barophile vs Non-barophile.                                                                                                                                                                                                                                                                                           | 63       |
| 2.8    | Dipeptide composition analysis in different extremophilic proteins – (A) Thermophile; (B) Psychrophile; (C) Acidophile; (D) Alkaliphile; (E) Halophile; and (F) Barophile.                                                                                                                                                                                                                                                                                                                                       | 64       |
| 2.9    | Unique ID system for extremophilic organisms and proteins.                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 65       |
| 3.1    | Relative abundance of statistically significant codons in the comparing datasets: (A) T-M, (B) P-M, (C) T-P, (D) A-B, (E) H-Nh and (F) B-Nb. Green colour bars represent positive contributors of main datasets and negative contributors of counter dataset whereas, dark blue colour bars represent positive contributors counter datasets and negative contributors of main dataset.                                                                                                                          | 79       |
| 3.2    | Graphical representation of ranking of statistically significant codons in the scale of 1-9 using a python script. Ranking of codons in the (A) T-M, (B) P-M, (C) T-P, (D) A-B, (E) H-Nh and (F) B-Nb.                                                                                                                                                                                                                                                                                                           | 82       |
| 3.3    | Nucleotide composition analysis considering two parameters - (A) % AT- or % GC-richness and (B) % A/T- or % G/C-ending at third wobble position in the preferred significant codons for six class of extremophiles.                                                                                                                                                                                                                                                                                              | 85       |
| 3.4    | Data-point analysis of most and least preferred codon w.r.t. extremophiles. Analysis of (A) AGG codon (most preferred w.r.t. thermophiles) of T-M dataset, (B) CAA codon (least preferred w.r.t. thermophiles) of T-M dataset, (C) CAA codon (most preferred w.r.t. psychrophiles) of P-M datasets (D) AGG codon (least preferred w.r.t. psychrophiles) of P-M datasets, (E) AGG codon (most preferred w.r.t. thermophiles) of T-P datasets (F) CAA codon (least preferred w.r.t. thermophiles) of T-P datasets, | 86       |



|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |     |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
|     | (G) TGG codon (most preferred w.r.t. acidophiles) of A-B dataset, (H) GAG codon (least preferred w.r.t. acidophiles) of A-B dataset, (I) GAC codon (most preferred w.r.t. halophiles) of H-Nh dataset (J) AGG codon (least preferred w.r.t. halophiles) of H-Nh dataset (K) AGG codon (most preferred w.r.t. barophiles) of B-Nb dataset (L) CAA codon (least preferred w.r.t. barophiles) of B-Nb dataset are represented.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |     |
| 3.5 | Positive contribution of codon features related to the codon harmony in extremophiles. The different types of extremophiles have been colour coded. The figure has been deduced from the relative abundance and codon ranking analysis applied on available datasets used in the present study.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 87  |
| 4.1 | The rationalized method for prediction of barostability of proteins. (I) The hierarchy of extreme-stability is composed of three tiers: A. goal, B. criteria and C. alternatives. (II) In Step 1, the dataset of extremophilic (P1....Pm) and non-extremophilic (Q1....Qm) protein pairs along with their protein features (F1....Fn) form a matrix. Primarily arithmetic means were calculated for each feature of both extremophilic (B) and non-extremophilic (NB) proteins, where m is the total number of proteins pairs; n is the total number of features; X1....Xn and Y1....Yn: numerical values of extremophilic and non- extremophilic protein attributes, respectively. In Step 2, the generated mean weights of each feature were normalized by taking ratio of extremophilic ( $\sum X_i/m$ ) and non- extremophilic ( $\sum Y_i/m$ ) protein feature, where $\xi_1....\xi_n$ : ratio values of extremophilic and non- extremophilic protein attributes. In Step 3, the normalized ratio was derived and converted into the 1-9 scale using python script (formula as given), where $\Omega_1....\Omega_n$ : ratio weight of individual feature. In Step 4, these 1-9 scale rank score was supplied to a pairwise reciprocal comparison matrix, followed by calculating the sum of the column (S1....Sn). In Step 5, calculation of priority vectors was performed by calculated column sum, where W1....Wn: Priority vector w.r.t. feature F1....Fn by taking ratio of $\Omega_i/S_i$ . In Step 6, Consistency index (CI) and consistency ratio (CR) were derived by calculating $\lambda_{max}$ which is the consistency measure of each feature that is calculated as the dot product of pairwise comparison matrix with the priority vector matrix which is then divided by n. | 107 |
| 4.2 | Enumerating relative abundance of statistically significant protein features for extremophilicity (A) T-M, (B) P-M, (C) T-P, (D) A-B, (E) H-Nh and (F) B-Nb. Green color bars represent positive contributors of main datasets and negative contributors of counter dataset whereas, dark blue color bars represent positive contributors counter datasets and negative contributors of main dataset                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 114 |
| 4.3 | Ranks generated by AHP (MCDM) for protein attributes rendering extreme-stability.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 128 |
| 5.1 | Flowchart exhibiting the methods employed for validation of chosen mutations.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 139 |
| 5.2 | Prediction of ranks by AHP generated models for different extreme-stabilization in lipase.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 143 |
| 5.3 | Analysis of residues in the most favorable region by Ramachandran plot.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 146 |
| 5.4 | Interaction of p-NPO with active site of wild type and mutant lipases                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 149 |
| 5.5 | Graphical representation of unique contacts formed in wild-type and mutants.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | 150 |
| 6.1 | Methodology of multi- site mutagenesis using QuikChange Lightning Multi Site-Directed mutagenesis and PCR conditions. (Adapted from Agilent Technologies, QuikChange Lightning Multi Site-Directed Mutagenesis Kit Instruction Manual).                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 158 |
| 6.2 | Restriction digestion of <i>Bacillus subtilis</i> 168 lipase wild type and mutants in pET28a (5.4 kb) was digested using BamHI and NdeI and released wild-type and mutant lipase genes was obtained at 639 bp.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 164 |
| 6.3 | SDS PAGE (12 %) analysis of purified wild type and mutant enzymes.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 164 |

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |     |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 6.4  | Comparative analysis of Michaelis-Menten kinetics studies of wild-type with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipase.                                                                                                                                                                                                                                                                                                                                                              | 166 |
| 6.5  | Comparative analysis of Lineweaver-Burk kinetics studies of wild-type with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipase                                                                                                                                                                                                                                                                                                                                                                | 167 |
| 6.6  | Graph showing percentage relative activity for identifying the optimum temperature (from temperatures 10 °C - 70 °C) of wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Green arrow represent optimum activity of wild-type lipase and blue arrow represents optimum activity of most stable mutant. Assays conducted at pH 8 using p-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates. | 171 |
| 6.7  | Graph showing percentage relative activity for identifying the temperature stability at 50 °C (except psychrophilic mutants, temperature stability at 25 °C) of wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Assays conducted at pH 8 using p-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.                                                                                      | 172 |
| 6.8  | Graph showing percentage relative activity for identifying the optimum pH (from pH 2 – pH 12) of wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Green arrow represent optimum activity of wild-type lipase and blue arrow represents optimum activity of most stable mutant. Assays conducted at 35 °C using p-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.                       | 174 |
| 6.9  | Graph showing percentage relative activity for identifying the pH stability at pH 9 (except acidophilic mutants, pH stability at pH 6.5) of wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Assays conducted at 35 °C using p-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.                                                                                                         | 175 |
| 6.10 | Graph showing percentage relative activity for identifying the optimum NaCl concentration (from 0 M - 3.5 M) of wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Assays conducted at 35 °C and pH 8 using p-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.                                                                                                                            | 177 |
| 6.11 | Graph showing percentage relative activity for identifying the halostability at 1 M NaCl for wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Assays conducted at 35 °C and pH 8 using p-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.                                                                                                                                               | 178 |
| 6.12 | The schematic of the rationale behind thermal shift assay.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 180 |
| 6.13 | Thermograph of wild-type and mutant lipases.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 184 |
| 6.14 | A graph representing amide I region from 1600 – 1700 cm <sup>-1</sup> fitted with peaks of secondary structures in protein of the rationale behind thermal shift assay. (Adopted from, <a href="http://nptel.ac.in/courses/102103044/module2/lec10/8.html">http://nptel.ac.in/courses/102103044/module2/lec10/8.html</a> ).                                                                                                                                                                                                                                                         | 185 |

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |     |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 6.15 | FTIR-ATR analysis for structural change in wild-type (A. bsl_wt) and mutant lipases (B. bsl_the1, C. bsl_the2, D. bsl_the3, E. bsl_psy1, F. bsl_psy2, G. bsl_acd1, H. bsl_acd2, I. bsl_alk1, J. bsl_alk2, K. bsl_hal1 and L. bsl_hal2).                                                                                                                                                                                                                             | 181 |
| 7.1  | Scheme of ZnO nanoparticle synthesis and lipase immobilization.                                                                                                                                                                                                                                                                                                                                                                                                     | 194 |
| 7.2  | Effect of various surfactant, detergents and oxidizing agents on lipase (before immobilization). All values are represented as mean $\pm$ sd of three replications.                                                                                                                                                                                                                                                                                                 | 202 |
| 7.3  | FESEM, TEM and DLS analysis of synthesized ZnO nanoparticles.                                                                                                                                                                                                                                                                                                                                                                                                       | 204 |
| 7.4  | TEM micrographs of naïve (as-synthesized), coated and lipase immobilized ZnO nanoparticles at higher resolution (scale of 10 nm).                                                                                                                                                                                                                                                                                                                                   | 205 |
| 7.5  | FESEM, TEM and DLS analysis of bsl_wt, bsl_the3, bsl_psy2, bsl_acd2, bsl_alk2 and bsl_hal2 on ZnO nanoparticles.                                                                                                                                                                                                                                                                                                                                                    | 206 |
| 7.6  | UV-Vis spectroscopic analysis of wild-type and mutant lipases.                                                                                                                                                                                                                                                                                                                                                                                                      | 207 |
| 7.7  | FT-IR spectroscopic analysis of wild-type and mutant lipases.                                                                                                                                                                                                                                                                                                                                                                                                       | 209 |
| 7.8  | Raman spectroscopic analysis of wild-type and mutant lipases.                                                                                                                                                                                                                                                                                                                                                                                                       | 211 |
| 7.9  | XRD spectrogram of wild-type and mutant lipases.                                                                                                                                                                                                                                                                                                                                                                                                                    | 212 |
| 7.10 | Comparative analysis of reusability and storage stability of immobilized wild-type lipases and (A and B) thermostable mutant lipase, ZnO-bsl_the3; (C and D) psychrostable mutant lipase, ZnO-bsl_psy2 (E and F) acidostable mutant lipase, ZnO-bsl_acd2; (G and H) alkalistable mutant, Zn-bsl_alk2; and (I and J) halostable mutant lipase, ZnO-bsl_hal2, respectively. All values in optimization graphs are represented as mean $\pm$ sd of three replications. | 216 |
| 7.11 | Effect of various surfactant, detergents and oxidizing agents on lipase after immobilization. All values in optimization graphs are represented as mean $\pm$ sd of three replications.                                                                                                                                                                                                                                                                             | 217 |
| 7.12 | Optimization of oil removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and thermostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean $\pm$ SD (error bars).                                                                                                                | 219 |
| 7.13 | Optimization of grease removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and thermostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean $\pm$ SD (error bars).                                                                                                             | 219 |
| 7.14 | Optimization of oil removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and psychrostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean $\pm$ SD (error bars).                                                                                                               | 220 |
| 7.15 | Optimization of grease removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and psychrostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean $\pm$ SD (error bars).                                                                                                            | 221 |
| 7.16 | Optimization of oil removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and acidostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All                                                                                                                                                                                                                            | 222 |

|      |                                                                                                                                                                                                                                                                                                                                                         |     |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
|      | experiments were conducted in triplicate and the data have been represented as mean $\pm$ SD (error bars).                                                                                                                                                                                                                                              |     |
| 7.17 | Optimization of grease removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and acidostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean $\pm$ SD (error bars).  | 222 |
| 7.18 | Optimization of oil removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and alkalistable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean $\pm$ SD (error bars).    | 223 |
| 7.19 | Optimization of grease removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and alkalistable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean $\pm$ SD (error bars). | 224 |
| 7.20 | Optimization of oil removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and halostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean $\pm$ SD (error bars).      | 225 |
| 7.21 | Optimization of grease removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and halostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean $\pm$ SD (error bars).   | 225 |

## LIST OF ABBREVIATIONS

| ABBREVIATION | FULL NAME                                                   |
|--------------|-------------------------------------------------------------|
| Å            | Angström                                                    |
| ADP          | Adenine Dinucleotide Diphosphate                            |
| AHP          | Analytical Hierarchy Process                                |
| ANN          | Artificial Neural Network                                   |
| APTES        | (3-Aminopropyl)triethoxysilane                              |
| ATR          | Attenuated Total Reflectance                                |
| BLAST        | Basic Local Alignment Search Tool                           |
| BSA          | Bovine Serum Albumin                                        |
| bsl_acd1     | <i>Bacillus subtilis</i> lipase acidostable mutant 1        |
| bsl_acd2     | <i>Bacillus subtilis</i> lipase acidostable mutant 2        |
| bsl_alk1     | <i>Bacillus subtilis</i> lipase alkalostable mutant 1       |
| bsl_alk2     | <i>Bacillus subtilis</i> lipase alkalostable mutant 2       |
| bsl_bar      | <i>Bacillus subtilis</i> lipase barostable mutant           |
| bsl_hal1     | <i>Bacillus subtilis</i> lipase halostable mutant 1         |
| bsl_hal2     | <i>Bacillus subtilis</i> lipase halostable mutant 2         |
| bsl_psy1     | <i>Bacillus subtilis</i> lipase psychrostable mutant 1      |
| bsl_psy2     | <i>Bacillus subtilis</i> lipase psychrostable mutant 2      |
| bsl_the1     | <i>Bacillus subtilis</i> lipase thermostable mutant 1       |
| bsl_the2     | <i>Bacillus subtilis</i> lipase thermostable mutant 2       |
| bsl_the3     | <i>Bacillus subtilis</i> lipase thermostable mutant 3       |
| bsl_wt       | <i>Bacillus subtilis</i> lipase wild type                   |
| CDS          | Coding DNA Sequence                                         |
| CHAID        | Chi-squared Automatic Interaction Detection                 |
| CI           | Consistency Index                                           |
| Cpsy         | <i>Colwellia psychrerythraea</i>                            |
| CR           | Consistency Ratio                                           |
| DBSCAN       | Density-Based Spatial Clustering of Applications with Noise |
| DLS          | Dynamic Light Scattering                                    |
| DNA          | Deoxyribonucleic Acid                                       |
| DSC          | Differential Scanning Calorimeter                           |
| DSF          | Differential Scanning Fluorimeter                           |
| DTGS         | Deuterated Triglycine Sulfate                               |
| E.C. No.     | Enzyme Commission Number                                    |
| EMC          | Expectation Maximization Clustering                         |
| ENA          | European Nucleotide Archive                                 |
| FESEM        | Field Emission Scanning Electron Microscope                 |
| FT-IR        | Fourier Transformation – Infrared Spectrometer              |
| hAQP10       | Human Aquaporin 10                                          |



|                      |                                                                                         |
|----------------------|-----------------------------------------------------------------------------------------|
| ID3                  | Iterative Dichotomize 3                                                                 |
| IPTG                 | Isopropyl $\beta$ -D-1-thiogalactopyranoside                                            |
| IUBMB                | International Union of Biochemistry and Molecular Biology                               |
| kcal/mol             | Kilocalorie per mole                                                                    |
| kDa                  | Kilo Dalton                                                                             |
| kJ mol <sup>-1</sup> | Kilojoule per mole                                                                      |
| KS test              | Kolmogorov – Smirnov Test                                                               |
| Mb                   | Megabases                                                                               |
| MD simulation        | Molecular Dynamics Simulation                                                           |
| MPa                  | Mega Pascal                                                                             |
| NTA                  | Nitrilotriacetic Acid                                                                   |
| PCA                  | Principle Component Analysis                                                            |
| PDB                  | Protein Data Bank                                                                       |
| PEI                  | Polyethylenimine                                                                        |
| Pfu                  | <i>Pyrococcus furiosus</i>                                                              |
| <i>p</i> -NPO        | <i>p</i> -Nitrophenyl Octanoate                                                         |
| RMSD                 | Root Mean Square Deviation                                                              |
| RNA                  | Ribonucleic Acid                                                                        |
| SDS-PAGE             | Sodium Dodecyl Sulphate – Polyacrylamide Gel Electrophoresis                            |
| SVC                  | Support Vector Clustering                                                               |
| SVM                  | Support Vector Machine                                                                  |
| TEM                  | Transmission Electron Microscope                                                        |
| T <sub>m</sub>       | Melting Temperature                                                                     |
| UV-Vis               | Ultraviolet Visible Spectrophotometer                                                   |
| XRD                  | X-Ray Diffractometer                                                                    |
| ZnO – bsl_acd1       | <i>Bacillus subtilis</i> lipase acidostable mutant 1 immobilized on ZnO nanoparticles   |
| ZnO – bsl_acd2       | <i>Bacillus subtilis</i> lipase acidostable mutant 2 immobilized on ZnO nanoparticles   |
| ZnO – bsl_alk1       | <i>Bacillus subtilis</i> lipase alkalostable mutant 1 immobilized on ZnO nanoparticles  |
| ZnO – bsl_alk2       | <i>Bacillus subtilis</i> lipase alkalostable mutant 2 immobilized on ZnO nanoparticles  |
| ZnO – bsl_bar        | <i>Bacillus subtilis</i> lipase barostable mutant immobilized on ZnO nanoparticles      |
| ZnO – bsl_hal1       | <i>Bacillus subtilis</i> lipase halostable mutant 1 immobilized on ZnO nanoparticles    |
| ZnO – bsl_hal2       | <i>Bacillus subtilis</i> lipase halostable mutant 2 immobilized on ZnO nanoparticles    |
| ZnO – bsl_psy1       | <i>Bacillus subtilis</i> lipase psychrostable mutant 1 immobilized on ZnO nanoparticles |
| ZnO – bsl_psy2       | <i>Bacillus subtilis</i> lipase psychrostable mutant 2 immobilized on ZnO nanoparticles |
| ZnO – bsl_the1       | <i>Bacillus subtilis</i> lipase thermostable mutant 1 immobilized on ZnO nanoparticles  |
| ZnO – bsl_the2       | <i>Bacillus subtilis</i> lipase thermostable mutant 2 immobilized on ZnO nanoparticles  |
| ZnO – bsl_the3       | <i>Bacillus subtilis</i> lipase thermostable mutant 3 immobilized on ZnO nanoparticles  |
| ZnO – bsl_wt         | <i>Bacillus subtilis</i> lipase wild type immobilized on ZnO nanoparticles              |

## LIST OF SOFTWARES, DATABASES AND WEBTOOLS

| Tools              | Salient Features                                                                                                         |
|--------------------|--------------------------------------------------------------------------------------------------------------------------|
| APACHE             | Web server host through HTTP                                                                                             |
| Autodock 4.2       | Designed to predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure |
| AUTO-MUTE          | Machine learning based, structure as input, single mutation                                                              |
| BRENDA Database    | Resource that comprises molecular and biochemical information on enzymes that have been classified by the IUBMB          |
| CHpredict          | Aromatic interactions                                                                                                    |
| CLUSS2             | Alignment-independent algorithm for clustering protein families with multiple biological functions                       |
| ClustalW           | Multiple Sequence Alignment                                                                                              |
| CMView 1.1.1       | Contact Map visualization and analysis tool                                                                              |
| COPid              | Calculate the composition of amino acids or dipeptides or particular chemical properties of amino acids                  |
| Cupsat             | Sequence as input, single amino acid mutations                                                                           |
| DiANNA             | Disulphide bonds                                                                                                         |
| Dinosolve          | Disulphide bonds                                                                                                         |
| Disulfind          | Disulphide bonds                                                                                                         |
| EMBL               | European Molecular Biology Laboratory – Nucleotide Sequence Database                                                     |
| EMBOSS             | Pairwise sequence alignment. Reads two input sequences and writes their optimal global sequence alignment to file        |
| ENCoM              | Neural Network based, single mutation                                                                                    |
| ERIS               | Structure as input, multiple mutations                                                                                   |
| ESBRI              | Salt bridges                                                                                                             |
| FASTA              | Modern bioinformatic programs that rely on the FASTA format expect the sequence headers to be preceded by ">"            |
| HaloDom            | New database of halophiles across all life domains                                                                       |
| HaloWeb            | Home page for complete genome sequences of haloarchaeal microorganisms studied by the DasSarma laboratory                |
| HBAT               | Hydrogen bond                                                                                                            |
| HBOND              | Hydrogen bond                                                                                                            |
| HBPLOT             | Hydrogen bond                                                                                                            |
| HotSpot Wizard 2.0 | Automated design of site-specific mutations and smart libraries in protein engineering                                   |
| I-Mutant2          | Support Vector machine based, both sequence and structure can be used, single mutation                                   |
| iPTREE-STAB        | Machine learning based, single mutation                                                                                  |
| iRDP               | Ensemble of servers                                                                                                      |
| iSTABLE            | Support vector machine based, structure/sequence as input, single mutation                                               |
| I-TASSER           | Iterative Threading Assembly Refinement – server for protein structure and function prediction                           |



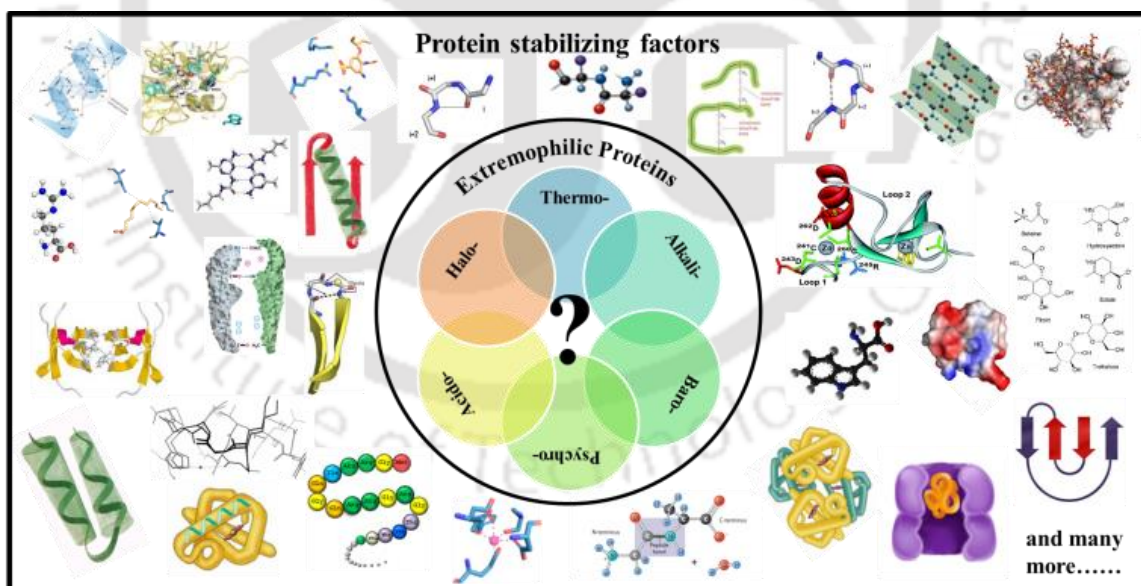
|                     |                                                                                                                                                                      |
|---------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| K-align             | An accurate and fast multiple sequence alignment algorithm                                                                                                           |
| KEGG Genome         | Database resource for understanding high-level functions and utilities of the biological system                                                                      |
| MATLAB              | Matrix laboratory is a multi-paradigm numerical computing environment                                                                                                |
| MUPRO               | Support Vector machine based, sequence as input, Single mutation                                                                                                     |
| MUSTAB              | Support Vector machine based, sequence as input, multiple mutations                                                                                                  |
| MySQL               | Open-source relational database management system (RDBMS)                                                                                                            |
| NCBI Genome         | Organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations                                                                 |
| NeEMO               | Machine learning based, structure as input,                                                                                                                          |
| Parallel PRRN       | Multiple sequence alignment by the best-first iterative refinement strategy with tree-dependent partitioning                                                         |
| PDB                 | Contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies                                                   |
| PEPSTATS            | Calculates statistics for your protein such as molecular weight, isoelectric point etc.                                                                              |
| PHP                 | Server-side scripting language designed for web development                                                                                                          |
| PIC                 | Protein Interaction Calculator recognizes various kinds of interactions; such as disulphide bonds, hydrophobic interactions, ionic interactions, hydrogen bonds etc. |
| PoPMuSiC            | Determines single mutation                                                                                                                                           |
| PROCHECK            | Allows checking of the stereochemical quality of a protein structure                                                                                                 |
| Promotif            | A program to identify and analyze structural motifs in proteins                                                                                                      |
| ProtDataTherm       | A database for thermostability analysis and engineering of proteins                                                                                                  |
| Protherm            | Thermodynamic database for proteins and mutants                                                                                                                      |
| ProtParam           | Allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL                                                |
| PSIPRED             | PSI-blast based secondary structure PREDiction (PSIPRED) is a method used to investigate protein structure                                                           |
| PubMed              | Free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM)         |
| PyMolV0.99          | A molecular visualization system                                                                                                                                     |
| RapidMiner v5.3.000 | Provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics                                      |
| RCSB                | The RCSB PDB curates and annotates PDB data according to agreed upon standards                                                                                       |
| ScrIDE              | IDE for the Python programming language                                                                                                                              |
| SDM                 | Sequence/structure as input, single mutation                                                                                                                         |
| UniProtKB           | Central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation                                                  |
| VADAR               | Hydrogen bond, accessible surface area                                                                                                                               |
| WET-STAB            | Machine learning based, multiple mutation                                                                                                                            |
| XAMPP               | Open source cross-platform web server solution stack package                                                                                                         |

## LIST OF PROTEIN ATTRIBUTES

| Amino Acid               | Amino acid classes | Intra-protein interaction                  | Secondary and tertiary structure attributes |
|--------------------------|--------------------|--------------------------------------------|---------------------------------------------|
| Alanine (Ala or A)       | Polar (Pol)        | Total hydrogen bonds (HB)                  | $\alpha$ -helix (AH)                        |
| Arginine (Arg or R)      | Non-polar (NPol)   | Main chain-main chain hydrogen bonds (MMH) | $\beta$ -strands (BST)                      |
| Asparagine (Asn or N)    | Small (Sml)        | Main chain-side chain hydrogen bonds (MSH) | $\beta$ -hairpins (BH)                      |
| Aspartic acid (Asp or D) | Tiny (Tiny)        | Side chain-side chain hydrogen bonds (SSH) | $\beta$ -sheets (BSH)                       |
| Cysteine (Cys or C)      | Aromatic (Aro)     | Hydrophobic interactions (HI)              | $\beta$ -turns (BT)                         |
| Glutamine (Gln or Q)     | Aliphatic (Ali)    | Disulfide bonds (DS)                       | $\gamma$ -turns (GT)                        |
| Glutamic acid (Glu or E) | Charged (Chrg)     | Ionic interactions (II)                    | Packing volume (PV)                         |
| Glycine (Gly or G)       | Basic (Bsc)        | Salt bridges (SB)                          | Charged accessible surface area (CASA)      |
| Histidine (His or H)     | Acidic (Acd)       | Aromatic-aromatic interactions (AAI)       | Polar accessible surface area (PASA)        |
| Isoleucine (Ile or I)    |                    | Aromatic-sulfur interactions (ASI)         | Non-polar accessible surface area (NPASA)   |
| Leucine (Leu or L)       |                    | Cation- $\pi$ interactions (CPI)           |                                             |
| Lysine (Lys or K)        |                    |                                            |                                             |
| Methionine (Met or M)    |                    |                                            |                                             |
| Phenylalanine (Phe or F) |                    |                                            |                                             |
| Proline (Pro or P)       |                    |                                            |                                             |
| Serine (Ser or S)        |                    |                                            |                                             |
| Threonine (Thr or T)     |                    |                                            |                                             |
| Tryptophan (Trp or W)    |                    |                                            |                                             |
| Tyrosine (Tyr or Y)      |                    |                                            |                                             |
| Valine (Val or V)        |                    |                                            |                                             |

## 1

## Introduction and literature review



*Multitude of factors, which one of them or combination of factors stabilize extreme-stable proteins?*

## Abstract

Research on extremophilic proteins has seen immense growth in the past decade owing to their industrial importance. Knowledge of attributes contributing to extreme-stability would assist the research community in attaining and using extremophilic proteins. Modern mechanistic approaches to engineer such proteins *in vitro* will have significant impact in industrial biotechnology economy. Developing *a priori* knowledge about the mechanism behind extreme-stability will nurture better understanding of pathways leading to molecular evolution of such candidate proteins. This chapter is a vivid compilation about all classes of extremophilic proteins and the attributes that lead to myriad of adaptations divulged after an extensive study of research articles belonging to extremophilic proteins. Along with detailing on the rationale behind extreme-stability of proteins, emphasis has been put on modern approaches that have been utilized to render proteins extremostable by protein engineering. It was reckoned that each protein shows different approaches to extreme-stability governed by minute differences in their biophysical properties and the milieu in which they exist. Thus, present research lays the foundation to further explore the biophysical pleiotropy of extreme-stable proteins in order to deduce a global prediction model for predicting the effect of mutations on protein stability.

## 1.1 Introduction

Organisms that are tolerant to extreme *milieu* are known as extremophiles and are present on earth since 4 billion years<sup>1</sup>. Such organisms can colonize in diverse range of environments such as hot springs, geothermal vents and mud holes, dry solfataric soil, area near nuclear reactor, deep marine settings, earth crust interiors, highly mineralized soils, extreme radiation surroundings, extreme aquatic chemical environments, etc.<sup>2</sup>. They are able to survive at extreme temperatures (psychrophiles and thermophiles), extreme pH (acidophiles and alkaliphiles), salinities (halophiles), high hydrostatic pressure (barophiles), low water activity (xerophiles), etc.<sup>3</sup>. Table 1.1 summarizes major classes of extremophiles with their growth condition and habitats.

Table 1.1: Classification of extremophiles and their condition for growth and habitats

| Extremophile             | Extreme Conditions                                                 | Environment                                            | Example                                      | Reference     |
|--------------------------|--------------------------------------------------------------------|--------------------------------------------------------|----------------------------------------------|---------------|
| Thermophile              | >45°C up to ~80°C                                                  | Hot springs, geothermal vents, solfataric fields, etc. | <i>Thermococcus celer</i> (88 °C)            | <sup>4</sup>  |
| Psychrophile (Cryophile) | -2 °C to ~15°C                                                     | Snow, lakewater, sediment, ice                         | <i>Vibrio</i> sp. (-17 °C)                   | <sup>5</sup>  |
| Hyper-thermophile        | >80°C; limit so far is about 115°C                                 | Submarine vent, terrestrial hot spring                 | <i>Pyrolobus fumarii</i> (113 °C)            | <sup>6</sup>  |
| Halophile                | >1M salt concentration                                             | Saline lakes and salt mashes                           | <i>Halobacterium</i> sp. (5.2 M NaCl)        | <sup>7</sup>  |
| Alkaliphile              | pH 8 – 12                                                          | Soda lakes                                             | <i>Clostridium paradoxum</i> (pH 12.5)       | <sup>8</sup>  |
| Acidophile               | pH 0 – 5                                                           | Dry solfataric soil                                    | <i>Ferroplasma acidarmanus</i> (pH 0 – 2.5)  | <sup>9</sup>  |
| Barophile (Piezophile)   | >150 MPa                                                           | Colonize deep sea habitats                             | <i>Pyrococcus yayanosii</i> (80 MPa)         | <sup>10</sup> |
| Xerophile                | Low water activity, grow in extremely dry, desiccating conditions  | Rock surfaces hypersaline, organic fluids              | <i>Xeromyces bisporus</i> ( <i>aw</i> =0.96) | <sup>11</sup> |
| Radioresistant           | Organisms resistant to high levels of ionizing radiation (like UV) | Soil, nuclear reactor water core, submarine vent       | <i>Deinococcus radiodurans</i> (500,000 rad) | <sup>12</sup> |
| Metallotolerant          | Capable of tolerating heavy metals, such as Cu, Cd, As, and Zn     | Acid mine drainage                                     | <i>Cupriavidus metallidurans</i>             | <sup>13</sup> |

Phylogenetically, extremophiles include members of all three domains of life that is bacteria, archaea, and eukarya but mainly belong to the domain archaea<sup>14</sup>. Extremophiles

have adapted specific mechanisms that help them to grow and survive in extreme environments by mechanisms such as (i) cell wall lacks peptidoglycan and cell membrane contains ether linkages; (ii) have characteristics of high GC content, positively super coiled DNA, smaller genome size and robust DNA repair system; (iii) unique codon and amino acid usage pattern in different frequencies depending on their extreme environment; (iv) have enzymes which are functional and enhanced stability under extreme conditions; (v) chaperone assisted protein folding; and (vi) have extremolytes to have reduced UV-induced DNA mutations, stabilization of enzymes, protect against oxidative damage and proteolytic cleavage.

Extremophiles and their proteins (enzymes) are potentially and efficiently used over normal mesophilic counterpart in many industries since they have enhanced catalysis rate, solubility and diminish microbial contamination in harsh industrial processes<sup>15</sup>. Most of the mesophilic enzymes cannot withstand extreme industrial processes and thus, this necessitates the development of extreme-stable biocatalysts to sustain such industrial processes. The industrially important enzymes extracted from extremophiles include amylases, cellulases, lipases, proteases and pectinases.

## 1.2 Extremophilic adaptations

Due to upsurge in research in the past decade about one of the “unsolved biological questions” of protein adaptability in extreme conditions, it has been rightly said that “reviews in rapidly growing fields of biochemistry have to be rewritten rather than written”<sup>16</sup>. By extreme, the physical variability of temperature, salt concentration, pH and pressure is comprehended. Table 1.2 provides some examples of extremophilic proteins. Extremophilic proteins not only tolerate but most of them require extreme conditions for their optimal functioning. Few exceptions exist, for example many extremely thermostable enzymes are able to catalyse their reaction also at ambient temperatures, although at a much lower rate. For example, *Pyrococcus furiosus* aldehyde: ferredoxin oxidoreductase is able to transfer electrons at room temperature and at 100 °C. Again, overlapping of extreme conditions, for example, thermophilic environment with high hydrostatic pressure, challenges evolution of proteins by incorporation of mutations that



encompasses stability features required for efficient protein functioning. Interestingly, extremophilic proteins may have more than one adaptation<sup>17</sup>. Therefore, the classification of extreme adaptations made on the basis of their source organisms may not be appropriate. Though, thermophilic, psychrophilic and halophilic stability are some of the primary branches of extreme adaptations with intersecting points<sup>17</sup>. This reflects that variability in adaptations is fuzzy and case specific as having few changes in the protein can lead to adaptations in a range of conditions<sup>17,18</sup>. For example, acid stable proteins can also be thermostable<sup>19</sup>; acidophilic thermostable endo-1,4- $\beta$ -mannanase of *Penicilliumoxalicum* GZ-2, is highly stable at 60 °C and pH 4.0<sup>20</sup>. Similarly, two thermostable and alkaliphilic xylanases, designated XylA and XylB from *Bacillus* sp. AR-009 have mixed thermophilic and alkaliphilic behaviours. Both enzymes are stable in a broad pH range and shows good stability when incubated at 60 °C and 65 °C in pH 8 and pH 9 buffers respectively<sup>21</sup>.

Further, complications arise in understanding the rationale behind extreme-stability of proteins as when extremophilic proteins belong to archaea, bacteria and eukaryotes. For example, thermostable proteins belong to all these groups such as glucose dehydrogenase from the archaeon *Thermoplasma acidophilum*<sup>22</sup>, glucokinase from the bacterium *Bacillus stearothermophilus*<sup>22</sup>, eukaryotic human aquaporin10 (hAQP10)<sup>23</sup> and thermostable protease CutA1 protein from human brain<sup>24</sup>. The point to ponder here is whether a universal protocol of stabilization is followed by extreme-stable proteins belonging to all these kingdoms of life. Another compelling question is why and how certain proteins are extremestable, even in the absence of a clear evolutionary drive. Researchers are also interested to understand extreme-stability of proteins due to their high priority in industries as they allow easy mixing, better substrate solubility, high mass transfer rate, and lowered risk of contamination<sup>25</sup>. To accomplish the aforementioned, the availability of extremophilic proteins for research is necessary, but it is difficult to obtain them from their natural sources. Therefore, attaining stable mutation through protein engineering has always been an evolving field. Hence, it is not only essential but also critical to understand molecular adaptations of extremophilic proteins to engineer efficient enzymes that are extreme-stable.



Table 1.2: Examples and applications of some extremophilic proteins.

| Proteins      | Optimum condition  | Example                                                       | Application                                  | Reference |
|---------------|--------------------|---------------------------------------------------------------|----------------------------------------------|-----------|
| Thermophilic  | Temperature >45 °C | Pfu polymerase from <i>Pyrococcus furiosus</i>                | Detergent, paper, organic synthesis, biofuel | 26        |
| Psychrophilic | Temperature <20 °C | 34H Cpsy-P1 from <i>Colwellia psychrerythraea</i>             | Organosynthesis, food processing             | 27        |
| Alkaliphilic  | pH >10             | Alkaline proteases from <i>Bacillus sp.</i>                   | Detergent                                    | 28        |
| Acidophilic   | pH <5              | $\alpha$ -amylase from <i>Alicyclobacillus acidocaldarius</i> | Organic synthesis, Food processing           | 17        |
| Barophilic    | >40 MPa            | P35 from <i>Thermococcus barophilus</i>                       | Food industry                                | 29        |
| Halophilic    | 2 – 5 M salt       | Malate dehydrogenase from <i>Haloarcula marismortui</i>       | Waste water treatment                        | 30        |

### 1.3 Multiple factors involved in extreme-stability of proteins

Currently, research focuses mainly on genomic and proteomic level adaptations of extremophiles and each level comprise of numerous attributes which requires further exploration<sup>31</sup>. The significant enhancement in our understanding of extremophilic adaptation has also been nurtured by availability of complete genome sequences and high-resolution, three dimensional structures of proteins. Most of the research work in this direction concluded that stability of extremophilic proteins is a resultant of cumulative effect of various stabilizing factors and each protein shows different approaches to extreme-stability governed by their biophysical properties. Figure 1.1 is a vivid illustration of the contribution of various factors for extreme-stability which have been reported till today.

Thermophilic adaptations can be observed to be the most versatile among all being capable to utilize majority of the listed stabilization features. As the bulk of the literature is about thermophilic adaptations, further exploration of other types of extremophilicity will aid in development of a better picture about the robustness in adaptations of the other extremophiles. It is also interesting to note that thermophiles and psychrophiles choose vice-versa features for enhancing their stability. For example, thermostability increases by increase in protein rigidity whereas psychophilicity increases by increase in flexibility. Similarly, alkaliphiles show enhanced basic amino acid residues whereas acidophiles

show an increase in acidic amino acid residues. Another complication in understanding extreme-stability is that certain proteins for example; psychrophilic proteins like lipases from *Candida antarctica* which is supposed to be psychrophilic is actually thermostable. They are stable due to rigidity and higher number of van der Waals forces, reduction in electrostatic interaction and increase in hydrogen bonding<sup>32</sup>. Therefore, further in depth knowledge about extreme-stability is necessary to draw a complete conclusion regarding the range of features that can classify different types of extreme-stable proteins.

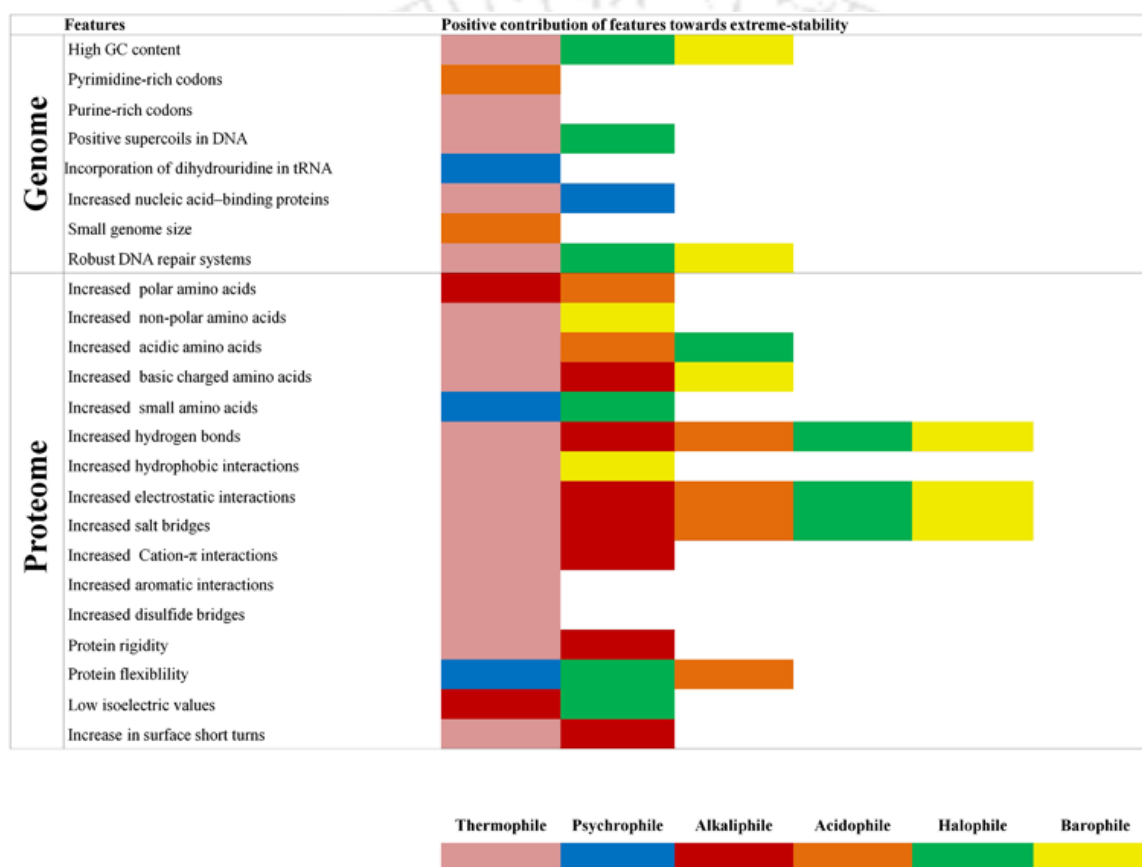


Figure 1.1: Positive contribution of features related to genome and proteome of extremophilic organisms towards extreme-stability<sup>17,33–36</sup>. The type of extremophile has been color coded. The figure has been deduced from the available data in literature. It is interesting to note that thermophiles are the most robust in adapting to elevated temperature conditions by employing the maximum number of versatile.

#### 1.4 Genomic adaptations of extremophiles: codon adaptability and GC-content

A large repertoire of extremophilic organisms has been identified. However, only few of them have been successfully sequenced. Search for genomes with the key words “thermophilic”, “hyperthermophilic”, “psychrophilic”, “halophilic”, “alkaliphilic”,

“acidophilic” and “barophilic” in NCBI genome search results in a total of 156 hits for thermophiles and hyperthermophiles, 14 hits for psychrophiles, 70 hits for halophiles, 13 hits for alkaliphiles, 29 hits for acidophiles and 8 hits for barophiles till 2017. Based on these genome search results it can be demonstrated that most of the work to understand extreme-stability has been carried out for thermophiles.

The genomes of extremophiles have evolved through the phenomenon of lateral gene transfer from bacteria by orthologous replacement or incorporation of paralogous genes<sup>37</sup>. Researchers compare the extremophiles on the basis of their genomic features like their genome size, order of genes, codon usage bias, and GC-content to determine what mechanisms could have produced the great variety of genomes that exist today<sup>38</sup>. Comparative analysis of the genomes of extremophiles and non-extremophiles reveals some common trends in extremophiles that cross the boundary of phylogenetic relationships and genomics such as genome size, DNA repair system, DNA stabilizers, genomic GC-content and genome super-coiling. Sequencing of extremophile genomes and their analysis has shed light on their genetic evolution.

It has been reported that variations in nucleotide composition can have very significant effects on the patterns of codon usage and thermostability<sup>39,40</sup>. Thermophiles and hyperthermophiles have been reported to have a high GC-content<sup>41,42</sup>. Correspondingly, the codons such as ATC, CGG, TTG, CGC, which are G/C ending and therefore, likely to give higher stability to the codon–anticodon interaction, have been reported to be higher in thermophiles than mesophiles<sup>43</sup>. Work carried out by Zeldovich et al. concluded that an increase in purine (A+G) of thermophilic bacterial genomes due to the preference for isoleucine, valine, tyrosine, tryptophan, arginine, glutamine, and leucine, which have purine-rich codon patterns, is responsible for the possible primary adaptation mechanism for thermophilicity<sup>44</sup>. These amino acid residues increase the content of hydrophobic and charged amino acids, enhancing thermostability. To the contrary, for acidophiles, pyrimidine-rich codons are preferred<sup>45</sup>. Thus, the genomes of thermoacidophiles such as *P. torridus* have evolved by lowering the purine-containing codons in long open reading frames<sup>46</sup>. The comparison of base composition, codon and amino acid usages of thermophilic *Aquifex aeolicus* with those of *Bacillus subtilis* shows that there is a

significant increase in purine content and GC composition in codon selection. This type of alteration in base compositions also influences the codon and amino acid usages of *A. aeolicus*<sup>47</sup>. Like the genome of hyperthermophiles and thermophiles, the halophiles also have high GC content of around 60 to 70% which can avoid ultraviolet-induced thymidine dimer formation and mutations<sup>48</sup>. At the DNA level, compared to non-halophilic genomes, the halophiles exhibit distinct dinucleotides (CG, GA/TC, and AC/GT) at the first and second codon positions, reflecting an abundance of aspartate, glutamine, threonine, and valine residues in halophile proteins, which leads to their stability<sup>49</sup>. The presence of high levels of CG dinucleotides leads to an increase in stacking energy and, thus, genome stability. By the above reports it can be concluded that the basis of extreme-stability selection purely based on GC content is fuzzy.

Further, it has also been reported that codon usage preferences might be based on an error-minimizing selection at the protein level. Mutation occurs in any codon for thermophilic adaptation, which results in similar amino acids that leads to the similar protein conformations which are relatively less deleterious compared to others. For example, Arg codon changing to Lys codon has less deleterious effect in protein conformations because of similar properties of both amino acids<sup>50</sup>. Arginine is coded by two sets of codons AGR (AGA and AGG) and CGN (CGU, CGC, CGA and CGG). The biasness is also seen in between these two sets of arginine codons AGR and CGN. Whenever there is an increase in lysine for thermophilic adaptations, AGR codons are more preferred codons for arginine coding<sup>51</sup>. Because a positive error minimization mechanism between arginine and lysine codons (AAA and AAG) occurs due to a single mutation at second position that could turn AGR arginine codons into codons for lysine (which can be harmful for thermostability). Whereas the decrease in CGN codon usage in organisms that live in high temperatures can be interpreted as a negative error minimization<sup>51</sup>. Since a single mutation at second position could turn CGN arginine codons into codons for histidine (CAU and CAC) and glutamine (CAA and CAG) which can be harmful for thermostability<sup>51</sup>.

Additionally, DNA was observed to possess positive supercoils resulting in greater stability<sup>52</sup>. It was also reported that genome size matters for extreme stability. The

genome size of acidophiles is smaller than that of neutrophiles. For example, the smallest genome belongs to thermoacidophilic *Thermoplasma* (<2 Mb)<sup>53</sup>. Tyson et al. reported a variety of genes in acidophiles involved in their membrane biosynthesis, indicative of acid tolerance capacity of these microorganisms. Similarly, alkali-stable genomes are adapted to maintained pH homeostasis. Recently, the complete genome of *Bacillus subtilis* and *Bacillus halodurans* C-125 has been sequenced. Genes responsible for the alkaliphily of *B. halodurans* C-125 and *Bacillus firmus* OF4 have been analyzed<sup>54</sup>. In their genome, several open reading frames for Na<sup>+</sup> /H<sup>+</sup> antiporters responsible for pH homeostasis in alkaliphiles have been characterized<sup>55</sup>. The *tupA* gene was identified in the *B. halodurans* genome, which is responsible for the synthesis of teichuronopeptide, a major structural component in the cell wall important for maintaining pH homeostasis<sup>54</sup>.

To the best of our knowledge less work on genomes from barophilic microorganisms has been reported. Reports on barophilic adaptations can be found in the work of Di Giulio who initiated his studies by comparison of homologous genomic sequences of barophilic *Pyrococcus abyssi* and non-barophilic *Pyrococcus furiosus*. He reported that GC-rich codons were significant in barophiles<sup>56</sup>. It has also been reported that in order to survive in such environments, barophilic microorganisms require robust DNA repair systems since high pressure can damage DNA and proteins, so survival necessitates avoidance of damage or high repair rates<sup>14,57</sup>. Additionally, pressure-regulated operons (*ompH*) have evolved in barophilic genomes which are responsible for high-pressure adaptations in barophiles<sup>58,59</sup>. Conclusively, it can be said by comparisons made between extremophile and non-extremophile genomes that difference in the patterns of codon usage results in a partial resistance towards the changes in the concentration of a given amino acid. This buffering capacity might explain the observed differences in codon usage trends in genes of two different set of organisms<sup>35</sup>. These studies help in understanding the use of preferred codon in an organism as powerful tools that can be utilized to improve function predictions and genome-environment mappings<sup>35</sup>. But there is still lack of knowledge of extremophile genomes and global codon usage. Further, codon usage studies were mainly done only for thermophiles but with no conclusive finding as which codon is preferred in thermophiles.

## 1.5 Proteomic adaptations of extremophiles: role in extreme-stabilization of proteins

The rapid progress in the sequencing of genomes has opened many new avenues of research. The complete genomic sequence of an extremophilic organism gives complete information about their proteomics as global amino acid compositions and amino acid usage frequency of all proteins. In particular, from genomic studies it can be observed that GC-content has a dominant effect for extremophilic species with the capability to change their global amino acid compositions and amino acid usage<sup>60</sup>. The differences in the preferences in amino acid composition in each group of extremophile are different depending on their surrounding environment. Enhancement of intra-protein interactions has been reported to be associated with increasing extreme-stability of proteins<sup>61–64</sup>. Such interactions and their possible contributions towards extreme stability of proteins have been elaborately discussed in the next sub-sections.

### 1.5.1 Role of amino acid in rendering proteins extreme-stabilization

Many researchers have attempted to study the effect of amino acids on protein thermostability<sup>65–68</sup>. At temperatures greater than 100 °C the thermostability of amino acids in proteins have been reported by Jaenicke and Böhm as follows (V,L) > I> Y> K> H> M> T> S> W> (D, E, R, C)<sup>69</sup>. The question that arises here is that whether the same rule is applicable to proteins that are stable at <100 °C, or the psychrophilic proteins like *Candida antarctica* lipase which show thermostability. An interesting rule was derived by Farias and Bonato who found that the E+K/Q+H ratio can distinguish hyperthermophiles (>4.5) from mesophiles (<2.5) and thermophiles (3.2-4.6)<sup>51</sup>. Trivedi et al. said that “as there are variations in preference for other amino acids between mesophiles, thermophiles and hyperthermophiles, it is apparent that these variations are not only organism specific but are also protein specific within the organism”<sup>64</sup>. A study of the existing literature does not lead to conclusive correlation between amino acid composition and extreme stability rather adding to confusion. Table 1.3 summarizes the reported features. It is observed that derivation of global rules based on the properties of amino acids has become puzzling for the stabilization of extremophilic proteins under different extreme environmental conditions. For instance, thermophiles include higher



hydrophobic residues, more charged amino acids (especially Glu, Arg and Lys), and fewer uncharged polar residues (Ser, Thr, Asn, and Gln)<sup>70</sup>. Likewise, halostable proteins have increased trend of negative surface charge due to increased acidic amino acid content (Asp and Glu) which compensates for the extreme ionic conditions<sup>17</sup>. Decreased hydrophobic amino acid frequency in halostable proteins is also reported<sup>45</sup>. Another example of *Pyrococcus abyssi*, a hyperthermophilic piezophile having an increase in small amino acids (Gly, Ala, Ser, Thr, Pro, Asp and Asn) across its proteome when compared to that of the related thermophilic non-piezophile archaeon, *Pyrococcus furiosus* have been found<sup>56</sup>. Adaptation of proteins at low pH seems to be attributed to the prevalence of acidic (negatively charged at a neutral pH) amino acids (Asp and Glu) whereas, adaptation of proteins at basic pH seems to be attributed to the prevalence of basic amino acids (Lys, Arg and His) on the surface of these enzymes and proteins<sup>71</sup>.

Table 1.3: Contribution of amino acid residues towards protein extreme-stability.

| Feature | Positive Contribution                                                  | Negative Contribution                                                                                   |
|---------|------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Ala     | High helix propensity, hydrophobic interaction, higher aliphatic index | Nil                                                                                                     |
| Cys     | Disulphide bond                                                        | Thermolabile                                                                                            |
| Glu     | Ionic interactions, highest capacity to bind water molecules.          | Nil                                                                                                     |
| Asp     | Nil                                                                    | Asp-Pro combination may be susceptible to hydrolysis of peptide bonds                                   |
| His     | Rigidity                                                               | Nil                                                                                                     |
| Phe     | Core hydrophobicity                                                    | Nil                                                                                                     |
| Gly     | Nil                                                                    | Makes cavity in inner part of protein structure                                                         |
| Ile     | Hydrophobic interaction, high aliphatic index                          | Nil                                                                                                     |
| Lys     | Ionic interactions                                                     | Nil                                                                                                     |
| Leu     | Hydrophobic interaction, high aliphatic index                          | Nil                                                                                                     |
| Met     | Nil                                                                    | Thermolabile                                                                                            |
| Asn     | Nil                                                                    | Thermolabile                                                                                            |
| Pro     | Rigidity, lowest conformational entropy                                | Nil                                                                                                     |
| Gln     | Nil                                                                    | Thermolabile                                                                                            |
| Arg     | Ionic interactions                                                     | Nil                                                                                                     |
| Ser     | Nil                                                                    | Best residue for interacting with the waters, at high temperature disturbance of interaction with water |
| Thr     | Hydrophobic interaction, high aliphatic index                          | Nil                                                                                                     |
| Val     | Hydrophobic interaction, high aliphatic index                          | Nil                                                                                                     |
| Trp     | Hydrophobic interaction, high aliphatic index                          | Nil                                                                                                     |
| Tyr     | Hydrophobic interaction, high aliphatic index                          | Nil                                                                                                     |



Protein destabilization at higher temperatures has been reported mainly due to the presence of thermolabile residues as they are prone to deamidation, cleavage of peptide bonds and oxidation of Cys and Met residues. Deamidation of Asn residues is a spontaneous and non-enzymatic reaction which can occur in acidic, neutral, or alkaline conditions. Therefore, proteins with high Asn residues are suspected to be less thermostable<sup>26,72</sup>. Furthermore, reports also suggest that Asn containing peptides are more prone to deamidation at a faster rate than glutamine counterparts<sup>73</sup>. The cleavage of a peptide can destabilize proteins by disrupting a protein chain. Shirley identified three modes of peptide bond cleavage. This can occur by (1) hydrolysis of peptide bonds under acidic conditions at Asp residues; (2) succinimide formation at Asn residues at physiological pH; and (3) proteolysis by enzymes. Literature survey shows that protein having high frequency of Asp and Asn residues were less thermostable<sup>26,72</sup>. Oxidation of Cys and Met often results in protein destabilization. Cys has been reported to be involved in protein stability of extracellular proteins by formation of disulphide bridges. Such covalent interactions have been found in many thermostable proteins. Formation of disulphide bond is dependent on the protein conformation. However, free Cys residues are prone to oxidation and  $\beta$ -elimination<sup>74</sup>. This has been shown to irreversibly inactivate ribonuclease and lysozyme at pH 6 - 8 and 90 - 100 °C<sup>75</sup>. Similarly, Met oxidation has been associated with the inactivation of proteins<sup>76</sup>. Many thermostable proteins have been reported to have low frequency of Met residues<sup>26,77</sup>.

Structural rigidity causes enhanced compactness with reduced radius of gyration. It plays a crucial role in attaining improved stability in thermophilic proteins by substituting conserved flexible regions in the counter homologous thermolabile proteins<sup>15,78</sup>. However, in contrary, the protein flexibility is the characteristic of cold-adaptive (psychrophilic) protein that help to retain structural stability at freezing temperatures<sup>79</sup>. In psychrophilic proteins, the number of contacts per residue are reduced that lead to decrease in protein compactness (or increase flexibility) and have increased radius of gyration. These biophysical factors are responsible for more or less packing of secondary structures in the three-dimensional structure of protein depending upon the temperature tolerance<sup>80</sup>. In psychrophilic proteins, the flexibility is achieved by decreasing the

number of Pro and Arg residues (which leads to the rigidity of proteins by restricting backbone rotations) and increasing the number of Gly residues in their sequences<sup>81</sup>. While comparing the psychrostable proteins with mesostable ones, it was found that they also comprise a significantly higher proportion of amino acids that contribute to higher protein flexibility in the coil regions of proteins, such as those with small or neutral side chains<sup>27,82</sup>. Metapally and Reddy also said that in psychrophilic proteins, Ser, Asp, Thr and Ala are overrepresented in the coil regions of secondary structures whilst, Glu and Leu are underrepresented in the helical regions. Comparing the amino acid properties, the psychrophilic proteins with aliphatic, basic, aromatic and hydrophilic side chains are underrepresented in the helical regions of proteins of psychrophiles<sup>27</sup>.

The important amino acid features of other extremophilic adaptations are thought provoking. Like halostable proteins are often also thermotolerant<sup>83</sup> but possesses less hydrophobicity, as at high salt concentrations. A decrease in large hydrophobic residues and an increase in smaller residues also contribute to the stability of halostable proteins. Proteins are destabilized by high hydrophobic interactions, because it leads to protein aggregation<sup>48</sup>. It has been noted that compared to non-halostable homologs, contain a greater proportion of negatively charged amino acids (Glu and Asp) relative to positively charged amino acids<sup>84,85</sup>. Acidic amino acids dominate their surface because they can hold the essential hydration shell layer for stability and catalysis intact at the surface of the protein. This is reflected in the low isoelectric values of halostable proteins<sup>48</sup>. The amino acid adaptation with the extreme pH organism is also dependent on their surrounding environment. Alkaliphiles are adapted to an increase in the surface exposure of acidic residues with respect to basic residues changes the net charge of the molecule towards negative. The gain of glutamate and arginine residues and the loss of aspartate and lysine residues are key players in alkaline adaptation of proteins<sup>48</sup>. Besides, the proteins of acidophiles were observed to be rich in acidic residues, which also show low solvent exposure. Stability is also obtained by replacement of charged amino acids by neutral polar amino acids in proteins which reduces the electrostatic repulsion that occurs between charged groups at low pH<sup>86</sup>. An increase in the isoleucine content of the proteins in acidophiles is another observed trend, which is assumed to contribute to acid

stability<sup>46</sup>. For extreme pressure adaptation, Di Giulio, reported that barophilic *P. abyssi* tends to substitute arginine (Arg) for all other amino acids in sequences homologous to non-barophilic *P. furiosus*<sup>56</sup>. He considered Arg to be the “barophilic amino acid”<sup>56</sup>. Yafremava et al. also shown that Arg is preferred in barostable proteins when compared to non-barostable proteins<sup>87</sup>.

Thus, it can be concluded that extremophiles have evolved ways to manipulate amino acid to optimize the protein or enzyme activity. Their adaptation to their environment for protein function and structure is due to the change in amino acid sequence<sup>17</sup>. Also amino acid usage profile of a gene is a reflection of the environment in which its protein has evolved. Amino acid profile is a powerful tool that can be utilized to improve function predictions and genome-environment mappings<sup>35</sup>. However, present knowledge about amino acid preferences being individualistic and overlapping among different classes of extremostable proteins does not suffice the requirement for developing a global predictor with universal rules for the purpose of classification and protein engineering. Therefore, this necessitates further insights into the biophysical feature space of such proteins. This can be achieved by analysis of their three dimensional structure.

### 1.5.2 The biophysical attributes responsible for extreme-stability of proteins

Enhancement of intra-protein interactions has been reported to be associated with increasing extreme-stability of proteins<sup>61–64</sup>. Such interactions and their possible contributions towards stabilizing proteins will be elaborately discussed in the next sub-sections. Table 1.4 presents the web tools and software that are available to predict such intra-protein interactions.

Table 1.4: Popular available tools to predict intra-protein interactions.

| Tools                        | Properties                              | Reference |
|------------------------------|-----------------------------------------|-----------|
| PIC web server               | Hydrogen bond, Ionic, hydrophobic, etc. | 88        |
| HBOND, HBAT, HBPLLOT         | Hydrogen bond                           | 89–91     |
| ESBRI                        | Salt bridges                            | 92        |
| VADAR                        | Hydrogen bond, accessible surface area  | 93        |
| DiANNA, Disulfind, Dinosolve | Disulphide bonds                        | 94–96     |
| CHpredict                    | Aromatic interactions                   | 97        |

### ***Importance of hydrogen bond as a major contributor of extreme-stability***

Hydrogen bonds are the most abundant type of non-covalent interactions, and 4–12 kJ mol<sup>-1</sup> is required to break a single bond<sup>18</sup>. Hydrogen bonding is the highest cited feature in literature for protein thermostabilization<sup>98</sup>. In thermostable proteins, hydrogen bonds shows an increase of 11.7 hydrogen bonds per chain per 10 °C rise in thermostability<sup>98</sup>. Hydrogen bonding partners in unfolded state of proteins are satisfied by hydrogen bonds to water. On protein folding, these protein-to-water H-bonds are broken, and only some are replaced by intra-protein H-bonds. For intra-protein hydrogen bonds, it is said that whenever two non-hydrogen atoms with opposite partial charges are found to be within a distance of 3.5 Å, a hydrogen bond can form<sup>26</sup>. Alternative sets of hydrogen bonds formation have also been implicated to enhance thermostability. It was reported that surface charged and polar side chains with high conformational mobility can form alternative hydrogen bonded donor-acceptor pairs<sup>99</sup>. The conclusion drawn was that residues located in the N- and C-terminal regions and in the extended loops that are capable of forming alternative longer range H-bonded pairs, leads to enhance the protein thermostability<sup>99</sup>.

Hydrogen bonds can be further divided into the following types depending on their donor and acceptor atoms as main chain-main chain, main chain-side chain and side chain-side chain hydrogen bonds. Main chain hydrogen bonds are crucial for proper positioning of ligands<sup>100</sup>. It has been shown that while only 1.3 % of backbone amino groups and 1.8 % of carbonyl groups in proteins fail to form H-bond (without any obviously compensating interactions), 80 % of main chain carbonyls fail to form a second hydrogen bond<sup>101</sup>. Backbone-backbone H-bonds are considered to have lower configurational entropy. The charge-transfer contribution to the hydrogen-bond energy increases and the angle decreases<sup>102</sup>. It has been said that lowering of configurational entropy stabilizes a protein. Comparatively, higher configurational entropy is assumed for two nearby residues that are not involved in backbone-backbone H-bonds<sup>103</sup>. Main chain to side chain hydrogen bonds are bonds involving side-chain acceptor/donor and main-chain donor/acceptor atoms. More than half the examples of such hydrogen bonds are found at the middle of alpha-helices rather than at their ends. They have not been observed to increase in

thermostable proteins. Besides, side chain to side chain hydrogen bonds are bonds involving side-chain acceptor/donor and a side-chain donor/acceptor atoms. They were observed to increase in thermostable monomeric proteins<sup>26</sup>. The structural features contributions of psychrophilic proteins have not been reported due to lack of 3D-structure. Alternatively, homology modeling may be employed to determine the structural features of proteins from psychrophilic microorganisms by comparative analysis with mesostable and thermostable proteins. However, the interpretation of hydrogen bonds in homology models is not much reliable as hydrogen bonds are highly directional. No hydrogen bond differences were identified from comparative analyses of the X-ray structures of  $\alpha$ -amylase, citrate synthase, and malate dehydrogenase<sup>104</sup>. The total difference in stabilization energy may be as little as 40–50 kJ mol<sup>-1</sup> between homologs from a psychrophile and mesophile, it is clear that this may be accounted for by a few critically placed hydrogen bonds. A decrease in the trend of number of hydrogen bonds at domain interfaces in psychrophiles as compared to mesophilic counterpart was reported earlier<sup>105–107</sup>. For example, cold adaptation of a triosephosphate isomerase was linked to an Ala replacement of a Ser; Ser is expected to confer thermostability by forming two additional intramolecular hydrogen bonds<sup>108</sup>.

The molecular basis of thermophilic, psychrophilic and halophilic adaptation is well-known but the barophilic, acidophilic and alkalophilic adaptation is still unclear. The three-dimensional structure of these proteins is mainly determined by the three types of biophysical interactions: hydrogen bond, electrostatic interactions and hydrophobic interactions. The investigation of any type of extremophilic protein using X-ray diffraction is turned to understanding of extreme-stability adaptations and its comparisons with mesostable protein in terms of structure and dynamics<sup>16,109</sup>. Considering available X-ray crystal structure and thermodynamic data of extremophilic proteins, a number of conclusions may be drawn regarding role of hydrogen bonds in stabilization of proteins such as: (a) increase in number of hydrogen bonding groups make suitable interactions in the folded structure, (b) it provides the organization for distinct folds, (c) it confers directionality and specificity to the other intramolecular interactions, and (d) it provides the selectivity in the inter-domain, protein–protein and protein–ligand interactions that



support molecular recognition<sup>110</sup>. Additionally, one concerning problem is that the very less number of extremophilic protein structures are available in database. Moreover, the homology modeling plays an important role that the 3D structure of a protein that can be estimated from homology modeling, which predicts structure based on experimental templates of proteins that are homologous in sequence to the protein of interest<sup>111–113</sup>. Most of the research has been carried for the target extremophilic protein structure prediction from the homologous mesophilic template protein<sup>72</sup>. The extremophilic proteins are frequent modeled through online available tools that partially fulfill the role of protein structure. For instance, the predicted homology-based structural model of thermostable 3-isopropylmalate dehydrogenase of *Thermus thermophilus* used template of *Escherichia coli* mesostable counterpart homology-based structural model of thermostable 3-isopropylmalate dehydrogenase with *Escherichia coli* mesostable counterpart. The generated model of *T. thermophilus* and the *E. coli* enzyme are very similar but only a very small difference was found in the normalized number of hydrogen bonds and hydrophobic interactions. The thermostable structure contains slightly more hydrogen bonds per atom than its mesostable counterpart which implicates the contribution of hydrogen bonds to the enhanced thermostability<sup>114</sup>.

The other types of hydrogen bonds that have been classified are charge-charge and charge-neutral hydrogen bonds. Charge-neutral hydrogen bonds are more stabilizing as desolvation energy making for an H-bond residue is lower than for an ion pair<sup>115</sup>. Moreover, binding energy of a charged-neutral H-bond is far larger than from neutral-neutral H-bonds, due to the charge-dipole interaction. A study of 16 protein families shows that thermostable proteins show a consistent increase in hydrogen bonds<sup>98</sup>. They can also be divided into: short strong hydrogen bonds: Distance 2 - 2.5 Å. They acquire covalent characteristics and are also known as low barrier hydrogen bonds. N–H...O, O–H...O, N–H...N hydrogen bonds are said to be higher in energy than other types of hydrogen bonds and biologically more important<sup>116,117</sup>. Recently, Srivastava et al. in 2014 showed that increase in hydrogen bond increases thermostability of *Bacillus subtilis* lipases as predicted through molecular dynamics simulations and network-based analysis<sup>118</sup>.



$\gamma$ -turns have also been reported to be a new factor responsible for thermostability of lipases<sup>119</sup>. A  $\gamma$ -turn consists of three consecutive residues at positions  $i$ ,  $i + 1$ ,  $i + 2$  and possess a short strong hydrogen bond between the CO group of  $(i)^{\text{th}}$  residue and NH group of  $(i + 2)^{\text{th}}$  residue<sup>120,121</sup>.  $\gamma$ -turns have been classified into classic and inverse based on the dihedral angle values of the  $(i + 2)^{\text{th}}$  residue<sup>120</sup>. The classic  $\gamma$ -turn gives rise to  $180^\circ$  chain-reversal in proteins and is often observed at loop end of  $\beta$ -hairpins<sup>122</sup>. The inverse  $\gamma$ -turns include a large proportion of weak hydrogen bonds according to the definition of hydrogen bonds<sup>123</sup>. Figure 1.2 shows the difference in classic and inverse  $\gamma$ -turns. It can be said here that tertiary structure which leads to an increase in hydrogen bonds may increase with the increment in protein thermal stability. GammaPred server is a neural network based server and can predict the  $\gamma$ -turn residues in the given protein sequence<sup>97</sup>. Promotif is the software which can predict  $\gamma$ -turns in protein structures<sup>124</sup>. Conclusively, increment in  $\gamma$ -turns can increase protein thermostability.

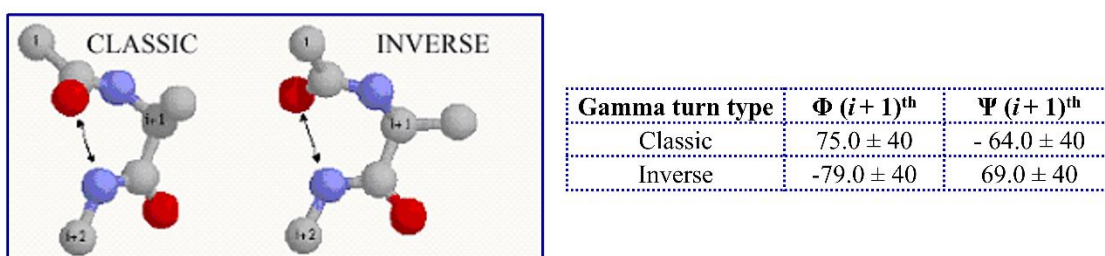


Figure 1.2: Types of  $\gamma$ -turns and the variation in their interval dihedral angle at  $(i + 1)^{\text{th}}$ .

### ***Electrostatic interactions: “key factor” behind protein stability***

Electrostatic interactions have long been implicated in the thermostability of thermophilic proteins<sup>66,68,69,98,125</sup>. Theoretical studies suggest that the stabilizing effect of electrostatic interactions increases with increasing temperature<sup>126–129</sup>. It was reported that thermostable proteins tend to have more salt bridges and surface charge residues<sup>130</sup>. Salt bridges are formed by spatially proximal pairs of oppositely charged residues in native protein structures. A salt bridge is constituted by a couple of oppositely charged groups, so in proteins it is recognized if at least one Asp or Glu side-chain carboxyl oxygen atom (i.e. OD in Asp or OE in Glu) and one side-chain nitrogen atom of Arg, Lys or His (i.e. NH in Arg, NZ in Lys or NE & ND in His) are within a distance of  $4.0 \text{ \AA}$ <sup>92</sup>. A single salt bridge

can contribute  $13 - 22 \text{ kJ mol}^{-1}$  to the free energy of folding<sup>131</sup> and, unlike hydrophobic interactions, they are relatively unaffected at extremely high temperatures<sup>132</sup>. Elcock proposed that salt-bridge should be more stabilizing at high temperatures because the unfavorable desolvation penalty and the entropic cost of fixing two charged side-chains would decrease with temperatures<sup>126</sup>. Furthermore, thermostability is achieved by upshifting or broadening the thermostability curve which is obtained by Differential Scanning Calorimeter (DSC). A smaller  $\Delta C_p$  can increase the maximum  $\Delta G_u$  where,  $\Delta G_u(T_s) \rightarrow \Delta H_m - \Delta C_p (T_m - T_s)$ , or in other words, the protein stability curve is up-shifted if  $\Delta H_m$  is increased or remains constant<sup>26,126</sup>. Salt bridges decrease  $\Delta C_p$  thus results in the upshift of the thermostability curve<sup>133</sup>. The effect of ionic interactions on thermostability has been studied by loss of function and gain of function mutations. Vetriani et al. reported that extensive ion-pair networks may provide a general strategy for manipulating enzyme thermostability of multisubunit enzymes<sup>134</sup>. They conclude this by studying structures of hexameric glutamate dehydrogenases (GluDHs) from the hyperthermophiles *Pyrococcus furiosus* and *Thermococcus litoralis*. Schmid and co-workers have implicated contributions of electrostatic interactions to the thermostability of thermophilic *Bacillus caldolyticus* cold shock protein.

Cation- $\pi$  interaction is another form of electrostatic interaction responsible for protein thermostability. They have been reported to be formed by the interactions between positively charged residues (Lys and Arg) and aromatic amino acids (Tyr, Trp and Phe). Gromiha et al. analyzed the influence of cation- $\pi$  interactions to enhance the stability from mesostable to thermostable proteins. Tyr has a greater number of such interactions with Lys in thermostable proteins. The influence of Phe in making cation- $\pi$  interactions is higher in mesophiles than in thermophiles<sup>135</sup>. Further, a network of cation- $\pi$  interactions is maintained by Lys in thermostable proteins, whereas Arg plays a major role in mesostable proteins. Moreover, atoms that have a substantial positive charge in both Lys and Arg make a more significant contribution for cation- $\pi$  interactions than do cationic group atoms<sup>136</sup>. The cation- $\pi$  interaction between Arg19 and Tyr93 in the protein indole-3-glycerol phosphate synthase from *Sulfolobus solfataricus* was reported to contribute towards stability<sup>137</sup>.

Several comparative studies based on X-ray structure data have reported about other type of extremophilic proteins. The psychrophilic proteins have flexibility in their native state, thus they require lower number of electrostatic interactions and salt bridges under cold-adapted conditions. For example, cold adapted subtilisin has two salt bridges compared with five and ten salt bridges in the homologs from mesophiles and thermophiles, respectively<sup>138</sup>. Similarly, Kim et al. in said that, a reduced number of intersubunit and ion-pair networks also appear to be important for the heat lability of a malate dehydrogenase from a psychrophile<sup>139</sup>.

For halophilic adaptation of protein electrostatic stabilization was suggested as the key determinant of their stability. However, contribution of specific electrostatic interactions (i.e. salt-bridges) to overall stability of halostable proteins is yet to be understood<sup>140</sup>. Therefore, the major focus of Physical Chemistry was in understanding the stability conferred by electrostatic interactions to render proteins stable in high salt concentration<sup>141</sup>. For example, Malate dehydrogenase from *Haloarcula marismortui* shows greater number of salt-bridges than its mesophilic counterpart which enhanced enzyme stability at high salt concentrations<sup>142</sup>. Again, the destabilization of halostable proteins at low-salt concentration was reported to be due to strong electrostatic repulsion<sup>143</sup>. Further, electrostatic interactions were suggested as a key factor of adaptation to extreme pH i.e. acidic and basic environment. Consequently, substitution of acidic by basic residues was used to improve the charge balance and stability at high pH, and vice versa<sup>144</sup>. In alkaline condition, it was also established that asparagines and glutamines deamidate and lead to destabilization of the protein structure<sup>145</sup>. It has been suggested to mutate these residues to improve stability at extreme alkaline conditions<sup>146</sup>.

### ***Implication of hydrophobic interactions on extreme-stability***

Hydrophobic interactions have been reported by many authors to play crucial role in protein folding. It is brought about by burial of solvated non-polar side chains. Each additional methyl group buried in the enzyme gives an increase in stability of 1.3 ( $\pm 0.5$ ) kcal mol<sup>-1</sup><sup>147</sup>. An enhanced hydrophobic effect is one of the reported reasons for the slow unfolding of thermostable proteins<sup>148</sup>. Rathi et al. studied a set of 130 pairs of

thermostable and mesostable proteins and reported hydrophobic interactions as “key factors” for protein thermostability<sup>149</sup>. Burg et al. increased thermostability of thermolysin-like neutral protease of *Bacillus stearothermophilus* by introducing Arg, Lys or bulky hydrophobic amino acids<sup>150</sup>. Through their experiments it was shown that surface hydrophobic contacts were the major determinants for protein thermostability. Unfortunately mutations attempting to fill cavities often were found to be not that stabilizing due to detrimental effects of unfavorable Van der Waals interactions and subsequent local rearrangements<sup>18</sup>. Furthermore it has been reported that hydrophobic interactions, which are entropic at room temperature but becomes enthalpic at higher temperature, reaches their maximum stabilizing effect at 75 °C<sup>151</sup>.

Core packing is often linked to increased hydrophobicity and stability<sup>152</sup>. An increase in hydrophobicity, given that it being buried will add to core stability due to increased Van der Waals interactions. Programs such as ROC, PROSE, PERLA and CORE can be used to redesign protein cores using molecular force fields. Many proteins were successfully redesigned using these methods. However sometimes over packing of the cores results in destabilization of the folded conformer<sup>153</sup>. Core packing also results in rigidity of proteins. Rigidity has also been related to enhance thermostability of proteins. Rigidity causes hyperthermostable enzymes to be often inactive at low temperatures<sup>18</sup>. Psychrophilic proteins enhance their flexibility by decreasing the number and strength of various interactions<sup>154,155</sup>. In particular, there is a significant decrease in the interaction between hydrophobic residues, and between hydrophobic and aromatic residues, as would be expected given the temperature dependence of the hydrophobic effect<sup>156</sup>. On the other hand, halostable proteins have very low bulky hydrophobic residues<sup>140,157</sup>. It has also been reported that weakening of the hydrophobic interactions in the protein core and conserved hydrophobic contacts results in protein stability in the presence of salts in preventing the proteins from aggregation at high salt concentrations<sup>157</sup>. In halostable proteins, lower content of bulky hydrophobic residues might indicate mandatory more polar protein interior than their mesostable counterparts<sup>140</sup>.

### *The potentiality of aromatic interactions in extremostability of proteins*

Hydrophobic interactions between aromatic groups of phenylalanine and tyrosine less than 7 Å distance results in  $\pi$ - $\pi$  stacking and enhances protein stability. Most of them are energetically favorable having potential energies between 0 and  $-2.0 \text{ kcal mol}^{-1}$ . These interactions have the polarity, partial negative on the face of the ring (caused by the  $\pi$ -electron) and partial positive charge on the C-H edges<sup>105</sup>. Since they have polarity, two types of aromatic interaction in proteins: aromatic-aromatic interactions (between aromatic rings at right angles to each other) and aromatic-amino interactions (between aromatic rings and the side chains of Arg and Lys). Aromatic interactions may therefore promote thermostabilization through an enthalpic contribution. Such interactions also link secondary protein structures leading to overall stability<sup>158</sup>. Surface exposed Tyr-Tyr and Phe-Phe pairs were observed to contribute  $-1.3 \text{ kcal mol}^{-1}$  toward thermostabilization in RNase from *Bacillus amyloliquefaciens*<sup>159</sup>. Although, few researchers were successful to engineer aromatic clusters that led to increase in thermostability of proteins<sup>158-160</sup>, such interactions are very hard to engineer because aromatic interactions are often found in networks and not as isolated units<sup>161</sup>.

Unlike thermostable proteins the psychrophilic proteins are flexible in nature, thus lack aromatic interactions in them<sup>138</sup>. For example, heat-labile subtilisin, from the Antarctic psychrophile *Bacillus TA41* has a general lack of aromatic interactions in contrast to 11 interactions identified on the surface of a thermophilic homolog); subtilisin Carlsberg from *Bacillus licheniformis*<sup>138</sup>. Similarly, in a  $\beta$ -lactamase of Antarctic psychrophile *Psychrobacter immobilis* A5 lack aromatic interactions when compared with homologous mesophilic counterpart<sup>105</sup>. Similar to psychrostable proteins, halostable proteins are reported to prefer less aromatic residues for their stabilization under high salt concentrations<sup>162</sup>. Longo and Blaber reported that “several lines of evidence indicate that aromatic amino acids were a late addition to the codon table and not part of the original “prebiotic” set comprising the earliest polypeptides.” For example, the aromatic substitutions in PV2 protein leads to the ability to move the folding properties from halophilic to mesophilic conditions<sup>162</sup>.

### ***Engineering Disulfide bonds for enhancing protein extreme-stability***

Covalent disulfide bonds between cysteine residues are an important tertiary structural feature that results in protein stability<sup>163,164</sup>. Thermostable proteins have been observed to possess such disulphide bridges. A disulphide bond leads to 2.5 - 3.5 kcal mol<sup>-1</sup> of stabilization. This depends on the distance of the bonds<sup>165</sup>. It has also been observed to lead to stability by reducing the entropy of the denatured state<sup>163</sup>. Disulfide bonds have been shown to play important role in oligomerization<sup>17</sup>. Oligomerization has been regarded as important determinant of thermostability<sup>166</sup>. They result in interlocking of monomeric chains conferring stability. Disulphide bonds have also been reported to reduce the entropy of denatured state. Effect on stability by insertion and deletion of such bonds were studied in *Cucurbita maxima* trypsin inhibitor-V stability<sup>68</sup>. It was concluded that disulphide bridging stabilizes both native and denatured state<sup>68</sup>. The difference in stabilization between the two states determines the state of protein stability<sup>68</sup>. Reduction of five disulphide bonds in *Aspergillus niger* phytase has been linked to its destability<sup>167</sup>. A sound example where insertion of disulphide bonds have been shown to increase stability was in Subtilisin E. Introduction of a disulphide bond resulted in 4.5 °C increase in melting temperature and a three-fold increase in its half life<sup>168</sup>. Irrespective of the aforesaid, not all thermostable proteins possess disulphide bridges. For example *Bacillus stearothermophilus* lipase (PDB Id: 1JI3) which is stable at temperatures greater than 80 °C lacks disulphide bonds. Further, disulfide mutants show decreased as well as increased stabilities<sup>169</sup>. This supports the fact that disulphide bonds are not the universal factor responsible for thermostabilizing proteins. They may play critical role in protein stability but are not signatures of thermostable proteins. Engineering disulphide bonds is also difficult as it is difficult to determining the position in proteins for introduction of a disulfide bond<sup>170</sup>. This is vouched for by the fact that stability enhancement by insertion of novel disulfide bonds have not always been successful<sup>68</sup>. Disulphide bonds have also been reported to have phylogenetic relationship being abundant in Crenarchaea and in the nonmethanogenic thermophilic Euryarchaea<sup>171</sup>. Disulphide bonds in psychrophilic *Shewanella violacea* cytochrome C have been reported to contribute to protein stability<sup>172</sup>. The effect was studied to be practically absent from the halophiles<sup>171</sup>. The



role of disulphide bonding in stabilizing other classes of extremophilic proteins is yet to be fully explored.

### 1.6 Approaches to attain extreme-stable proteins by protein engineering

During the past decade, many researchers have attempted to develop an optimized enzyme that possesses high activity and stability suitable for industrial processes. Enzyme stability is an important parameter in biocatalytic applications, and there is a strong need for efficient methods to generate robust enzymes. *In vitro* evolution strategies along with temperature as selection pressure are mostly employed to increase stability (Table 1.3). The alternative approach is identification of thermostabilising features in stable enzymes. This is utilized to engineer thermostable enzymes by site-directed mutagenesis techniques. A wide variety of such thermostable enzymes have been cloned and successfully expressed in mesophilic organisms. However one drawback is the randomness of the strategies is that such approaches result in compromise of stability and flexibility with the functional state of proteins<sup>173</sup>. One might also argue that finding critical regions or nucleation seeds of unfolding is more complicated than designing stabilizing mutations once these regions are found. Directed evolution strategies applied are still partially successful rational concepts. This approach is practically limited as it requires ample amount of time and is also labor intensive. Furthermore it involves application of selective pressure to achieve mutagenesis followed by colony screening for the desired trait<sup>174</sup>. Additionally repeated rounds of random mutagenesis are the key to success for this strategy. The lacuna is that it may lead to neutral and deleterious mutations<sup>174</sup>. Correspondingly, Reetz et al. proposed an innovative method for improving protein thermostability of mesostable enzymes without compromising the activity<sup>175</sup>. The method chooses the appropriate sites at which rigidity is to be increased in the wild type mesostable enzyme on the basis of atomic displacement parameters obtained from X-ray data, namely the B-Factors<sup>176,177</sup>. Their strategy constitutes a combination of rational design and directed evolution. The method requires the availability of the wild type protein structure and illustrated that iterative cycles of saturation mutagenesis were performed at high B-Factors sites of protein<sup>178</sup>. The B-factor based site-directed mutagenesis has been exploited to enhance the thermostability of *Serratia plymuthica*

sucrose isomerase<sup>179</sup>, *Candida antarctica* lipase B<sup>180</sup>, *Rhizomucor miehei* lipase<sup>181</sup>, *Coprinus cinereus* peroxidase<sup>182</sup> and *Bacillus subtilis* lipase<sup>175</sup>. Some of these recipes cease to exist which can render proteins thermostable through protein engineering approaches. Moreover the predictive power of these concepts is rather limited, the targeted mutations are random and require to be tested individually by site-directed mutagenesis<sup>183</sup>. Thus to overcome the demerits of such approaches there is a need to develop a guided method which can predict mutations as extremostabilizing so that such mutations can be carried out *in vitro* through site directed mutagenesis.

### 1.7 Available theoretical prediction models for predicting stabilizing mutations

To overcome the demerits of directed evolution approaches numerous *in silico* algorithms have been proposed which can predict whether conceptualized mutations will be extreme-stabilizing. These models have been developed by investigating protein features by comparing extremostable proteins with non-extremostable proteins at different hierarchies of protein organization: from the nucleotide codons in their genes, their amino acid preferences in their protein sequence to their tertiary structures. The algorithms available to date with the capability of distinguishing thermostabilizing mutants are mostly knowledge based<sup>184</sup>. Few are support vector machine (SVM) based<sup>185</sup> and others are based on molecular dynamics<sup>186</sup>. For example, ROSETTA is a powerful modeling package with modules for protein structure prediction, modeling, docking and design. At present, methods are available to predict stability changes only when the atomic structure of the protein is available<sup>185</sup>. Thus, the methods to predict extreme-stability have focused on the three dimensional structures of proteins. Table 1.5 presents the software and tools required to predict and design mutations that can stabilize proteins.

Table 1.5: Popular software that predict stability changes after mutations.

| Tools       | Salient Features                                                                                             | References |
|-------------|--------------------------------------------------------------------------------------------------------------|------------|
| ROSETTA     | A dynamic and evolving macromolecular modeling suite addressing biomolecular structure prediction and design | 184        |
| I-Mutant    | Support Vector machine based, both sequence and structure can be used, single mutation                       | 185        |
| Cupsat      | Sequence as input, single amino acid mutations                                                               | 73         |
| MUPRO       | Support Vector machine based, sequence as input, Single mutation                                             | 187        |
| ERIS        | Structure as input, multiple mutations                                                                       | 188        |
| iPTREE-STAB | Machine learning based, single mutation                                                                      | 189        |

|           |                                                                            |     |
|-----------|----------------------------------------------------------------------------|-----|
| PoPMuSiC  | Single mutation                                                            | 190 |
| WET-STAB  | Machine learning based, multiple mutation                                  | 191 |
| MUSTAB    | Support Vector machine based, sequence as input, multiple mutations        | 192 |
| AUTO-MUTE | Machine learning based, structure as input, single mutation                | 193 |
| SDM       | Sequence/structure as input, single mutation                               | 194 |
| iSTABLE   | Support vector machine based, structure/sequence as input, single mutation | 195 |
| NeEMO     | Machine learning based, structure as input,                                | 196 |
| ENCoM     | Neural Network based, single mutation                                      | 197 |
| iRDP      | Ensemble of servers                                                        | 116 |

Although a lot of work has been done for identifying stabilizing mutations, protein engineering methods utilized to achieve them are still random and success rate is probabilistic. It can be said here that the accurate prediction of the thermodynamic consequences caused by mutations through *in silico* algorithms remains challenging<sup>198</sup>. Khan and Vihinen recently evaluated and compared 11 online stability predictors and found that the predictions were only moderately accurate<sup>199</sup>. Limitations are that majority of them require complex computational power and proficiencies. Another drawback is that they are based on calculations of features from protein sequences and can consider only single point mutations at a time and also requires several empirical parameters or heuristics such as patterning of residues for their calculations. Moreover statistical analysis based on  $T_m$  values (the midpoint of the thermal transition), suffers the fact that it is available only for a few proteins in a high resolution protein structural dataset. This limits the ability to examine correlations in a significant way<sup>26</sup>. Molecular dynamic simulations of mutation are several orders of magnitude complicated than that with a knowledge-based scoring function<sup>198</sup>. The other concern is that, only few algorithms can predict the effect of multiple mutations. Multi-site mutations are expected to have more complex effect on protein thermostability than from single point mutations<sup>200</sup>. For example, a predictive model WET was developed. It is a weighted decision table method for predicting protein thermostability change upon double mutation from amino acid sequences (Huang and Gromiha 2009). However, the accuracy drops to 0.57 when it is tested on the hypothetical reverse mutations<sup>201</sup>. The other model PROTS-RF is based on Random Forest algorithm and achieves an accuracy of 78.7% for multiple mutations<sup>201</sup>. The accuracy achieved until date creates limitation when greater than two mutations are to be performed. Molecular dynamics simulation (MD simulation) has been a recent and

popular methodology to understand the rationale behind protein thermostability. To design thermostable proteins by approaches of MD simulation one can look at the root mean square fluctuation, root mean square deviation, Radius of gyration graphs of the wild type protein at high temperatures. Such graphs represent the residue flexibility, backbone rigidity and compactness. MD simulation can be carried out using packages like GROMACS, NAMD and DESMOND. Mutating highly flexible residues with ones that are more rigid can enhance their thermal stability. Manjunath and Sekar carried out 50 ns MD simulation of SAICAR synthetase from mesophilic and hyperthermophilic sources<sup>202</sup>. They concluded that the thermostable proteins were more rigid. Long distance interactions are lost in mesostable proteins in contrast to that observed in the thermostable counterparts. Paul et al. performed 10 ns simulations each at 300 and 350 K, and 20 ns each at 400 and 450 K for chemotaxis protein from *Thermotoga maritima* and its mesophilic counterpart *Salmonella enteric*. They observed the mesostable protein to have greater flexibility at higher temperatures<sup>49</sup>. In another work, 16 thermostable mutants of *Bacillus subtilis* lipase A were studied and a direct correlation was derived for structural rigidity and thermostability<sup>149</sup>. Irrespective of the success of the method, the only drawback is that it is case sensitive and specific. It can guide in mutating a particular protein to enhance its temperature stability but do not lay foundations for a universal approach.

Additionally the cumulative effect of all the mutations on the physicochemical features or structural changes associated with the mutations cannot be as such predicted using the aforementioned algorithms. Also another lacuna is that all these methods give multiple choices of possible stabilizing mutations and do not conclude whether they will actually lead to thermostability. Moreover in doing so, they also fail to select as to which point mutation (single, multiple) or which combination of mutations will actually lead to thermostability of proteins. In short they are unable to rank or prioritize the plausible mutations based on their effect on stability on proteins. Therefore a new method is needed that can prioritize features according to their importance in rendering proteins thermostable at a desired temperature. This will give rise to a guided approach to extreme stabilize proteins.

### 1.8 Salient features deduced from engineered extreme-stable mutants to better understand extreme-stability of proteins

Mutants have been generated either to understand extreme-stability of proteins or to obtain proteins for industrial applications. As it is important to understand the effect of mutation on protein function, only those mutants that have their structures deposited in RCSB PDB have been presented in Table 1.5. The extremophilic proteins are better tolerated to destabilizing mutations have a higher chance to survive the selection pressure<sup>203</sup>. As a result, extra stability provides a strong advantage in evolution and promotes evolving ability<sup>44</sup>. Thus, investigating the mechanisms of protein stability appears to be important not only for studying enzyme evolution and understanding structure-function relationship, but also for the engineering of novel enzymes<sup>71</sup>. For industrial needs, enzyme evolution were extensively studied by mutating the significant attribute to understand the tradeoff between activity-stability and to engineer more tolerant variants for adapt to extreme environmental conditions<sup>204</sup>. These studies focused mainly on the thermal stability of the enzymes as most of the industrial processes involve reactions at elevated temperatures for improved productivity.

To better understand the mechanism behind extreme-stability, both gain of function and loss of function mutations have been presented in Table 1.5. It can be observed that the mutations have been attained by random mutagenesis experiments. Elucidation of the crystal structures of the mutants revealed that most of the stabilizing mutations are located on the loops. The loops have been previously reported to be the mutational hot spots for engineering stabilizing mutations. The effect of mutations on the biophysical properties also can be comprehended to be cumulative as well as individualistic differing from one protein to another. Further analysis of the mutated structures revealed that multiple point mutations scattered all over the protein structural space have profound effect than single point mutations on extreme-stability. These observations lays the foundation to further explore the biophysical feature space of extreme-stable proteins in order to deduce a global prediction model for predicting the effect of mutations on protein stability.



Table 1.5: Engineered mutants of extreme-stable proteins.

| Mutated protein            | Organism                              | No. of Mutants in PDB | Method                                        | Property of mutant                                                                                       | Region of mutation  | Reason                                                                         |
|----------------------------|---------------------------------------|-----------------------|-----------------------------------------------|----------------------------------------------------------------------------------------------------------|---------------------|--------------------------------------------------------------------------------|
| <b>Thermostabilizing</b>   |                                       |                       |                                               |                                                                                                          |                     |                                                                                |
| Lysozyme                   | Bacteriophage T4                      | 104                   | Directed evolution, site-directed mutagenesis | Enhanced thermostability                                                                                 | Loop, helix, strand | Enhanced core repacking, aromatic interactions                                 |
| <b>Psychro stabilizing</b> |                                       |                       |                                               |                                                                                                          |                     |                                                                                |
| Superoxide dismutase       | <i>Pseudoalteromonas haloplanktis</i> | 3                     | Site-directed mutagenesis                     | T <sub>m</sub> increased from 48 to 52 °C and an enhanced stability against GdnHCl                       | Loop                | Increased electrostatic interactions                                           |
| <b>Halostabilizing</b>     |                                       |                       |                                               |                                                                                                          |                     |                                                                                |
| NAD-dependent DNA ligase   | <i>Finegoldia magna</i>               | 1                     | Site-directed mutagenesis                     | Improves the halophilic adaptation                                                                       | Helix               | Salt-dependent stabilizing effect in the mesophilic proteins                   |
| <b>Acidostabilizing</b>    |                                       |                       |                                               |                                                                                                          |                     |                                                                                |
| Carboxyl-esterase          | <i>Sulfolobus tokodaii</i>            | 3                     | Site-directed mutagenesis                     | At pH 3 and 4, the T <sub>m</sub> values is reduced due to disruption of the H-bond or ionic interaction | Loop                | Disruption of Hydrogen bond or ionic interaction                               |
| <b>Alkalistabilizing</b>   |                                       |                       |                                               |                                                                                                          |                     |                                                                                |
| Glycosidase                | <i>Trichoderma reesei</i>             | 1                     | PCR overlap extension method                  | pH optimum is more alkaline in the mutant                                                                | Loop                | Histidine residue are establishing the correct electrostatic micro-environment |
| <b>Barostabilizing</b>     |                                       |                       |                                               |                                                                                                          |                     |                                                                                |
| IgG binding-Protein G      | <i>Streptococcus</i> sp.              | 2                     | Site-directed mutagenesis                     | Hydrogen bonds shorten by $0.022 \pm 0.108$ Å over 2 kbar (i.e., $\sim 0.01$ Å/kbar)                     | Loop                | High pressure causes distortion in the secondary structure                     |

### 1.9 Stabilizing proteins by non-protein engineering methods

Other than the intrinsic properties of a protein there exist other extrinsic methods for attaining extreme-stability. The engineering of glycosylation is one of the most common posttranslational modifications that stabilize proteins. Its effect on the thermodynamics and kinetics of proteins is poorly understood. The protein stability increases as the degree of glycosylation increases and, to a much smaller extent, with the size of the polysaccharides<sup>205</sup>. The stabilization effect depends on the position of the glycans; thus,



the same degree of glycosylation may result in a different thermal effect depending on the location of the sugars. Importantly, thermodynamic stabilization is accompanied by kinetic stabilization, with the unfolding barrier increased by  $\approx 20\%$  for the highly glycosylated protein variants. It results in a higher enthalpy<sup>206</sup>. Olsen and Thomsen observed the effect of glycosylation on thermostability of two *Bacillus*  $\beta$ -glucanases that were expressed in *E. coli* and in *S. cerevisiae*. The one that expressed in *S. cerevisiae* was strongly kinetically stabilized by glycosylation at 70 °C, and its optimum temperature for activity was higher<sup>207</sup>. These examples suggest that glycosylation could represent an alternative method for protein thermostabilization. Similarly, methylation of lysine is also protein stabilization step in posttranslational modification. Edmondson and co-workers reported the monomethylated Lys 5 and Lys7 of native small DNA binding protein Sac7d *Sulfolobus acidocaldarius* cause stabilization up to 100 °C<sup>208</sup>.

Inorganic salts are also stabilizing proteins in two ways. First is through a specific effect, where a metal ion interacts with the protein in a conformational manner, and second is through a general salt effect, which mainly affects the water activity. Correspondingly, the general effects of inorganic salt include chaotropic and kosmotropic effects. The 'kosmotrope' (order-maker) solutes stabilize proteins by reducing the solubility of hydrophobic aggregates whereas 'chaotrope' (disorder-maker) solutes destabilize the proteins by disrupting hydrophobic aggregates and increase their solubility leading to unfolding of proteins<sup>209</sup>. Protein-stabilizing solutes (kosmotropes) increase the extent of hydrogen bonding. Example includes trimethylamine N-oxide, proline, ectoine,  $\alpha,\alpha$ -trehalose, glycine betaine and 3-dimethylsulfoniopropionate. Ectoine and its derivative hydroxyectoine are well-known kosmotropic solutes that are widely employed by many of the extremophiles as protein stabilizers<sup>210,211</sup> and osmoprotectants<sup>212,213</sup>. Interestingly, it is reported in some of the extremophile that the ectoine and hydroxyectoine biosynthesis is enhanced by up-regulation of ectoine synthesizing enzymes by extremes in growth temperature, pH and salts environment<sup>214–217</sup>. Interestingly, Thauer and coworkers reported that five *Methanopyrus kandleri* methanogenic enzymes have become thermostable and their activity increases by higher concentrations of inorganic salts<sup>218</sup>.

The physical factor like pressure is also involved in the stabilization of proteins. Like, hyperthermophiles survive under high-temperature environments and also survive under high-pressure environments. Because of these environments, hyperthermophiles cannot avoid pressure and temperature (such as *Thermococcus barophilus*). All their macromolecular cell components have to be adapted to high pressures<sup>219</sup>. In such microorganisms, the enzymes are stabilized and activated by high pressures (e.g., *M. jannaschii* protease and hydrogenase). The reason behind their stabilization by pressure says that the pressure favors the structure with the smallest volume and high percentage of hydrophobic interactions<sup>219</sup>. Since pressure played an important role in stabilization of enzymes, it has great potential benefits for activation in biocatalysis. Michels and Clark found that increase in pressure to 500 atm resulted in 3.4-fold and 2.7-fold increase in activity and thermostability, of *M. jannaschii* protease and hydrogenase, respectively<sup>219</sup>.

In addition, enzyme activity and thermostability increased with pressure: raising the pressure to 500 atm increased the reaction rate at 125 °C 3.4-fold and the thermostability 2.7-fold. Spin labeling of the active-site serine revealed that the active-site geometry of the *M. jannaschii* protease is not grossly different from that of several mesostable proteases; however, the active-site structure may be relatively rigid at moderate temperatures. The barostable and thermostable behavior of the enzyme is consistent with the barophilic growth of *M. jannaschii* observed previously<sup>219</sup>.

#### **1.10 Vivid role of chaperones and extremolytes in stabilizing proteins in extreme milieu**

Even though the proteins of thermophilic archaea are adapted and tolerant to high temperatures, additionally they are protected by a set of specific proteins inside the thermophiles only. These proteins are called molecular chaperones and their role is to prevent high temperature or heat shock conditions the native protein structure from denaturation under. These molecular chaperones are ubiquitous, constitutive proteins which couple ATP hydrolysis with the ability to assist correct protein folding and to rescue denaturated proteins<sup>220</sup>. The response to thermal stress tolerance in the extremophilic archaea involves chaperones, chaperonins and other heat shock (stress)

proteins. The thermoresistant features of these molecules as yet incompletely understood<sup>221</sup>. In the last two decades, archaeal thermal stress genes and proteins belonging to the families Hsp70, Hsp60 (chaperonins), Hsp40 (DnaJ), and small heat-shock proteins (sHsp) have been discovered<sup>221</sup>. The molecular chaperone of the Hsp60 and Hsp70 family are the protein hallmark of the heat-shock response in the extreme thermophilic microorganisms of Archaea domain<sup>222</sup>. Till now six kinds of chaperones have been identified. They are group II chaperonins, prefoldin, small heat-shock proteins, peptidyl-prolyl cis-trans isomerases, AAA proteins, and NAC protein<sup>223</sup>.

The molecular chaperone systems of thermophilic eubacteria *Thermus thermophilus* HB8<sup>224</sup> and thermophilic archaea *Thermoplasma acidophilum*<sup>225</sup> were deliberated as first discovered, characterized and best-studied<sup>223</sup>. These molecular chaperone proteins are also involved in regulation of protein activities, refolding/degradation, prevention of protein aggregation under stress conditions. The mechanism of action of DnaK chaperone system of mesophilic eubacteria *Escherichia coli* and thermophilic eubacteria *Thermus thermophilus* were reported as similar. Both the systems comprises of high homologies proteins like DnaK (Hsp70), DnaJ (Hsp40) and GrpE. The DnaK chaperone system is found in all the three phylogenetic domains: eubacteria, eukarya, and some archaea<sup>226</sup>. As all chaperone systems are ATP-dependent, DnaK is also an ATP-dependent chaperone protein and usually present as ATP-bound state. Its ATP-bound state has low affinity with fast exchange rates for substrates whereas ADP-bound undergoes an internal conformational change that results in comparatively high affinity with slow exchange rates for substrates<sup>227</sup>. The conformational change is coupled to ATP hydrolysis and ADP/ATP exchange, which is due to the action of closely related co-chaperones DnaJ (Hsp40) and nucleotide exchange factor (GrpE)<sup>227</sup>. The whole mechanism of action involves DnaK ATPase activity which is facilitated by DnaJ in synergy with substrate binding and further ADP dissociation from DnaK is accelerated by a nucleotide exchange factor GrpE<sup>226,228</sup>. This resulted in efficient ATP rebinding to DnaK and substrate release from DnaK<sup>229</sup>. The two similar kind of DnaK chaperone systems of *E. coli* and *T. thermophilus* HB8 show similar biochemical and biophysical properties and also show significant differences in thermal stability and temperature-dependent conformational

changes, which are tightly correlated with the optimal growth temperature of each organism<sup>230</sup>. It is interesting to report that degree of chaperone induced folding of thermostable protein is lower than mesostable counterparts, for example DnaKJ refolding of the *P. luminescens* thermostable luciferase were marked degree lower as compared with the *V. fischeri* mesostable luciferase<sup>231</sup>. Thus, there may be a more general need for chaperones in thermophilic organisms, as protein unfolding or misfolding may be more prone at higher temperatures. No doubt, the extremophilic archaea are intriguing organisms that offer an opportunity to find novel molecules and mechanisms that will, most likely, enhance our understanding of the stress response and the protein folding and refolding processes in the three phylogenetic domains<sup>221</sup>. Thus, it is probable that by studying archaea one will unveil molecular mechanisms also present in bacteria and/or eukaryotes but modified by evolution, and not as easily detectable in the latter two groups of organisms.

Like molecular chaperones, cold-shock proteins assist in protein renaturation during the growth of psychrophiles<sup>232</sup>. They also increase translation efficiency by destabilizing secondary structures in mRNA<sup>233</sup>. The presence of unique antifreeze proteins such as the proteins from *Marinomonas primoryensis* and *Pseudomonas putida* GR12-2 lowers the freezing point of cellular water<sup>81</sup>. On the proteome level, thermophiles and psychrophiles differ markedly from neutrophiles by synthesizing chaperones and antifreeze proteins, respectively, which leads to refolding of denatured proteins at such extremes. For acidophiles, chaperones are also involved in protein refolding. They are expressed strongly in acidophiles<sup>46</sup>.

Many questions remain unanswered about the investigation and elucidation of the mechanism of action of the molecular chaperone complex at different temperatures for the analysis of protein folding and refolding in thermophilic archaea, and determination of the molecular mechanisms involved in stress regulation in other extremophilic archaeal species that thrive under widely different conditions (temperature, pH, osmolarity, and barometric pressure)<sup>221</sup>. Other than chaperones extremolytes, which are biological molecules present naturally in the cells of extremophilic microbes aid in extreme-stability of proteins. They can be either accumulated by uptake from the milieu

in which such microbes are present or synthesized by the microbes<sup>234</sup>. Extremolytes can be amino acids, sugars, phosphorus-containing compounds, glyceric acid derivatives, and polyols<sup>235</sup>. For example hydroxyacetone, ectoine and mannosylglycerate (pyrimidine derivatives) are widespread extremolyte among thermophilic organisms<sup>217</sup>.

### 1.11 Conclusion

This chapter is an extensive study of 6495 articles belonging to extremophilic proteins providing a wealth of information which challenges current ideas regarding engineering proteins for extreme-stability. Each protein shows different approaches to extreme-stability governed by minute differences in their biophysical properties and the milieu in which they exist. Further, it was noted that proteins show mixotrophic stability at various extreme conditions such as temperature, pH, salt concentration and pressure. Any general rules could not be drawn regarding adaptive mechanisms in extreme environments. Among all the extremophilic adaptations, thermophilic adaptation was understood to be the most explored and reported to be most versatile among all. After analyzing literature on extremophilic genomes, it was understood that though large repertoire of extremophilic organisms has been identified, only few genomes have been successfully sequenced. The least sequenced among all being barostable genomes, as it results in only 2 hits in NCBI genome search. Comparative analysis of the genomes of extremophiles and non-extremophiles revealed some common trends that go above the boundary of phylogenetic relationships and genomics such as genome size, DNA repair system, DNA stabilizers, genomic GC-content and genome super-coiling. On the level of proteome, and restricting ourselves to the role of canonical amino acids, it was noted that the differences in the preferences in amino acid composition in each group of extremophile share subtle differences often showing superimposing behavior. It was clear that derivation of global rules based on the properties of amino acids has become puzzling for the stabilization of extremophilic proteins. Elaborate discussion about all the reported protein biophysical attributes like hydrogen bonding, disulphide bridges and electrostatic interactions led to the conclusion that the global balance of these attributes renders proteins extreme-stable in different conditions. The range of balance is specific and shows remarkable, yet unidentifiable equations among different classes of extreme-stable proteins. Further

mechanisms of stability of intracellular and extracellular proteins were reviewed to differ. Other than the intrinsic properties that aid in extreme-stability, it was brought out that molecular chaperones and extremolytes also aid in attaining extreme-stability. Further, it was divulged that extreme-stability is the result of customization of multifactorial level of molecular adaptations. This knowledge led to the development of methods like directed evolution to engineer stable proteins. The method follows either ancestral or structure guided approaches, resulting in random mutagenesis protocols followed by selection of desired trait after application of selection pressure. Further assistance to these methods are provided by the available 14 software and web tools which can calculate effect of single point mutations at a time by enumerating the effect of mutation basically on the free energy of stabilization. The lacuna that was noted is that all these methods give multiple choices of possible stabilizing mutations and do not conclude whether they will actually lead to extreme-stability additionally creating confusion by failing to predict the class of extreme-stability. Further, to better understand the mechanism behind extreme-stability, both gain of function and loss of function engineered mutants were reviewed. Analysis of the mutated structures revealed that multiple point mutations scattered all over the protein structural space have profound effect than single point mutations on extreme-stability. Other than the intrinsic properties of a protein it was reviewed that there exist extrinsic methods for attaining extreme-stability. The engineering of glycosylation is one of the most common posttranslational modifications that stabilize proteins.

Conclusively it can be said that, the review raises crucial questions about extreme-stability of proteins. The most intriguing being whether a universal protocol of stabilization is followed by extreme-stable proteins belonging to all three kingdoms of life and whether the stability parameters are enzyme class specific? If such a rule does not exist is it an implication that natural mutations are non-random and happen in particular regions in the protein structure like the loops, purposefully to confer stability at different extreme conditions? If this be the case then it may be possible to deduce such non-random stabilizing mutations that will aid in faster recombinant production of the industrially important extremophilic proteins. Another compelling question is why and



how certain proteins are extreme-stable, even in the absence of a clear evolutionary drive. Other questions that need to be further researched upon are development of an algorithm that can rationalize the approach of attaining mutations for extreme-stabilizing proteins. Further experiments are required to be performed to compare a better and develop an easier method to extreme-stabilize proteins, i.e., whether by intrinsic property modifications through mutations or extrinsic engineering by incorporation of glycosylation or addition of extremolytes. We hope that this work will be instrumental and interesting to researchers who are dedicated towards developing extreme-stable proteins through protein engineering approaches.

### 1.12 Origin of the research work

The origin of this research work illustrates that random protein engineering approaches are still utilized for extremophilic proteins from mesophilic sources with limited success. Thus there is a need to develop a rationalized prepredicted protocol which will employ prepredicted site-directed mutagenesis to attain desired mutations with process economy and less time. It is evident from the aforementioned that though protein extremophilic stability has been extensively studied, however, no universal rules to render proteins extremophilic-stable exist. The questions that need to be answered are:

- What are the statistically significant factors for extreme-stabilizing proteins?
- Can such factors be prioritized?
- Which of the factors can be exploited for attaining prepredicted mutations – a rationalized platform for protein engineering?
- Can the developed platform take care of multiple mutations?
- Can a holistic explanation be given for the developed rationalized platform?

### 1.13 Thesis organization

The present thesis entitled “Exploring Molecular adaptations of extremophilic proteins: A platform for protein engineering” is an effort in developing consummate understanding of protein extreme-stability. It has been divided into eight chapters.

***Chapter 1: Introduction and literature review***

In this chapter a thorough literature review that provides an insight into research pertaining to the multifactorial level of extreme-stability of proteins has been comprehensively reported. Protein stability can be affected at three hierarchies of protein organization – DNA level, RNA level and protein level (including primary sequence features, various intra-protein interactions, secondary and tertiary structural features). This chapter put forth the question whether these factors can be exploited for protein engineering to fill the caveat of attaining pre-determined mutations. The chapter lays foundation for the entire work which has been reported in this thesis.

***Chapter 2: Extreme-stable Protein Database (ExProtDB)***

Extremophilic adaptations though a burgeoning topic, a curated database for extremophiles and their proteins is still lacking. Thus, this chapter is a compilation of all the collected genomic and proteomic data of extremophiles into a manually curated database named as “Extreme-stable Protein Database (ExPDb)”. Architecture of the database has been built using a standard platform based on the Linux-Apache-MySQL-PHP (LAMP). It is a knowledgebase about all the information available in the World Wide Web about extremophiles and their extreme-stable proteins. The novelty of the database lies in the fact that it gives consolidated and comprehensive information of all types of extremophilic proteins and organisms.

***Chapter 3: Understanding the specific codon usage patterns in extremophiles***

The collected genomic and proteomic data were then utilized for comparative analyses of extremophiles and non-extremophiles in the proceeding chapters. This chapter throws light on primary level i.e. genomic adaptations of extremophilic proteins. It deciphers the rationale behind the specific codon usage patterns in extremophiles. Codon usage patterns exhibited harmony among different extremophiles and has been detailed. However, the codon attribute preferences and their selectivity of extremophiles varied in comparison to non-extremophiles. Although codon study is being partially deciphered in thermophiles, this is probably the first report of codon preferences pertaining to all extremophile classes.

#### ***Chapter 4: Analysis of attributes contributing to extreme-stability of proteins***

After the codon contribution further work in chapter 4 encompassed on prediction and ranking models for analyzing protein sequence and structure attributes contributing towards extreme-stability through machine learning and multi-criteria decision making (also known as Analytic Hierarchy Process, AHP) approaches. The positively contributing attributes have been further exploited to design mutations in a mesophilic (non-extremophilic) protein in such a way as to increase the content of these attributes to endorse the prediction of generated models. This chapter reports additional contributing factors for extreme-protein stabilization, which were not reported earlier.

#### ***Chapter 5: In silico prediction of mutations to attain extreme-stability of proteins and their validation***

This chapter gives a pertinent account of *in silico* prediction of plausible mutations and their validation in *Bacillus subtilis* lipase (a chosen mesophilic enzyme) by substituting the preferred protein attributes to attain extreme-stability of proteins by designing mutants. *In silico* validation of mutant was done by AHP generated ranking models, various mutation prediction servers (HotSpot Wizard, I-Mutant2, Cupsat, iPTREE-STAB, WET-STAB and ERIS webserver), Substrate Docking (Autodock 4.2 version), Ramachandran analysis (PROCHECK) and Contact Map analysis (CMView 1.1.1). Interestingly, results showed that the designed mutations did not affect the catalytic activity of the lipase and thus, mutation did not disturb the catalytic properties.

#### ***Chapter 6: In vitro characterization and validation of predicted mutations***

This chapter deals with experimentation employed for *in vitro* cloning, mutagenesis, expression, purification and characterization of mutated lipases. Results revealed that the optimum activity and stability increased from 35 °C to 55 °C in bsl\_the3, pH 8 to pH 10 in bsl\_alk2 but reduced to pH 6.5 in bsl\_acd2 w.r.t. wild type lipase (bsl\_wt). Further, the bsl\_wt was halo-sensitive (0.5M NaCl) but in bsl\_hal2 halostability was increased to 1.5M NaCl. The content of secondary structure enhanced extreme-stability (with increased  $T_m$  value) which was confirmed through FT-IR (ATR mode) and Differential Scanning Fluorimetry (Thermal shift assay).

### **Chapter 7: Employing predicted extreme-stable mutants for applications**

This chapter employs generated extreme-stable mutants for applications. Detergent formulations was the chosen application here. To further increase the enzyme stability, bsl\_wt and engineered lipases (bsl\_the3, bsl\_psy2, bsl\_acd2, bsl\_alk2 and bsl\_hal2 have better extreme-stability) were immobilized on synthesized ZnO nanoparticles as it is cheap, biocompatible and environmental friendly. The order of performance in oil and grease stain removal from the stained cotton fabric pieces by immobilized lipases on ZnO nanoparticle- bsl\_alk2 > bsl\_the3 > bsl\_hal2 > bsl\_psy2 > bsl\_acd2 > bsl\_wt. Results also showed that the maximum removal of oil and grease stains were up to 90 % and 82 %, respectively.

### **Chapter 8: Conclusion and future perspectives**

It summarizes the findings of the present investigations and states the future scopes and upcoming challenges. It can be concluded that thesis work has been successful to develop a single design strategy to mutate amino acid residues which can render proteins stable at different extremophilic range. Conclusively, this work reveals a promising perspective of using low-cost nanoparticle conjugated biocatalysts which was produced through rationalized protein engineering approach and was instrumented for detergent formulation.

#### **1.14 References**

1. Di Giulio, M. The Universal Ancestor and the Ancestor of Bacteria Were Hyperthermophiles. *J. Mol. Evol.* **57**, 721–730 (2003).
2. Delong, E. F. Everything in moderation: Archaea as ‘non-extremophiles’. *Current Opinion in Genetics and Development* (1998). doi:10.1016/S0959-437X(98)80032-4
3. Riessen, S. & Antranikian, G. Isolation of Thermoanaerobacter keratinophilus sp. nov., a novel thermophilic, anaerobic bacterium with keratinolytic activity. *Extremophiles* **5**, 399–408 (2001).
4. Zillig, W., Holz, I., Janekovic, D., Schäfer, W. & Reiter, W. D. The Archaeobacterium Thermococcus celer Represents, a Novel Genus within the Thermophilic Branch of the Archaeobacteria. *Syst. Appl. Microbiol.* **4**, 88–94 (1983).
5. Carpenter, E. J., Lin, S. & Capone, D. G. Bacterial Activity in South Pole Snow. *Appl. Environ. Microbiol.* **66**, 4514–4517 (2000).
6. Blochl, E. et al. Pyrolobus fumarii, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113°C. *Extremophiles* **1**, 14–21 (1997).
7. Oren, A. Life at High Salt Concentrations. in *The Prokaryotes* 421–440 (Springer Berlin Heidelberg, 2013). doi:10.1007/978-3-642-30123-0\_57
8. Jones, B. E., Grant, W. D., Duckworth, A. W. & Owenson, G. G. Microbial diversity of soda lakes. *Extremophiles* **2**, 191–200 (1998).

9. Baker-Austin, C. *et al.* Extreme arsenic resistance by the acidophilic archaeon 'Ferroplasma acidarmanus' Fer1. *Extremophiles* **11**, 425–434 (2007).
10. Horikoshi, K. Barophiles: deep-sea microorganisms adapted to an extreme environment. *Curr. Opin. Microbiol.* **1**, 291–295 (1998).
11. Leong, S. L., Pettersson, O. V., Rice, T., Hocking, A. D. & Schnürer, J. The extreme xerophilic mould *Xeromyces bisporus* — Growth and competition at various water activities. *Int. J. Food Microbiol.* **145**, 57–63 (2011).
12. Battista, J. R. AGAINST ALL ODDS: The Survival Strategies of *Deinococcus radiodurans*. *Annu. Rev. Microbiol.* **51**, 203–224 (1997).
13. Erb, T. J., Kiefer, P., Hattendorf, B., Gunther, D. & Vorholt, J. A. GFAJ-1 Is an Arsenate-Resistant, Phosphate-Dependent Organism. *Science* (80-. ). **337**, 467–470 (2012).
14. Rothschild, L. J. & Mancinelli, R. L. Life in extreme environments. *Nature* **409**, 1092–1101 (2001).
15. Yu, H. & Huang, H. Engineering proteins for thermostability through rigidifying flexible sites. *Biotechnology Advances* (2014). doi:10.1016/j.biotechadv.2013.10.012
16. Jaenicke, R. Protein folding: local structures, domains, subunits, and assemblies. *Biochemistry* **30**, 3147–61 (1991).
17. Reed, C. J., Lewis, H., Trejo, E., Winston, V. & Evilia, C. Protein adaptations in archaeal extremophiles. *Archaea* **2013**, 373275 (2013).
18. Vieille, C. & Zeikus, G. J. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* **65**, 1–43 (2001).
19. Sharma, A., Kawarabayasi, Y. & Satyanarayana, T. Acidophilic bacteria and archaea: Acid stable biocatalysts and their potential applications. *Extremophiles* **16**, 1–19 (2012).
20. Liao, H. *et al.* A new acidophilic thermostable endo-1,4- $\beta$ -mannanase from *Penicillium oxalicum* GZ-2: cloning, characterization and functional expression in *Pichia pastoris*. *BMC Biotechnol.* **14**, 90 (2014).
21. Gessesse. Purification and properties of two thermostable alkaline xylanases from an alkaliphilic bacillus sp. *Appl. Environ. Microbiol.* **64**, 3533–5 (1998).
22. D'Auria, S., Rossi, M. & Lakowicz, J. R. Glucose-sensing proteins from mesophilic and thermophilic bacteria as new tools in diabetes monitoring. in *Advances in Fluorescence Sensing Technology V* (eds. Lakowicz, J. R. & Thompson, R. B.) **2**, 21–31 (2001).
23. Öberg, F. *et al.* Glycosylation increases the thermostability of human aquaporin 10 protein. *J. Biol. Chem.* **286**, 31915–23 (2011).
24. Bagautdinov, B. *et al.* Thermodynamic analysis of unusually thermostable CutA1 protein from human brain and its protease susceptibility. *J. Biochem.* **157**, 169–176 (2015).
25. Turner, P., Mamo, G. & Karlsson, E. N. Potential and utilization of thermophiles and thermostable enzymes in biorefining. *Microb. Cell Fact.* **6**, 9 (2007).
26. Kumar, S., Tsai, C.-J. & Nussinov, R. Factors enhancing protein thermostability. *Protein Eng. Des. Sel.* **13**, 179–191 (2000).
27. Metpally, R. P. R. & Reddy, B. V. B. Comparative proteome analysis of psychrophilic versus mesophilic bacterial species: Insights into the molecular basis of cold adaptation of proteins. *BMC Genomics* **10**, 11 (2009).
28. Fujinami, S. & Fujisawa, M. Industrial applications of alkaliphiles and their enzymes—past, present and future. *Environ. Technol.* **31**, 845–56 (2010).
29. Kato, C., Li, L., Tamaoka, J. & Horikoshi, K. Molecular analyses of the sediment of the 11,000-m deep Mariana Trench. *Extremophiles* **1**, 117–23 (1997).
30. Madern, D. & Ebel, C. Influence of an anion-binding site in the stabilization of halophilic malate dehydrogenase from *Haloarcula marismortui*. *Biochimie* **89**, 981–987 (2007).
31. Chakravorty, D. & Patra, S. Attaining Extremophiles and Extremolytes: Methodologies and Limitations. in *Extremophiles* 29–74 (John Wiley & Sons, Inc., 2012). doi:10.1002/9781118394144.ch2
32. D'Amico, S., Collins, T., Marx, J. C., Feller, G. & Gerday, C. Psychrophilic microorganisms: Challenges for life. *EMBO Reports* **7**, 385–389 (2006).
33. Debamitra, C., Kumar, S. A., Sarath, B. V. R. & Sanjukta, P. Molecular Evolution of Extremophiles. *Extremophiles* (2012). doi:doi:10.1002/9781118394144.ch1
34. Stetter, K. O. Extremophiles and their adaptation to hot environments. *FEBS Letters* (1999).



- doi:10.1016/S0014-5793(99)00663-8
35. Goodarzi, H., Torabi, N., Najafabadi, H. S. & Archetti, M. Amino acid and codon usage profiles: Adaptive changes in the frequency of amino acids and codons. *Gene* **407**, 30–41 (2008).
  36. Kumar, A. *et al.* Protein adaptations in extremophiles: An insight into extremophilic connection of mycobacterial proteome. in *Seminars in cell & developmental biology* (Elsevier, 2018).
  37. Allers, T. & Mevarech, M. Archaeal genetics - The third way. *Nature Reviews Genetics* **6**, 58–73 (2005).
  38. Gibson, G. & Muse, S. V. *Précis de génomique*. (De Boeck Supérieur, 2004).
  39. Frank, A. C. & Lobry, J. R. Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms. *Gene* **238**, 65–77 (1999).
  40. Grocock, R. J. & Sharp, P. M. Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene* **289**, 131–139 (2002).
  41. Bao, S. *et al.* Evaluation of next-generation sequencing software in mapping and assembly. *J. Hum. Genet.* **56**, 406–414 (2011).
  42. Saunders, N. F. W. *et al.* Mechanisms of thermal adaptation revealed from the genomes of the antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. *Genome Res.* **13**, 1580–1588 (2003).
  43. Singer, G. A. C. & Hickey, D. A. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. in *Gene* **317**, 39–47 (2003).
  44. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput. Biol.* **3**, 0062–0072 (2007).
  45. Paul, S., Bag, S. K., Das, S., Harvill, E. T. & Dutta, C. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol.* **9**, R70 (2008).
  46. Baker-Austin, C. & Dopson, M. *Life in acid: pH homeostasis in acidophiles*. *Trends in Microbiology* **15**, 165–171 (2007).
  47. Basak, S., Banerjee, T., Gupta, S. K. & Ghosh, T. C. Investigation on the causes of codon and amino acid usages variation between thermophilic *aquifex aeolicus* and mesophilic *bacillus subtilis*. *J. Biomol. Struct. Dyn.* **22**, 205–214 (2004).
  48. Siddiqui, K. S. & Thomas, T. *Protein adaptation in extremophiles*. (Nova Publishers, 2008).
  49. Paul, M., Hazra, M., Barman, A. & Hazra, S. Comparative molecular dynamics simulation studies for determining factors contributing to the thermostability of chemotaxis protein ‘CheY’. *J. Biomol. Struct. Dyn.* **32**, 928–49 (2014).
  50. Najafabadi, H. S., Goodarzi, H. & Torabi, N. Optimality of codon usage in *Escherichia coli* due to load minimization. *J. Theor. Biol.* **237**, 203–9 (2005).
  51. Farias, S. T. & Bonato, M. C. M. Preferred amino acids and thermostability. *Genet. Mol. Res.* **2**, 383–393 (2003).
  52. Madigan, M. T. Extremophilic bacteria and microbial diversity. *Ann. Missouri Bot. Gard.* **3**–12 (2000).
  53. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
  54. Takami, H. & Horikoshi, K. Analysis of the genome of an alkaliphilic *Bacillus* strain from an industrial point of view. *Extremophiles* **4**, 99–108 (2000).
  55. Horikoshi, K. Alkaliphiles: some applications of their products for biotechnology. *Microbiol. Mol. Biol. Rev.* **63**, 735–50, table of contents (1999).
  56. Di Giulio, M. A comparison of proteins from *Pyrococcus furiosus* and *Pyrococcus abyssi*: barophily in the physicochemical properties of amino acids and in the genetic code. *Gene* **346**, 1–6 (2005).
  57. Abe, F., Kato, C. & Horikoshi, K. Pressure-regulated metabolism in microorganisms. *Trends Microbiol.* **7**, 447–453 (1999).
  58. Kato, C., Sato, T. & Horikoshi, K. Isolation and properties of barophilic and barotolerant bacteria from deep-sea mud samples. *Biodivers. Conserv.* **4**, 1–9 (1995).
  59. Kato, C. & Bartlett, D. H. D. H. The molecular biology of barophilic bacteria. *Extremophiles* **1**, 111–116 (1997).
  60. Kreil, D. P. P. & Ouzounis, C. A. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res.* **29**, 1608–1615 (2001).
  61. Chang, C., Park, B. C., Lee, D.-S. & Suh, S. W. Crystal structures of thermostable xylose isomerases



- from *Thermus caldophilus* and *Thermus thermophilus*: possible structural determinants of thermostability 1 Edited by D. Rees. *J. Mol. Biol.* **288**, 623–634 (1999).
62. Delboni, L. F. *et al.* Crystal structure of recombinant triosephosphate isomerase from *Bacillus stearothermophilus*. An analysis of potential thermostability factors in six isomerases with known three-dimensional structures points to the importance of hydrophobic interactions. *Protein Sci.* **4**, 2594–2604 (1995).
63. Reetz, M. T. Biocatalysis in Organic Chemistry and Biotechnology: Past, Present, and Future. *J. Am. Chem. Soc.* **135**, 12480–12496 (2013).
64. Trivedi, S., Gehlot, H. S. & Rao, S. R. Protein thermostability in Archaea and Eubacteria. *Genet. Mol. Res.* **5**, 816–27 (2006).
65. Spassov, V. Z., Karshikoff, A. D. & Ladenstein, R. The optimization of protein-solvent interactions: Thermostability and the role of hydrophobic and electrostatic interactions. *Protein Sci.* **4**, 1516–1527 (1995).
66. Vogt, G. & Argos, P. Protein thermal stability: hydrogen bonds or internal packing? *Fold. Des.* **2**, S40–S46 (1997).
67. Karshikoff, A. & Ladenstein, R. Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. *Protein Eng. Des. Sel.* **11**, 867–872 (1998).
68. Szilágyi, A. & Závodszky, P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: Results of a comprehensive survey. *Structure* **8**, 493–504 (2000).
69. Jaenicke, R. & Böhm, G. The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.* **8**, 738–48 (1998).
70. Haney, P. J. *et al.* Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 3578–83 (1999).
71. Suplatov, D. *et al.* Computational design of a pH stable enzyme: understanding molecular mechanism of penicillin acylase's adaptation to alkaline conditions. *PLoS One* **9**, e100643 (2014).
72. Chakravarty, S. & Varadarajan, R. Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Lett.* **470**, 65–69 (2000).
73. Parthiban, V., Gromiha, M. M. & Schomburg, D. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* **34**, W239–42 (2006).
74. Whitaker, J. R., Feeney, R. E. & Sternberg, M. M. Chemical and physical modification of proteins by the hydroxide ion. *CRC Crit. Rev. Food Sci. Nutr.* **19**, 173–212 (1983).
75. Ahern, T. J. & Klibanov, A. M. Analysis of processes causing thermal inactivation of enzymes. *Methods Biochem. Anal. Vol.* **33**, 91–128 (1988).
76. Swaim, M. W. & Pizzo, S. V. Methionine sulfoxide and the oxidative regulation of plasma proteinase inhibitors. *J. Leukoc. Biol.* **43**, 365–79 (1988).
77. Mattos, C. Protein–water interactions in a dynamic world. *Trends Biochem. Sci.* **27**, 203–208 (2002).
78. Radestock, S. & Gohlke, H. Protein rigidity and thermophilic adaptation. *Proteins Struct. Funct. Bioinforma.* **79**, 1089–1108 (2011).
79. Pack, S. P. & Yoo, Y. J. Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. *J. Biotechnol.* **111**, 269–77 (2004).
80. Lobanov, M. Y., Bogatyreva, N. S. & Galzitskaya, O. V. Radius of gyration as an indicator of protein structure compactness. *Mol. Biol.* **42**, 623–628 (2008).
81. Feller, G. & Gerday, C. Psychrophilic enzymes: hot topics in cold adaptation. *Nat. Rev. Microbiol.* **1**, 200–8 (2003).
82. Creighton, T. E. *Protein structure: a practical approach*. (Oxford University Press, 1997).
83. Mathabatha, E. S. Diversity and industrial potential of hydrolase-producing halophilic/halotolerant eubacteria. *African J. Biotechnol.* **9**, 1555–1560 (2010).
84. Tokunaga, H., Arakawa, T. & Tokunaga, M. Engineering of halophilic enzymes: two acidic amino acid residues at the carboxy-terminal region confer halophilic characteristics to *Halomonas* and *Pseudomonas* nucleoside diphosphate kinases. *Protein Sci.* **17**, 1603–10 (2008).
85. Lanyi, J. K. Salt-dependent properties of proteins from extremely halophilic bacteria. *Bacteriol. Rev.* **38**, 272–90 (1974).
86. Norris, P. R., Burton, N. P. & Foulis, N. A. Acidophiles in bioreactor mineral processing.

- Extremophiles* **4**, 71–6 (2000).
87. Yafremava, L. S., Di Giulio, M. & Caetano-Anollés, G. Comparative Analysis of Barophily-Related Amino Acid Content in Protein Domains of *Pyrococcus abyssi* and *Pyrococcus furiosus*. *Archaea* **2013**, 1–9 (2013).
  88. Tina, K. G., Bhadra, R. & Srinivasan, N. PIC: Protein Interactions Calculator. *Nucleic Acids Res.* **35**, W473–6 (2007).
  89. Bikadi, Z., Demko, L. & Hazai, E. Functional and structural characterization of a protein based on analysis of its hydrogen bonding network by hydrogen bonding plot. *Arch. Biochem. Biophys.* **461**, 225–234 (2007).
  90. Petsko, G. A. & Ringe, D. *Protein structure and function*. (New Science Press, 2004).
  91. Tiwari, A. & Panigrahi, S. K. HBAT: a complete package for analysing strong and weak hydrogen bonds in macromolecular crystal structures. *In Silico Biol.* **7**, 651–61 (2007).
  92. Costantini, S., Colonna, G. & Facchiano, A. M. ESBRI: a web server for evaluating salt bridges in proteins. *Bioinformatics* **3**, 137–8 (2008).
  93. Willard, L. *et al.* VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.* **31**, 3316–9 (2003).
  94. Ceroni, A., Passerini, A., Vullo, A. & Frasconi, P. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res.* **34**, W177–81 (2006).
  95. Ferrè, F. & Clote, P. DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine classification. *Nucleic Acids Res.* **34**, W182–5 (2006).
  96. Yaseen, A. & Li, Y. Dinosolve: a protein disulfide bonding prediction server using context-based features to enhance prediction accuracy. *BMC Bioinformatics* **14 Suppl 1**, S9 (2013).
  97. Kaur, H. & Raghava, G. P. S. A neural-network based method for prediction of  $\gamma$ -turns in proteins from multiple sequence alignment. *Protein Sci.* **12**, 923–929 (2003).
  98. Vogt, G., Woell, S. & Argos, P. Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **269**, 631–43 (1997).
  99. Khechinashvili, N. N. *et al.* Side Chain Dynamics and Alternative Hydrogen Bonding in the Mechanism of Protein Thermostabilization. *J. Biomol. Struct. Dyn.* **24**, 255–262 (2006).
  100. Aparna, V., Rambabu, G., Panigrahi, S. K., Sarma, J. A. R. P. & Desiraju, G. R. Virtual screening of 4-anilinoquinazoline analogues as EGFR kinase inhibitors: importance of hydrogen bonds in the evaluation of poses and scoring functions. *J. Chem. Inf. Model.* **45**, 725–38 (2005).
  101. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–93 (1994).
  102. Kollman, P. A. Theory of hydrogen bond directionality. *J. Am. Chem. Soc.* **94**, 1837–1842 (1972).
  103. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *J. Mol. Biol.* **320**, 369–387 (2002).
  104. Russell, R. J., Gerike, U., Danson, M. J., Hough, D. W. & Taylor, G. L. Structural adaptations of the cold-active citrate synthase from an Antarctic bacterium. *Structure* **6**, 351–61 (1998).
  105. Feller, G., Arpigny, J. L., Narinx, E. & Gerday, C. Molecular adaptations of enzymes from psychrophilic organisms. *Comp. Biochem. Physiol. Part A Physiol.* **118**, 495–499 (1997).
  106. Marshall, C. J. Cold-adapted enzymes. *Trends Biotechnol.* **15**, 359–364 (1997).
  107. Russell, N. J. Molecular adaptations in psychrophilic bacteria: potential for biotechnological applications. *Adv. Biochem. Eng. Biotechnol.* **61**, 1–21 (1998).
  108. Alvarez, M. *et al.* Triose-phosphate isomerase (TIM) of the psychrophilic bacterium *Vibrio marinus*. Kinetic and structural properties. *J. Biol. Chem.* **273**, 2199–206 (1998).
  109. Zuber, H. Temperature adaptation of lactate dehydrogenase. Structural, functional and genetic aspects. *Biophys. Chem.* **29**, 171–9 (1988).
  110. Hubbard, R. E. & Kamran Haider, M. Hydrogen Bonds in Proteins: Role and Strength. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Ltd, 2010). doi:10.1002/9780470015902.a0003011.pub2
  111. Wu, S., Skolnick, J. & Zhang, Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol.* **5**, 17 (2007).
  112. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40 (2008).
  113. Brunk, E. *et al.* Systems biology of the structural proteome. *BMC Syst. Biol.* **10**, 26 (2016).
  114. Magyar, C., Szilágyi, A. & Závodszy, P. Relationship between thermal stability and 3-D structure in a

- homology model of 3-isopropylmalate dehydrogenase from *Escherichia coli*. *Protein Eng.* **9**, 663–70 (1996).
115. Tanner, J. J., Hecht, R. M. & Krause, K. L. Determinants of Enzyme Thermostability Observed in the Molecular Structure of *Thermus aquaticus* d -Glyceraldehyde-3-phosphate Dehydrogenase at 2.5 Å Resolution †, ‡. *Biochemistry* **35**, 2597–2609 (1996).
116. Panigrahi, P., Sule, M., Ghanate, A., Ramasamy, S. & Suresh, C. G. Engineering Proteins for Thermostability with iRDP Web Server. *PLoS One* **10**, e0139486 (2015).
117. Watson, J. D. *Molecular biology of the gene*. (2008).
118. Srivastava, A. & Sinha, S. Thermostability of in vitro evolved *Bacillus subtilis* lipase A: a network and dynamics perspective. *PLoS One* **9**, e102856 (2014).
119. Chakravorty, D., Parameswaran, S., Dubey, V. K. & Patra, S. In silico characterization of thermostable lipases. *Extremophiles* **15**, 89–103 (2011).
120. Rose, G. D., Gierasch, L. M. & Smith, J. A. Turns in peptides and proteins. *Adv. Protein Chem.* **37**, 1–109 (1985).
121. Toniolo, C. & Benedetti, E. Intramolecularly Hydrogen-Bonded Peptide Conformation. *Crit. Rev. Biochem.* **9**, 1–44 (1980).
122. Milner-White, E., Ross, B. M., Ismail, R., Belhadj-Mostefa, K. & Poet, R. One type of gamma-turn, rather than the other gives rise to chain-reversal in proteins. *J. Mol. Biol.* **204**, 777–82 (1988).
123. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–637 (1983).
124. Hutchinson, E. G. & Thornton, J. M. PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci.* **5**, 212–220 (1996).
125. Petsko, G. A. [34] Structural basis of thermostability in hyperthermophilic proteins, or “There’s more than one way to skin a cat”. in *Methods in enzymology* **334**, 469–478 (2001).
126. Elcock, A. H. The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. *J. Mol. Biol.* **284**, 489–502 (1998).
127. Xiao, L. & Honig, B. Electrostatic contributions to the stability of hyperthermophilic proteins 1 Edited by J. M. Thornton. *J. Mol. Biol.* **289**, 1435–1444 (1999).
128. Panja, A. S., Bandopadhyay, B. & Maiti, S. Protein Thermostability Is Owing to Their Preferences to Non-Polar Smaller Volume Amino Acids, Variations in Residual Physico-Chemical Properties and More Salt-Bridges. *PLoS One* **10**, e0131495 (2015).
129. Matsuura, Y. *et al.* Thermodynamics of protein denaturation at temperatures over 100 °C: CutA1 mutant proteins substituted with hydrophobic and charged residues. *Sci. Rep.* **5**, 15545 (2015).
130. Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M. & Nishikawa, K. Unique amino acid composition of proteins in halophilic bacteria. *J. Mol. Biol.* **327**, 347–357 (2003).
131. Jaenicke, R., Schurig, H., Beaucamp, N. & Ostendorp, R. Structure and stability of hyperstable proteins: glycolytic enzymes from hyperthermophilic bacterium *Thermotoga maritima*. *Adv. Protein Chem.* **48**, 181–269 (1996).
132. Privalov, P. L. & Gill, S. J. Stability of Protein Structure and Hydrophobic Interaction. in *Advances in protein chemistry* **39**, 191–234 (1988).
133. Chan, C.-H., Yu, T.-H. & Wong, K.-B. Stabilizing salt-bridge enhances protein thermostability by reducing the heat capacity change of unfolding. *PLoS One* **6**, e21624 (2011).
134. Vetriani, C. *et al.* Protein thermostability above 100 degreesC: a key role for ionic interactions. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 12300–5 (1998).
135. Gromiha, M. M., Thomas, S. & Santhosh, C. Role of cation-pi interactions to the stability of thermophilic proteins. *Prep. Biochem. Biotechnol.* **32**, 355–62 (2002).
136. Gromiha, M. M., Santhosh, C. & Ahmad, S. Structural analysis of cation-pi interactions in DNA binding proteins. *Int. J. Biol. Macromol.* **34**, 203–11 (2004).
137. Knöchel, T. R. *et al.* The crystal structure of indole-3-glycerol phosphate synthase from the hyperthermophilic archaeon *Sulfolobus solfataricus* in three different crystal forms: effects of ionic strength. *J. Mol. Biol.* **262**, 502–15 (1996).
138. Davail, S., Feller, G., Narinx, E. & Gerday, C. Cold adaptation of proteins. Purification, characterization, and sequence of the heat-labile subtilisin from the antarctic psychrophile *Bacillus TA41*. *J. Biol. Chem.* **269**, 17448–53 (1994).
139. Kim, S.-Y. *et al.* Structural basis for cold adaptation sequence, biochemical properties, and crystal



- structure of malate dehydrogenase from a psychrophile *Aquaspirillum arcticum*. *J. Biol. Chem.* **274**, 11761–11767 (1999).
140. Nayek, A., Sen Gupta, P. S., Banerjee, S., Mondal, B. & Bandyopadhyay, A. K. Salt-Bridge Energetics in Halophilic Proteins. *PLoS One* **9**, e93862 (2014).
141. Pieper, U., Kapadia, G., Mevarech, M. & Herzberg, O. Structural features of halophilicity derived from the crystal structure of dihydrofolate reductase from the Dead Sea halophilic archaeon, *Haloferax volcanii*. *Structure* **6**, 75–88 (1998).
142. Dym, O., Mevarech, M. & Sussman, J. L. Structural features that stabilize halophilic malate dehydrogenase from an archaeobacterium. *Science* **267**, 1344–6 (1995).
143. Elcock, A. H. & McCammon, J. A. Electrostatic contributions to the stability of halophilic proteins1. *J. Mol. Biol.* **280**, 731–748 (1998).
144. Yang, H. *et al.* Structure-based engineering of histidine residues in the catalytic domain of  $\alpha$ -amylase from *Bacillus subtilis* for improved protein stability and catalytic efficiency under acidic conditions. *J. Biotechnol.* **164**, 59–66 (2013).
145. Robinson, N. E. Protein deamidation. *Proc. Natl. Acad. Sci.* **99**, 5283–5288 (2002).
146. Palmer, B., Angus, K., Taylor, L., Warwicker, J. & Derrick, J. P. Design of stability at extreme alkaline pH in streptococcal protein G. *J. Biotechnol.* **134**, 222–30 (2008).
147. Pace, C. N. Contribution of the hydrophobic effect to globular protein stability. *J. Mol. Biol.* **226**, 29–35 (1992).
148. Okada, J. *et al.* Evolution and thermodynamics of the slow unfolding of hyperstable monomeric proteins. *BMC Evol. Biol.* **10**, 207 (2010).
149. Rath, P. C., Jaeger, K.-E. & Gohlke, H. Structural Rigidity and Protein Thermostability in Variants of Lipase A from *Bacillus subtilis*. *PLoS One* **10**, e0130289 (2015).
150. Van den Burg, B. *et al.* Protein stabilization by hydrophobic interactions at the surface. *Eur. J. Biochem.* **220**, 981–5 (1994).
151. Makhatadze, G. I. & Privalov, P. L. Energetics of Protein Structure. in *Advances in Protein Chemistry* **47**, 307–425 (1995).
152. Schumann, J., Böhm, G., Schumacher, G., Rudolph, R. & Jaenicke, R. Stabilization of creatinase from *Pseudomonas putida* by random mutagenesis. *Protein Sci.* **2**, 1612–20 (1993).
153. Ventura, S. *et al.* Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat. Struct. Biol.* **9**, 485–93 (2002).
154. Gianese, G., Argos, P. & Pascarella, S. Structural adaptation of enzymes to low temperatures. *Protein Eng.* **14**, 141–8 (2001).
155. Violot, S. *et al.* Structure of a full length psychrophilic cellulase from *Pseudoalteromonas haloplanktis* revealed by X-ray diffraction and small angle X-ray scattering. *J. Mol. Biol.* **348**, 1211–24 (2005).
156. Goldstein, R. A. Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: Insights from the quasi-chemical approximation. *Protein Sci.* **16**, 1887–1895 (2007).
157. Siglioccolo, A., Paiardini, A., Piscitelli, M. & Pascarella, S. Structural adaptation of extreme halophilic proteins through decrease of conserved hydrophobic contact surface. *BMC Struct. Biol.* **11**, 50 (2011).
158. Burley, S. & Petsko, G. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science (80-. )*. **229**, 23–28 (1985).
159. Serrano, L., Bycroft, M. & Fersht, A. R. Aromatic-aromatic interactions and protein stability. *J. Mol. Biol.* **218**, 465–475 (1991).
160. Georis, J. *et al.* An additional aromatic interaction improves the thermostability and thermophilicity of a mesophilic family 11 xylanase: Structural basis and molecular study. *Protein Sci.* **9**, 466–475 (2008).
161. Tatko, C. D. & Waters, M. L. Selective aromatic interactions in beta-hairpin peptides. *J. Am. Chem. Soc.* **124**, 9372–3 (2002).
162. Longo, L. M. & Blaber, M. Prebiotic protein design supports a halophile origin of foldable proteins. *Front. Microbiol.* **4**, 418 (2014).
163. Betz, S. F. Disulfide bonds and the stability of globular proteins. *Protein Sci.* **2**, 1551–1558 (1993).
164. Matsumura, M. & Matthews, B. W. Stabilization of functional proteins by introduction of multiple disulfide bonds. *Methods Enzymol.* **202**, 336–56 (1991).
165. Braxton, S. Protein engineering for stability. *Protein Eng. Wiley-Liss, New York, NY* 299–316 (1996).
166. Tanaka, Y. *et al.* How Oligomerization Contributes to the Thermostability of an Archaeon Protein. *J.*

- Biol. Chem.* **279**, 32957–32967 (2004).
167. Wang, X.-Y., Meng, F.-G. & Zhou, H.-M. The role of disulfide bonds in the conformational stability and catalytic activity of phytase. *Biochem. Cell Biol.* **82**, 329–334 (2004).
168. Takagi, H. *et al.* Enhancement of the thermostability of subtilisin E by introduction of a disulfide bond engineered on the basis of structural comparison with a thermophilic serine protease. *J. Biol. Chem.* **265**, 6874–6878 (1990).
169. Johnson, C. M., Oliveberg, M., Clarke, J. & Fersht, A. R. Thermodynamics of denaturation of mutants of barnase with disulfide crosslinks. *J. Mol. Biol.* **268**, 198–208 (1997).
170. Anthony, L. C., Dombkowski, A. A. & Burgess, R. R. Using disulfide bond engineering to study conformational changes in the beta'260-309 coiled-coil region of Escherichia coli RNA polymerase during sigma(70) binding. *J. Bacteriol.* **184**, 2634–41 (2002).
171. Jorda, J. & Yeates, T. O. Widespread Disulfide Bonding in Proteins from Thermophilic Archaea. *Archaea* **2011**, 1–9 (2011).
172. Ogawa, K. *et al.* Roles of a short connecting disulfide bond in the stability and function of psychrophilic Shewanella violacea cytochrome c (5). *Extremophiles* **11**, 797–807 (2007).
173. Jaenicke, R. Do ultrastable proteins from hyperthermophiles have high or low conformational rigidity? *Proc. Natl. Acad. Sci.* **97**, 2962–2964 (2000).
174. Lehmann, M. & Wyss, M. Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr. Opin. Biotechnol.* **12**, 371–5 (2001).
175. Reetz, M. T., Carballera, J. D. & Vogel, A. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed. Engl.* **45**, 7745–51 (2006).
176. Karplus, P. A. & Schulz, G. E. Prediction of chain flexibility in proteins. *Naturwissenschaften* **72**, 212–213 (1985).
177. Parthasarathy, S. & Murthy, M. R. Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng.* **13**, 9–13 (2000).
178. Schlessinger, A., Yachdav, G. & Rost, B. PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics* **22**, 891–893 (2006).
179. Duan, Y. *et al.* The role of leucine and its metabolites in protein and energy metabolism. *Amino Acids* **48**, 41–51 (2016).
180. Xie, Y. *et al.* Enhanced enzyme kinetic stability by increasing rigidity within the active site. *J. Biol. Chem.* **289**, 7994–8006 (2014).
181. Zhang, J. *et al.* High-throughput screening of B factor saturation mutated Rhizomucor miehei lipase thermostability based on synthetic reaction. *Enzyme Microb. Technol.* **50**, 325–30 (2012).
182. Kim, S. J. *et al.* The development of a thermostable CiP (Coprinus cinereus peroxidase) through in silico design. *Biotechnol. Prog.* **26**, 1038–46 (2010).
183. Spector, S. *et al.* Rational modification of protein stability by the mutation of charged surface residues. *Biochemistry* **39**, 872–9 (2000).
184. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).
185. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **33**, (2005).
186. Benedix, A., Becker, C. M., de Groot, B. L., Caflisch, A. & Böckmann, R. A. Predicting free energy changes using structural ensembles. *Nat. Methods* **6**, 3–4 (2009).
187. Cheng, J., Randall, A. & Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* **62**, 1125–32 (2006).
188. Yin, S., Ding, F. & Dokholyan, N. V. Eris: an automated estimator of protein stability. *Nat. Methods* **4**, 466 (2007).
189. Huang, L.-T., Gromiha, M. M. & Ho, S.-Y. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics* **23**, 1292–1293 (2007).
190. Dehouck, Y. *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* **25**, 2537–2543 (2009).
191. Huang, L.-T. & Gromiha, M. M. Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. *Bioinformatics* **25**, 2181–2187 (2009).
192. Teng, S., Srivastava, A. K. & Wang, L. Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics* **11 Suppl 2**, S5 (2010).

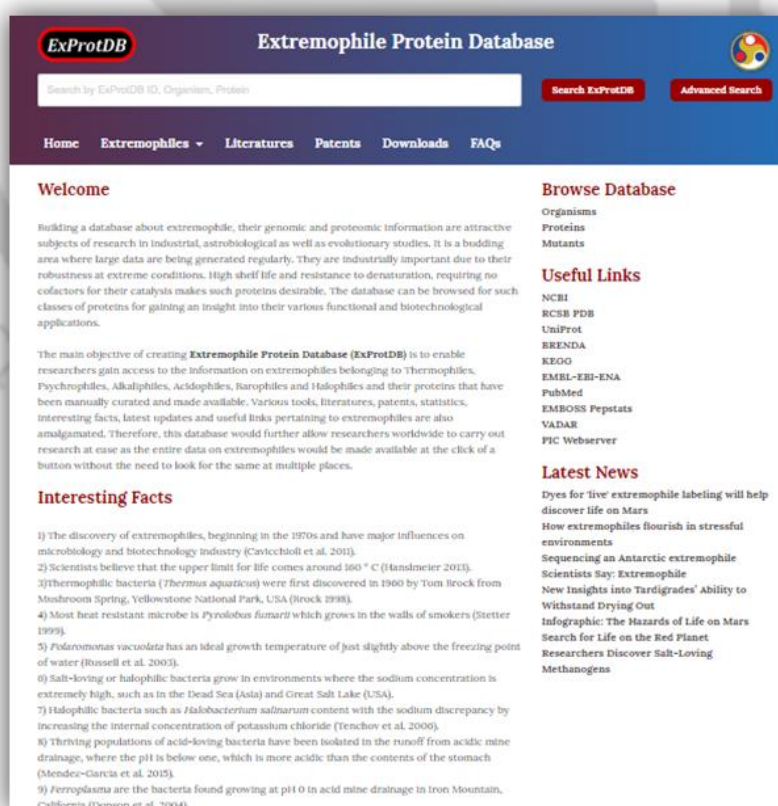
193. Masso, M. & Vaisman, I. I. A structure-based computational mutagenesis elucidates the spectrum of stability-activity relationships in proteins. in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 3225–3228 (IEEE, 2011). doi:10.1109/IEMBS.2011.6090877
194. Worth, C. L., Preissner, R. & Blundell, T. L. SDM--a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res.* **39**, W215–22 (2011).
195. Chen, C.-W., Lin, J. & Chu, Y.-W. iStable: off-the-shelf predictor integration for predicting protein stability changes. in *BMC bioinformatics* **14**, S5 (BioMed Central, 2013).
196. Giollo, M., Martin, A. J. M., Walsh, I., Ferrari, C. & Tosatto, S. C. E. NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. *BMC Genomics* **15**, S7 (2014).
197. Frappier, V. & Najmanovich, R. J. A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. *PLoS Comput. Biol.* **10**, e1003569 (2014).
198. Seeliger, D. & de Groot, B. L. Protein thermostability calculations using alchemical free energy simulations. *Biophys. J.* **98**, 2309–16 (2010).
199. Khan, S. & Vihinen, M. Performance of protein stability predictors. *Hum. Mutat.* **31**, 675–84 (2010).
200. Tian, J., Wu, N., Chu, X. & Fan, Y. Predicting changes in protein thermostability brought about by single- or multi-site mutations. *BMC Bioinformatics* **11**, 370 (2010).
201. Li, Y. & Fang, J. PROTS-RF: A Robust Model for Predicting Mutation-Induced Protein Stability Changes. *PLoS One* **7**, e47247 (2012).
202. Manjunath, K. & Sekar, K. Molecular dynamics perspective on the protein thermal stability: a case study using SAICAR synthetase. *J. Chem. Inf. Model.* **53**, 2448–61 (2013).
203. Wagner, A. Neutralism and selectionism: a network-based reconciliation. *Nat. Rev. Genet.* **9**, 965–974 (2008).
204. Eijsink, V. G. H. *et al.* Rational engineering of enzyme stability. *J. Biotechnol.* **113**, 105–120 (2004).
205. Shental-Bechor, D. & Levy, Y. Effect of glycosylation on protein folding: A close look at thermodynamic stabilization. *Proc. Natl. Acad. Sci.* **105**, 8256–8261 (2008).
206. Britton, K. L. *et al.* Insights into thermal stability from a comparison of the glutamate dehydrogenases from *Pyrococcus furiosus* and *Thermococcus litoralis*. *Eur. J. Biochem.* **229**, 688–95 (1995).
207. Olsen, O. & Thomsen, K. K. Improvement of bacterial -glucanase thermostability by glycosylation. *J. Gen. Microbiol.* **137**, 579–585 (1991).
208. Edmondson, S. P., Qiu, L. & Shriver, J. W. Solution Structure of the DNA-Binding Protein Sac7d from the Hyperthermophile *Sulfolobus acidocaldarius*. *Biochemistry* **34**, 13289–13304 (1995).
209. Marcus, Y. Effect of Ions on the Structure of Water: Structure Making and Breaking. *Chem. Rev.* **109**, 1346–1370 (2009).
210. Eiberweiser, A., Nazet, A., Kruchinin, S. E., Fedotova, M. V. & Buchner, R. Hydration and Ion Binding of the Osmolyte Ectoine. *J. Phys. Chem. B* **119**, 15203–11 (2015).
211. Widderich, N. *et al.* Biochemical properties of ectoine hydroxylases from extremophiles and their wider taxonomic distribution among microorganisms. *PLoS One* **9**, e93809 (2014).
212. GALINSKI, E. A., PFEIFFER, H.-P. & TRUPER, H. G. 1,4,5,6-Tetrahydro-2-methyl-4-pyrimidinecarboxylic acid. A novel cyclic amino acid from halophilic phototrophic bacteria of the genus *Ectothiorhodospira*. *Eur. J. Biochem.* **149**, 135–139 (1985).
213. Pastor, J. M. *et al.* Ectoines in cell stress protection: Uses and biotechnological production. *Biotechnol. Adv.* **28**, 782–801 (2010).
214. Kuhlmann, A. U., Bursy, J., Gimpel, S., Hoffmann, T. & Bremer, E. Synthesis of the Compatible Solute Ectoine in *Virgibacillus pantothenticus* Is Triggered by High Salinity and Low Growth Temperature. *Appl. Environ. Microbiol.* **74**, 4560–4563 (2008).
215. Hoefft, S. E. *et al.* *Alkalilimnicola ehrlichii* sp. nov., a novel, arsenite-oxidizing haloalkaliphilic gammaproteobacterium capable of chemoautotrophic or heterotrophic growth with nitrate or oxygen as the electron acceptor. *Int. J. Syst. Evol. Microbiol.* **57**, 504–12 (2007).
216. Calderón, M. I. *et al.* Complex regulation of the synthesis of the compatible solute ectoine in the halophilic bacterium *Chromohalobacter salexigens* DSM 3043T. *Microbiology* **150**, 3051–63 (2004).
217. Küsel, K., Dorsch, T., Acker, G. & Stackebrandt, E. Microbial reduction of Fe(III) in acidic sediments: Isolation of *Acidiphilium cryptum* JF-5 capable of coupling the reduction of Fe(III) to the oxidation of



- glucose. *Appl. Environ. Microbiol.* **65**, 3633–3640 (1999).
218. Ma, K., Linder, D., Stetter, K. O. & Thauer, R. K. Purification and properties of N5,N10-methylenetetrahydromethanopterin reductase (coenzyme F420-dependent) from the extreme thermophile *Methanopyrus kandleri*. *Arch. Microbiol.* **155**, 593–600 (1991).
219. Michels, P. C. & Clark, D. S. Pressure-enhanced activity and stability of a hyperthermophilic protease from a deep-sea methanogen. *Appl. Environ. Microbiol.* **63**, 3985–91 (1997).
220. Gupta, R. S. Evolution of the chaperonin families (HSP60, HSP 10 and TCP-1) of proteins and the origin of eukaryotic cells. *Mol. Microbiol.* **15**, 1–11 (1995).
221. Macario, A. J. L. & Conway de Macario, E. The pathology of cellular anti-stress mechanisms: a new frontier. *Stress* **7**, 243–9 (2004).
222. Klumpp, M. & Baumeister, W. The thermosome: archetype of group II chaperonins. *FEBS Lett.* **430**, 73–77 (1998).
223. Sahlan, M. & Yohda, M. Molecular chaperones in thermophilic eubacteria and arch. in *Thermophilic Microbes in Environmental and Industrial Biotechnology: Biotechnology of Thermophiles* 375–394 (2013). doi:10.1007/978-94-007-5899-5\_14
224. Taguchi, H., Konishi, J., Ishii, N. & Yoshida, M. A chaperonin from a thermophilic bacterium, *Thermus thermophilus*, that controls refoldings of several thermophilic enzymes. *J. Biol. Chem.* **266**, 22411–8 (1991).
225. Holden, J., Adams, M. W. W. & Baross, J. A. Heat-shock response in hyperthermophilic microorganisms. *Microb. Biosyst. New Front. Press. Acadia Univ. Press. Halifax, Canada* (2000).
226. Bukau, B., Weissman, J. & Horwich, A. Molecular Chaperones and Protein Quality Control. *Cell* **125**, 443–451 (2006).
227. Bukau, B. & Horwich, A. L. The Hsp70 and Hsp60 Chaperone Machines. *Cell* **92**, 351–366 (1998).
228. Packschies, L. *et al.* GrpE Accelerates Nucleotide Exchange of the Molecular Chaperone DnaK with an Associative Displacement Mechanism †. *Biochemistry* **36**, 3417–3422 (1997).
229. Mally, A. & Witt, S. N. GrpE accelerates peptide binding and release from the high affinity state of DnaK. *Nat. Struct. Biol.* **8**, 254–7 (2001).
230. Nakamura, A., Takumi, K. & Miki, K. Crystal structure of a thermophilic GrpE protein: insight into thermosensing function for the DnaK chaperone system. *J. Mol. Biol.* **396**, 1000–11 (2010).
231. Manukhov, I. V., Eroshnikov, G. E., Vyssokikh, M. Y. & Zvilgelsky, G. B. Folding and refolding of thermolabile and thermostable bacterial luciferases: the role of DnaKJ heat-shock proteins. *FEBS Lett.* **448**, 265–268 (1999).
232. Berger, F., Morellet, N., Menu, F. & Potier, P. Cold shock and cold acclimation proteins in the psychrotrophic bacterium *Arthrobacter globiformis* SI55. *J. Bacteriol.* **178**, 2999–3007 (1996).
233. Podar, M. & Reysenbach, A.-L. L. New opportunities revealed by biotechnological explorations of extremophiles. *Curr. Opin. Biotechnol.* **17**, 250–255 (2006).
234. Lentzen, G. & Schwarz, T. Extremolytes: natural compounds from extremophiles for versatile applications. *Appl. Microbiol. Biotechnol.* **72**, 623–634 (2006).
235. Chakravorty, D., Parameswaran, S., Dubey, V. K. & Patra, S. Unraveling the rationale behind organic solvent stability of lipases. *Appl. Biochem. Biotechnol.* **167**, 439–461 (2012).

## 2

## Extremophile Protein Database (ExProtDB)



## Abstract

Extremophilic adaptations though a burgeoning topic, a curated database for extremophiles and their proteins is still lacking. Thus, this chapter is a compilation of all the collected genomic and proteomic data of extremophiles into a manually curated database named as 'Extremophile Protein Database (ExProtDB)'. Architecture of the database has been built using a standard platform based on the Linux-Apache-MySQL-PHP (LAMP). It contains information of 536 organisms (175 archaea, 255 bacteria, 100 eukaryotes and 6 viruses), 795 protein entries (375 thermophilic proteins, 115 psychrophilic proteins, 67 acidophilic proteins, 94 alkaliphilic proteins, 100 halophilic proteins and 43 barophilic proteins) and 252 engineered extreme-stable proteins. It is integrated with a user friendly search and browse interface where search can be performed by user entered organism name; protein name; engineered extreme-stable proteins; literature and patents; E.C. number; physicochemical properties (temperature, pH, pressure, salt concentrations, etc.). Additionally, the amino acid compositions and intra-protein interaction data have been integrated with each protein entry in the database. The tools for calculation of amino acid compositions and intra-protein interaction were also incorporated as the download option for users. Therefore, it is a knowledgebase about all the information available in the World Wide Web ([www.exprotdb.com](http://www.exprotdb.com)) about extremophiles and their extreme-stable proteins. The novelty of the database lies in the fact that it gives consolidated and comprehensive information of all types of extremophiles.

## 2.1 Introduction

Extremophiles are delved deep into the search of biocatalysts functioning in hostile environments. In the process, interest is accorded towards the plethora of information which is escalating with time. However, the task of extracting information from literature resources is tedious and time consuming. One is often required to browse different databases for extracting different type of information *e.g.* NCBI Genome for genomic information, UniProtKB and NCBI protein for protein sequence, RCSB-PDB for structural information, BRENDA for enzyme information, and KEGG for pathway information. Regardless of the pivotal role of extremophiles in industries, no effort has yet been made to compile their unique properties on a single platform<sup>1</sup>.

In this regard, database provides a good platform for such data organization, easy information enabling retrieval of cross-references and querying interfaces. The database system differs from the traditional file-based system as the data definition is part of the application programs. Databases have been used since long for organizing and integrating complex data<sup>2</sup>. It resembles a well-organized electronic filing system in which the software, known as a database management system (DBMS) manages the contents. Database management systems (DBMS) performs the fundamental database tasks of storing and organizing information, and provides consistent results to multiple queries simultaneously. Thus, database creation facilitates extracting meaningful information as well as leads to better communication with other databases.

So far, there are some reports of biological databases with respect to extremophiles like the ProtDataTherm, a database for thermostability analysis and engineering of proteins; HaloDom, database of halophiles across all life domains; HaloWeb, providing genome information and complete genome sequences<sup>3</sup>. ProtDataTherm database for analysis and engineering of protein thermostability contains more than 14 million protein sequences categorized based on their thermal stability and protein family. HaloDom is a database hosting information on more than 1000 halophilic species whereas, HaloWeb contains the newly sequenced genomes, including up-to-date annotation data. However, till date, to

the best of our knowledge, there are no database on all extremophiles and extremophilic proteins.

Thus, this chapter reports a compilation of all the collected genomic and proteomic data of extremophiles into a manually curated database named as “Extremophile Protein Database” and abbreviated as ExProtDB.

## 2.2 Methodology

### 2.2.1 Servers and software used

Database management system (Linux, Apache, MySQL and PHP), Dreamweaver, UniProtKB, RCSB PDB, NCBI PubMed, NCBI Genome, KEGG Genome, BRENDA enzyme database, Protherm, XAMPP, PROTPARAM, COPid, VADAR, PIC webserver, VADA), ESBRI and PROMOTIF<sup>4-7</sup>.

### 2.2.2 Organism and genome information

The database incorporates extremophilic organisms and their genomic information which were collected from NCBI genome and KEGG genome database. The information includes organism name; phylogenetic domain; extremophile type (thermophile, psychrophile, acidophile, alkaliphile, halophile and barophile or combination of them); taxonomic lineage; optimum growth conditions (temperature, pH, salt, pressure); genome size; GC%; number of chromosomes and their lengths; number of plasmids and their lengths; number of gene (number of protein genes, mRNA genes, rRNA genes, tRNA genes); protein in ExProtDB and NCBI PubMed ID.

### 2.2.3 Collection of protein data and their refinement

Sequence and structures of all the available extremophilic proteins (such as thermophilic proteins; psychrophilic protein; acidophilic protein; alkaliphilic proteins; halophilic proteins and barophilic proteins) were collected from UniProtKB and RCSB PDB, respectively. As the main motive of this research was to correlate sequence and structural features to protein extreme-stability, sequences that do not have their crystallized

structures in PDB were excluded from this study. The mutant engineered proteins structures were also collected from the PDB and the Protherm database. Redundant information was discarded. The sequences were collected in FASTA format and the structures in .pdb format. Out of the total proteins in our database, protein structures with a resolution worse than 2.5 Å were removed. The mesostable homologous counterparts were obtained through a BLAST search with all other structures in PDB and optimal parameters. The top ranking (by E-value) protein was chosen from each BLAST search and only wild type thermophilic and mesophilic proteins were retained and the collected homologues were further confirmed by CLUSS2.

#### 2.2.4 Amino acid composition analysis of all thermostable proteins

The FASTA sequences of the thermostable proteins were collected from UniProtKB and EMBOSS: pepstats tool was used to generate percentage of amino acid composition (G, P, A, V, L, I, M, C, F, Y, W, H, K, R, Q, N, E, D, S and T); amino acid class contents (Acid, Bsc, Chrg, Ali, Aro, Tiny, Sml, Pol, Npol). The tool can be accessed through [https://www.ebi.ac.uk/Tools/seqstats/emboss\\_pepstats/](https://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/). The dipeptide composition was also enumerated with the aid of COPid server<sup>8</sup>. The server can be accessed through <http://www.imtech.res.in/raghava/copid/>. It gives average of percentage of occurrence of a dipeptide in the protein sequences. This aided in shedding light to the amino acid composition of all extremophilic protein types and a correlation was drawn with previous citations regarding the same. These features were collected and normalized based on the count of their frequency of occurrence in protein sequences and expressed as percentage score.

#### 2.2.5 Structural analysis of all extremophilic proteins and feature generation

The PDB structures were used for feature generation. An algorithm was written for the calculation of intra-protein interaction from the 3-D coordinates obtained from the PDB files intra-protein interactions. Protein Interaction Calculator (PIC) webserver, Volume Area Dihedral Angle Reporter (VADAR), ESBRI and Promotif were used for calculation of these intra-protein interactions. These features were collected and normalized based on



the count of their frequency of occurrence in protein structures and expressed as percentage score. Thus, the overall information for protein information included protein name; extreme-stable behavior (thermostable, psychrostable, acidostable, alkalistable, halostable and barostable); source organism (phylogenetic domain and extremophile type); EC number; protein type (enzyme or structural protein); optimum condition (temperature, pH, salt, pressure); protein sequence; sequence length; molecular weight; UniprotKB ID and PDB ID (structure resolution, PubMed ID, authors, journal, title, volume, year).

### 2.2.6 Tools and downloads

Two tools have been integrated in the database and they can be downloaded to calculate the composition of aforementioned features of protein sequences (amino acid calculator) and structures (intra-protein calculator).

### 2.2.7 Literature and patent

The collected literature information included year of publication; title; authors; journal name; NCBI PubMed ID and Patent ID. Similarly, collected patent information included year of publication; title; authors and Patent ID from Google Patent.

### 2.2.8 Data integration

Extremophile and their protein database was created using MySQL, APACHE and PHP platforms and web interface was created using Dreamweaver.

## 2.3 Features of ExProtDB

Extremophile Protein Database (ExProtDB) was built on an entity relationship model. It is the curated database for all types of extremophiles, their genomic, proteins and enumerated biophysical attributes. It hosts literature and patent information available regarding extremophiles and extremophilic proteins. It has a user friendly search, advanced search and browsing options. The main search can be performed by ExProtDB ID, organism name, protein name and mutants whereas, advanced search can be

performed by extremophile, temperature stability, pH stability, salt tolerance, pressure tolerance, EC number, PDB ID, UniProt ID, Mutant ID, etc. (Figure 2.1).

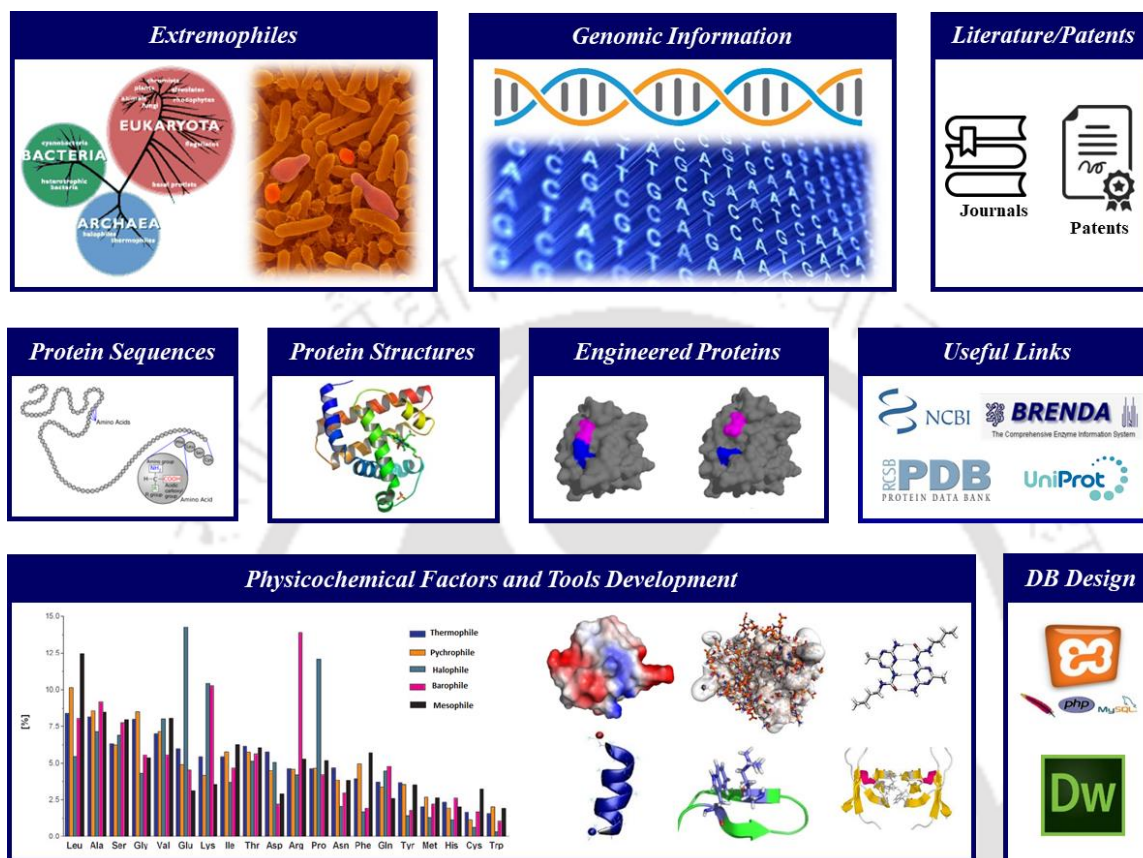


Figure 2.1: Representative features of database.

It has information about engineered extreme-stable protein mutants. The amino acid composition and intra-protein interaction data have been integrated with each protein entry and can be searched for each individual entry in the database. Tools for calculation of amino acid compositions and intra-protein interaction were also integrated are freely available for download for academic purposes. The database has also links to other related databases, webserver and tools such as, NCBI, PubMed, KEGG, PDB, UniProt KB, BRENDA, etc.

### 2.3.1 Information about extremophiles, their proteins and contributing genomic and proteomic factors

The database also incorporated the reading material of extremophiles, their proteins and contributing genomic and proteomic factors responsible for extreme-stability, interesting facts about extremophiles and latest news.

### 2.3.2 Genomic information

The data pertaining to collected extremophiles were extracted from NCBI Genome database and KEGG Genome database. It comprises of organism name; phylogenetic domain; extremophile type (thermophile, psychrophile, acidophile, alkaliphile, halophile and barophile or combination of them); taxonomic lineage; optimum growth conditions (temperature, pH, salt, pressure); genome size; GC%; number of chromosomes and their lengths; number of plasmids and their lengths; number of gene (number of protein genes, mRNA genes, rRNA genes, tRNA genes); protein in ExProtDb and PubMed ID.

### 2.3.3 Protein information

The data pertaining to the collected extremophilic protein were extracted from UniprotKB and RCSB PDB. It comprises of protein name; extreme-stable behavior (thermostable, psychrostable, acidostable, alkalistable, halostable and barostable); source organism (phylogenetic domain and extremophile type); EC number; protein type (enzyme or structural protein); optimum condition (temperature, pH, salt, pressure); protein sequence; sequence length; molecular weight; UniprotKB ID and PDB ID (structure resolution, PubMed ID, authors, journal, title, volume, year).

### 2.3.4 Mutant information

The data pertaining to the collected extreme-stable mutants comprises of protein name; point mutations; PDB ID of mutant and their wild type proteins (structure resolution, PubMed ID, authors, journal, title, volume, year); source organism; EC number; protein length and optimum conditions (temperature, pH, salt, pressure).

### 2.3.5 Physicochemical protein factors

Each collected protein and mutant of ExProtDB have been integrated with amino acid composition and intra-protein interaction data. These factors includes amino acid composition (G, P, A, V, L, I, M, C, F, Y, W, H, K, R, Q, N, E, D, S and T); amino acid class contents (Acid, Basic, Charged, Aromatic, Aliphatic, Tiny, Small, Polar, Non-polar); dipeptide analysis and intra-protein interactions (Main chain to main chain hydrogen bonds; Main chain to side chain hydrogen bonds; Side chain to side chain hydrogen bonds; Hydrophobic interactions; Ionic interactions; Aromatic-sulphur interaction; Cation- $\pi$  interactions; Aromatic-aromatic interactions; Salt bridge). The composition of amino acids and their classes were enumerated by Emboss Pepstat; Dipeptide analysis by COPid; Protein secondary structures by PROMOTIF; Intra-protein interactions by PIC webserver, VADAR and ESBRI.

### 2.3.6 Literature and patents

Separate search engines were developed for searching literature and patent. The information comprises of year of publication; title; authors; journal name; PubMed ID (for journals) and Patent ID.

### 2.3.7 Tools

The database is also integrated with two downloadable tools. One is for enumeration of percentage composition of amino acids in a protein sequence and was developed by coding in MATLAB. The input protein sequence file should be in the text format (.txt). The percentage count of each of the 20 standard amino acids (A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, and V) and 9 amino acid classes such as tiny (A+C+G+S+T), small (A+C+D+G+N+P+S+T+V), aliphatic (A+I+L+V), aromatic (F+H+W+Y), non-polar (A+C+F+G+I+L+M+P+V+W+Y), polar (D+E+H+K+N+Q+R+S+T), charged (B+D+E+H+K+R), basic (H+K+R) and acidic (D+E) are calculated by the tool. The ambiguous amino acid characters (B, Z, or X), gaps (-) and end terminators (\*) are ignored by default. The output of the analysis is produced and stored in excel spreadsheet format (.xls).

The other tool calculates intra-protein interactions which was written in Python. It is capable of calculating percentage number of main chain to main chain hydrogen bonds, main chain to side chain hydrogen bonds, side chain to side chain hydrogen bonds, hydrophobic interactions, ionic interactions, aromatic-sulphur interaction, cation- $\pi$  interactions, aromatic-aromatic interactions and salt bridge. These interactions are calculated based on empirical distance between the bond forming atoms. The input should be in the Protein Data Bank (.pdb) format and the output will be generated and stored in comma-separated values (.csv) file.

## 2.4 Database architecture and integration

Finally, with all these informations in hand, an Extremophile Protein Database (ExProtDB) was created using MySQL, APACHE and PHP platforms. A web interface was created using Dreamweaver. It will be accessed through [www.exprotdb.com](http://www.exprotdb.com). Figure 2.2 is the web interface snapshot of the HTML look of the database. The scheme has been presented in Figure 2.3.

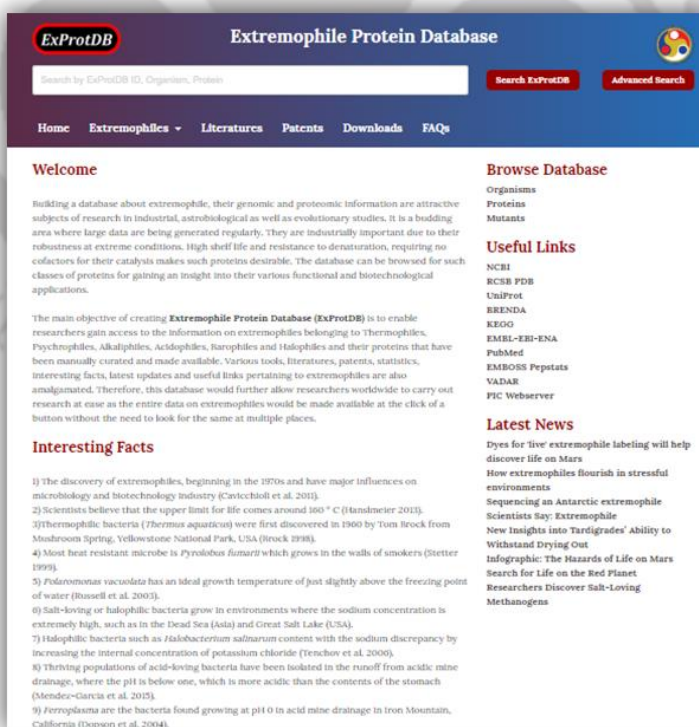
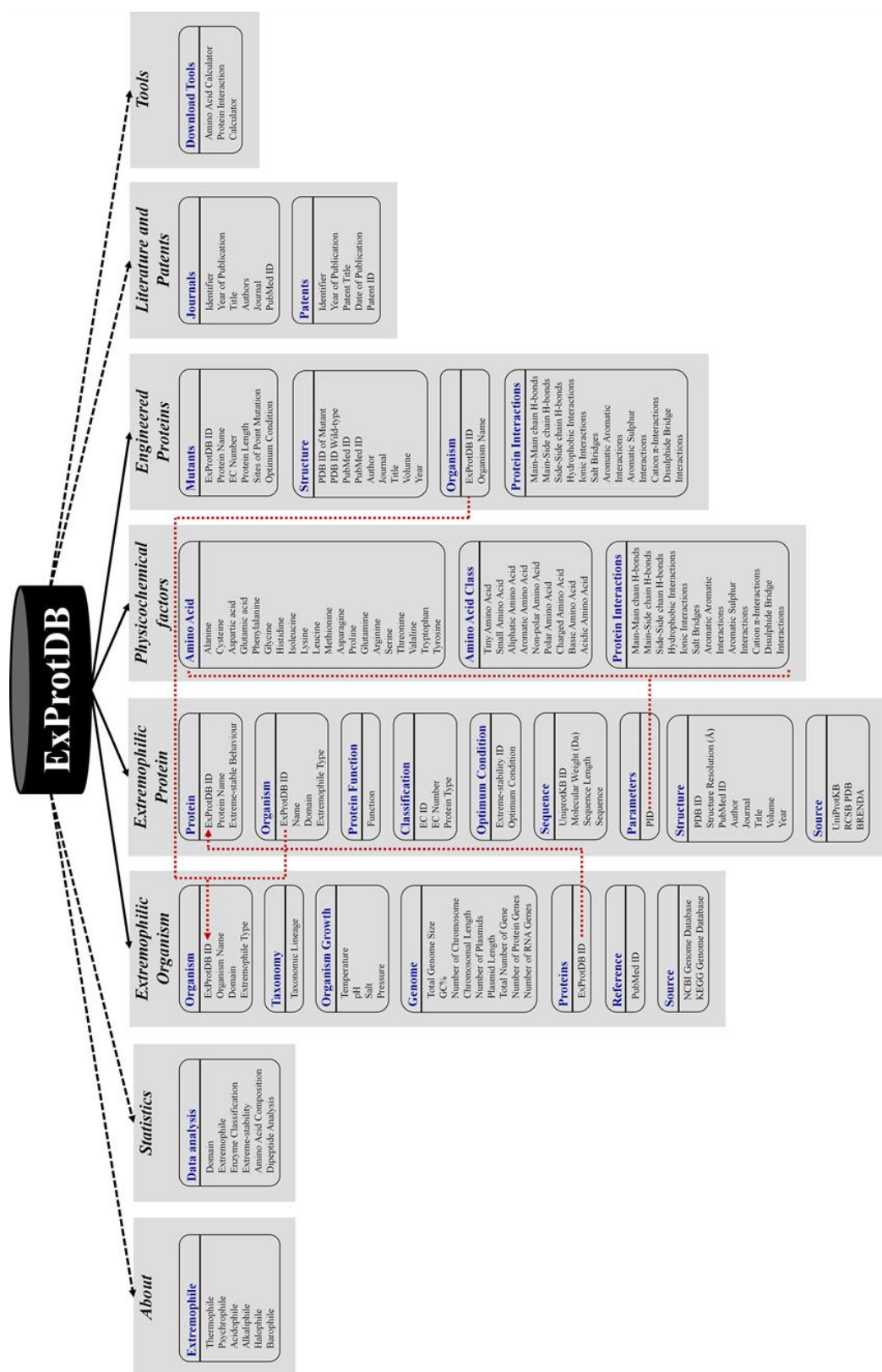


Figure 2.2: Graphical unit interface of ExProtDB.





**Figure 2.3:** Schematic illustration of detailed integration of components associated with ExProtDB. It depicts the major aspects of ExProtDB are extremophilic organisms, extremophilic proteins, mutants and physicochemical features



## 2.5 Data collection and analysis: statistics

ExProtDB contains the data pertaining to six types of extremophiles, their protein (sequences, structures), genomic information (GC content, genome size, number of chromosomes and plasmid, number of genes, RNA and proteins, etc.), mutants, physicochemical factors, literature and patents. Table 2.1 shows the collected data for thermophiles, psychrophiles, acidophiles, alkaliphiles, halophiles and barophiles regarding their number of organisms, mutants, literature and patents. A total number of 535 organisms, 794 proteins, 252 mutants, 19894 journal papers and 5711 patents related to these extremophiles were collected. The genomic data was collected from NCBI and KEGG Genome database and protein data was collected from UniProt, PDB and BRENDA. As structure and function of protein is driven by physicochemical factors, protein sequence and structure were collected as an input to enumerate these factors by Pepstat; PROMOTIF; VADAR tools, etc. Literature were collected from NCBI PubMed and patents were collected from Google patent.

Table 2.1: Summary of collected data from each type of extremophiles.

| Entries      | Organisms  | Proteins   | Mutants    | Literatures  | Patents     |
|--------------|------------|------------|------------|--------------|-------------|
| Thermophile  | 267        | 375        | 220        | 14483        | 1461        |
| Psychrophile | 148        | 115        | 10         | 934          | 631         |
| Acidophile   | 84         | 67         | 5          | 935          | 876         |
| Alkaliphile  | 61         | 94         | 6          | 583          | 701         |
| Halophile    | 64         | 100        | 7          | 1923         | 876         |
| Barophile    | 7          | 43         | 4          | 121          | 686         |
| Other        | 62         | -          | -          | 925          | 480         |
| <b>Total</b> | <b>535</b> | <b>794</b> | <b>252</b> | <b>19894</b> | <b>5711</b> |

The statistics of the collected data distributes extremophiles and their proteins into phylogenetic domains, extremophiles types, enzyme classes and the protein extreme-stability. The number of bacteria compared to other domains; thermophiles compared to other extremophile type; and hydrolase enzymes were found to be highest in the collected data. Figure 2.4, 2.5 and 2.6 shows the distribution of extremophiles into domains and types; extremophilic proteins into enzyme classes; extremophilic proteins into extreme-stability, respectively.

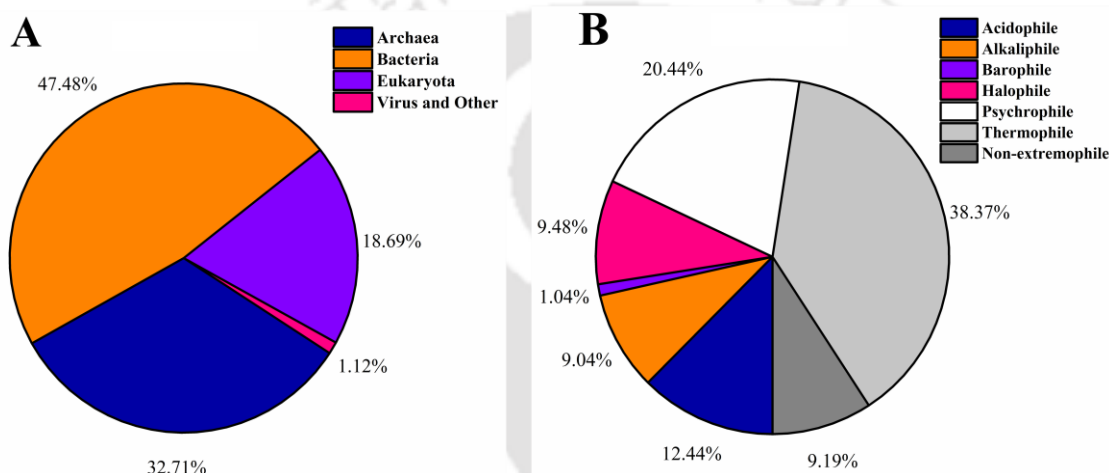


Figure 2.4: Distribution of extremophilic organism into – (A) Domains – archaeae, bacteria and eukaryota, viruses and others; (B) Extremophile types – alkaliphile, barophile, halophile, psychrophile and thermophiles.

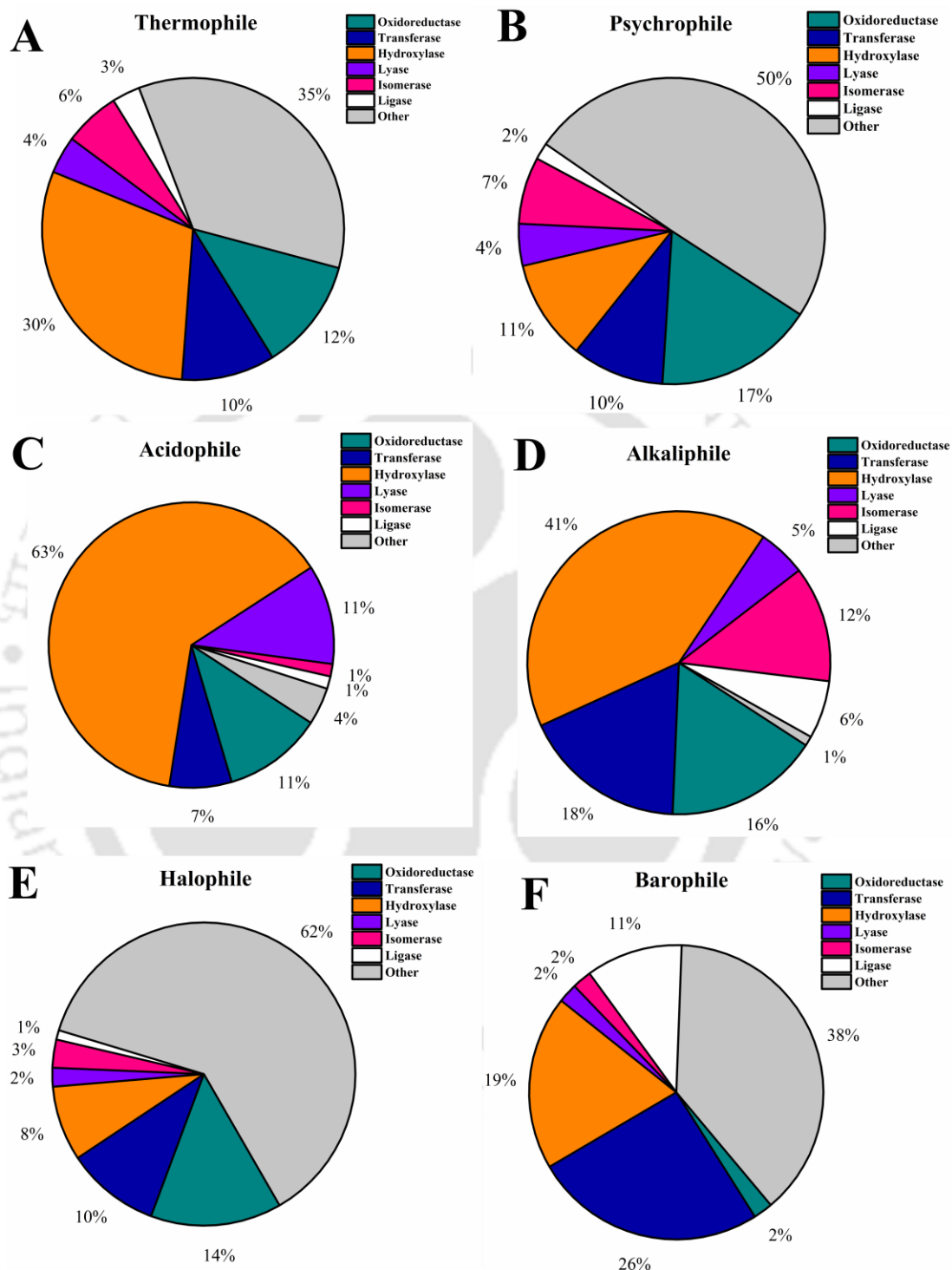


Figure 2.5: Distribution of extremophilic proteins (A) thermophilic proteins, (B) psychrophilic proteins, (C) acidophilic proteins, (D) alkaliphilic proteins, (E) halophilic protein and (F) barophilic proteins into enzyme classes – oxidoreductase, transferase, hydrolase, lyase, isomerase, ligase and other proteins.

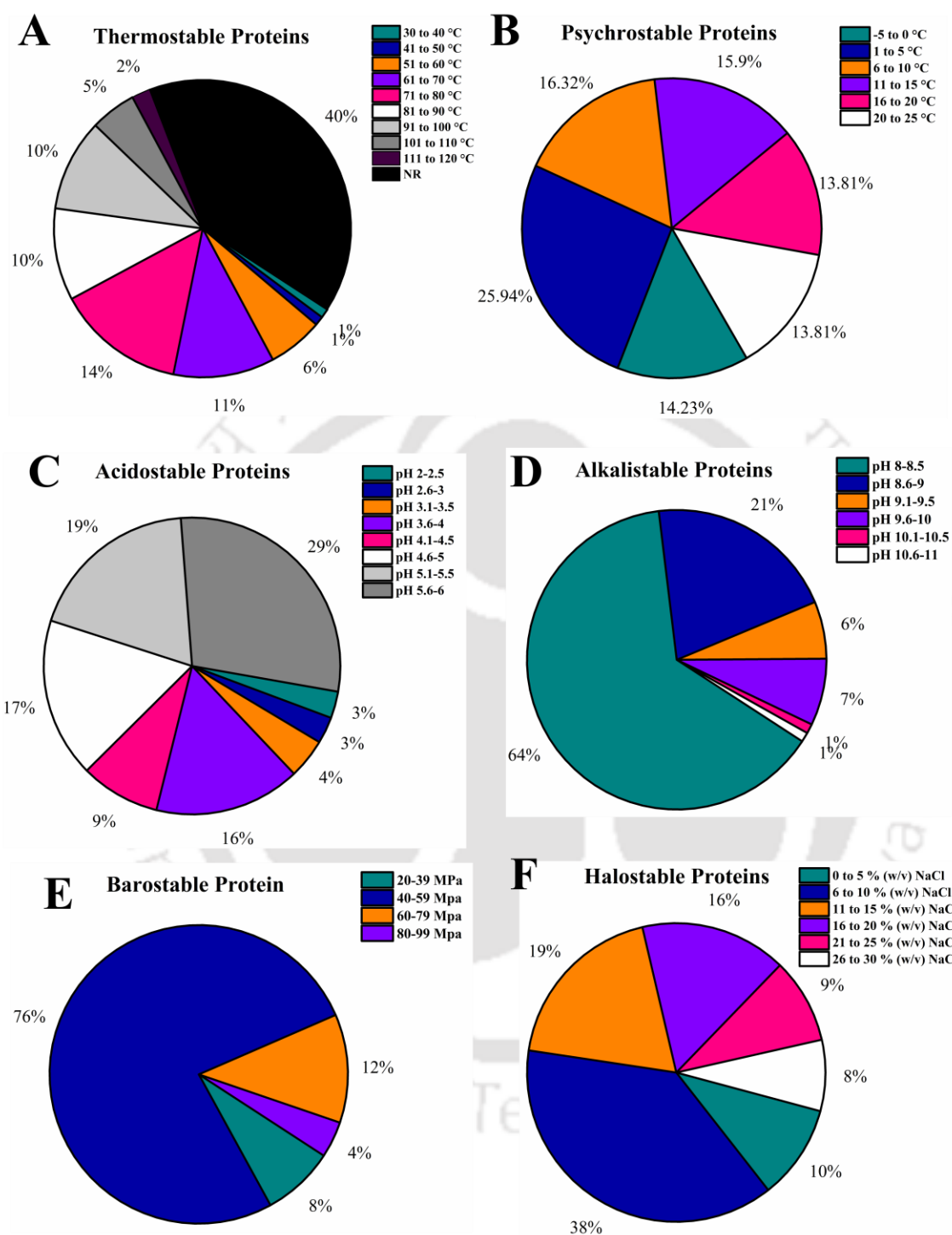


Figure 2.6: Distribution of extremophilic protein into extreme-stability – (A) thermostability (30 to 120 °C), (B) psychroability (-5 to 25 °C), (C) acidostability (pH 2 to pH 6), (D) alkalistability (pH 8 to pH 11), (E) halostability (0 to 30 % (w/v) NaCl) and (F) barostability (20 to 99 MPa).

The amino acid composition was further carried out since the primary structure of protein contains sequence of amino acids which plays a crucial role in protein stability determination<sup>9</sup>. The analysis was performed in the collected datasets of different extremophilic proteins and revealed the percentage of glutamic acid (E) in thermophiles and alkaliphiles; alanine (A) in psychrophiles; asparagine (N) in acidophiles; aspartic acid (D) in halophiles and arginine (R) in barophiles showed a significant difference in their compositions. These results are in accordance with earlier reports who reported adaptive changes in the frequency of amino acids for extremostability of proteins<sup>10</sup>. Figure 2.7 shows the comparative analysis of amino acid composition of different extremophilic proteins with their non-extremophilic counterparts.

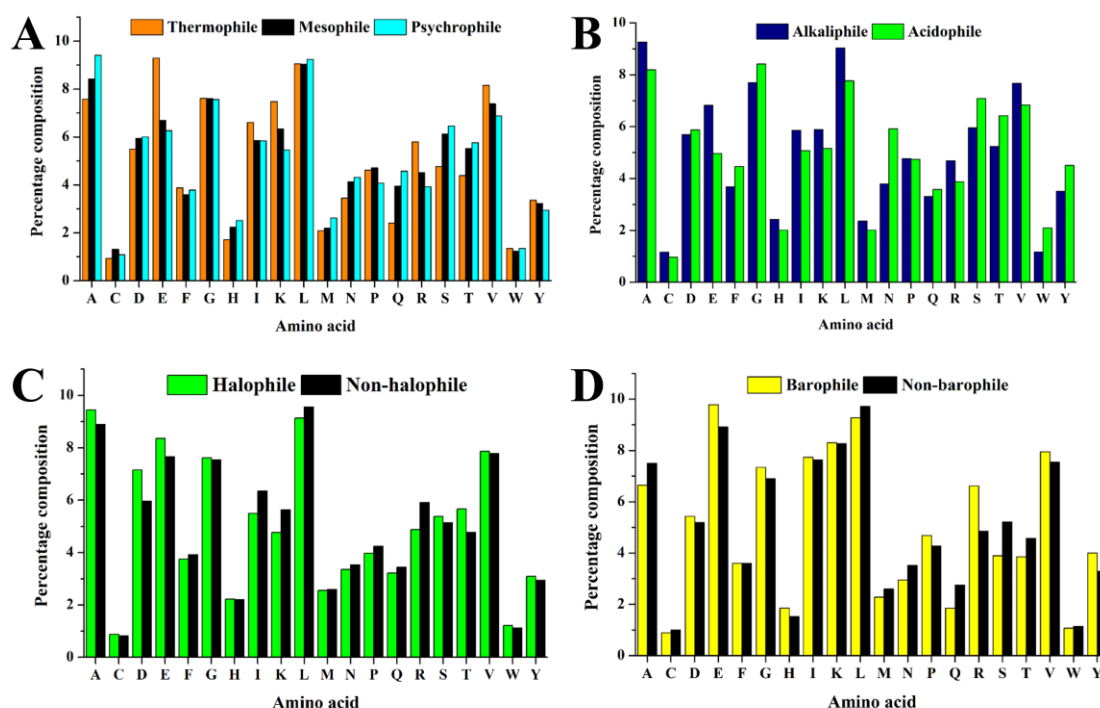


Figure 2.7: Amino acid composition analysis in different extremophilic proteins – (A) Thermophile vs Mesophile vs Psychrophile; (B) Acidophile vs Alkaliphile; (C) Halophile vs Non-halophile; and (D) Barophile vs Non-barophile.

Moreover, researchers also reported dipeptides as indicators of protein stability but restricted only to thermophiles<sup>11</sup>. Thus, the present study was accorded to analyse the dipeptides for all the classes of extremophilic proteins. The composition of all the 400 dipeptides of each protein sequence was computed. Figure 2.8 illustrates the preference

of dipeptides in different extremophilic proteins and demonstrated only those dipeptide which had preference score greater than 0.5 (chosen as threshold value). The resultant higher priority dipeptides showed significant variability among different extremophiles. The overall analysis of dipeptide composition analysis revealed that the frequency of alanine, leucine, and glutamic acid containing dipeptides in all extremophilic proteins were higher. For instance, Ala-Ala dipeptide in thermophilic proteins, acidophilic, alkaliphilic proteins and halophilic proteins; Leu-Ala dipeptide in psychrophile; and, Glu-Glu dipeptide in barophilic proteins had highest frequency. The present outcome is reported for the first time and they clearly indicated that the dipeptide features contribute significantly to discriminate extremophilic proteins from non-extremophilic ones.

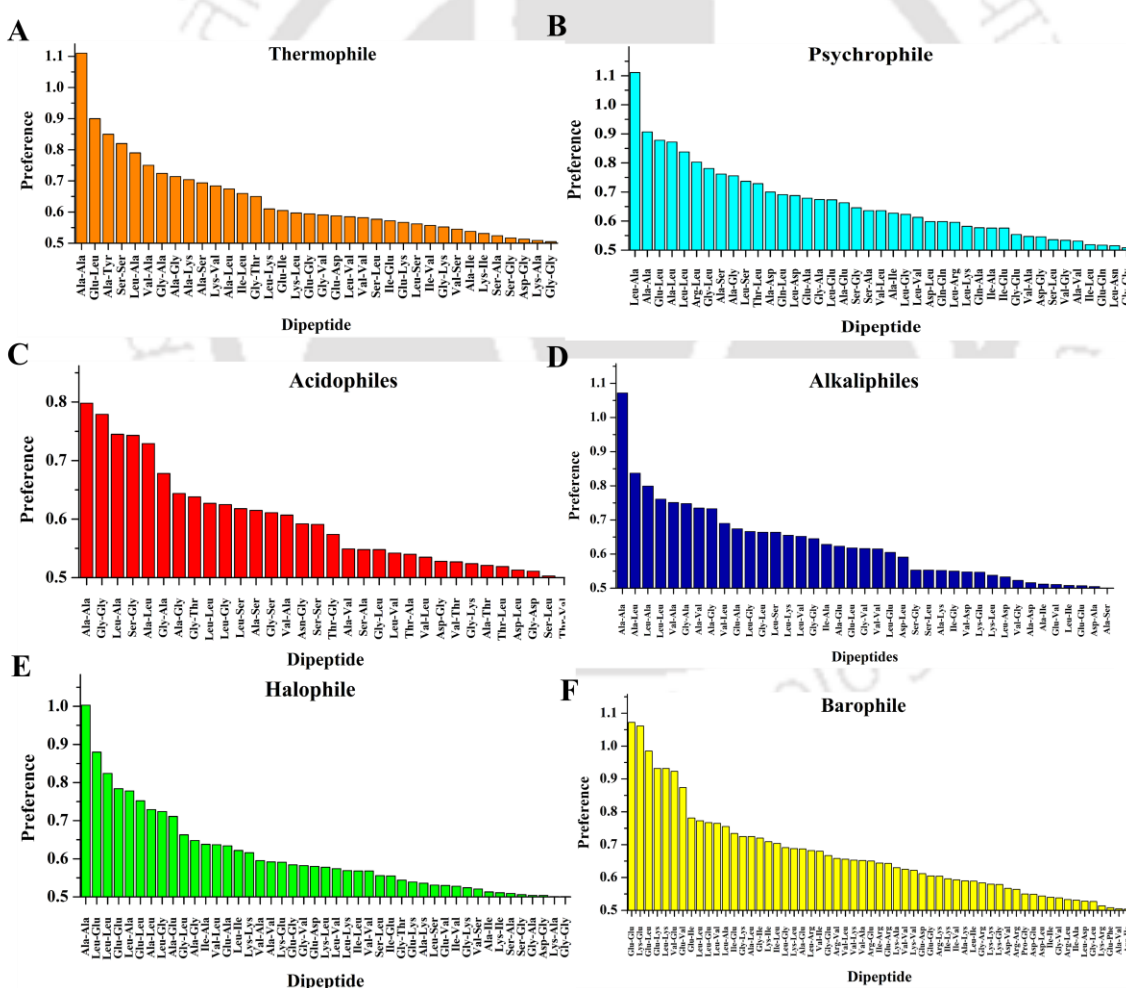


Figure 2.8: Dipeptide composition analysis in different extremophilic proteins – (A) Thermophile; (B) Psychrophile; (C) Acidophile; (D) Alkaliphile; (E) Halophile; and (F) Barophile.



## 2.6 Unique ID generation for collected extremophiles and their proteins

The unique ID are provided to individual extremophilic organisms and extremophilic protein. The significance of unique ID is that one can easily understand the organism and protein belongs from which classification domain, extremophile type, and extreme-stable type. Figure 2.9 illustrates the ExProtDB unique ID system for organisms and proteins.

An individual unique ID was provided to each entry of organism and protein. Organism ID is segregated into domain; extremophile type and organism series. Similarly, protein ID is the extension of the organism ID with extreme-stable protein series.

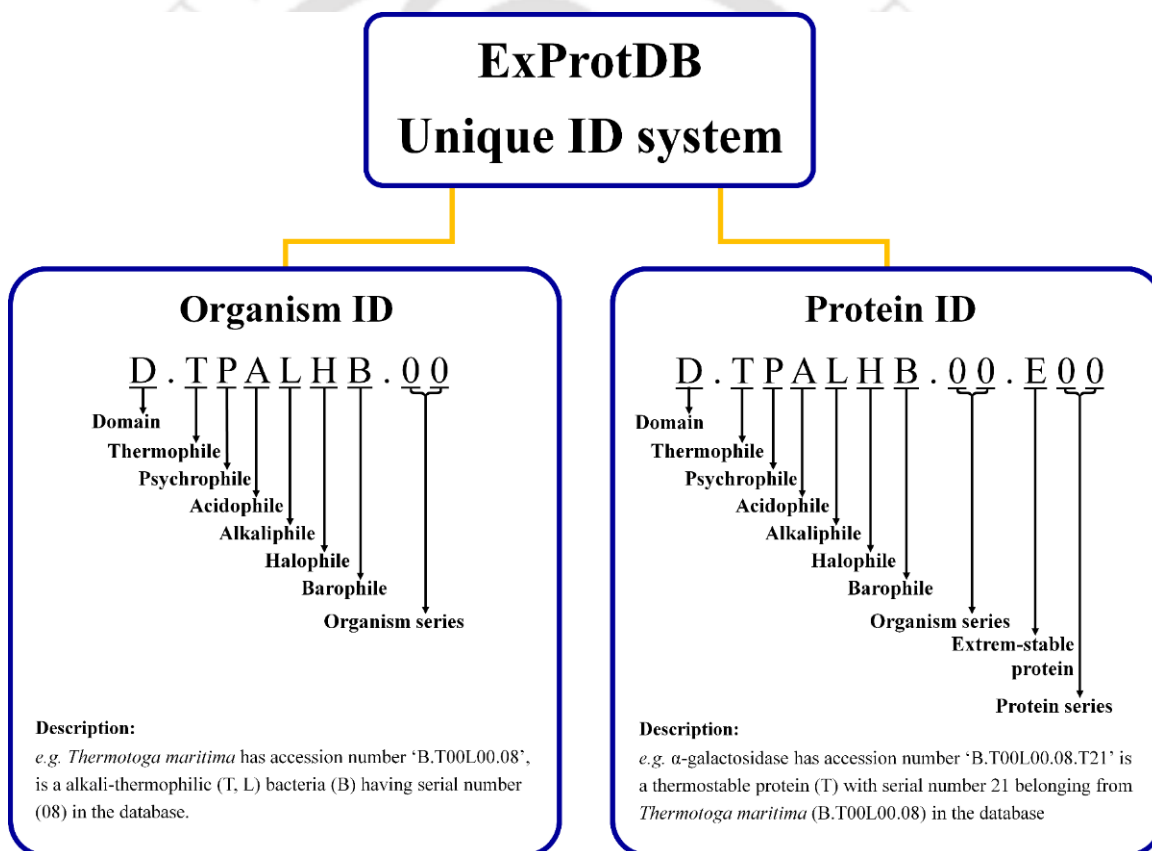


Figure 2.9: Unique ID system for extremophilic organisms and proteins.

## 2.7 Conclusion

Till date, there is a lack of a curated database for extremophiles and their proteins. Thus, this chapter presents a dedicated manually curated database having the compilation of collected genomic and proteomic data of extremophiles named as 'Extremophile Protein

Database (ExProtDB)' abbreviated as ExProtDB that can be accessed at [www.exprotodb.com](http://www.exprotodb.com). ExProtDB has been created with simple graphical user interface using a standard platform based on the Linux-Apache-MySQL-PHP (LAMP). It consists of 536 organisms (175 archaea, 255 bacteria, 100 eukaryotes and 6 viruses), 795 protein entries (375 thermophilic proteins, 115 psychrophilic proteins, 67 acidophilic proteins, 94 alkaliphilic proteins, 100 halophilic proteins and 43 barophilic proteins) and 252 engineered extreme-stable proteins along with bibliography and patents information. It is integrated with a user friendly search and browse interface for searching organism name; protein name; engineered extreme-stable proteins; literature and patents; E.C. No.; physicochemical properties; the amino acid compositions and intra-protein interaction data. The tools for calculation of amino acid compositions and intra-protein interaction can also be downloaded. ExProtDB is thus, a consolidated and comprehensive information platform for all types of extremophiles that will facilitate in data analysis to find potential applications of extremophiles in basic and applied research.

## 2.8 References

1. Majhi, M. C. *et al.* ExtremeDB: A Unified Web Repository of Extremophilic Archaea and Bacteria. *PLoS One* **8**, 1–6 (2013).
2. Shekhar, S., Chang-Tien Lu, Chawla, S. & Ravada, S. Efficient join-index-based spatial-join processing: a clustering approach. *IEEE Trans. Knowl. Data Eng.* **14**, 1400–1421 (2002).
3. Loukas, A., Kappas, I. & Abatzopoulos, T. J. HaloDom: a new database of halophiles across all life domains. *J. Biol. Res.* **25**, 2 (2018).
4. Willard, L. *et al.* VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.* **31**, 3316–9 (2003).
5. Tina, K. G., Bhadra, R. & Srinivasan, N. PIC: Protein Interactions Calculator. *Nucleic Acids Res.* **35**, W473–6 (2007).
6. Costantini, S., Colonna, G. & Facchiano, A. M. ESBRI: a web server for evaluating salt bridges in proteins. *Bioinformatics* **3**, 137–8 (2008).
7. Figueroa-Montiel, A. *et al.* In Silico Identification of Protein Disulfide Isomerase Gene Families in the De Novo Assembled Transcriptomes of Four Different Species of the Genus *Conus*. *PLoS One* **11**, e0148390 (2016).
8. Kumar, M., Thakur, V. & Raghava, G. P. S. COPid: composition based protein identification. *In Silico Biol.* **8**, 121–128 (2008).
9. Dosztányi, Z., Csizmók, V., Tompa, P. & Simon, I. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.* **347**, 827–839 (2005).
10. Goodarzi, H., Torabi, N., Najafabadi, H. S. & Archetti, M. Amino acid and codon usage profiles: Adaptive changes in the frequency of amino acids and codons. *Gene* **407**, 30–41 (2008).
11. Nakariyakul, S., Liu, Z.-P. & Chen, L. Detecting thermophilic proteins through selecting amino acid and dipeptide composition features. *Amino Acids* **42**, 1947–1953 (2012).

# 3

## Abstract

Protein stability is affected at different hierarchies – gene, amino acid sequence and structure. Gene is the first level which contributes via varying codon compositions. Codon selectivity of an organism differs with normal and extremophilic *milieu*. The aim of the present work was to elucidate the codon usage pattern in six extremophilic classes and generate ranking and classification models which differentiate extremophiles from non-extremophiles by codons. For that, homologous gene datasets of thermophile-mesophile, psychrophile-mesophile, thermophile-psychrophile, acidophile-alkaliphile, halophile-nonhalophile and barophile-nonbarophile were analysed for filtering statistically significant attributes. Relative abundance analysis, 1-9 scale ranking, nucleotide compositions, attribute weighting and machine learning algorithms were employed to arrive at findings. AGG in thermophiles and barophiles, CAA in mesophiles and psychrophiles, TGG in acidophiles, GAG in alkaliphiles and GAC in halophiles had highest preference. Preference of GC-rich and G/C-ending codons were observed in halophiles and barophiles whereas a decreasing trend was reflected in psychrophiles and alkaliphiles. GC-rich codons were found to decrease and G/C-ending codons increased in thermophiles whereas, acidophiles showed equal contents of GC-rich and G/C-ending codons. Codon usage patterns exhibited harmony among different extremophiles and has been detailed. However, the codon attribute preferences and their selectivity of extremophiles varied in comparison to non-extremophiles. The finding can be instrumental in codon optimization application for heterologous expression of extremophilic proteins.

### 3.1 Introduction

The genetic codes (or codons) serve as coding units for translation of nucleic acid information into protein sequences. In 1966, Francis Crick stated Wobble Hypothesis which describes that an amino acid of a coding protein can be coded by multiple synonymous codons (degenerate codons) during the translation process<sup>1</sup>. It reveals the third base is not spatially confined for the amino acid than the first two bases<sup>2</sup>. Thus, queries arise “why nature went for wobble hypothesis and why do different organisms prefer different codons?” Due to the degeneracy of the genetic code, many codons are synonymous for the same amino acid<sup>3,4</sup>. It reduces the diversity of cognate tRNAs which lead to reduction in the metabolic load that is beneficial for the rapid growth of an organism<sup>5</sup>. Thus, nature went for Wobble Hypothesis. Moreover, it is known that organisms show different trends of codon usage, between the genomes of different organisms, between genes of same genome and within a particular gene<sup>6</sup>. Therefore, different organisms prefer different codons. Additionally, Selection-Mutation-Drift theory further explains the variations in the usage of synonymous codons that play a key role in the co-evolution of codon usage bias and cytosolic tRNA contents, as the selection force for further optimization of gene translation for the maximized growth of an organism<sup>3,5</sup>.

Microorganisms are well-known for their exclusive ability to adapt to variable extreme environments like thermophiles, psychrophiles, halophiles, etc.<sup>7</sup> Extremophiles have developed molecular mechanisms for physicochemical adaptations towards their extreme *milieu*. Currently, research focuses mainly on genomic and proteomic level adaptations of extremophiles and each level comprise of numerous attributes which requires further exploration<sup>8,9</sup>. It was done usually through comparing their genomic features which broadly includes genome size, sequence and order of genes in the genomes, regulated expression of genes, genomic compositions and their codon usage pattern to determine the mechanisms that have produced diversity of genomes existing in different extreme *milieu*<sup>10–13</sup>. These genomic features of extremophiles will nurture better indulgent of physicochemical adaptations towards their extreme *milieu*. Moreover, the trends of codon usage in microorganisms often bring the signals of their surrounding environment. The present study is accorded towards codon preferences and their adaptations in various extreme *milieus*.

Additionally, the evolutionary adaptation to extreme *milieus* utilizes codon bias which resulted into suitable amino acid substitution for identifying molecular adaptations to particular environmental conditions<sup>14</sup>. Thus, the adaptability of proteome of extremophiles predominantly depend upon codon usage bias to adjust to the extreme conditions of the environment<sup>14,15</sup>. For translation of extremophilic proteins, those preferred codons are chosen which can combat and subsequently adapt surrounding extremophilic *milieus* for allowing the extremophile to rapidly grow in that *milieu*<sup>16,17</sup>. The adaptability of extremophilic proteins under extreme conditions is predominantly achieved by genomic adaptations via codon usage pattern<sup>16</sup>. For example, the AGR (AGG and AGA) codons are preferred since, they codes for arginine which is involved in improving the protein thermostability by enhancing number of ionic interactions and salt bridges on the protein surfaces<sup>18,19</sup>. Zeldovich et al. revealed that the codon usage pattern creates a direct link between principles of proteins stability and evolutionary mechanisms of extremophilic adaptation<sup>15</sup>.

Till date, research has been done on genomic level adaptations of extremophiles especially, thermophiles. From genomic studies it was observed that the heightened GC-content leads to DNA and protein stability under different extreme *milieus* such as in thermophiles, hyperthermophiles<sup>10,11</sup>, halophiles<sup>20</sup> and barophiles<sup>21</sup>. Most of the previous studies reported that (i) the genomes of thermophiles preferred purine-rich and G/C-ending codon due to their higher priority for Arg, Glu, Ile, Leu, Trp, Tyr and Val which is responsible for thermostability of proteins<sup>12,15</sup>; (ii) the genome of thermoacidophiles (*Picrophilus torridus*) have been evolved by increasing the pyrimidine-rich codons<sup>22</sup>; (iii) the genome of alkaliphilic bacterium *Bacillus halodurans* was observed of having less GC-content and so, they have poor usage of GC-rich codons<sup>23</sup>; (iv) GC-rich codons are significant in piezophiles<sup>21</sup>, and (v) halophiles exhibit higher content of CG dinucleotides at the first and second codon positions leading to genomic DNA stability by increasing the base stacking energy and cause an abundance of Asp, Glu, Thr and Val residues in halostable proteins<sup>20</sup>. Therefore, it is indicated that the codon usage profiles of each extremophiles showed variable adaptability and identification of codon preferences are still not understood.

The present work investigates four key questions, (i) codon usage patterns in extremophiles are significantly similar or dissimilar to that of non-extremophiles, (ii)



can relative abundance of contributing codons to extremophilicity be ranked to comprehend codon usage pattern, (iii) is there harmony in codon preference among different groups of extremophiles, and (iv) can prediction models be generated for classification of extremophiles based on their contributing codons. To investigate these questions, codon compositions was studied for comparative analysis of coding DNA sequences (CDS) of extremophiles with non-extremophiles. To further elucidate the codon usage patterns, various *in silico* approaches were employed to generate prediction models for classification of extremophilic genomes.

### 3.2 Methodology

#### 3.2.1 Creation of comparative datasets and enumeration of statistically significant codons

To study codon usage patterns, gene CDS of extremophiles were comparatively analysed with their non-extremophilic homologous counterparts. Six groups of extremophiles were searched in the available literature with keywords “thermophile”, “psychrophile”, “halophile”, “acidophile”, “alkaliphile” and “barophile” in PubMed-NCBI. The data pertaining to their protein sequences were collected from UniprotKB. But acidophilic proteins (with optimum  $\text{pH} \leq 6$ ) and alkaliphilic proteins (with optimum  $\text{pH} \geq 8$ ) were searched from BRENDA enzyme database. The homologous non-extremophilic counter protein sequences were chosen by BLAST, ClustalW, K-align and Parallel PRRN and CLUSS2 for comparative analysis of their CDS. Further, the manual curation of data resulted in non-redundant and homologous proteins. The CDS of these homologous proteins were derived for the six comparative datasets namely, T-M (thermophiles-mesophiles), P-M (psychrophiles-mesophiles), T-P (thermophiles-psychrophiles), B-Nb (Barophiles-Nonbarophiles), H-Nh (Halophiles-Nonhalophiles) and A-B (Acidophiles-Alkaliphiles) datasets (T-M, P-M, T-P, A-B, H-Nh and B-Nb) from EMBL-ENA database (Table A1.1-A1.6 of Appendix I). The percentage of codon compositions of all the 64 codons from collected CDS were normalized with their sequence length using a python script.

#### 3.2.2 Statistical analysis of significant codons in different extremophiles

The normalized codon weights were subjected to non-parametric two-sample Kolmogorov-Smirnov (KS) test of significance at 95% confidence level. This test was

performed in MATLAB. Codon with  $p$ -value less than 0.05 were retained for further analyses and considered as statistically significant. The two-sample KS statistic or D-statistic quantifies a significant distance between the empirical distribution functions of two samples<sup>24</sup>. Thus, the final dataset contained only statistically significant codons generated after KS-test which were used for further analyses.

### 3.2.3 Relative abundance analysis of codons

The generated statistically significant codons could not give the knowledge of preference of codons towards any of the comparing datasets as they might have preference to either extremophiles or non-extremophiles. The individual dataset was utilized for enumeration of relative abundance of significant codons for understanding the preferential occurrence of significant codons in comparing groups of extremophiles and non-extremophiles. This statistical enumeration was performed by comparing the weighted average of same significant codons in comparing counterparts of a datasets. It can be defined as the ratio of their weighted average difference of a chosen codon to the codon having the maximum weighted average difference in the entire set. The weighted average differences were first calculated for each significant codon corresponding to extremophilic *w.r.t.*, non-extremophilic datasets that were found to be either positive or negative. The relative abundance of a codon can be calculated using a derived equation (1):

$$\beta_{rel} = \frac{\overline{\alpha_e} - \overline{\alpha_{ne}}}{\alpha_{max}} \quad (3.1)$$

where,  $\beta_{rel}$ , relative abundance of a codon in a comparing extremophiles and non-extremophiles datasets;  $\overline{\alpha_e}$ , weighted average of a codon in extremophilic dataset;  $\overline{\alpha_{ne}}$ , weighted average of the same codon in non-extremophilic dataset;  $\alpha_{max}$ , maximum of weighted average differences in all the statistically significant codons.

Then, the derived mathematical expressions for  $\overline{\alpha_e}$ ,  $\overline{\alpha_{ne}}$  and  $\alpha_{max}$  were incorporated in the following equations (2), (3), (4):

$$\overline{\alpha_e} = \frac{\sum_{i=1}^N (\alpha_e)_i}{N} \quad (3.2)$$

$$\overline{\alpha_{ne}} = \frac{\sum_{i=1}^N (\alpha_{ne})_i}{N} \quad (3.3)$$

$$\alpha_{max} = \max \{ \alpha_j \}_{0 \leq j \leq M} \quad (3.4)$$

where,  $(\alpha_e)_i$ , statistically significant codon of  $i^{th}$  genes in extremophilic dataset;  $(\alpha_{ne})_i$ , statistically significant codon (same) of  $i^{th}$  genes in non-extremophilic dataset;  $N$ , total protein pairs in the comparing datasets;  $\alpha_j$ , weighted average difference of codon from extremophilic dataset and non-extremophilic dataset and  $M$ , total number of significant codons in the comparing datasets.

### 3.2.4 Prioritizing the codons to understand their preference in extremophiles

The most significant codons in each extremophile class were prioritized or ranked in 1-9 scale according to their contribution towards extremophilicity by using a python script (Table A1.7 of Appendix I). The generated weighted average of each codon was normalized through taking ratio of same codon of extremophile dataset *w.r.t.* their non-extremophile counterpart. The ratio weights were considered as normalized weight and were further used for deriving their 1 to 9 interval scale weight. All the ratio weights were scaled down to a 1-9 interval scale weight by a python script which uses the following equation (5):

$$W_i = \left[ \frac{(\xi_i - \alpha)}{(\beta - \alpha)} \times 8 \right] + 1 \quad (3.5)$$

where  $W_i$  is the derived weight in the 1 to 9 scale of any  $i^{th}$  significant codon in any of the comparing dataset,  $i = 1, \dots, n$  where  $n$  is the number of statistically significant codon;  $\xi_i$  is the value of the weight for  $i^{th}$  significant codon,  $\alpha$  is the minimum value in the weight for codon feature and  $\beta$  is the maximum value in the weight of codon feature. This gave the relative importance of each feature *w.r.t.* each other.

### 3.2.5 Analysis of AT- or GC-rich and A/T- or G/C-ending codons

In section “Relative abundance analysis of codons”, the codons showing positive weighted average difference showed higher preference towards extremophile and were taken up for analysing AT- or GC-rich codons and A/T- or G/C-ending codons. The percentage of AT-rich or GC-rich codons and A/T- or G/C-ending codons were estimated and normalized by total number of significant codons having positive weighted average difference. AT-rich or GC-rich codons were calculated by counting the nucleotides (A, T, G or C) in all the three positions of a codon as they should have

at least two A or T and G or C nucleotide in the codons, respectively. The analysis of A/T- or G/C-ending codons in the comparing dataset was estimated by analysing nucleotides (A, T, G or C) at third positions of the codon. Such statistical analysis is imperative for expanding the understanding of nucleotide composition of codon usage patterns and codon adaptations in different classes of extremophiles.

### 3.2.6 Analysing data-points of highest and lowest ranked codon

In section “Prioritizing the codons to understand their preference in extremophiles”, the resulted highest and lowest ranked significant codons of each datasets were used for data-points analysis by plotting their percentage score in their respective CDS. The data-points analysis was carried out for the highest and lowest ranked codon. It was estimated by normalizing with the data-points having the maximum value to have scores in the range of 0-1. Further, the data-points of highest and lowest ranked codons were separately graphically represented for each comparing extremophile datasets.

### 3.2.7 Finding codon harmony among different classes of extremophiles

To find the common codon usage pattern in extremophile types, the harmony in codon usage among six studied groups of extremophiles was analysed. On the basis of relative abundance and 1-9 scale ranking of significant codons, the positively contributing codons from the datasets were classified among six types of extremophile types to figure out codon harmonies. This analysis was considered as a conclusive output of all the previous analyses.

### 3.2.8 Generation of machine learning models to classify and predict extremophilic codons

Machine learning algorithms were used to predict, classify and generate models for extremophilic codon usages by attribute weighting, unsupervised and supervised machine learning algorithms. The datasets were subjected to test these algorithms using RapidMiner (version 5.3.000). The prediction of these algorithms were validated by split validation (70% testing and 30% training datasets)<sup>25</sup>. The employed approaches classified binary extremophile/non-extremophile datasets on the basis of their discriminating codons. Eleven different algorithms such as SVM, PCA, Correlation, Deviation, Chi squared statistic, Gini index, Information gain, Information gain ratio,

Uncertainty, Relief and Rule algorithms were applied independently on the datasets and weight the codons in a range of 0-1. The codon attributes with weight  $\geq 0.5$  were selected for analysing codon preference in extremophiles. The datasets were further subjected to unsupervised and supervised learning algorithms since attribute weighting is insufficient in generating models for codon usage pattern in extremophiles. The unsupervised clustering algorithms group the similar data-points and dissimilar data-points into separate clusters according to various criteria<sup>26</sup>. Six unsupervised clustering algorithms were applied separately on datasets such as *k*-Means, *k*-Means (kernel), *k*-Medoids, SVC, DBSCAN and EMC. Usually, the unsupervised methods fail to correctly cluster data-points into their correct and thus, to get to the most accurate model, it is necessary to apply supervised algorithms. Therefore, in supervised learning, training instances are labelled with the correct result, which gives the feedbacks in the learning process and have pre-determined categorizations. Lazy modelling (*k*-NN, Naïve Bayes), logistic regression, SVM, decision trees and ANN were applied. Logistic Regression and SVM models were generated through kernel function parameters such as dot, radial, polynomial, sigmoid and anova kernels. Four tree induction models such as Decision Tree, Decision Stump, Random Tree and Random Forest (generate trees up to 500) were applied for classification of datasets using four criteria (Gini Index, Information Gain, Gain Ratio and Accuracy)<sup>27</sup>. Additionally, CHAID, ID3 and weight-based parallel decision tree model was also run with aforementioned 11 different attribute weighting criteria. Finally, best tree induction models with highest prediction accuracy were selected for interpretation of adaptable codons in various extremophiles. Furthermore, the feed-forward neural networks were employed on the datasets that were trained by a back propagation algorithm (such as multi-layer perceptron). The parameters described for neural networks are the size of all hidden layers. The neural networks were validated with 10 times cross validation and were carried out to test the model on all patterns. The number of nodes and neurons were chosen to be more than one value with an interval of 10 specified as hidden layer size. The accuracy of prediction was obtained for each supervised learning method for categorization of codon features into two labelled attributes of extremophile and their counter non-extremophile datasets.



### 3.3 Results and Discussion

The evolution of extremophile's genomes were through lateral gene transfer from bacteria by replacing ortholog genes or incorporating paralog genes<sup>28</sup>. To understand the adaptations of extremophiles and genetic evolution, their genomes are increasingly evaluated and characterized through advanced genomic technologies and sequencing. In past 20 years, various extremophilic genomes have been studied and characterized which indicated that the selection of synonymous codons on a genome-wide scale is by mutational bias, dominant effect of nucleotide composition and dependent upon surrounding *milieus*<sup>12,29</sup>. The codon usage patterns and selectivity of an organism are correlated with its natural habitats<sup>30</sup>.

To achieve the aforementioned hypotheses, the experimental strategies were designed as (i) creation of six datasets that compared codon usage of extremophiles and non-extremophiles, followed by filtering statistically significant codons; (ii) finding relative abundance of codons among them; (iii) ranking codons in 1-9 scale; (iv) analysing nucleotide compositions of codons by computing AT- or GC-rich and A/T- or G/C-ending codons in various extremophile types; (v) investigating codon harmony among extremophiles; and (vi) generation of machine learning based classification models that differentiate extremophiles from non-extremophiles by codons. Thus, the present study employed various algorithms and software to comprehend the codon usage pattern in extremophiles.

#### 3.3.1 Dataset creation and enumeration of statistically significant codons for extremophiles

The present study commenced with the data collection of gene sequences (coding DNA sequences, CDS) of homologous extremophilic and non-extremophilic proteins. The homology search was first carried out by multiple sequence alignment approach such as BLAST, ClustalW, K-align and Parallel PRRN, but their homology resulted in poor alignment with many gap penalties due to variation in the protein sequences. Such variations may create difficulties in measuring similarity between sequences and consequently complicate the clustering task<sup>31</sup>. Therefore, non-alignment based method such as CLUSS2 (version 1.2) was chosen. Its homology search measures Substitution Matching Similarity and is especially designed for application to non-



aligned protein sequences<sup>31</sup>. The homology search leading to the formation of clusters which had similar extremophilic and non-extremophilic protein pairs were chosen as homologous pairs for six dataset creation namely, T-M (thermophiles-mesophiles), P-M (psychrophiles-mesophiles), T-P (thermophiles-psychrophiles), B-Nb (Barophiles-Nonbarophiles), H-Nh (Halophiles-Nonhalophiles) and A-B (Acidophiles-Alkaliphiles) datasets. Our group previously reported collection of the homologous T-M proteins by BLAST analysis on the basis of homology (above 70%) and the collected homologues were further confirmed by CLUSS2<sup>25</sup>. The full-length CDS of these homologous protein pairs were collected from EMBL-ENA. After data cleansing through manual curation, the resultant non-redundant homologous CDS pairs were retained in the final datasets. Therefore, the collected homologous CDS pairs in each dataset included 116 in T-M, 110 in P-M, 110 in T-P, 112 in A-B, 100 in H-Nh and 40 in B-Nb (Table 3.1 and detailed in Table A1.1-A1.6 of Appendix I). The collected CDS were used to compute percentage frequency of codons and then, filtered through non-parametric two-sample Kolmogorov–Smirnov (KS) test which measured the deviation between the distributions of two groups of a comparing dataset (Young 1977). The codon having  $p$ -value  $<0.05$  (at 95% confidence level) were considered statistically significant. The analysis revealed that out of 64 codons, 33 codons in T-M, 26 codons in P-M, 44 codons in T-P, 49 codons in A-B, 40 codons in H-Nh and 23 codons in B-Nb were found to be statistically significant (Table 3.1). In contrary to  $t$ -test and Wilcoxon rank-sum test, the KS-test performs better and is considered as preferred method to enumerate meaningful statistical difference between two populations and is used to reasonably assume whether two populations may come from the same distribution<sup>32,33</sup>. Finally, all the statistical and *in silico* analyses were performed using the final datasets having only statistically significant codons.

Table 3.1 List of collected gene CDS from homologous extremophilic and non-extremophilic proteins and enumerated statistically significant codon features obtained after KS test (with  $p < 0.05$ ).

| Comparing datasets | Number of genes (CDS) | Number of source organisms from which the CDS collected* | Data collection criteria and homology search method used                                                                                                        | Enumerated statistically significant codons by KS test (with $p$ -value $< 0.05$ , out of total 64 codons)                                                                                                                                               |
|--------------------|-----------------------|----------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| T-M                | 116 pairs             | 37 thermophiles and 51 mesophiles                        | BLAST (>70% homology) and CLUSS 2 (alignment-free algorithm)                                                                                                    | 33 (ATT, ATA, CTT, CTC, CTA, CTG, TTA, TTG, GTT, TGT, GCT, GCA, GGT, GGC, CCT, CCA, ACT, ACC, TCT, TCA, AGT, TAT, CAA, CAG, AAT, CAT, GAA, GAT, CGT, CGC, CGA, AGA, AGG)                                                                                 |
| P-M                | 110 pairs             | 27 psychrophiles and 50 mesophiles                       | CLUSS 2 (alignment-free algorithm)                                                                                                                              | 26 (AAG, AAT, AGA, AGG, AGT, ATA, ATG, CAA, CAG, CAT, CGT, CTC, CTG, GAC, GAT, GCA, GCG, GCT, GGA, GGT, GTA, TCC, TTA, TTC, TTG, TTT)                                                                                                                    |
| T-P                | 110 pairs             | 36 thermophiles and 27 psychrophiles                     | CLUSS 2 (alignment-free algorithm)                                                                                                                              | 44 (AAG, AAT, ACC, ACT, AGA, AGG, AGT, ATA, ATG, ATT, CAA, CAT, CCC, CCT, CGA, CGC, CGT, CTA, CTC, CTG, CTT, GAA, GAC, GAT, GCA, GCT, GGA, GGC, GGT, GTA, GTC, GTG, TAA, TAC, TAT, TCA, TCC, TCT, TGA, TGT, TTA, TTC, TTG, TTT)                          |
| A-B                | 112 pairs             | 73 acidophiles and 85 alkaliphiles                       | CDS of those proteins having extreme optimum pH were collected (Acid stable, pH $\leq 6$ and Alkaline stable, pH $\geq 8$ ); CLUSS 2 (alignment-free algorithm) | 49 (TTT, TTC, TTG, CTT, CTC, CTG, ATA, ATC, ATG, GTT, TGA, TCC, TCA, TCG, CCT, CCC, CCA, ACT, ACC, ACA, ACG, GCT, GCG, TAT, TAC, TAA, CAA, CAC, CAG, AAA, AAT, AAC, AAG, GAT, GAA, GAC, GAG, TGG, CGA, CGC, CGG, CTA, AGT, AGC, AGA, GGT, GGC, GGA, GTG) |
| H-Nh               | 100 pairs             | 19 halophiles and 12 non-halophiles                      | CLUSS 2 (alignment-free algorithm)                                                                                                                              | 40 (TTT, TTA, TTG, CTT, CTG, ATT, ATC, ATA, GTT, GTC, GTA, GTG, TCT, TCA, TCG, CCT, CCC, CCA, ACT, ACC, ACA, ACG, GCT, GCC, GCA, TAT, CAT, CAG, AAT, AAC, AAA, GAT, GAC, GAA, GAG, CGA, CGG, AGA, AGG, GGT)                                              |
| B-Nb               | 40 pairs              | 6 barophiles and 5 non-barophiles                        | CLUSS 2 (alignment-free algorithm)                                                                                                                              | 23 (TTT, TTC, TTA, ATT, ATA, GTC, GTA, ACT, ACA, ACG, GCA, GCG, TAC, CAA, AAT, AAA, AAG, GAA, GAG, AGT, AGA, AGG, GGG)                                                                                                                                   |

\*Not all the organisms are extremophiles but the proteins having extremophilic physicochemical behavior were also included and thus, their CDS are collected in the datasets.

### 3.3.2 Analysing relative abundance of codons among extremophiles and non-extremophiles

Relative abundance of statistically significant codons was calculated to understand codon frequency. The relative abundance was either positive or negative (Figure 3.1A-F). The positive relative abundance of codons showed higher preference towards extremophiles whereas negative relative abundance showed lower preference towards extremophile. It revealed positive relative abundance of 8 codons for thermophiles in T-M; 16 for psychrophiles in P-M; 16 for thermophiles and 26 for psychrophiles in T-P; 28 for acidophiles and 21 for alkaliphiles in A-B; 18 for halophiles in H-Nh; and 10 for barophiles in B-Nb.

Comprehensively, in T-M dataset, the codons like ATA (Ile), AGG (Arg), CTC (Leu), AGA (Arg), GAA (Glu), CTT (Leu), etc. had higher abundance and CAA (Gln) had lowest abundance in thermophiles. In P-M dataset, CAA (Gln) had highest abundance and GAC (Asp) had lowest abundance in psychrophiles. In T-P dataset, the codons like AAG (Lys), GAG (Glu), CTC (Leu), AGG (Arg), ATA (Ile), GAA (Glu), etc. had higher abundance and CAA (Gln) had lowest abundance in thermophiles and *vice versa* for psychrophiles. This clearly depicted that the codons for charged amino acids as AGG (Arg), AAG (Lys), GAG (Glu) and GAA (Glu) and aliphatic hydrophobic amino acids as ATA (Ile) and CTC (Leu) had higher frequencies in coding thermophilic proteins. CAA (Gln) had highest frequency in coding mesophilic and psychrophilic proteins. In A-B dataset, the codons like AAC (Asn), TGG (Trp), TAC (Tyr), ACT (Thr), TTC (Phe), ACC (Thr), AAT (Asn), TCC (Ser), TAT (Tyr), CAA (Gln), etc. had higher abundance and GAG (Glu), GAA (Glu), CTG (Leu), AAA (Lys), GTG (Val), GCG (Ala), AAG (Lys), CGC (Arg), etc. had lowest abundance in acidophiles and *vice versa* for alkaliphiles. Codons for small, polar and aromatic amino acid had higher frequency in acidophilic proteins whereas, charged and aliphatic amino acid codons had higher frequencies in alkaliphilic proteins. Goodarzi et al. (2008) evaluated the codon and amino acid usage in acidophile/non-acidophile and alkaliphile/non-alkaliphile showing positive and negative correlations, respectively with their surrounding environment suggesting variation in codon usage patterns in different extremophiles<sup>19</sup>. In H-Nh, GAC (Asp), GTC (Val), GAG (Glu), etc. showed abundance in

halophiles which depicted that corresponding acidic amino acids had higher frequencies in halophilic proteins. In B-Nb, AGG (Arg), GAG (Glu), AAG (Lys), ATA (Ile), TAC (Tyr), TTC (Phe), etc. had higher abundance. The results of barophiles and thermophiles were similar. Both preferred codons for charged and hydrophobic amino acids.

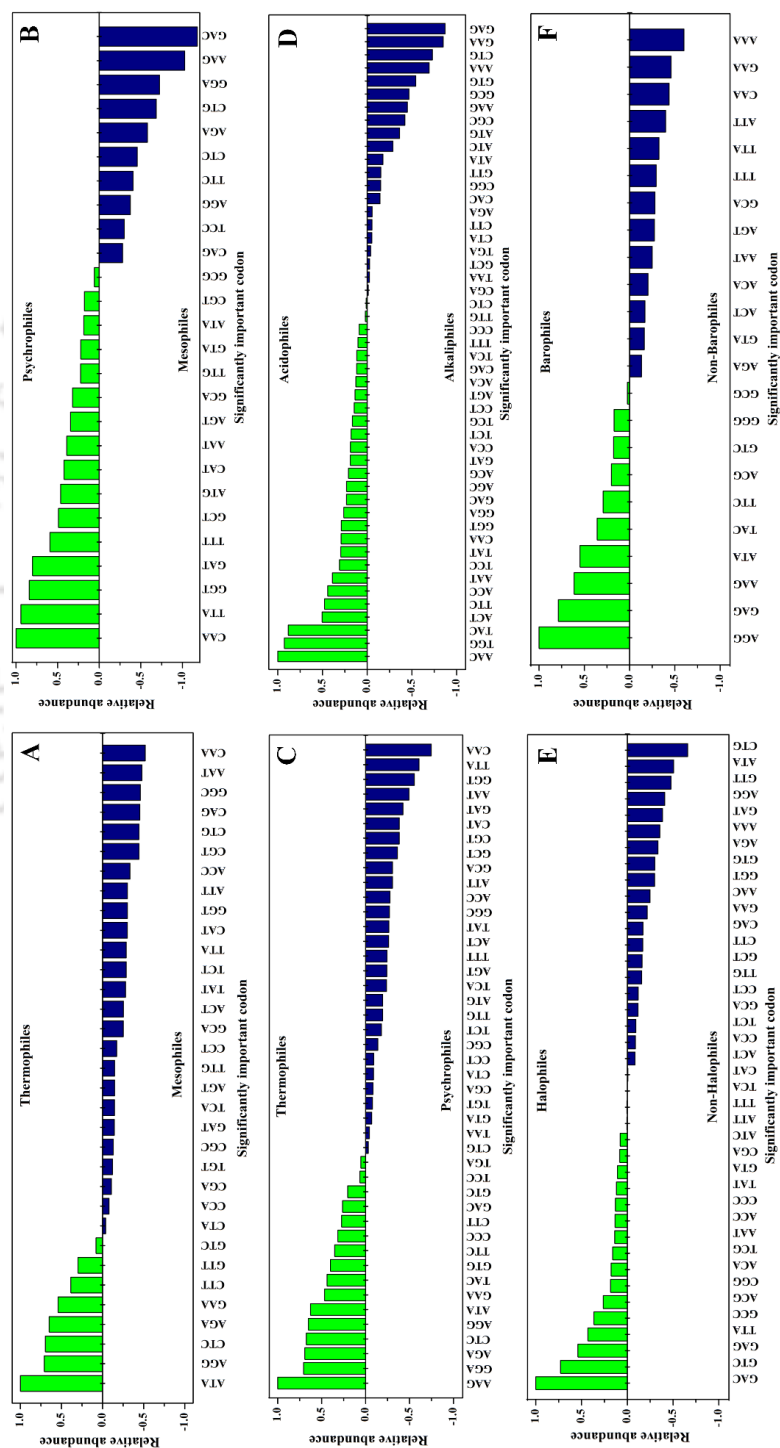


Figure 3.1: Relative abundance of statistically significant codons in the comparing datasets: (A) T-M, (B) P-M, (C) T-P, (D) A-B, (E) H-Nh and (F) B-Nb. Green colour bars represent positive contributors of main datasets and negative contributors of counter dataset whereas, dark blue colour bars represent positive contributors counter datasets and negative contributors of main dataset.

### 3.3.3 Understanding codon preferences in extremophiles by ranking them in 1-9 scale

The statistically significant codons were grouped out into 1 to 9 ranks according to their contribution towards extremophilicity in the comparing datasets, where 9 is the highest rank and 1 is the lowest rank (Figure 3.2A-F). The highest ranked and lowest ranked codons for thermophilic preference were AGG and CAA, respectively and *vice versa* for mesophiles in T-M dataset. Likewise in P-M dataset, CAA codon had highest rank in psychrophiles and lowest rank in mesophile whereas, AGG codon had lowest rank in psychrophiles and highest rank mesophiles. Correspondingly, in T-P dataset, the highest ranked and lowest ranked codons for thermophilic preference were AGG and CAA, respectively and *vice versa* for psychrophiles. The overall ranking predictions in T-M, P-M and T-P datasets showed an increasing trend of CAA codon and decreasing trend of AGG codon from thermophiles to mesophiles to psychrophiles. Correspondingly, the similar results were seen in B-Nb dataset that the highest ranked and lowest ranked codons for barophilic preference were AGG and CAA, respectively and *vice versa* for non-barophiles. These results indicated that CAA and AGG codon usage have some relevance for optimum growth temperature and pressure. It means, high pressure and temperature tolerant organisms have similar type codon adaptations for making their protein machinery thermostable and barostable. On contrary to thermophiles and barophiles, the AGG codon has lowest rank in halophiles and highest rank in non-halophiles whereas, GAC (Asp) codon had ranked highest. New finding was obtained in the A-B, as TGG (Trp) and TAC (Tyr) codon ranked highest in acidophiles and *vice versa* for alkaliphiles whereas, GAG (Glu) codon got highest rank in alkaliphiles and *vice versa* for acidophiles.

To discriminate between codon usage patterns of extremophiles and non-extremophiles, the relative abundance analysis and 1-9 scale ranking were initially employed. The results of all these analyses corroborated with one another. The comparative analysis of codon usage between thermophiles, mesophiles and psychrophiles using their respective datasets (T-M, P-M and T-P) revealed a decreased preference of AGG (Arg) codon from thermophiles to mesophiles and least to psychrophiles as well as a decreased preference of CAA (Gln) codon from psychrophiles to mesophiles and least to thermophiles. Previously, decrease usage of CGN (Arg) codons in thermophiles was reported<sup>34</sup>. This is because, if

the second nucleotide 'G' of CGN is mutated to 'A' then it codes for histidine (CAT and CAC) and glutamine (CAA and CAG) which is harmful for thermostability of proteins<sup>18</sup>. The preference of CAA codon showed detrimental effects since it codes for thermolabile residue *i.e.* glutamine are prone to deamidation spontaneously which resulted into cleavage of peptide bonds at elevated temperature<sup>35</sup>. This suggests that CAA codon is significantly preferred in psychrophiles and mesophiles rather than thermophiles. Therefore, nature selects an alternative approach to sustain thermostability by AGR (AGA and AGG) codon bias for arginine. The AGR codons have roles in protecting the protein thermostability and making the Arg usage for thermostable proteins<sup>18,34,36</sup>. Further, Liu et al. also reported that purine-rich codons usage such as AGR (Arg) codons have positive correlation with optimum growth temperature of organism and have role in protein thermostability<sup>37</sup>. The present work also determined the preference of codon other than AGG such as ATA (Ile), CTC (Leu), AGA (Arg), GAA (Glu), CTT (Leu), etc. which had higher ranks and abundance in thermophiles than mesophiles and psychrophiles, thus contribute to protein thermostability by translating to amino acids that enhances hydrophobic interactions and surface charges<sup>38</sup>. The variations in nucleotide composition lead to change in patterns of codon usage that indirectly affect thermostability<sup>39,40</sup>.



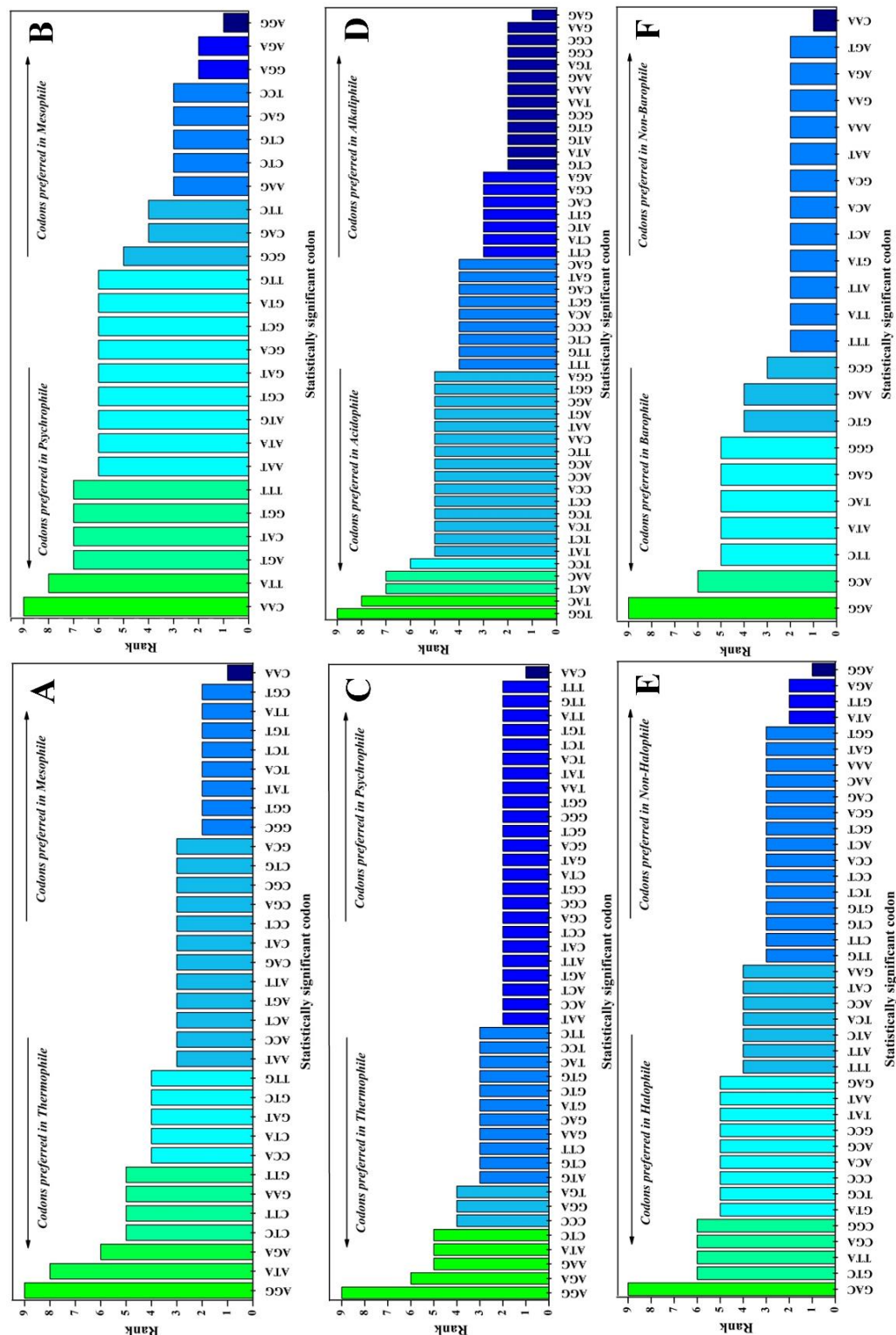


Figure 3.2: Graphical representation of ranking of statistically significant codons in the scale of 1-9 using a python script. Ranking of codons in the (A) T-M, (B) P-M, (C) T-P, (D) A-B, (E) H-Nh and (F) B-Nb.

### 3.3.4 Analysis of AT- or GC-rich and A/T- or G/C-ending codons

A review on contributing genomic factors in the different extremophiles which extensively revealed factors like high GC-rich codons, positively supercoiled DNA and robust DNA repair in thermophiles, halophiles and barophiles; purine-rich codons in thermophiles; pyrimidine-rich codons and small genome size in acidophiles showed positive contribution towards extreme-stability<sup>17</sup>. Further, the adaptability of the protein sequences, its structure and functions are primarily dependent upon their coding gene compositions<sup>16,19</sup>. Zhao et al. revealed that codon usage regulates protein stability and function<sup>41</sup>. The present study deciphers the rationale behind extremophilic codon adaptability that indirectly affects protein stability in various extreme *milieus*. Among all the extremophilic adaptations, thermophilic adaptation was understood to be the most explored and reported to be most versatile among all. For that, the nucleotide composition of significant codons were analysed to know the characteristics of extremophilic codons. Subsequently, the codon usage bias and nucleotide composition of preferred codons across extremophilic species has not been reported till date. The aforementioned codons having positive relative abundance values contributed to extremophilicity were taken into account for analysing AT- or GC-rich and A/T- or G/C-ending codons *w.r.t.* six groups of extremophiles. These preferred codons in various groups of extremophiles were counted for these nucleotide composition analysis from their respective datasets for AT-rich or GC-rich codon and wobble base analysis at third position of codon for A/T- or G/C-ending codons. The present study also analyzed nucleotide composition in extremophile codons which revealed that AT-rich codons are more significant in both thermophiles and psychrophiles but thermophiles preferred G/C-ending codons whereas, psychrophiles again preferred A/T-ending codons at wobble position. The most discriminating codons AGG is GC-rich codon in thermophiles whereas, CAA is AT-rich codons in psychrophiles and mesophiles and is systematically avoided<sup>42</sup>. Basak et al. (2004) revealed codon selection influences enhanced usage of purine-rich and GC-rich in thermophiles<sup>43</sup>. Interestingly, high G/C-base at third codon position in thermophiles corroborates the work of Singer and Hickey that G/C-ending codon in the third position are more significant than mesophiles<sup>36</sup>. This suggested that the thermophiles have AT-rich bases at first two base positions of codons not at third

position and third position is usually occupied by G/C-base. Also, it was noticed that the increased usage of GC-rich and GC-ending codons were detrimental for alkaline adaptability of proteins but did not affect acidic behaviour of the proteins at all. Chakravorty et al. deduced the GC-content jumble found in different classes of extremophilic genomes<sup>17</sup>. By all the above mentioned result it can be concluded that the basis of extreme-stability selection purely based on GC-content is ambiguous.

The statistical analysis of relative nucleotide composition of codons showed that the AT-rich codons were relatively more significant in psychrophiles (63.3%) followed by alkaliphiles (57.2%) and thermophiles (52.9%) whereas, GC-rich codons were relatively more significant in barophiles (60%) and then, halophiles (55.5%). Acidophiles showed equal proportion of AT-rich and GC-rich codons (Figure 3.3A). The A/T- or G/C-ending codons analysis revealed that psychrophiles and alkaliphiles preferred A/T ending codons whereas thermophiles, halophiles and barophiles preferred G/C-ending codons (Figure 3.3B). Similar to AT-rich and GC-rich codons analysis, acidophiles had also showed equal proportion of A/T-ending and G/C-ending codons. The results of AT- or GC-rich and A/T- or G/C-ending codons of all groups of extremophiles corroborated with each other except that of thermophiles. Contrarily, thermophiles have higher priorities of AT-rich codons but they prefer upto 60% of G/C-base at wobble position of codon. Such statistical analysis is imperative for expanding the understanding of nucleotide composition of codon usage patterns and codon adaptations in different classes of extremophiles. Correspondingly, Gunbin (2009), compared the genomic sequence through codon usage pattern of three species of the *Pyrococcus* genus: *P. horikoshii*, *P. furiosus* and *P. abyssi* which revealed great variation in the usage of GC-rich and AT-rich codon frequencies<sup>29</sup>. Wan et al. revealed that the synonymous codon usage bias was related only with the G/C-base at third position of codons in barophiles<sup>44</sup>.

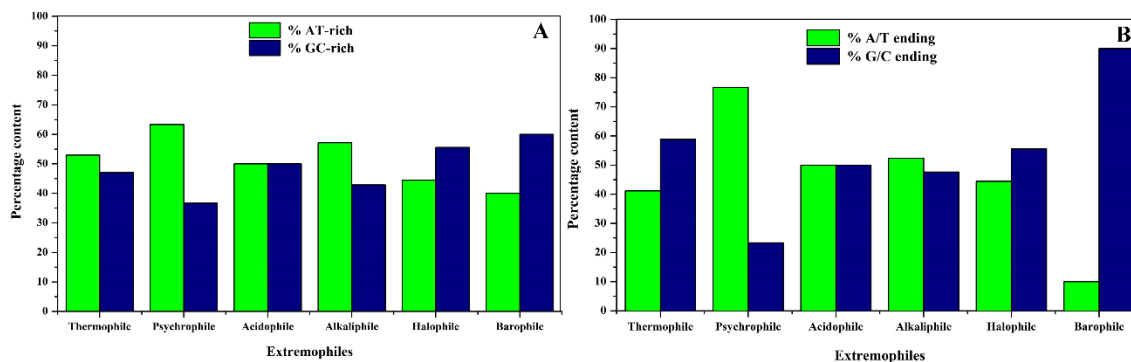
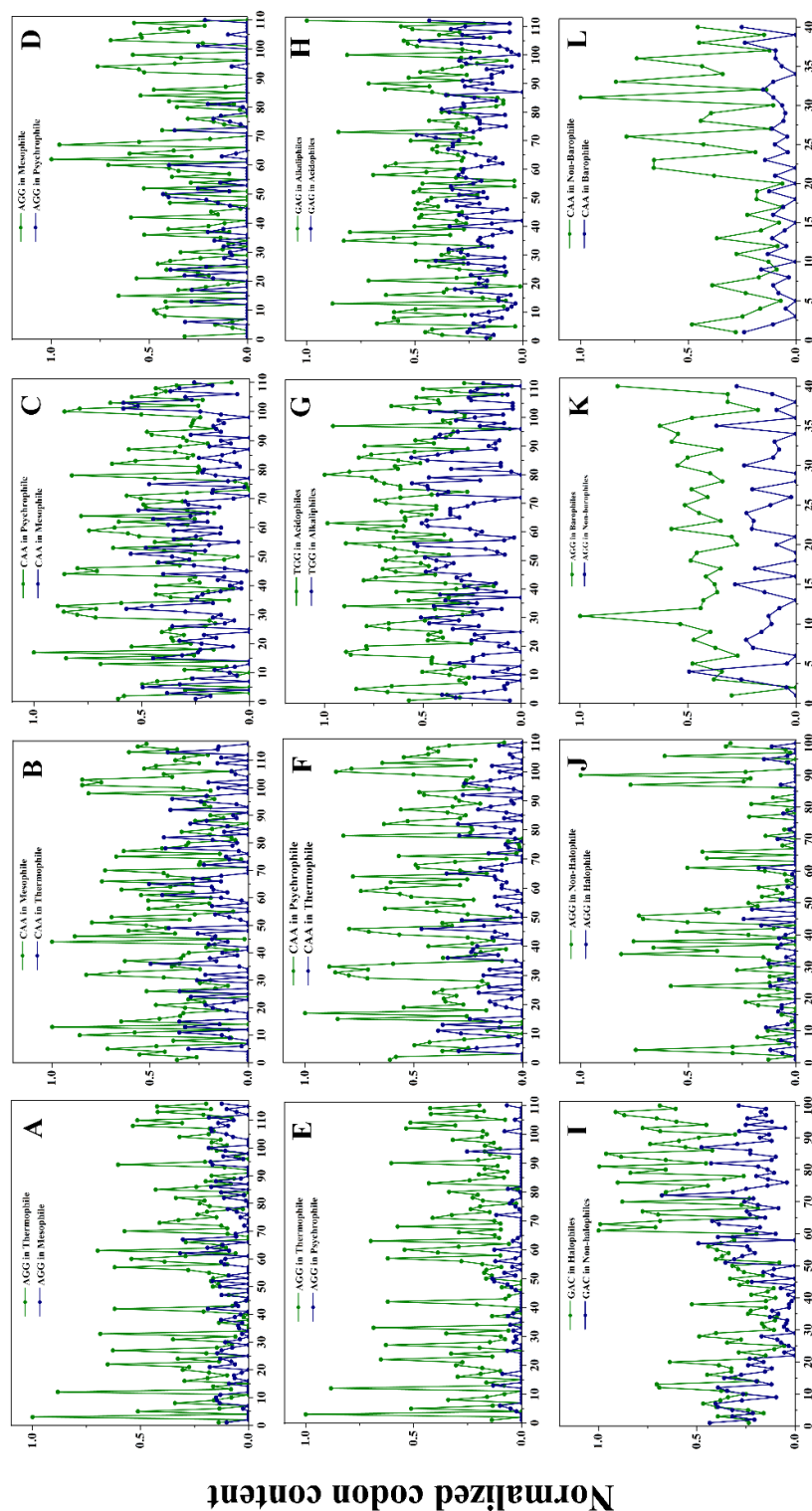


Figure 3.3: Nucleotide composition analysis considering two parameters - (A) % AT- or % GC-richness and (B) % A/T- or % G/C-ending at third wobble position in the preferred significant codons for six class of extremophiles.

### 3.3.5 Analysis of variation in the normalized data-point of highest and lowest ranked codons w.r.t. extremophiles

Synonymous codons are not used equally and the variations of codon usage are reported to be different from one gene to another. The result of 1-9 ranking analysis of codons identified highest (*i.e.* rank 9) and lowest (*i.e.* rank 1) ranked codons in the comparing datasets. Out of all the statistically significant codons, only highest and lowest ranked codons were chosen to find the variation in their percentage scores of the collected homologous CDS pair of a comparing dataset. The normalized data-points of highest and lowest ranked codons were plotted separately against their homologous CDS pairs (Figure 3.4). The analysis showed significant variability in the highest and the lowest ranked codons in all extremophile types. The ranking results presented that CAA, TGG, GAC codons were ranked highest in P-M, A-B, H-Nh, respectively and AGG codon was commonly ranked highest in T-M, T-P and B-Nb. Similarly, CAA (in T-M, T-P, B-Nb), GAG (in A-B) and AGG (in P-M, H-Nh) were ranked lowest. In Figure 3.4, the green coloured data-points represent highest ranked codons with respect to either extremophiles or non-extremophiles whereas, dark blue coloured data-points represent lowest ranked codons with respect to either extremophiles or non-extremophiles. This significant difference in normalized score of codons was because of difference in their values.



**Number of homologous gene CDS pairs**

Figure 3.4: Data-point analysis of most and least preferred codon w.r.t. extremophiles. Analysis of (A) AGG codon (most preferred w.r.t. thermophiles) of T-M dataset, (B) CAA codon (least preferred w.r.t. thermophiles) of T-M dataset, (C) CAA codon (most preferred w.r.t. psychrophiles) of P-M datasets (D) AGG codon (least preferred w.r.t. psychrophiles) of P-M datasets, (E) AGG codon (most preferred w.r.t. thermophiles) of T-P datasets (F) CAA codon (least preferred w.r.t. thermophiles) of T-P datasets, (G) TGG codon (most preferred w.r.t. acidophiles) of A-B dataset, (H) GAG codon (least preferred w.r.t. acidophiles) of A-B dataset, (I) GAC codon (most preferred w.r.t. halophiles) of H-Nh dataset (J) AGG codon (least preferred w.r.t. halophiles) of H-Nh dataset (K) AGG codon (most preferred w.r.t. barophiles) of B-Nb dataset (L) CAA codon (least preferred w.r.t. barophiles) of B-Nb dataset are represented.





### 3.3.6 Exploration of codon harmony among various extremophiles

Recent studies have shown that GAC (Asp) codon had ranked highest whereas, AGG codon has lowest rank in halophiles which conflicts to the finding of thermophiles and barophiles. Additionally, the other codons like GTC (Val), GAG (Glu), TTA (Leu), CGA (Arg), etc. also had preference in halophiles which clearly depicted that the codons code for acidic charged and aliphatic amino acids leading to higher frequencies in halophilic proteins. As seen in the preferred codons, most of them were relatively more GC-rich as well as GC-ending codons in halophiles as thermophiles and barophiles but their preferences of codons were conflicting with them. Siddiqui and Thomas also confirmed the high GC-content (60-70%) in halophiles. The advantage of increase GC-content is to avoid mutations caused by UV-induced such as thymidine dimer formation<sup>45</sup>. These results corroborated with the finding of Paul et al. revealing that the halophiles exhibit codons of distinct dinucleotides such as GA, TC, AC, GT and CG at the first and second codon positions. It reflects that such codon are abundance for expression in halophilic genes and codes for aspartate, glutamine, threonine, and valine residues in halophile proteins which leads to halostability<sup>20</sup>. The presence of such dinucleotides resulted in base stacking energy and, thus, genome stability in halophiles.

The adaptability of codons in various types of extremophiles showed commonality in codon usage patterns and is termed as codon harmony. The relative abundance analysis of codons showed 12 codons in thermophiles; 30 codons in psychrophiles; 13 codons in acidophiles; 21 codons in alkaliphile; 18 codons in halophiles and 10 codons in barophiles contributed positively and were explored for finding codon harmony among various extremophiles (Figure 3.5). The codon harmony analysis revealed GCG (Ala), CGA (Arg), GAG (Glu), TTT (Phe), CCC (Pro) and GTC (Val) codons were found to be positively contributing to maximum of three extremophiles whereas, AAC (Asn), TGC (Cys), GGA (Gly), CCG (Pro), AGC (Ser) and TAG (Stop) codons were not favoured in any extremophiles and thus, regarded as non-contributing codons. The remaining codons were preferred in two and one extremophile types. The overall analysis of all the 64 codons patterns for finding codon harmony among different extremophiles also deciphers non-extremophilic codons which neither contributes to any of the extremophile classes like

AAC (Asn), TGC (Cys), GGA (Gly), CCG (Pro), AGC (Ser) and TAG (Stop) codons. These codons are insignificant for extremophiles. Finally, the resultant higher priority codons were analysed through codon variability in the collected CDS from extremophiles and non-extremophiles proved a significant difference in the codons compositions.

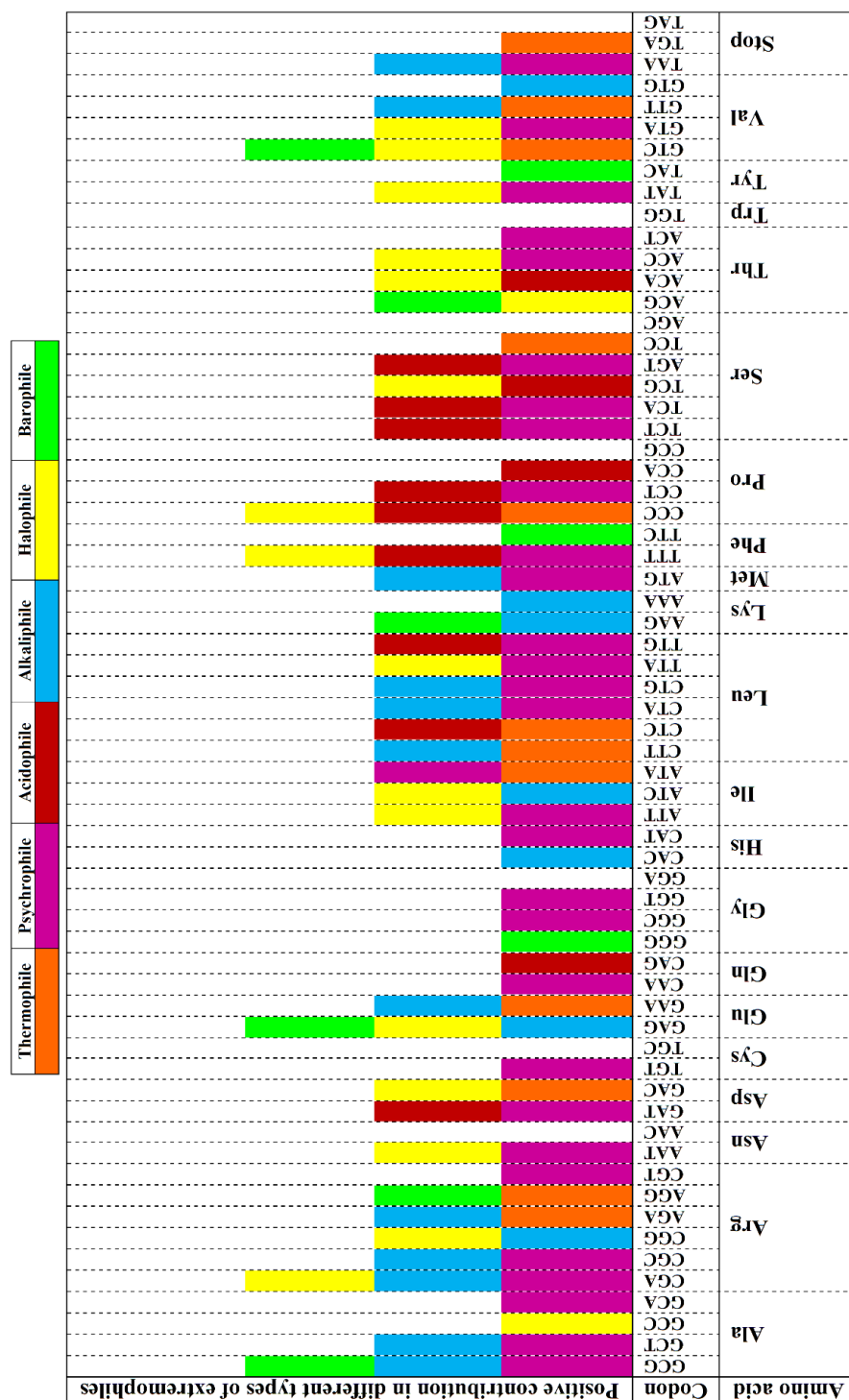


Figure 3.5: Positive contribution of codon features related to the codon harmony in extremophiles. The different types of extremophiles have been colour coded. The figure has been deduced from the relative abundance and codon ranking analysis applied on available datasets used in the present study.

### 3.3.7 Generation of machine learning models to classify and predict extremophilic codons

To find out the most essential codons which contributes to protein extreme-stability various machine learning approaches were employed. Such approaches have many applications such as analysing gene expression levels, discriminating proteins structure and function, exploring codon usage bias, predicting proteins stability, solvent accessibility, protein folding rates, etc.<sup>26,46</sup>. Unlike unsupervised clustering algorithms, the supervised algorithms predict and classify codons of extremophiles with minimum of 75% prediction accuracy of classification. Among all the employed algorithms, the study showed SVM and ANN for T-M and T-P; *k*-NN, logistic regression, ANN and decision tree in T-P; SVM in A-B; ANN in H-Nh; and, *k*-NN, SVM and decision tree in B-Nb performed outstanding in codon classification of extremophiles. Correspondingly, Ebrahimi et al. applied various machine learning algorithm to classify the 2090 protein sequences with 800 amino acid attributes which accurately differentiates proteins into correct groups of thermostable and mesostable proteins<sup>27</sup>. It was concluded that all the employed supervised machine learning classifiers had very good prospect of successful classification and prediction of codon preferences in extremophiles using training codon datasets of labelled CDS.

The present study proposes to build machine learning models which are capable of classifying and predicting statistically significant codons depending on their usability in extremophiles by comparing them with non-extremophilic counterparts. It is difficult to know the importance, usability, biasness of codons for expression of a gene in an organism *in vivo*, which is a costly and time-consuming process. Here, the machine learning models were exploited for efficient *in silico* identifications of codon usage patterns in extremophiles using available data. To distinguish the importance of codons in extremophiles, attribute weighting, unsupervised and supervised machine learning algorithms were applied, as they gave predictions with high accuracy. In attribute weighting analysis, the datasets were independently subjected to 11 different attribute weighting algorithms (Table 3.2). The analysis was performed to enumerate the number of weighting algorithms that weighted the statistically significant codons  $\geq 0.5$ , since the individual codon was weighted in the range of 0 to 1 by these algorithms. For instance, the

CAA codon of T-M dataset is weighted by 10 algorithms out of 11; it means that only one algorithm is unable to weight the CAA codon above 0.5 score. Thus, the statistically significant codons might be weighted by 0, 1, 2 or by all of the 11 algorithms. Similarly, AGA codon in P-M dataset was weighted by 8 algorithms; CAA codon in T-P dataset was weighted by 8 algorithms; TGG codon in A-B dataset was weighted by 9 algorithms; GAC codon in H-Nh dataset was weighted by 11 algorithms; and, AGG and AAG codons in B-Nb dataset were weighted equally by 10 algorithms. These most weighted codons have indicated some significance for extremophilicity but did not exhibit any preference either towards extremophiles or non-extremophiles. The results corroborated with relative abundance and 1-9 scale ranking analysis for most weighted codons.

Further, the datasets were subjected to unsupervised and supervised learning algorithms since attribute weighting was insufficient in generating models for codon usage pattern in extremophiles. The applied unsupervised clustering algorithms performed a task of dividing the labelled CDS into extremophile and non-extremophile clusters or groups based on their contributing codons in such a way that the labelled CDS in the same clusters are more similar than the labelled CDS in the other clusters. The drawback of using these models is that it often suffers from overfitting. Each comparing dataset contains two types of labelled CDS, one belongs to extremophilic group and the other to non-extremophilic counterpart (Table 3.3). The clustering analysis of *k*-means, *k*-means (kernel), *k*-medoids and EMC could partly cluster labelled CDS into distinct group. For example, T-M dataset was analysed by *k*-means algorithm and it contained 232 CDS (or 116 pairs) were distributed to cluster 0 (179 CDS) and cluster 1 (53 CDS). The 179 CDS of cluster 0 were classified as 94 thermophilic and 79 mesophilic. The remaining 53 CDS in cluster 1 were classified as 22 thermophilic and 31 mesophilic. Similar kind of results was obtained in other datasets when *k*-means, *k*-means (kernel), *k*-medoids and EMC were applied. On the other hand, DBSCAN and SVC were completely unsuccessful in clustering labelled CDS of all the comparing datasets and they clustered both the extremophilic CDS and non-extremophilic CDS into a single group *i.e.* cluster 0. The reason for failure in classification was that the minimum number of data-points required to form a dense region is not chosen appropriately for all clusters<sup>47</sup>.

Table 3.2: Summary of results obtained by using 11 different attribute weighting algorithms on different extremophilic datasets.

| T-M            |                               | P-M            |                               | T-P            |                               | A-B            |                               | H-Nh           |                               | B-Nb           |                               |
|----------------|-------------------------------|----------------|-------------------------------|----------------|-------------------------------|----------------|-------------------------------|----------------|-------------------------------|----------------|-------------------------------|
| Codon features | Algorithms weighted above 0.5 | Codon features | Algorithms weighted above 0.5 | Codon features | Algorithms weighted above 0.5 | Codon features | Algorithms weighted above 0.5 | Codon features | Algorithms weighted above 0.5 | Codon features | Algorithms weighted above 0.5 |
| CAA            | 10                            | AGA            | 8                             | CAA            | 10                            | TGG            | 9                             | GAC            | 11                            | AGG            | 10                            |
| TAT            | 9                             | AAG            | 7                             | AGA            | 9                             | AAC            | 8                             | GTC            | 9                             | AAG            | 10                            |
| CGT            | 9                             | TTA            | 7                             | CGT            | 9                             | TAC            | 8                             | TTT            | 8                             | AAA            | 8                             |
| TCT            | 7                             | GGA            | 7                             | AGG            | 9                             | GCT            | 4                             | TTA            | 6                             | AGT            | 7                             |
| AAT            | 6                             | AGG            | 7                             | GGA            | 8                             | TAT            | 4                             | AAA            | 6                             | GAG            | 5                             |
| ATA            | 5                             | GCG            | 5                             | ATA            | 7                             | GAT            | 4                             | ACG            | 6                             | ATA            | 4                             |
| ACC            | 5                             | GAC            | 4                             | AAG            | 7                             | GGT            | 4                             | AGG            | 6                             | GAA            | 4                             |
| GCT            | 5                             | GGT            | 4                             | TTA            | 6                             | GGC            | 4                             | GAG            | 5                             | CAA            | 4                             |
| CAG            | 5                             | CAA            | 4                             | GGT            | 5                             | TTC            | 3                             | AAT            | 5                             | TAC            | 4                             |
| CTG            | 5                             | TCC            | 4                             | AAT            | 4                             | CTG            | 3                             | ATA            | 4                             | TTC            | 4                             |
| GCA            | 4                             | GAT            | 3                             | GAA            | 4                             | GTT            | 3                             | ATT            | 4                             | ACG            | 3                             |
| TGT            | 4                             | ATG            | 2                             | CTG            | 4                             | TCA            | 3                             | CTG            | 4                             | ATT            | 3                             |
| TCA            | 4                             | CTG            | 2                             | TAC            | 4                             | CCA            | 3                             | AGA            | 3                             | TTA            | 2                             |
| AGT            | 4                             | CGT            | 2                             | GAT            | 3                             | ACT            | 3                             | GAT            | 3                             | TTT            | 2                             |
| CGA            | 4                             | AGT            | 1                             | CTC            | 3                             | GCC            | 3                             | GAA            | 3                             | GCA            | 1                             |
| ACT            | 4                             | CAT            | 1                             | CAT            | 3                             | CAA            | 3                             | GCC            | 3                             | GTC            | 1                             |
| TTG            | 4                             | TTT            | 1                             | TTT            | 2                             | CAG            | 3                             | GTT            | 3                             | GGG            | 1                             |
| AGA            | 4                             | AAT            | 1                             | GCT            | 2                             | AAT            | 3                             | GTA            | 2                             |                |                               |
| GGT            | 4                             | GCA            | 1                             | GAC            | 2                             | TTT            | 2                             | TAT            | 2                             |                |                               |
| CCT            | 3                             | CAG            | 1                             | CCC            | 2                             | TTG            | 2                             | TTG            | 2                             |                |                               |
| GGT            | 3                             | CAT            | 1                             | GGC            | 2                             | CTT            | 2                             | ACA            | 2                             |                |                               |
| CAT            | 3                             | ATT            | 3                             | ATT            | 2                             | CTC            | 2                             | ACC            | 2                             |                |                               |
| CTT            | 3                             | GTG            | 2                             | ACC            | 2                             | ATC            | 2                             | GTG            | 2                             |                |                               |
| GTT            | 3                             | AGT            | 1                             | AGT            | 2                             | ATG            | 2                             | GCA            | 2                             |                |                               |
| CCA            | 3                             | ACT            | 1                             | ACT            | 1                             | TGC            | 2                             | CCT            | 2                             |                |                               |
| GAT            | 3                             | TTC            | 1                             | TTC            | 1                             | CCG            | 2                             | AAC            | 2                             |                |                               |
| CTC            | 2                             | ATG            | 1                             | ATG            | 1                             | ACC            | 2                             | ATC            | 1                             |                |                               |
| GGC            | 2                             | TCC            | 1                             | TCC            | 1                             | ACA            | 2                             | CAT            | 1                             |                |                               |
| AGG            | 2                             | GCA            | 1                             | GCA            | 1                             | ACG            | 2                             | TCT            | 1                             |                |                               |
| GAA            | 2                             | TAT            | 1                             | TAT            | 1                             | GCA            | 2                             | CAG            | 1                             |                |                               |
| CTA            | 1                             | AGG            | 1                             | AGG            | 1                             | AAG            | 2                             | CCA            | 1                             |                |                               |
| TTA            | 1                             | TTA            | 1                             | AGC            | 2                             | GAC            | 2                             | TCG            | 1                             |                |                               |
|                |                               |                |                               | AGA            | 2                             | AGT            | 2                             | CGA            | 1                             |                |                               |
|                |                               |                |                               | AGC            | 2                             | AGC            | 2                             | CGG            | 1                             |                |                               |
|                |                               |                |                               | AGA            | 2                             | AGA            | 2                             | GGT            | 1                             |                |                               |
|                |                               |                |                               | TCT            | 1                             | TCT            | 1                             |                |                               |                |                               |
|                |                               |                |                               | TCG            | 1                             | TCG            | 1                             |                |                               |                |                               |
|                |                               |                |                               | CCT            | 1                             | CCT            | 1                             |                |                               |                |                               |
|                |                               |                |                               | CCC            | 1                             | CCC            | 1                             |                |                               |                |                               |
|                |                               |                |                               | CAT            | 1                             | CAT            | 1                             |                |                               |                |                               |
|                |                               |                |                               | CGA            | 1                             | CGA            | 1                             |                |                               |                |                               |
|                |                               |                |                               | CGG            | 1                             | CGG            | 1                             |                |                               |                |                               |

Table 3.3: Results of unsupervised clustering for classification of extremophiles on the basis of codon usage.

| Datasets                               | Unsupervised clustering algorithms used |                                      |                                       |                                        |                                       |                                        |
|----------------------------------------|-----------------------------------------|--------------------------------------|---------------------------------------|----------------------------------------|---------------------------------------|----------------------------------------|
|                                        | k-means                                 | k-means kernel                       | k-medoid                              | SVC                                    | EMC                                   | DBSCAN                                 |
| <b>T-M</b><br>(Total<br>CDS =<br>232)  | Cluster 0 = 173<br>T = 94<br>M = 85     | Cluster 0 = 201<br>T = 96<br>M = 105 | Cluster 0 = 143<br>T = 45<br>M = 98   | Cluster 0 = 232<br>T = 116<br>M = 116  | Cluster 0 = 220<br>T = 111<br>M = 109 | Cluster 0 = 232<br>T = 116<br>M = 116  |
|                                        | Cluster 1 = 59<br>T = 22<br>M = 37      | Cluster 1 = 31<br>T = 20<br>M = 11   | Cluster 1 = 89<br>T = 71<br>M = 18    |                                        | Cluster 1 = 12<br>T = 5<br>M = 7      |                                        |
|                                        | Cluster 0 = 155<br>P = 88<br>M = 67     | Cluster 0 = 106<br>P = 54<br>M = 52  | Cluster 0 = 140<br>P = 47<br>M = 93   | Cluster 0 = 220<br>P = 110<br>M = 110  | Cluster 0 = 178<br>P = 91<br>M = 87   | Cluster 0 = 220<br>P = 110<br>M = 110  |
| <b>P-M</b><br>(Total<br>CDS =<br>220)  | Cluster 1 = 65<br>P = 22<br>M = 43      | Cluster 1 = 134<br>P = 56<br>M = 54  | Cluster 1 = 80<br>P = 63<br>M = 17    |                                        | Cluster 1 = 42<br>P = 19<br>M = 23    |                                        |
|                                        | Cluster 0 = 183<br>T = 93<br>P = 90     | Cluster 0 = 104<br>T = 53<br>P = 51  | Cluster 0 = 1<br>T = 0<br>P = 1       | Cluster 0 = 220<br>T = 110<br>P = 110  | Cluster 0 = 176<br>T = 91<br>P = 85   | Cluster 0 = 220<br>T = 110<br>P = 110  |
|                                        | Cluster 1 = 37<br>T = 17<br>P = 20      | Cluster 1 = 116<br>T = 57<br>P = 59  | Cluster 1 = 119<br>T = 110<br>P = 109 |                                        | Cluster 1 = 44<br>T = 19<br>P = 25    |                                        |
| <b>A-B</b><br>(Total<br>CDS =<br>224)  | Cluster 0 = 146<br>A = 70<br>B = 76     | Cluster 0 = 98<br>A = 46<br>B = 52   | Cluster 0 = 151<br>A = 41<br>B = 110  | Cluster 0 = 224<br>A = 112<br>B = 112  | Cluster 0 = 180<br>A = 84<br>B = 96   | Cluster 0 = 224<br>A = 112<br>B = 112  |
|                                        | Cluster 1 = 78<br>A = 42<br>B = 36      | Cluster 1 = 126<br>A = 66<br>B = 60  | Cluster 1 = 73<br>A = 71<br>B = 2     |                                        | Cluster 1 = 44<br>A = 28<br>B = 16    |                                        |
|                                        | Cluster 0 = 131<br>H = 62<br>Nh = 69    | Cluster 0 = 91<br>H = 45<br>Nh = 46  | Cluster 0 = 140<br>H = 49<br>Nh = 91  | Cluster 0 = 200<br>H = 100<br>Nh = 100 | Cluster 0 = 175<br>H = 82<br>Nh = 93  | Cluster 0 = 200<br>H = 100<br>Nh = 100 |
| <b>H-Nh</b><br>(Total<br>CDS =<br>200) | Cluster 1 = 69<br>H = 38<br>Nh = 31     | Cluster 1 = 109<br>H = 55<br>Nh = 54 | Cluster 1 = 60<br>H = 51<br>Nh = 9    |                                        | Cluster 1 = 25<br>H = 18<br>Nh = 7    |                                        |
|                                        | Cluster 0 = 49<br>B = 15<br>Nb = 34     | Cluster 0 = 38<br>B = 17<br>Nb = 21  | Cluster 0 = 35<br>B = 31<br>Nb = 4    | Cluster 0 = 80<br>B = 40<br>Nb = 40    | Cluster 0 = 28<br>B = 22<br>Nb = 6    | Cluster 0 = 80<br>B = 40<br>Nb = 40    |
|                                        | Cluster 1 = 31<br>B = 25<br>Nb = 9      | Cluster 1 = 42<br>B = 23<br>Nb = 19  | Cluster 1 = 45<br>B = 9<br>Nb = 36    |                                        | Cluster 1 = 52<br>B = 18<br>Nb = 34   |                                        |



The supervised learning analysis showed all the model generation algorithms gave different accuracy of prediction in different extremophile datasets (Table 3.4). In T-M, SVM and ANN gave the highest prediction accuracy of 87.67%; in P-M, SVM and ANN gave the highest prediction accuracy of 80.88%; in T-P, *k*-NN, Logistic regression, ANN and Random Forest gave the highest prediction accuracy of 92.65%; in A-B, SVM gave the highest prediction accuracy of 81.23%; in H-Nh, *k*-NN and ANN gave the highest prediction accuracy of 91.61%; in B-Nb, *k*-NN, SVM and Random Forest gave the highest accuracy of prediction of 96.55% for codon classification. Interesting observation was that most of the algorithms gave accuracy of prediction for codon classification above 75% which is statistically good. In lazy modelling, *k*-NN (with  $k = 10$ ) performed well with T-M, T-P, A-B, H-Nh and B-Nb whereas, Naïve Bayes performed well only with P-M. Logistic regression with anova kernel type algorithm gave good results in T-M, P-M, A-B, H-Nh, B-Nb except T-P. T-P dataset was classified better by dot kernel type. Likewise, for performing SVM, the SVM (linear- using kernels), libSVM, c-SVC and nu-SVC were exploited for classification tasks. SVM with anova kernel gave good accuracy of prediction (87.61%) in T-M whereas, SVM with dot kernel type performed well in T-P and A-B for codon classification. LibSVM (with both c-SVC and nu-SVC type) performed well in P-M, B-Nb and H-Nh for classifying codons. The advantage of using SVMs was to classify codons depending upon extremophile types due to their high accuracy prediction, ability to deal with high-dimensional and large datasets, and their flexibility in modelling diverse sources of data<sup>48</sup>. In ANN, two hidden layers with 20 neurons in each layer achieved the highest accuracy of 87.61% in T-M whereas, in P-M and T-P two hidden layers (with 40 neurons in each layer) and one hidden layer (with 10 neurons) gave accuracy of 80.88% and 92.65%, respectively. The A-B, H-Nh and B-Nb were classified with best accuracy of 89.66% (2 hidden layers with 20 neurons in each layer), 78.85% (2 hidden layers with 30 neurons in each layer) and 91.67% (3 hidden layers with 30 neurons in each layer), respectively. The advantages of using ANN based model predictions was that (i) the dataset is processed several times during the training process of a network, as the connection weights are ever refined, and (ii) in between the input and the output layers, hidden layers of neurons may exist, which as a rule enhances the computational power of the ANN<sup>49</sup>. The main drawback with ANN based prediction is hard to adjust and debug to ensure it learns well<sup>49</sup>.

Table 3.4: Predicted accuracy of supervised learning for classification and model generation for different extremophiles on the basis of codon usage.

| Model                       | Criterion used and their percentage accuracy of prediction (%) |       |                                               |       |                                |       |                                               |       |                                               |       |                                               |       |
|-----------------------------|----------------------------------------------------------------|-------|-----------------------------------------------|-------|--------------------------------|-------|-----------------------------------------------|-------|-----------------------------------------------|-------|-----------------------------------------------|-------|
|                             | T-M                                                            |       | P-M                                           |       | T-P                            |       | A-B                                           |       | H-Nh                                          |       | B-Nb                                          |       |
| Lazy modeling               | k-NN (k=10)                                                    | 82.86 | Naïve Bayes                                   | 76.47 | k-NN (k=10)                    | 92.65 | k-NN (k=10)                                   | 71.15 | k-NN (k=10)                                   | 91.67 | k-NN (k=10)                                   | 96.55 |
|                             | Anova kernel type                                              | 78.08 | Anova kernel type                             | 75.00 | Dot kernel type                | 92.65 | Anova kernel type                             | 78.08 | Anova kernel type                             | 83.33 | Anova kernel type                             | 86.21 |
| SVM                         | Anova kernel type                                              | 87.61 | libSVM (C-SVC and nu-SVC type)                | 80.88 | Dot kernel type                | 91.81 | Dot kernel type                               | 81.23 | libSVM (c-SVC and nu-SVC type)                | 90.00 | libSVM (c-SVC and nu-SVC type)                | 96.55 |
| ANN                         | 2 hidden layer with 20 neurons in each layer                   | 87.61 | 2 hidden layers with 40 neurons in each layer | 80.88 | 1 hidden layer with 10 neurons | 92.65 | 3 hidden layers with 30 neurons in each layer | 78.85 | 2 hidden layers with 30 neurons in each layer | 91.67 | 2 hidden layers with 20 neurons in each layer | 89.66 |
| Decision Tree/Random Forest | Information Gain                                               | 78.57 | Information Gain                              | 75.00 | Information Gain               | 92.65 | Gini Index                                    | 80.77 | Gain Ratio                                    | 85.00 | Gini Index                                    | 96.55 |

The extremophile datasets were finally evaluated for model generation through various decision tree algorithms that classifies statistically significant codons into the groups of extremophile and non-extremophile attribute. Besides, a special type of decision tree prediction method was employed, known as Random Forest, which ensembles a group of decision trees by naturally combining selections and interactions of significant codons in the learning process of tree induction<sup>50</sup>. The advantages of using Random Forest were that, it is non-parametric, interpretable and efficient for learning any predictive problem in biology. The results revealed that Decision Tree and Random Forest with four classification criteria better classified codon datasets with good accuracy percentage. However, CHAID, ID3 and weight-based parallel decision tree model failed to classify codon datasets, since they generated trees without roots and leaves and thus, were discarded. Therefore, the best and most accurate trees were selected and their discrimination rules are shown in Table 3.5 and detailed in Figure A2.1-A2.6 of Appendix II. The selected tree for T-M, P-M and T-P classified codons of thermophile, mesophile and psychrophile genes was achieved using information gain criterion with 78.57%, 75.00% and 92.65% performance accuracy, respectively. In T-M and P-M, CAA (Gln) is the selection criterion for mesophiles and psychrophiles when their percentage are above 1.866% and 4.092%, respectively. Correspondingly, CAA is >1.056% (in combination with CGT >1.314%) in T-P comparison is the selection criterion for psychrophiles. The T-M tree also depicted that if the combination of percentage occurrence of CAA (Gln) is  $\leq 1.866$ , ATA (Ile) is >1.866%, CGC (Arg) is >1.866 and % CTT (Leu) is >2.823, it will fall into the thermophilic category. Also, in T-P decision tree showed a combination of percentage occurrence of CAA is  $\leq 1.056$ %, CGT (Arg) is  $\leq 1.029$ % having preferences for genes of thermophiles. Therefore, genes with high CAA contents are suspected to express mesophilic and psychrophilic proteins or less thermophilic proteins<sup>51,52</sup>. Further, in A-B dataset, the Random Forest (Gini index) gave performance accuracy of 80.77% for classification of codons of acidophiles and alkaliphiles. The tree depicted the occurrence percentage of GAG (Glu) is >4.202% and AAG (Lys) is >5.007% in a gene which codes for alkaliphilic proteins whereas, the occurrence percentage of GAG (Glu) is  $\leq 4.202$ %, CTC (Leu) >2.705% and GAT (Asp) is  $\leq 5.524$ % in a gene which codes for acidophilic proteins. In H-Nh, Decision Tree (gain ratio) gave highest accuracy of 85.00% and showed

that GAC (Asp) is the selection criterion when its occurrence frequency is greater than 8.861% for halophilic genes whereas, the combination of percentage occurrence of GAC is  $\leq 8.861\%$  and AGG (Arg) is  $> 1.441\%$  for non-halophilic genes. Finally, in B-Nb, the Random Forest (gini index) gave the highest accuracy of 96.55% for codon classification prevalent in barophiles and non-barophiles. It depicted that when composition of AGG (Arg) is  $> 3.007\%$  and ATA (Ile) is  $> 3.553\%$  in a gene, it codes for barophilic proteins, while when the composition of AGG (Arg) is  $\leq 3.007\%$ , TAC (Tyr) is  $\leq 2.105\%$  and AGT (Ser) is  $> 1.200\%$ , it codes for non-barophilic proteins. Here, the major advantages of using decision tree prediction was that it (i) reduces ambiguity in decision-making, (ii) gives alternatives for any courses of action, and (iii) is easy to interpret.

### 3.4 Conclusion

The present study shows the importance of using different computational approaches were imperative for expanding the understanding of codon usage patterns and codon adaptations in different classes of extremophiles. The advantage of the present study are that it can (i) help in understanding the codon usage patterns for functional prediction, (ii) evaluate the abundance of the cognate tRNAs in cytosolic pools of an extremophile for its optimum growth under extreme *milieu*, (iii) develop into a tool for prediction of codon and amino acid usage profiles of an organism, and (iv) used for codon optimization application for optimum selection of suitable codons in heterologous expression of extremophilic proteins. The expression of such engineered extremophilic proteins in heterologous system will further test their potentials for various industrial applications.

Table 3.5 Summary of decision tree prediction on extremophile datasets with their criteria chosen and best discriminatory rule for classification of codons.

| Dataset | Tree induction method | Criterion (algorithm) chosen | Number of models generated | Best possible discriminatory rule                                                                                                                                                               | Accuracy of prediction (%) |
|---------|-----------------------|------------------------------|----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------|
| T-M     | Decision Tree         | Information Gain             | 1                          | If % CAA ( $\leq 1.866$ ) and % ATA ( $> 1.866$ ) and % CGC ( $> 1.866$ ) and % CTT ( $> 2.823$ ) $\rightarrow$ <b>Thermophile</b><br>If % CAA ( $> 1.866$ ) $\rightarrow$ <b>Mesophile</b>     | 78.57                      |
| P-M     | Random Forest         | Information Gain             | 500 internal trees         | If % CAA ( $> 4.092$ ) $\rightarrow$ <b>Psychrophile</b><br>If % CAA ( $\leq 4.092$ ) and % GCG ( $> 0.659$ ) and % GGT ( $\leq 2.791$ ) $\rightarrow$ <b>Mesophile</b>                         | 75.00                      |
| T-P     | Random Forest         | Information Gain             | 100 internal trees         | If % CAA ( $\leq 1.056$ ) and % CGT ( $\leq 1.029$ ) $\rightarrow$ <b>Thermophile</b><br>If % CAA ( $> 1.056$ ) and % CGT ( $> 1.314$ ) $\rightarrow$ <b>Psychrophile</b>                       | 92.65                      |
| A-B     | Random Forest         | Gini Index                   | 500 internal trees         | If % GAG ( $\leq 4.202$ ) and % CTC ( $> 2.705$ ) and % GAT ( $\leq 5.524$ ) $\rightarrow$ <b>Acidophile</b><br>If % GAG ( $> 4.202$ ) and % AAG ( $> 5.007$ ) $\rightarrow$ <b>Alkaliphile</b> | 80.77                      |
| H-Nh    | Decision Tree         | Gain Ratio                   | 1                          | If % GAC ( $> 8.861$ ) $\rightarrow$ <b>Halophile</b><br>If % GAC ( $\leq 8.861$ ) and % AGG ( $> 1.441$ ) $\rightarrow$ <b>Non-halophile</b>                                                   | 85.00                      |
| B-Nb    | Random Forest         | Gini Index                   | 500 internal trees         | If % AGG ( $> 3.007$ ) and % ATA ( $> 3.553$ ) $\rightarrow$ <b>Barophile</b><br>If % AGG ( $\leq 3.007$ ) and % TAC ( $> 2.105$ ) and % AGT ( $> 1.200$ ) $\rightarrow$ <b>Non-barophile</b>   | 96.55                      |

### 3.5 References

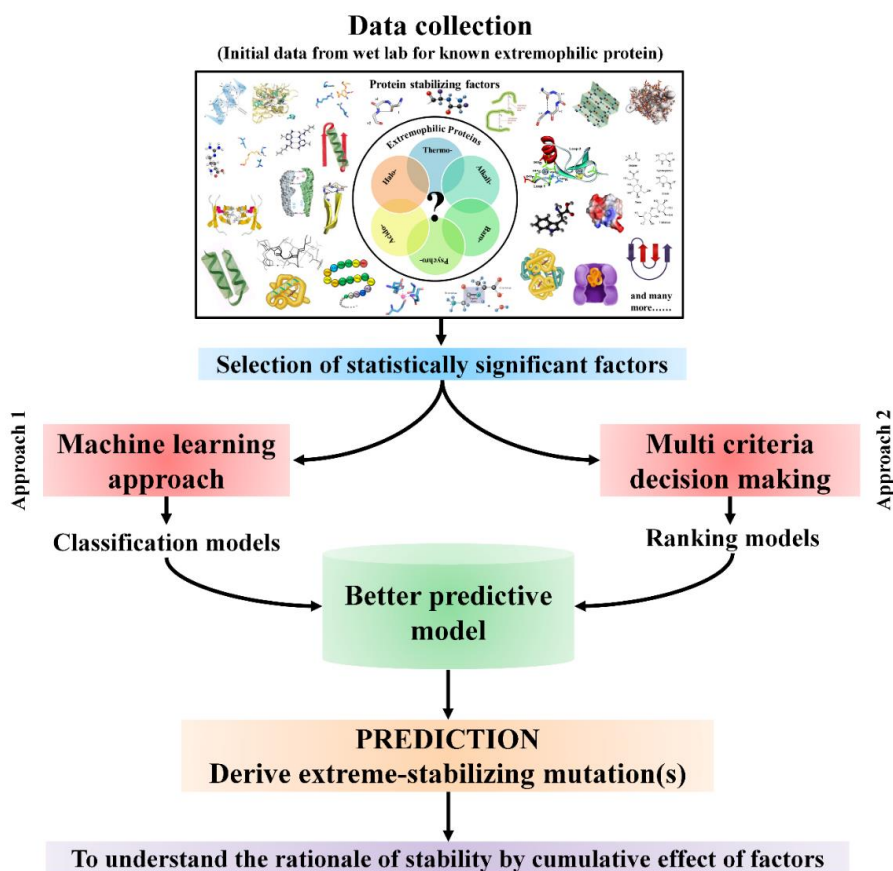
1. Crick, F. H. C. Codon—anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* **19**, 548–555 (1966).
2. Ogle, J. M., Murphy, F. V., Tarry, M. J. & Ramakrishnan, V. Selection of tRNA by the Ribosome Requires a Transition from an Open to a Closed Form. *Cell* **111**, 721–732 (2002).
3. Rocha, E. P. C. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* **14**, 2279–2286 (2004).
4. Shah, P. & Gilchrist, M. A. Effect of Correlated tRNA abundances on translation errors and evolution of codon usage bias. *PLoS Genet.* **6**, (2010).
5. Quax, T. E. F., Claassens, N. J., Söll, D. & van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell* **59**, 149–161 (2015).
6. Supek, F. & Vlahoviček, K. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics* **6**, (2005).
7. Dutta, C. & Paul, S. Microbial Lifestyle and Genome Signatures. *Curr. Genomics* **13**, 153–162 (2012).
8. Tehei, M. & Zaccai, G. Adaptation to extreme environments: Macromolecular dynamics in complex systems. *Biochimica et Biophysica Acta - General Subjects* **1724**, 404–410 (2005).
9. Debamitra, C., Kumar, S. A., Sarath, B. V. R. & Sanjukta, P. Molecular Evolution of Extremophiles. *Extremophiles* (2012). doi:doi:10.1002/9781118394144.ch1
10. Bao, Q. *et al.* A complete sequence of the *T. tengcongensis* genome. *Genome Res.* **12**, 689–700 (2002).
11. Saunders, N. F. W. *et al.* Mechanisms of thermal adaptation revealed from the genomes of the antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. *Genome Res.* **13**, 1580–1588 (2003).
12. Hickey, D. A. & Singer, G. A. C. Genomic and proteomic adaptations to growth at high temperature. *Genome Biol.* **5**, 117 (2004).
13. Oh, D. H., Dassanayake, M., Bohnert, H. J. & Cheeseman, J. M. Life at the extreme: Lessons from the genome. *Genome Biology* **13**, (2012).
14. Campanaro, S., Treu, L. & Valle, G. Protein evolution in deep sea bacteria: An analysis of amino acids substitution rates. *BMC Evol. Biol.* **8**, (2008).
15. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput. Biol.* **3**, 0062–0072 (2007).
16. Kreil, D. P. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res.* **29**, 1608–1615 (2001).
17. Chakravorty, D., Khan, M. F. & Patra, S. Multifactorial level of extremostability of proteins: can they be exploited for protein engineering? *Extremophiles* **21**, 419–444 (2017).
18. Farias, S. T. & Bonato, M. C. M. Preferred amino acids and thermostability. *Genet. Mol. Res.* **2**, 383–393 (2003).
19. Goodarzi, H., Torabi, N., Najafabadi, H. S. & Archetti, M. Amino acid and codon usage profiles: Adaptive changes in the frequency of amino acids and codons. *Gene* **407**, 30–41 (2008).
20. Paul, S., Bag, S. K., Das, S., Harvill, E. T. & Dutta, C. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol.* **9**, R70 (2008).
21. Michoud, G. & Jebbar, M. High hydrostatic pressure adaptive strategies in an obligate piezophile *Pyrococcus yamanoi*. *Sci. Rep.* **6**, (2016).
22. Baker-Austin, C. & Dopson, M. *Life in acid: pH homeostasis in acidophiles*. *Trends in Microbiology* **15**, 165–171 (2007).
23. Takami, H. *et al.* Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Res.* **28**, 4317–31 (2000).
24. Marsaglia, G., Tsang, W. W. & Wang, J. Evaluating Kolmogorov's Distribution. *J. Stat. Softw.* **8**, (2003).
25. Chakravorty, D., Khan, M. F. & Patra, S. Thermostability of Proteins Revisited Through Machine Learning Methodologies: From Nucleotide Sequence to Structure. *Current Biotechnology* **6**, 39–49 (2017).
26. Tarca, A. L., Carey, V. J., Chen, X., Romero, R. & Drăghici, S. Machine Learning and Its Applications to Biology. *PLoS Comput. Biol.* **3**, e116 (2007).
27. Ebrahimi, M., Lakizadeh, A., Agha-Golzadeh, P., Ebrahimie, E. & Ebrahimi, M. Prediction of Thermostability from Amino Acid Attributes by Combination of Clustering with Attribute Weighting: A New Vista in Engineering Enzymes. *PLoS One* **6**, e23146 (2011).
28. Allers, T. & Mevarech, M. Archaeal genetics - The third way. *Nature Reviews Genetics* **6**, 58–73 (2005).



29. Gunbin, K. V., Afonnikov, D. A. & Kolchanov, N. A. Molecular evolution of the hyperthermophilic archaea of the *Pyrococcus* genus: Analysis of adaptation to different environmental conditions. *BMC Genomics* **10**, (2009).
30. Bahir, I., Fromer, M., Prat, Y. & Linial, M. Viral adaptation to host: A proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.* **5**, (2009).
31. Kelil, A., Wang, S. & Brzezinski, R. CLUSS2: an alignment-independent algorithm for clustering protein families with multiple biological functions. *Int. J. Comput. Biol. Drug Des.* **1**, 122–140 (2008).
32. Filion, G. J. The signed Kolmogorov-Smirnov test: Why it should not be used. *GigaScience* **4**, (2015).
33. Pasieka, T. J., Woolson, R. F. & Grose, C. Viral induced fusion and syncytium formation: Measurement by the Kolmogorov-Smirnov statistical test. *J. Virol. Methods* **111**, 157–161 (2003).
34. Van Der Linden, M. G. & De Farias, S. T. Correlation between codon usage and thermostability. *Extremophiles* **10**, 479–481 (2006).
35. Nosoh, Y. & Sekiguchi, T. Protein engineering for thermostability. *Trends in Biotechnology* **8**, 16–20 (1990).
36. Singer, G. A. C. & Hickey, D. A. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. in *Gene* **317**, 39–47 (2003).
37. Liu, L., Wang, L., Zhang, Z., Wang, S. & Chen, H. Effect of codon message on xylanase thermal activity. *J. Biol. Chem.* **287**, 27183–27188 (2012).
38. Sælensminde, G., Halskau, Ø. & Jonassen, I. Amino acid contacts in proteins adapted to different temperatures: Hydrophobic interactions and surface charges play a key role. *Extremophiles* **13**, 11–20 (2009).
39. Frank, A. C. & Lobry, J. R. Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms. *Gene* **238**, 65–77 (1999).
40. Grocock, R. J. & Sharp, P. M. Synonymous codon usage in *Pseudomonas aeruginosa* PA01. *Gene* **289**, 131–139 (2002).
41. Zhao, F., Yu, C. & Liu, Y. Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx501
42. Lobry, J. R. & Nècşulea, A. Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene* **385**, 128–136 (2006).
43. Basak, S., Banerjee, T., Gupta, S. K. & Ghosh, T. C. Investigation on the causes of codon and amino acid usages variation between thermophilic *aquifex aeolicus* and mesophilic *bacillus subtilis*. *J. Biomol. Struct. Dyn.* **22**, 205–214 (2004).
44. Wan, X. F., Xu, D., Kleinhofs, A. & Zhou, J. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol. Biol.* **4**, (2004).
45. Siddiqui, K. S. & Thomas, T. *Protein adaptation in extremophiles*. (Nova Publishers, 2008).
46. Michael Gromiha, M. & Huang, L.-T. Machine Learning Algorithms for Predicting Protein Folding Rates and Stability of Mutant Proteins: Comparison with Statistical Methods. *Curr. Protein Pept. Sci.* **12**, 490–502 (2011).
47. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. in *Advances in Knowledge Discovery and Data Mining* 160–172 (2013). doi:10.1007/978-3-642-37456-2\_14
48. Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B. & Rätsch, G. Support vector machines and kernels for computational biology. *PLoS Comput. Biol.* **4**, (2008).
49. Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **49**, 1225–1231 (1996).
50. Qi, Y. Random forest for bioinformatics. in *Ensemble Machine Learning: Methods and Applications* 307–323 (2012). doi:10.1007/9781441993267\_10
51. Chakravarty, S. & Varadarajan, R. Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Lett.* **470**, 65–69 (2000).
52. Kumar, S., Tsai, C.-J. & Nussinov, R. Factors enhancing protein thermostability. *Protein Eng. Des. Sel.* **13**, 179–191 (2000).

## 4

## Analysis of attributes contributing to extreme-stability of proteins



## Abstract

After the codon contribution analysis further work in this chapter encompasses divulging the attributes contributing to extreme-stability of proteins. The questions addressed are: which is the most important among them? Can all such factors be ranked according to their role in contributing towards extreme-stability? Can these factors be exploited for classifying extremostable proteins from their normal counterparts? Can the divulged factors be exploited to arrive at pre predicted mutations leading to extremostability?

The present chapter is an effort in this direction. Extremostable behavior is relative to normal. Hence, dataset of homologous extremophilic and non-extremophilic protein pairs were comparatively analyzed for factors contributing to extreme-stability. This was followed by converging upon the statistically significant ones using Kolmogorov-Smirnov test. Further, the preferences and priorities of attributes toward extreme-stability were enumerated through relative abundance analysis, machine learning and multi-criteria decision making (Analytic Hierarchy Process, AHP) approaches. At sequence level, the analysis revealed charged amino acid (Glu, Arg, Lys) in thermophilic, alkaliphilic and barophilic proteins; small, tiny amino acids and Gln in psychrophilic proteins; aromatic and polar amino acids in acidophilic proteins; and acidic amino acids (Asp) in halophilic protein were preferred. At structural level, gamma turns, charge accessible surface area, electrostatic interactions, salt bridge and hydrophobic interactions in thermophilic, alkaliphilic and barophilic proteins; polar accessible surface area in psychrophilic proteins; and aromatic-aromatic interactions in acidophilic protein were preferred. These positively contributing attributes have been further exploited to design mutations in a mesophilic (or non-extremophilic) protein to increase the content of these attributes to endorse the prediction of generated models. This chapter reports novel contributing factors for extreme-protein stabilization, which were not reported earlier.

## 4.1 Introduction

Protein stability and energetics are the main concern of structural biology which conveys a vast array of biochemical functions dependent of the biophysical pleiotropy of protein. Protein shows different approaches to extreme-stability governed by minute differences in their biophysical properties and the *milieu* in which they exist. Most research work directed towards the significant enhancement in the understanding of extremophilic adaptation concluded that stability of extremostable proteins is a resultant of cumulative effect of various stabilizing factors and each protein shows different approaches to extreme-stability governed by their biophysical properties<sup>1</sup>. The point of ponderance is the relativity of extremophilic protein attributes to the normal counterparts and how these differences can be measured. Differences in attributes contributing towards extremostability vary amongst the different classes/types of extremophiles. There have been several reports about contribution of different factors for extreme stability, most of them being the sequence based approaches. The rapid progress in the sequencing of genomes has opened new avenues. The complete genomic sequence of an extremophilic organism furnishes information about their proteomics in terms of global amino acid compositions and amino acid usage frequency of all proteins. Various researchers have attempted studying the effect of amino acids on protein thermostability<sup>2-5</sup> found that at temperatures greater than 100 °C, the thermostability of amino acids in proteins are (V, L) > I > Y > K > H > M > T > S > W > (D, E, R, C). Further, while comparing the psychrostable proteins to mesostable counterparts, a significant higher proportion of amino acids contributing to higher protein flexibility in the coil regions of proteins, such as those with tiny/small or neutral side chains were reported<sup>6,7</sup>. It has also been noted that compared to non-halostable homologs, halostable proteins contain a more negatively charged (Glu and Asp) amino acids relative to positively charged amino acids<sup>8,9</sup>. Protein adaptation with the extreme pH organism is also dependent on their surrounding environment. The alkaliphiles are adapted to an increase in the surface exposure of acidic residues changing the net charge of the molecule towards negative.

It can be concluded that in order to optimize the protein or enzyme activity, changes of amino acids sequences and amino acid usage profile have been evolved in

extremophiles<sup>10,11</sup>. Almost all of the above reports are sequence based studies. Sequence based studies miss out the various contributing attributes which work on the primary sequence to generate the secondary and tertiary structure rendering functionality to the protein. This necessitates further insights into the biophysical features of such proteins. This can be achieved by looking into the importance of hydrogen bond, electrostatic interactions, hydrophobic interactions, disulfide bonds as a major contributor of extreme-stability. It was found that charged accessible surface area, hydrogen bonds, hydrophobic interactions, electrostatic interactions, salt bridge makes the structure of proteins more robust and stable in most of the extremophiles such as thermophiles, alkalophiles, barophiles and halophiles<sup>12–15</sup>.

Extremostable behavior is relative to normal. Hence the present chapter is accorded towards collection of the relative datasets of homologous extremophilic and non-extremophilic proteins having available 3D structures followed by differentiating the key attributes. Further, extremophilic and non-extremophilic proteins were classified through various supervised and unsupervised machine learning approaches and the attributes contributing to extreme-stability were ranked through multi-criteria decision making approach.

## 4.2 Materials and methods

The methodology begins with data collection of protein sequences and structures for feature generations from extremophilic microorganisms and its homologous non-extremophilic counterparts. It was followed by searching statistically significant features and relative abundance analysis of them in extremophilic proteins. Then, classification model were generated for prediction of extreme-stability contributing features by employing various machine learning algorithms. Further, the statistically significant features rendering proteins extreme-stability were ranked using Analytic Hierarchical Process (AHP) analysis.



#### 4.2.1 Data collection and enumeration of statistically significant features

Homologous extremophilic and non-extremophilic proteins sequences and structures were collected from UniProtKB and RCSB PDB, respectively. Manual curation of collected data resulted in non-redundancy of proteins with removal of partial and putative proteins for the analysis. Further, to choose non-extremophilic counterparts for the extremophilic protein dataset, homology was carried out by using alignment based methods (ClustalW, K-align, Parallel PRRN) and non-alignment based method (CLUSS2 version 1.2). The homologous counterparts of non-extremophilic proteins were chosen and datasets were prepared that contained homologous extremophilic and non-extremophilic proteins. The homology search leading to the formation of clusters which had similar extremophilic and non-extremophilic protein pairs were chosen as homologous pairs for six dataset creation namely, T-M (thermophiles-mesophiles), P-M (psychrophiles-mesophiles), T-P (thermophiles-psychrophiles), B-Nb (Barophiles-Nonbarophiles), H-Nh (Halophiles-Nonhalophiles) and A-B (Acidophiles-Alkaliphiles) datasets. The data pertaining to amino acid sequences and protein structures were used for feature generation using various software and tools such as PEPSTATS server, PIC webserver, VADAR, ESBRI and Promotif<sup>16</sup>. These features were collected and normalized based on the count of their frequency of occurrence in protein sequences and structures and expressed as percentage score.

#### 4.2.2 Statistical analysis of feature selection using Kolmogorov-Smirnov test

The collected normalized features were subjected to non-parametric two-sample Kolmogorov-Smirnov (KS) test for selection of statistically significant one. This test was run on MATLAB platform. Protein features with  $p$ -value less than 0.05 (at 95 % confidence level) were retained and considered statistically significant features. The two-sample KS statistic or D-statistic quantifies a significant distance between the empirical distribution functions of two samples<sup>17</sup>. KS test have the advantages that (i) the distribution of statistic does not depend on cumulative distribution function being tested, and (ii) the test is exact than other statistical test and can be adapted to give the correct  $p$ -value. They have the disadvantage that they are more sensitive to deviations near the



center of the distribution than at the tails<sup>18</sup>. Thus, the final datasets comprised of only statistically significant features generated after KS-test. To understand the differentiating protein features to extreme-stability, the selected features were further analyzed for classification and ranking model generation through various machine learning algorithms and AHP analysis. Since feature selection through KS-test is a preprocessing step to machine learning analysis for effective reduction in dimensionality, remove irrelevant data, increase learning accuracy and improve result comprehensibility<sup>19</sup>.

The protein datasets were segregated into amino acid dataset (AA) and protein structure dataset (ST) in T-P and A-B. But for P-M, B-Nb and H-Nh, both the amino acid and structural datasets were merged and represented by (AA+ST) due to lesser number of attributes obtained in AA and ST dataset in KS-test analysis. Thus, there were three types of levels such as AA, ST and AA+ST.

#### 4.2.2 Finding relative abundance of selected significant features

The enumerated statistically significant features could not give the knowledge of preference neither towards extremophilic proteins nor towards non-extremophilic proteins. To clear this dilemma, relative abundance of significant features was calculated for understanding their preference in these two comparing datasets *w.r.t.* extreme-stability of proteins. This statistical enumeration was performed by comparing the weighted mean of same significant features in their respective comparing datasets. For the same, the weighted mean differences were first calculated for each feature corresponding to extremophilic *w.r.t.*, non-extremophilic protein datasets which was followed by calculation of ratio of their weighted means difference to the feature having the highest weighted mean difference in the entire set of significant features. Mathematically, it was calculated by derived equation (4.1):

$$\beta_{rel} = \frac{\bar{\alpha}_e - \bar{\alpha}_{ne}}{\alpha_{max}} \quad (4.1)$$

where,  $\beta_{rel}$ , relative abundance of an attribute in a comparing extremophiles and non-extremophiles datasets;  $\bar{\alpha}_e$ , weighted average of an attribute in extremophilic dataset;

$\overline{\alpha_{ne}}$ , weighted average of the same attribute in non-extremophilic dataset;  $\alpha_{max}$ , maximum of weighted average differences in all the statistically significant attributes.

Then, the derived mathematical expressions for  $\overline{\alpha_e}$ ,  $\overline{\alpha_{ne}}$  and  $\alpha_{max}$  were incorporated in the following equations (4.2), (4.3), (4.4):

$$\overline{\alpha_e} = \frac{\sum_{i=1}^N (\alpha_e)_i}{N} \quad (4.2)$$

$$\overline{\alpha_{ne}} = \frac{\sum_{i=1}^N (\alpha_{ne})_i}{N} \quad (4.3)$$

$$\alpha_{max} = \max \{ \alpha_j \}_{0 \leq j \leq M} \quad (4.4)$$

where,  $(\alpha_e)_i$ , statistically significant attribute of  $i^{th}$  genes in extremophilic dataset;  $(\alpha_{ne})_i$ , statistically significant attribute (same) of  $i^{th}$  genes in non-extremophilic dataset;  $N$ , total protein pairs in the comparing datasets;  $\alpha_j$ , weighted average difference of attribute from extremophilic dataset and non-extremophilic dataset and  $M$ , total number of significant attributes in the comparing datasets. It was followed by mathematical model generation to further classify the features contributing to protein extreme-stability by subjecting the final datasets for various machine learning algorithms.

#### 4.2.3 Classification of extremophilic proteins by employing machine learning analysis using statistically significant attributes

The machine learning algorithms were used to predict, categorize and classify the features which signify their importance for extreme-stability. For the same, the final datasets were imported into RapidMiner version 5.3.000 software<sup>15,20</sup> and then subjected to various attribute weighting, unsupervised and supervised machine learning algorithms. In T-M, thermophilic (categorized as T) and mesophilic (categorized as M); in P-M psychrophilic (categorized as P) and mesophilic (categorized as M); in T-P, thermophilic (categorized as T) and psychrophilic (categorized as P); in A-B, acidophilic (categorized as A) and alkaliphilic (categorized as B); in H-Nh, halophilic (categorized as H) and non-halophilic (categorized as N); and in B-Nb, Barophilic (categorized as B) and non-barophilic (categorized as NB) proteins were set as the label attributes.

### ***Attribute weighting analysis***

Attribute weighting was used to find out important factors that contribute towards protein extreme-stability. For performing the same, 11 different algorithms of attribute weightings were applied independently on datasets that include weight by Correlation, Information gain, Information gain ratio, Rule, Deviation, Chi squared statistic, Gini index, Uncertainty, Relief, Support Vector Machine (SVM) and Principle Component Analysis (PCA) were calculated. Only importance of attributes has been identified through the application of attribute weighting but it also could not give the knowledge of preference of attributes towards either to extremophilic proteins or to non-extremophilic proteins. Therefore, attribute weighting alone is not sufficient in generating models for protein extreme-stability. Thus, the datasets were further subjected to unsupervised and supervised learning algorithms and the generated models were validated with split validation of 30 % dataset for non-overlapping training and 70 % testing.

### ***Generation of classification models by unsupervised clustering approach for extreme-stability***

Depending on various criteria, the unsupervised clustering algorithm groups the similar objects into a cluster and dissimilar objects in other clusters<sup>21</sup>. In unsupervised learning methods, the data is unlabeled and the algorithms learn to inherent structure from the input data. For performing the same, six different types of unsupervised clustering algorithms were applied separately on the final datasets. *k*-Means, *k*-Means (kernel), *k*-Medoids, Support Vector Clustering (SVC), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Expectation Maximization Clustering (EMC) were employed. The main disadvantages of unsupervised classification are that such methods are sometimes unable to cluster the attributes into their correct classes and clusters the attributes inefficiently with less accuracy<sup>16</sup>. Thus, it is essential to employ supervised machine learning algorithms to get more accurate model. For that reason, the final dataset was further subjected to supervised learning methods.

### ***Generation of classification models by supervised learning approach for extreme-stability***

Unlike unsupervised clustering, in supervised learning algorithms, all data is labeled and the algorithms learn to predict the output from the input data and such algorithms are also referred as model based clustering method<sup>22</sup>. In supervised learning, other than finding groups of data-points, they find characteristic descriptions for each group representing a concept or class<sup>23</sup>. The most frequently used induction methods comprising of  $k$ -nearest neighbor ( $k$ -NN), Naïve Bayes, logistic regression, SVM, decision trees and artificial neural networks (ANN) were employed in the present study. Similarly, as prediction accuracies may differ based on datasets, all these supervised algorithms were also employed. For tree induction, four types of algorithms were independently applied namely, Decision tree, Random tree, Random forest, Decision stump for classification of extremophilic features. Each tree induction model was run with four different criteria of Gain Ratio, Information Gain, Gini Index and Accuracy. Additionally, Iterative Dichotomiser 3 (ID3), Chi-squared Automatic Interaction Detection (CHAID) and a weight-based parallel decision tree model were also run with 11 different weighting criteria (correlation, information gain, information gain ratio, rule, deviation, chi squared statistics, gini index, uncertainty, relief, SVM, PCA). Comprehensively, the final datasets were analyzed through 40 different types of tree induction methods and total generated trees for each dataset were around ~2500. Out of them, the best accuracy prediction tree was selected for interpretation of the study outcomes. Furthermore, feed forward neural networks were applied on both the datasets. In neural network, 10 times cross validation was carried out to test the model on all patterns. The learning algorithm in all networks was back propagation. The accuracy of classification for all the employed machine learning algorithms was obtained. Followed by machine learning prediction, the features were ranked for their significance in accordance to their role in stabilizing extremophilic proteins. Therefore, the final datasets were further subjected to multi-criteria decision making approach *i.e.* Analytic Hierarchical Process (AHP).

#### 4.2.4 Ranking of attributes contributing to extreme-stability through multi-criteria decision making approach

Analytic Hierarchical Process (AHP) is a multi-criteria decision making approach, was first introduced by Professor Thomas L. Saaty in 1990. AHP is a decision support tool which can be used to solve complex decision problems. In the present study, the approach was applied as a prediction method to rank the protein extreme-stabilizing features, because it is a complex problem to rank or prioritize the multiple protein features that lead to extreme-stability to understand the characteristics of extremophilic proteins. In accordance to AHP principles, a hierarchical representation of the problem in hand was constructed<sup>24</sup>. The steps involved in ranking the features have been illustrated in Figure 4.1.

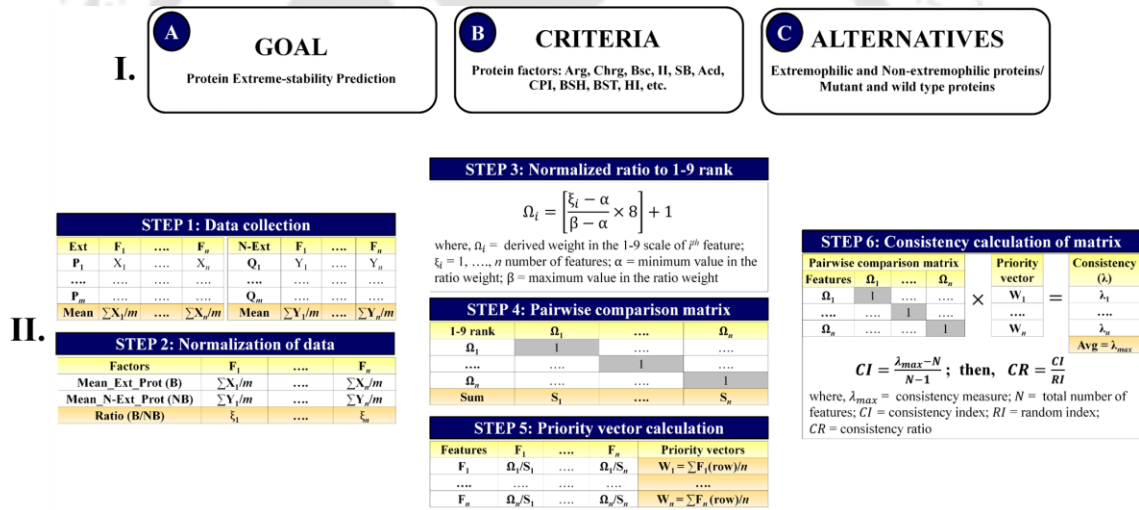


Figure 4.1: The rationalized method for prediction of barostability of proteins. (I) The hierarchy of extreme-stability is composed of three tiers: A. goal, B. criteria and C. alternatives. (II) In **Step 1**, the dataset of extremophilic (P1...P<sub>m</sub>) and non-extremophilic (Q1...Q<sub>m</sub>) protein pairs along with their protein features (F1...F<sub>n</sub>) form a matrix. Primarily arithmetic means were calculated for each feature of both extremophilic (B) and non-extremophilic (NB) proteins, where m is the total number of proteins pairs; n is the total number of features; X<sub>1</sub>...X<sub>n</sub> and Y<sub>1</sub>...Y<sub>n</sub>: numerical values of extremophilic and non-extremophilic protein attributes, respectively. In **Step 2**, the generated mean weights of each feature were normalized by taking ratio of extremophilic (ΣX<sub>i</sub>/m) and non-extremophilic (ΣY<sub>i</sub>/m) protein feature, where ξ<sub>1</sub>...ξ<sub>n</sub>: ratio values of extremophilic and non-extremophilic protein attributes. In **Step 3**, the normalized ratio was derived and converted into the 1-9 scale using python script (formula as given), where Ω<sub>1</sub>...Ω<sub>n</sub>: ratio weight of individual feature. In **Step 4**, these 1-9 scale rank score was supplied to a pairwise reciprocal comparison matrix, followed by calculating the sum of the column (S<sub>1</sub>...S<sub>n</sub>). In **Step 5**, calculation of priority vectors was performed by calculated column sum, where W<sub>1</sub>...W<sub>n</sub>: Priority vector w.r.t. feature F<sub>1</sub>...F<sub>n</sub> by taking ratio of Ω<sub>i</sub>/S<sub>i</sub>. In **Step 6**, Consistency index (CI) and consistency ratio (CR) were derived by calculating λ<sub>max</sub> which is the consistency measure of each feature that is calculated as the dot product of pairwise comparison matrix with the priority vector matrix which is then divided by n.

As illustrated in Figure 4.1(A), AHP approaches comprises of three hierarchies that include, goal, criteria and alternatives. The goal was to predict whether a protein is extremophilic when compared to its non-extremophilic homologue on the basis of selected significant features, ranking of features and generation of models that can predict extreme-stabilizing mutations for each datasets. The criteria level represented all the statistically significant features. The alternative level comprised of the proteins to be categorized as extremophilic and non-extremophilic proteins based on feature ranks. The AHP involves six major steps to generate a model that can predict extreme-stabilizing mutations, as illustrated in Figure 4.1(B). Primarily, arithmetic means were calculated for each significant feature of extremophile and non-extremophile datasets (such as T-M, P-M, T-P, A-B, H-Nh and B-Nb) of both extremophilic and non-extremophilic proteins. Then, normalization of mean weights was done by taking ratio of mean weights of the same attributes from extremophilic dataset and non-extremophilic dataset. The normalized mean weights were further scaled into 1-9 rank using python script. This scaling was needed to derive a positive reciprocal pairwise comparison matrix. This step of AHP is considered to be crucial. The matrix was further exploited for consistency measurement and priority vector calculations. The relative priority vector or eigenvectors were calculated for finding comparative preference of each feature in both extremophilic and non-extremophilic proteins. For this, column wise normalization of the matrix was first done by taking column sum followed by dividing each column element with the same column sum. This generates another matrix called priority vector calculation matrix. Here, priority vector or eigenvectors of each feature were calculated as the mean of row of generated matrix. The relative priority vectors were indicators of the relative importance of each feature over the others for their positive contribution towards extremostability of proteins. Conclusively, it can be pointed out here that higher the value the more is its impact towards rendering proteins to be extreme-stable. The next step was to calculate the consistency of the matrix. The purpose for doing this was to make sure that the original weights given to the features were consistent.  $\lambda_{\max}$  is the consistency measure of each row in the subsequent matrix which is calculated as the dot product of reciprocal pairwise comparison matrix with the calculated priority eigenvectors which is then divided by number of features (N). Finally, the Consistency Index (CI) and



Consistency Ratio (CR) was calculated using average consistency ( $\lambda_{\max}$ ) and Random Index (RI), respectively. According to Alonso and Lamata, RI is a random index and depends upon number of feature. A matrix is accepted as consistent if and only if  $CR < 0.1^{25}$ .

#### 4.2.5 Testing the performance of AHP generated ranking model by blind test

To validate the performance of AHP models, the blind tests were carried out. In order to test the protein whether, it is extremophilic or non-extremophilic, the values of significant features were imported into the model. Here, 100 extremophilic and non-extremophilic protein pairs were chosen randomly as alternatives from the existing extremophile and non-extremophile dataset. A second blind test was carried out with a set of 50 pairs of non-extremophilic proteins with another non-extremophilic counterpart chosen randomly from the existing dataset. These were also tested through the same model to derive at the ranks. The proteins were assigned ranks and the mean rank value to comparing class of the protein (e.g. extremophilic or non-extremophilic). The difference in rank values were calculated for the extremophilic/non-extremophilic set and non-extremophilic/non-extremophilic sets. Finally, the final performance were graded by accuracy percentage as how accurately the model test the protein whether it is extremophilic or non-extremophilic according to their assigned ranks. Thus, they are the global representation of performance of the method or they shows the number of sample for which class of extremophile or extremophile are classified correctly. They were calculated using False Positive (FP), False Negative (FN), True Positive (TP) and True Negative (TN) values.

$$Accuracy \% = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (4.5)$$

where, TP is the number of extremophilic proteins that our proposed model identifies them correctly as extremophilic proteins. TN is the number of non-extremophilic proteins that are identified correctly as non-extremophilic proteins. FP is the number of non-extremophilic proteins that the model identifies them wrongly as extremophilic proteins. FN is the number of extremophilic proteins that system identifies them wrongly as non-extremophilic proteins.

## 4.3 Results and discussion

Thorough literature search revealed that multiple features from both the sequence and structure levels of protein organization can be responsible for protein extreme-stability<sup>15,16,20,26</sup>. To understand the rationale behind extreme-stability, it is important to classify extremophilic proteins and non-extremophilic proteins and further prioritize the extreme-stabilizing features pertaining to sequence and structure that are responsible for their stability. Thus, various *in silico* approaches employed to predict the rationale behind extreme-stability of proteins (such as thermostability, psychrostability, acidostability, alkalistability, halostability and barostability of proteins). There are executed five independent approaches that were subjected to distinguish extremophilic proteins and non-extremophilic proteins based on their contributing attributes in order to characterize the extremophilic protein. First, enumerate the statistically significant features for predicting protein extreme-stability by KS-test. Second, compute relative abundance of statistically significant attributes. Third, predicting which protein feature is important for extreme-stability of proteins by applying various machine learning algorithms. Fourth, rank the statistically significant features in accordance to their contribution towards different extreme-stability by AHP analysis. Fifth, validation of generated ranking model through randomly choosing extremophilic and non-extremophilic as alternatives from the existing datasets. Such an extensive research on extremophilic proteins leads to expand the understanding of molecular characteristic and physicochemical adaptation of extremophilic proteins.

### 4.3.1 Data collection and statistically significant feature generation

The collected homologous protein sequences and structure (same homologous proteins were utilized as used in Chapter 3) were exploited for data collected from six classes of extremophilic proteins and their counterpart non-extremophilic proteins such as The homology search leading to the formation of clusters which had similar extremophilic and non-extremophilic protein pairs were chosen as homologous pairs for six dataset creation namely, T-M (thermophiles-mesophiles), P-M (psychrophiles-mesophiles), T-P (thermophiles-psychrophiles), B-Nb (Barophiles-Nonbarophiles), H-Nh (Halophiles-

Nonhalophiles) and A-B (Acidophiles-Alkaliphiles) datasets (details of datasets are shown in Appendix Table A1.1 – A1.6). Our group previously reported collection of the homologous T-M proteins by BLAST analysis on the basis of homology (above 70%) and the collected homologues were further confirmed by CLUSS2<sup>1</sup>. The amino acid sequences and proteins structures were collected from UniProtKB and RCSB PDB, respectively. For amino acid (AA) and protein structure (ST), datasets were created which had 29 amino acid features and 21 structure features, respectively and were collected for each extremophile/non-extremophile dataset (T-M, P-M, T-P, A-B, H-Nh and B-Nb). The ST of T-M dataset was previously studied by our group, thus ST is not included in the present thesis (Chakravorty et al. 2017a). If ST had less number of attributes then, the AA attributes were clubbed into ST and named as AA+ST. Table 4.1 enlists the collected features for AA and ST of different extremophile/non-extremophile dataset. The numerical features were generated using various software and tools which includes PEPSTATS server (for amino acid percentage), PIC Webserver, VADAR, ESBRI and Promotif (for protein structural features percentage). The tools and software employed in the study are enlisted in Table 4.2. All the amino acid and structure features were normalized with respect to the length of the protein sequence and number of atoms, respectively.

Table 4.1: Attributes used for predication of protein extreme-stability (29 AA and 21 ST).

| Protein dataset         | Attribute type                                   | Collected attributes                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|-------------------------|--------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Amino acid (AA) dataset | Standard amino acids (20)                        | Alanine (Ala or A), Arginine (arg or R), Asparagine (Asn or N), Aspartic acid (Asp or D), Cysteine (Cys or C), Glutamine (Gln or Q), Glutamic acid (Glu or E), Glycine (Gly or G), Histidine (His or H), Isoleucine (Ile or I), Leucine (Leu or L), Lysine (Lys or K), Methionine (Met or M), Phenylalanine (Phe or F), Proline (Pro or P), Serine (Ser or S), Threonine (Thr or T) Tryptophan (Trp or W) and Tyrosine (Tyr or Y) and Valine (Val or V) |
|                         | Amino acid classes (9)                           | Polar (Pol), Non-polar (NPol), Small (Sml), Tiny (Tiny), Aromatic (Aro), Aliphatic (Ali), Charged (Chrg), Basic (Bsc) and Acidic (Acd) amino acids                                                                                                                                                                                                                                                                                                      |
| Structure (ST) dataset  | Intra-protein interactions (11)                  | Total hydrogen bonds (HB), main-main chain hydrogen bonds (MMH), main-side chain hydrogen bonds (MSH), side-side chain hydrogen bonds (SSH), Hydrophobic interactions (HI), disulfide bonds (DS), Ionic interactions (II), Salt bridges (SB), aromatic-aromatic interactions (AAI), aromatic-sulfur interactions (ASI) and cation- $\pi$ interactions (CPI)                                                                                             |
|                         | Secondary and tertiary structure attributes (10) | $\alpha$ -helix (AH), $\beta$ -strands (BST), $\beta$ -hairpins (BH), $\beta$ -sheets (BSH), $\beta$ -turns (BT), $\gamma$ -turns (GT), packing volume (PV), charged accessible surface area (CASA), polar accessible surface area (PASA) and non-polar accessible surface area (NPASA)                                                                                                                                                                 |

Table 4.2: Software, tools and algorithms used for prediction studies.

| Stepwise Methods                                       | Tools and their applications                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|--------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Data collection</b>                                 | <b>UniProtKB:</b> amino acid (AA) data<br><b>RCSB PDB:</b> structural (ST) data                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| <b>Homology search</b>                                 | <b>NCBI:</b> Blastp<br><b>CLUSS version 1.2:</b> Clustering homologous proteins                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| <b>Numerical features</b>                              | <b>Python script:</b> Codon composition analysis<br><b>EMBOSS Pepstat:</b> Amino acid composition analysis<br><b>PIC, ESBRI, PDBsum or VADAR webserver:</b> Interactions and structural attribute calculations                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <b>Feature selection</b>                               | <b>MATLAB:</b> Kolmogorov–Smirnov test of significance                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Data mining and classification model generation</b> | <b>RapidMiner:</b> Attribute weighting [such as Correlation, Information gain, Information gain ratio, Rule, Deviation, Chi squared statistic, Gini index, Uncertainty, Relief, Support Vector Machine (SVM) and Principle Component Analysis (PCA)], unsupervised [such as <i>k</i> -Means, <i>k</i> -Means (kernel), <i>k</i> -Medoids, Support Vector Clustering (SVC), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Expectation Maximization Clustering (EMC)]; and supervised machine learning applications [such as -nearest neighbor ( <i>k</i> -NN) and Naïve Bayes, logistic regression, SVM, decision trees and artificial neural networks (ANN)] |
| <b>Ranking model generation</b>                        | <b>Multi-criteria Decision Making:</b> Analytic hierarchy process; 1-9 scale ranking (using python script)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |

The collected amino acid and protein structure features were filtered identifying the statistically significant features by non-parametric two-tail KS-test of significance at 95% confidence level (or  $p$ -value  $<0.05$ ). Features with  $p$ -value  $<0.05$  were retained and considered significant. This statistical test measures the discrepancy between the distributions of two comparing samples. It is used to reasonably assume whether two samples come from the same distribution<sup>27</sup>. In contrast to  $t$ -tests and the usual Wilcoxon rank-sum tests, the KS-test is a better and preferred method to enumerate whether differences in distribution between two populations are meaningful statistically<sup>28</sup>. The KS statistics is usually referred for determining an appropriate estimate of the overall variability in bipartite data<sup>27</sup>. Here, the KS-statistical analysis has been performed on MATLAB platform. The datasets were subjected to two-sample KS-test of significance at 95% confidence level. The protein features with  $p$ -value  $<0.05$  were retained and considered significant. Table 4.3 summarizes the results obtained in KS-test analysis for statistically significant protein features. The analysis revealed that the significant feature generated by 2-tail KS test were then analyzed through three approaches, (a) relative abundance analysis of attributes for extremophilic proteins and extremophilic proteins; (b) machine learning for classification model prediction; and (c) multi-criteria decision

making approach (Analytic hierarchy process, AHP) for ranking model prediction to finally be able to devise ranking models (for each extremophilic/extremophilic proteins dataset) for categorizing future proteins and prioritize the attributes.

Table 4.3: Enumerating statistically significant protein features by KS test.

| Dataset | Total features | Statistically significant codon features |                                                                                                                                                             |
|---------|----------------|------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| T-M     | 29             | 19                                       | <b>AA dataset:</b> Ala, Cys, Asp, Glu, His, Ile, Lys, Asn, Gln, Arg, Ser, Thr, Val, Tiny, Sml, Pol, Chrg, Bsc, Acd                                          |
| P-M     | 49             | 15                                       | <b>AA + ST dataset:</b> Ala, His, Met, Asn, Ser, Thr, Trp, Tiny, Ali, Aro, Bsc, Acd, NPASA, CASA, GT                                                        |
| T-P     | 29             | 21                                       | <b>AA dataset:</b> Ala, Cys, Asp, Glu, His, Lys, Met, Asn, Gln, Arg, Ser, Thr, Val, Trp, Tiny, Sml, Aro, Pol, Chrg, Bsc, Acd                                |
|         | 20             | 11                                       | <b>ST dataset:</b> GT, HI, MMH, MSH, II, CPI, SB, NPASA, PASA, CASA, PV                                                                                     |
| A-B     | 29             | 27                                       | <b>AA dataset:</b> Ala, Phe, Asp, Glu, His, Ile, Lys, Leu, Met, Asn, Pro, Gln, Arg, Ser, Thr, Val, Trp, Tyr, Tiny, Sml, NPol, Pol, Ali, Aro, Chrg, Bsc, Acd |
|         | 20             | 16                                       | <b>ST dataset:</b> II, MMH, MSH, SSH, HB, SB, PASA, CASA, PV, AAI, ASI, CPI, HI, DS, AH, BSH                                                                |
| H-Nh    | 49             | 17                                       | <b>AA+ST dataset:</b> Asp, Ile, Asn, Arg, Thr, Tiny, Sml, NPol, Pol, Chrg, Bsc, Acd, HI, CPI, NPASA, SB                                                     |
| B-Nb    | 49             | 13                                       | <b>AA+ST dataset:</b> Gln, Arg, Ser, Tiny, Chrg, Bsc, Acd, HI, II, CPI, SB, BSH, BST                                                                        |

#### 4.3.2 Relative abundance of protein feature in extremophiles

The relative abundance of features was calculated for understanding their priorities or preferences in extremophilic and non-extremophilic proteins. The features ranking for relative abundance was derived through weighted arithmetic mean followed by calculating the difference of weighted mean of same attribute from extremophilic *w.r.t.* non-extremophilic protein datasets. The calculated mean differences were either positive or negative. Therefore, the positive difference showed more preference towards extreme-stability of proteins whereas negative difference showed less preference towards extreme-stability of proteins. Further, the weighted mean differences were ordered from highest to lowest values. Finally, for obtaining relative abundance of a feature their weighted mean difference was divided by highest weighted mean difference. The relative abundance analysis was employed either to ST or AA+ST of the comparing extremophilic/non-extremophile dataset. The results are presented graphically for each dataset in Figure 4.2.

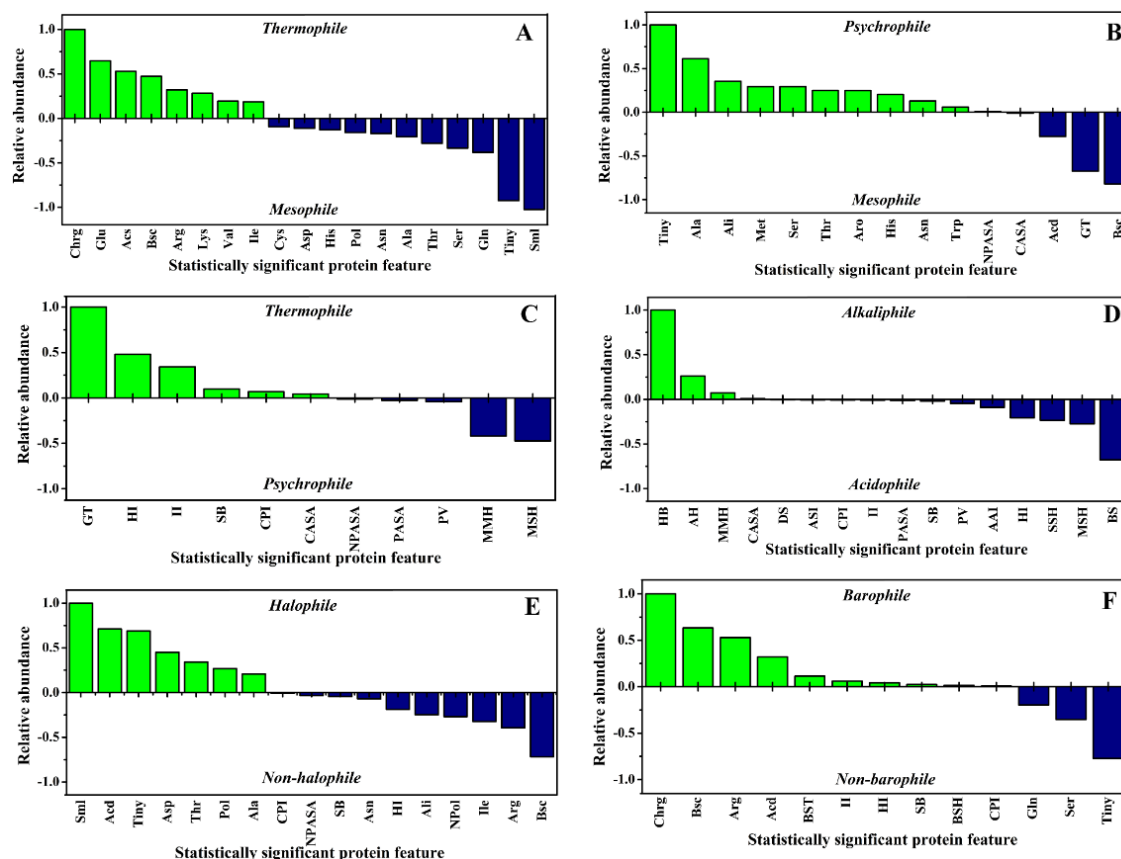


Figure 4.2: Enumerating relative abundance of statistically significant protein features for extremophilicity (A) T-M, (B) P-M, (C) T-P, (D) A-B, (E) H-Nh and (F) B-Nb. Green color bars represent positive contributors of main datasets and negative contributors of counter dataset whereas, dark blue color bars represent positive contributors counter datasets and negative contributors of main dataset.

The statistical analysis of relative abundance revealed: in T-M dataset – Chrg, Glu, Acd, Bsc, Arg, Lys, etc. are abundant in thermophilic proteins whereas Sml, Tiny, Gln, Ser, etc. are abundant in mesophilic proteins; in P-M dataset – Tiny, Ala, Ali, Met, Ser, etc. are abundant in psychrophilic proteins whereas, Bsc, GT, Acd, CASA, etc. are abundant in mesophilic proteins; in T-P dataset – GI, HI, HB, SB, etc. are abundant in thermophilic proteins whereas, MSH, MMH, PV, PASA, etc. are abundant in psychrophilic proteins; in A-B dataset – HB, AH, MMH, CASA, etc. are abundant in alkaliphilic proteins whereas, BS, MSH, SSH, HI, AAI, etc. are abundant in acidophilic proteins; in H-Nh dataset, Sml, Acd, Tiny, Asp, Thr, etc. are abundant in halophilic proteins whereas, Bsc, Arg, Ile, Npol, Ile, Ali, etc. are abundant in non-halophilic proteins; in B-Nb dataset, Chrg, Bsc, Arg, Acd, BST, II, HI, SB, etc. are abundant in barophilic proteins whereas, Tiny, Sml, Gln, etc. are abundant in non-barophilic proteins. Previous reports also



corroborate the present result showed that charged amino acid such as Glu, Arg, Lys, etc. showed higher frequency in thermophilic, alkaliphilic and barophilic proteins<sup>7,10,29</sup>, whereas, small amino, polar and acidic acids such as Ser, Asp, Gln, Thr and Ala showed higher frequency in psychrophilic and halophilic proteins<sup>7,20</sup> at sequence level. The psychrophilic proteins also comprise a significantly higher proportion of amino acids that contribute to higher protein flexibility in the coil regions of proteins, such as those with tiny/small or neutral side chains<sup>7</sup>. On contrary, another example of *Pyrococcus abyssi*, a hyperthermophilic piezophile having an increase in small amino acids (Gly, Ala, Ser, Thr, Pro, Asp and Asn) across its proteome when compared to that of the related thermophilic archaeon but non-piezophile, *Pyrococcus furiosus* have been found<sup>13</sup>. Adaptation of proteins at low pH seems to be attributed to the prevalence of acidic (negatively charged at a neutral pH) amino acids (Asp and Glu) whereas adaptation of proteins at basic pH seems to be attributed to the prevalence of basic amino acids (Lys, Arg and His) on the surface of these enzymes and proteins<sup>30</sup>. Correspondingly, Reed et al., was also reported the basic amino acids such as arginine is beneficial for stabilization of proteins under extreme temperature as well as extreme pressure condition<sup>10</sup>. In contrary to barophilic proteins, Metapally and Reddy said that in psychrophilic proteins, small amino acids such as Ser, Asp, Gln, Thr and Ala are overrepresented in the coil regions of secondary structures, whilst charged and basic side chains such as Arg and Glu are underrepresented in the helical regions<sup>7</sup>.

At structural level, hydrogen bonds, electrostatic interactions, hydrophobic interaction, charge accessible surface area, gamma turns, etc. are the most prominent factors related to the enhanced thermal stability of proteins and make the protein structure robust that can sustain extreme-conditions<sup>2,16,31</sup>. Increment in main chain-main chain hydrogen bonds and ionic interactions has been reported to increase thermostability of proteins<sup>32</sup>. The electrostatic interactions, salt bridges are among the most prominent factors related to the enhanced thermal stability of proteins from extremophiles. It was also reported that barophilic *P. abyssi* tends to substitute arginine (Arg) for all other amino acids in sequences homologous to non-barophilic *P. furiosus*<sup>13</sup>. At sequence level, this corroborates with the previous results that arginine is frequently occurred amino acid in

most of the barophilic proteins at primary levels<sup>13</sup>. One of the major finding of this work is that gamma turns increases in thermostable lipases as compared to their mesostable counterparts<sup>1</sup>.

#### 4.3.3 Machine learning model generation for extremophilic proteins

In bioinformatics, predicting protein features by machine learning algorithms have wide range of applications such as discrimination of proteins based on their structure and function, prediction of stability of mutant proteins, protein secondary structures, solvent accessibility, protein folding rates, etc.<sup>33</sup>. Previously, the multiple machine learning was applied for discriminating extremophilic proteins such as thermophilic and halophilic proteins from mesophilic proteins on the basis of their protein features<sup>1,15,20</sup>. In the present work, machine learning models were built at all the both the hierarchies of protein organization and their accuracies were compared. The aim was to understand that data at which level of protein organization can lead to the best extreme-stability predictor, which can in turn learn to differentiate proteins of these two classes (extremophilic and non-extremophilic proteins) and also characterize mutation(s) as extreme-stabilizing or not. For that, the datasets (AA, ST or AA+ST) with their features were imported into RapidMiner (version 5.3.000), and in T-M, thermophilic and mesophilic proteins (categorised as T and M); in P-M psychrophilic and mesophilic proteins (categorised as P and M); in T-P, thermophilic and psychrophilic proteins (categorised as T and P); in A-B, acidophilic and alkaliphilic proteins (categorised as A and B); in H-Nh, halophilic and halophilic proteins (categorised as H and N); and in B-Nb barophilic and non-barophilic proteins (categorised as B and NB) were set as the label attribute. The attribute weighting, unsupervised and supervised machine learning algorithms were applied to the final dataset of extremophilic and non-extremophilic proteins, as they are better at classification, prediction and give high accuracy results.

#### *Application of attribute weighting to enumerate important extreme-stabilizing features*

Interesting observations could be made when attributes were weighted by the 11 different weighting algorithms. The protein attributes with resulted weight  $\geq 0.5$  by these

algorithms were only selected and counted for interpreting the weightage of each attribute in regard to either of extremophilic or non-extremophilic proteins and the results have been summarized in Table 4.4. The result corroborate with the relative abundance analysis that most of the attribute in these two analysis are common and most preferred. In T-M dataset – Chrg, Gln, Glu, etc.; in P-M dataset – GT, Bsc, Met, etc.; in T-P dataset – Glu, Gln, Chrg, etc. (at AA level) and GT, CASA, PASA, etc. (at ST level); in A-B dataset – Trp, Aro, Chrg, etc. (at AA level) and AAI, PASA, BST, etc. (at ST level); in H-Nh dataset – Acd, Asp, Bsc, Arg, etc.; in B-Nb dataset – Chrg, Bsc, Arg, etc. were weighted most of the applied algorithms. It has been reported that charged residues increase the salt bridge formation in proteins hence stabilizes them at a higher temperature<sup>4,34</sup>. A recent study also corroborates our results as it reported *Thermus thermophilus* thermophilic proteins showed an increase in non-polar, tiny, and charged amino acids<sup>35</sup>. Gln has been reported to be a thermolabile residue and their frequency of occurrence is low in thermophilic proteins<sup>4</sup>. In contradiction to the aforesaid, researchers have reported that Gln increases hydrophilicity of thermophilic proteins and thus its frequency tends to increase<sup>15,36</sup>. It has also been reported recently that arginine is substituted by glutamine and lysine in *T. thermophilus* HB27 for enhancing its stability at elevated temperatures<sup>34</sup>. Similarly, for extreme pressure adaptation, Di Giulio reported that barophilic *P. abyssi* tends to substitute arginine (Arg) for all other amino acids in sequences homologous to non-barophilic *P. furiosus*. He considered Arg to be the “barophilic amino acid”<sup>13</sup>. Yafremava et al. also shown that Arg is preferred in barostable proteins when compared to non- barostable proteins<sup>37</sup>. The gain of glutamate and arginine residues and the loss of aspartate and lysine residues are key players in thermal, alkaline and pressure adaptation of proteins<sup>38</sup>. Again surface exposed charged and polar amino acids and buried hydrophobic and non-polar amino acids can contribute to the increase intraprotein interaction and make the protein extreme-stable. For example, higher frequency of charged residues results in enhancement of ionic interactions and charged surface area of thermostable proteins. Conclusively, the weighting of protein amino acid and structural features through machine learning approaches is as per many previous observations related to different types of extreme-stable proteins. The attribute weighting alone is not sufficient in developing a guided protocol for enhancing protein extreme-

stability and generating predictive models for protein extreme-stability. Thus, the datasets were further subjected to unsupervised and supervised learning algorithms.

### ***Unsupervised clustering to generate model for protein extreme-stability***

In order to derive an accurate classification model for extreme-stability, the extremophilic-nonextremophilic protein datasets were subjected to various unsupervised clustering algorithms ( $k$ -means,  $k$ -means (kernel),  $k$ -medoids, SVC, DBSCAN and EMC). The results have been presented in Table 4.5. The clustering analysis of  $k$ -means,  $k$ -means (kernel),  $k$ -medoids and EMC in T-M, P-M, T-P, A-B, H-Nh and B-Nb dataset revealed that the comparing protein pairs were partly distributed into two distinct clusters *i.e.* cluster 0 and cluster 1 whereas, SVC and DBSCAN completely failed to distribute comparing proteins into separated two clusters but collected into a single group *i.e.* in cluster 0. For instance, it was seen in  $k$ -mean clustering of T-M dataset, both the cluster 0 and cluster 1 had both T and M proteins. Similar results were obtained with other tested unsupervised clustering algorithms in different dataset. Thus, these clustering methods failed to correctly cluster proteins into separate classes. These algorithms showed large differences in densities of data points due to which minimum number of data points formed dense region in both the clusters of comparative proteins<sup>39</sup>. They performed well with large-sized datasets that shows better clustering, reduces noise, non-sensitive to outliers and gives higher accuracy of clustering<sup>15,40</sup>. For instance, EMC algorithm clustered more accurately a dataset of 2090 thermostable and mesophilic proteins with 800 amino acid attributes. With larger datasets, the unsupervised learning algorithms are trying to find the connection between two comparing datasets by increasing learning task exponentially in the number of steps and clustered data with deep hierarchies<sup>41</sup>. Thus, it can be concluded that unsupervised clustering algorithms are biased towards big datasets. This necessitates employing other machine learning methodologies to get to the most accurate model.

Table 4.4: Attribute weighting analysis of protein features.

| T-M              |                               | P-M              |                               | T-P              |                               | A-B              |                               | H-Nh             |                               | B-Nb             |                               |
|------------------|-------------------------------|------------------|-------------------------------|------------------|-------------------------------|------------------|-------------------------------|------------------|-------------------------------|------------------|-------------------------------|
| Protein features | Algorithms weighted above 0.5 | Protein features | Algorithms weighted above 0.5 | Protein features | Algorithms weighted above 0.5 | Protein features | Algorithms weighted above 0.5 | Protein features | Algorithms weighted above 0.5 | Protein features | Algorithms weighted above 0.5 |
| AA dataset       |                               | AA+ST dataset    |                               | AA dataset       |                               | AA dataset       |                               | AA+ST dataset    |                               | AA+ST dataset    |                               |
| Chrg             | 8                             | GT               | 11                            | Glu              | 9                             | Trp              | 8                             | Acid             | 9                             | Chrg             | 9                             |
| Gln              | 8                             | Bsc              | 7                             | Gln              | 8                             | Aro              | 8                             | Asp              | 8                             | Arg              | 8                             |
| Glu              | 7                             | Met              | 6                             | Chrg             | 8                             | Chrg             | 8                             | Bsc              | 8                             | Bsc              | 7                             |
| Tiny             | 6                             | Ala              | 4                             | Tiny             | 7                             | Bsc              | 7                             | Arg              | 7                             | Tiny             | 7                             |
| Acid             | 5                             |                  |                               | Bsc              | 7                             | Asn              | 6                             | Sml              | 6                             | Acid             | 6                             |
|                  |                               |                  |                               | ST dataset       |                               | ST dataset       |                               |                  |                               |                  |                               |
|                  |                               |                  |                               | GT               | 9                             | AAI              | 9                             | Ile              | 4                             | II               | 4                             |
|                  |                               |                  |                               | CASA             | 4                             | PASA             | 7                             | NPol             | 4                             | SB               | 3                             |
|                  |                               |                  |                               | PASA             | 3                             | BS               | 7                             | HI               | 3                             |                  |                               |
|                  |                               |                  |                               |                  |                               | MSH              | 4                             |                  |                               |                  |                               |
|                  |                               |                  |                               |                  |                               | CASA             | 3                             |                  |                               |                  |                               |
|                  |                               |                  |                               |                  |                               | SB               | 2                             |                  |                               |                  |                               |

Table 4.5: Results of unsupervised clustering for classification of extremophiles on the basis of codon usage.

| Datasets                                      | Unsupervised clustering algorithms used |                                      |                                       |                                        |                                      |                                        |
|-----------------------------------------------|-----------------------------------------|--------------------------------------|---------------------------------------|----------------------------------------|--------------------------------------|----------------------------------------|
|                                               | k-means                                 | k-means kernel                       | k-medoid                              | SVC                                    | EMC                                  | DBSCAN                                 |
| <b>T-M</b><br>(Total number of protein = 232) | Cluster 0 = 159<br>T = 84<br>M = 75     | Cluster 0 = 181<br>T = 86<br>M = 95  | Cluster 0 = 123<br>T = 35<br>M = 88   | Cluster 0 = 232<br>T = 116<br>M = 116  | Cluster 0 = 200<br>T = 101<br>M = 99 | Cluster 0 = 232<br>T = 116<br>M = 116  |
|                                               | Cluster 1 = 73<br>T = 32<br>M = 41      | Cluster 1 = 71<br>T = 40<br>M = 31   | Cluster 1 = 109<br>T = 81<br>M = 28   |                                        | Cluster 1 = 32<br>T = 15<br>M = 17   |                                        |
|                                               | Cluster 0 = 135<br>P = 78<br>M = 57     | Cluster 0 = 66<br>P = 34<br>M = 32   | Cluster 0 = 120<br>P = 37<br>M = 83   | Cluster 0 = 220<br>P = 110<br>M = 110  | Cluster 0 = 158<br>P = 81<br>M = 77  | Cluster 0 = 220<br>P = 110<br>M = 110  |
|                                               | Cluster 1 = 85<br>P = 32<br>M = 53      | Cluster 1 = 174<br>P = 76<br>M = 74  | Cluster 1 = 100<br>P = 73<br>M = 27   |                                        | Cluster 1 = 62<br>P = 29<br>M = 33   |                                        |
| <b>T-P</b><br>(Total number of protein = 220) | Cluster 0 = 163<br>T = 83<br>P = 80     | Cluster 0 = 84<br>T = 43<br>P = 41   | Cluster 0 = 2<br>T = 1<br>P = 1       | Cluster 0 = 220<br>T = 110<br>P = 110  | Cluster 0 = 156<br>T = 81<br>P = 75  | Cluster 0 = 220<br>T = 110<br>P = 110  |
|                                               | Cluster 1 = 57<br>T = 27<br>P = 30      | Cluster 1 = 136<br>T = 67<br>P = 69  | Cluster 1 = 118<br>T = 109<br>P = 109 |                                        | Cluster 1 = 64<br>T = 29<br>P = 35   |                                        |
| <b>A-B</b><br>(Total number of protein = 224) | Cluster 0 = 126<br>A = 60<br>B = 66     | Cluster 0 = 78<br>A = 36<br>B = 42   | Cluster 0 = 131<br>A = 31<br>B = 100  | Cluster 0 = 224<br>A = 112<br>B = 112  | Cluster 0 = 160<br>A = 74<br>B = 86  | Cluster 0 = 224<br>A = 112<br>B = 112  |
|                                               | Cluster 1 = 98<br>A = 52<br>B = 46      | Cluster 1 = 146<br>A = 76<br>B = 70  | Cluster 1 = 93<br>A = 81<br>B = 12    |                                        | Cluster 1 = 64<br>A = 38<br>B = 26   |                                        |
| <b>H-Nh</b> (Total number of protein = 200)   | Cluster 0 = 111<br>H = 52<br>Nh = 59    | Cluster 0 = 71<br>H = 35<br>Nh = 36  | Cluster 0 = 120<br>H = 39<br>Nh = 81  | Cluster 0 = 200<br>H = 100<br>Nh = 100 | Cluster 0 = 155<br>H = 72<br>Nh = 83 | Cluster 0 = 200<br>H = 100<br>Nh = 100 |
|                                               | Cluster 1 = 89<br>H = 48<br>Nh = 41     | Cluster 1 = 129<br>H = 65<br>Nh = 64 | Cluster 1 = 80<br>H = 61<br>Nh = 19   |                                        | Cluster 1 = 45<br>H = 28<br>Nh = 17  |                                        |
| <b>B-Nb</b> (Total number of protein = 80)    | Cluster 0 = 29<br>B = 5<br>Nb = 24      | Cluster 0 = 18<br>B = 7<br>Nb = 11   | Cluster 0 = 15<br>B = 11<br>Nb = 4    | Cluster 0 = 80<br>B = 40<br>Nb = 40    | Cluster 0 = 28<br>B = 22<br>Nb = 6   | Cluster 0 = 80<br>B = 40<br>Nb = 40    |
|                                               | Cluster 1 = 51<br>B = 35<br>Nb = 19     | Cluster 1 = 62<br>B = 33<br>Nb = 29  | Cluster 1 = 65<br>B = 29<br>Nb = 36   |                                        | Cluster 1 = 52<br>B = 18<br>Nb = 34  |                                        |



### ***Supervised learning to generate model for protein extreme-stability***

Unlike unsupervised clustering which could not yield a classification model with high prediction performance accuracy; the final datasets were analyzed through supervised methods. The supervised learning methods classify samples in a given dataset using a set of attributes (or features) with a set of rules that prescribe assignments of samples to classes based solely on values of features<sup>42</sup>. Both unlabeled and labeled data are more promptly classified by supervised learning algorithms. For performing the same, lazy modeling ( $k$ -NN and Naïve Bayes), logistic regression, SVM, decision tree and ANN were employed to generate model for protein extreme-stability (Table 4.6).

#### **• Lazy modeling to generate model for protein extreme-stability**

The two lazy modeling algorithms were applied to the datasets which included  $k$ -Nearest Neighbour ( $k$ -NN) and Naïve Bayes.  $k$ -NN classification can give highly competitive results and its output is a class membership depending upon  $k$ . The  $k$  is a positive integer and is user-defined constant. On the other hand, Naïve Bayes based modeling has been considered to be very simple and more accurate classifying doing a bunch of counts<sup>43</sup>. The advantage of Naïve Bayes classifier classifies attributes using Bayes' theorem<sup>43</sup>. It is referred as a probabilistic classifier which assumes a particular attribute and are independent of all other attributes. The results of both the lazy modeling ( $k$ -NN and Naïve Bayes) have been presented in Table 4.6. As observed the accuracies of the models varied among the different extremophile-nonextremophile datasets. Highest accuracy of Naïve Bayes classification of protein was obtained at 92.65 %, 91.30 % and 82.67 % when applied on P-M, T-P and T-M, respectively. Interestingly, this algorithm classifying proteins more accurately of low, moderate and higher temperature tolerance. Additionally, it also classify the B-Nb proteins with highest accuracy of 85.71 % on account of significant attributes. Literature also shows low accuracy of prediction when Naïve Bayes was used as the prediction model<sup>44</sup>. On the other hand, A-B and H-Nh proteins more accurately classify by  $k$ -NN ( $k = 10$ ) with 80.08 % and 75.00 %. To further increase the accuracy of the models the datasets were subjected to supervised learning methods.

- **Logistic regression, support vector machines and artificial neural networks to generate model for protein extreme-stability**

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The logistic regression was applied independently with five kernel parameter dot, radial, polynomial, neural and anova. The best model was obtained with dot and anova kernel. The dot kernel is defined by  $k(x, y) = x * y$  i.e. it is inner product of  $x$  variable and  $y$  variable. The anova kernel is defined by  $k(x, y) = (x * y + 1)^d$  where  $d$  is degree of summation of  $\exp(-g(x - y))$  and  $g$  is gamma. Gamma and degree are adjusted by the kernel gamma and kernel degree parameters, respectively<sup>45</sup>. In the present study, the binary extremophile-nonextremophile datasets were tested with logistic regression and the prediction accuracy of classification have been shown in Table 4.6. Result revealed that logistic regression with anova kernel type criteria performed better in T-M, P-M, A-B and H-Nh with classification accuracy of 81.33%, 85.29%, 82.76% and 73.33%, respectively whereas the criteria dot kernel type gave accuracy of 86.96% and 78.57% in T-P and B-Nb dataset.

Support vector machines is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs<sup>46</sup>. The SVM has the ability to classify both the linear or non-linear patterns<sup>47</sup>. An SVM with a linear kernel is not very different from a logistic regression classifier. Linear patterns are easily distinguishable but non-linear patterns are not distinguishable. So, manipulation is required for a linear pattern<sup>48</sup>. However, the best SVM classification model was generated through linear kernel type. In the present research, LibSVM, c-SVC, nu-SVC, SVM (linear), dot and anova kernel type criteria were employed and considered as better classifier for classification of comparing extremophilic and non-extremophilic proteins. The accuracy of prediction for each dataset have been shown in Table 4.6. The highest accuracy was achieved in T-P i.e. 94.20% with dot kernel type and SVM (linear). This shows the importance of testing

various models on different datasets as the performance of the models varies with the composition of the datasets.

Furthermore, in this study feed forward neural network (Neural net and Perceptron) were applied on extremophilic/non-extremophilic proteins datasets. In neural network, 10 times cross validation was carried out to test the model on all patterns. The learning algorithm in all networks was back propagation. The accuracy for true, false and total accuracy was obtained. Neural network algorithm represents each cluster by a neuron. The input data is also represented by neurons, which are connected to the prototype neurons. Each such connection has a weight, which is learned adaptively during learning<sup>23</sup>. Feed forward neural network without any back propagation with varied hidden layers and neuron in each layer for each dataset achieved the higher prediction accuracies of classification. A maximum of three hidden layer with maximum of 50 neurons in each hidden layer were tested on these datasets. The results of ANN (artificial neural networks) prediction have been shown in Table 4.6.

Conclusively, it can be said that one has to do a lot of experimenting to find the best combination of parameters and in a broader sense models by avoiding overfitting of data to achieve at a model which performs best in prediction of different type of extreme-stability of proteins. The dataset comprising of amino acid composition as well as structural attributes information outperformed all the datasets when used for prediction of protein extreme-stability. Supervised clustering performs better than unsupervised clustering and lazy modeling on datasets for predicting protein extreme-stability.

#### • Decision Trees to generate model for protein thermostability

Decision tree is a special type of predictor which is trained by an iterative selection of individual features that are the most salient at each node in the tree<sup>42</sup>. The datasets were then analyzed through decision trees which are simple to read and understand. The topmost node in the tree is the root node, each internal node denotes an attribute test, each branch represents an outcome of the test, and each leaf node represents classes<sup>44</sup>. Therefore, the tree induction models with four weighting criteria, ID3, CHAID and a

weight based decision tree with 11 weighting criteria were independently run on the T-M, P-M, T-P, A-B, H-Nh and B-Nb datasets. Most of the decision trees were without roots and leaves and thus were discarded. Thus, the multiple prediction trees were induced and best accuracy prediction trees were chosen for the interpretation of difference type of extreme-stability for enumerating the contribution of attributes at sequence (AA) and structure level (ST). It was observed that T-P dataset gave the highest accuracy *i.e.* 92.75, when Random Forest tree induction models was applied with criteria of information gain in 100 tree models. The prediction accuracy of other datasets are enlisted in Table 4.7. Random Forest is an ensemble of decision trees and it naturally incorporates selection and interactions of features in the learning process of tree induction<sup>49</sup>. It is non-parametric, interpretable and efficient classifier for learning any predictive problem. It has good prediction accuracy for several types of data<sup>49,50</sup>.

The analysis revealed Glu, (>7.910%) and GT (>0.346%) for thermophiles; Gln (>2.306%) for psychrophiles; Aro (>12.474%) and AAI (>0.350%) for acidophiles; CASA (>0.185%) for alkaliphiles; Acd (>17.436%) and % Asp (>5.564%) for halophiles; and, Arg (>6.062%), SB (>0.460%) and Chrg (>30.768%) for barophiles are the best possible discriminatory rule. Our results corroborate previous study which reported such bonds to be higher in extreme-stable proteins<sup>10,16,51</sup>. Glu is a charged residue and have been implicated to be involved in ionic interactions which results in protein thermostabilization<sup>1</sup>. Gamma turns (GT) have been reported to be higher in thermostable lipases as they stabilize the loops in protein structure by formation of short strong main chain to main chain hydrogen bonds<sup>52</sup>. This study further extends the aforementioned observation to be true for extreme-stabilizing proteins from all classes. Aromatic aromatic interactions (AAI) stabilize proteins as a typical aromatic-aromatic interaction has contributes between -0.6 and -1.3 kcal mol<sup>-1</sup> to protein stability<sup>53,54</sup>. Conclusively, in all these analysis, extreme-stability has always been attributed to be the cumulative effect of all such factors. Therefore a single recipe ceases to exist which can render proteins extreme-stable through protein engineering approaches.

Table 4.6: Classification models by machine learning approaches and their prediction accuracies

| Model                            | Criterion used and their percentage accuracy of prediction (%) |          |                                               |          |                                                     |          |                                                     |          |                                                  |          |
|----------------------------------|----------------------------------------------------------------|----------|-----------------------------------------------|----------|-----------------------------------------------------|----------|-----------------------------------------------------|----------|--------------------------------------------------|----------|
|                                  | T-M                                                            |          | P-M                                           |          | T-P                                                 |          | A-B                                                 |          | H-Nh                                             |          |
| Model                            | Criteria                                                       | Accuracy | Criteria                                      | Accuracy | Criteria                                            | Accuracy | Criteria                                            | Accuracy | Criteria                                         | Accuracy |
| <b>Lazy modeling</b>             | Naïve Bayes                                                    | 82.67    | Naïve Bayes                                   | 92.65    | Naïve Bayes                                         | 91.30    | k-NN (k=10)                                         | 80.08    | k-NN (k=10)                                      | 75.00    |
| <b>Logistic regression</b>       | Anova kernel type                                              | 81.33    | Anova kernel type                             | 85.29    | Dot kernel type                                     | 86.96    | Anova kernel type                                   | 82.76    | Anova kernel type                                | 73.33    |
| <b>Support vector machine</b>    | Lib SVM, nu-SVC                                                | 90.97    | LibSVM, C-SVC and nu-SVC                      | 88.24    | Dot kernel type and SVM (Linear)                    | 94.20    | LibSVM nu-SVC type                                  | 84.67    | Anova kernel type                                | 80.00    |
| <b>Artificial Neural Network</b> | 3 hidden layer 10, 20, 30 neuron in each                       | 77.60    | hidden layer 10, 20, 30 neuron in each        | 88.24    | 2 hidden layer with 30 neurons in each layer        | 91.30    | hidden layers with 10 neurons                       | 82.38    | 1 hidden layers with 40 neurons                  | 70.00    |
| <b>Decision Tree</b>             | Random Forest (gain ratio in 500 tree models)                  | 86.67    | Random Forest (gini index in 100 tree models) | 89.71    | Random Forest (information gain in 100 tree models) | 92.75    | Random Forest (Information gain in 500 tree models) | 77.01    | Gain Ratio                                       | 81.67    |
|                                  |                                                                |          |                                               |          |                                                     |          |                                                     |          | Random Forest (by gini index in 500 tree models) | 78.57    |

Table 4.7: Decision tree prediction and chosen best possible discriminatory rule for extreme-stability

| Comparison  | Protein dataset | Tree induction method | Criterion (algorithm) chosen | Number of models generated | Best possible discriminatory rule                                                                                                                                                                                      | Accuracy of prediction (%) |
|-------------|-----------------|-----------------------|------------------------------|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------|
| <b>T-M</b>  | AA              | Random Forest         | Gain Ratio                   | 500 internal trees         | If % Glu ( $>7.910$ ] and % Arg ( $>6.743$ ] $\rightarrow$ <b>Thermophilic protein</b>                                                                                                                                 | 86.67                      |
|             |                 |                       |                              |                            | If % Glu ( $\leq 7.910$ ] and % Cys ( $>1.167$ ] and % Acd ( $>10.040$ ] $\rightarrow$ <b>Mesophilic proteins</b>                                                                                                      |                            |
| <b>P-M</b>  | AA+ST           | Random Forest         | Gini Index                   | 100 internal trees         | If % GT ( $\leq 0.219$ ] and % Met ( $\leq 2.825$ ] and % Met ( $>1.246$ ] $\rightarrow$ <b>Psychrophilic proteins</b>                                                                                                 | 89.71                      |
|             |                 |                       |                              |                            | If % GT ( $>0.219$ ] $\rightarrow$ <b>Mesophilic proteins</b>                                                                                                                                                          |                            |
| <b>T-P</b>  | AA              | Random Forest         | Gini Index                   | 100 internal trees         | If % Gln ( $\leq 2.306$ ] and % Lys ( $>7.974$ ] $\rightarrow$ <b>Thermophilic proteins</b>                                                                                                                            | 88.73                      |
|             |                 |                       |                              |                            | If % Gln ( $>2.306$ ] and % Asp ( $>4.480$ ] and % Gln ( $>4.782$ ] $\rightarrow$ <b>Psychrophilic proteins</b>                                                                                                        |                            |
|             | ST              | Random Forest         | Gain Ratio                   | 500 internal trees         | If % GT ( $>0.346$ ] $\rightarrow$ <b>Thermophilic proteins</b><br>If % GT ( $\leq 0.346$ ] and % CAsA ( $\leq 0.285$ ] and % HI ( $\leq 12.514$ ] and % PASA ( $>0.155$ ] $\rightarrow$ <b>Psychrophilic proteins</b> | 92.75                      |
| <b>A-B</b>  | AA              | Random Forest         | Information Gain             | 500 internal trees         | If % Arg ( $>12.474$ ] $\rightarrow$ <b>Acidophilic proteins</b>                                                                                                                                                       | 80.46                      |
|             |                 |                       |                              |                            | If % Arg ( $\leq 12.474$ ] and % Chrg ( $>21.474$ ] $\rightarrow$ <b>Alkaliphilic proteins</b>                                                                                                                         |                            |
|             | ST              | Random Forest         | Information Gain             | 500 internal trees         | If % AAI ( $>0.350$ ] $\rightarrow$ <b>Acidophilic proteins</b><br>If % CAsA ( $>0.185$ ] $\rightarrow$ <b>Alkaliphilic proteins</b>                                                                                   | 82.00                      |
| <b>H-Nh</b> | AA+ST           | Random Forest         | Gini Index                   | 100 internal trees         | If % Acd ( $>17.436$ ] and % Asp ( $>5.564$ ] $\rightarrow$ <b>Halophilic proteins</b>                                                                                                                                 | 81.67                      |
|             |                 |                       |                              |                            | If % Acd ( $>17.436$ ] and % Asp ( $\leq 5.564$ ] $\rightarrow$ <b>Non-halophilic proteins</b>                                                                                                                         |                            |
| <b>B-Nb</b> | AA+ST           | Random Forest         | Gini Index                   | 500 internal trees         | If % Arg ( $>6.062$ ] and % SB ( $>0.460$ ] and % Chrg ( $>30.768$ ] $\rightarrow$ <b>Barophilic proteins</b>                                                                                                          | 78.57                      |
|             |                 |                       |                              |                            | If % Arg ( $>6.062$ ] and % SB ( $\leq 0.460$ ] and % Arg ( $\leq 6.518$ ] $\rightarrow$ <b>Non-barophilic proteins</b>                                                                                                |                            |



#### 4.3.4 Generation of ranking models for protein extreme attributes by AHP

The analytical hierarchical process (AHP) helps decision makers find one that best suits their goal and their understanding of the problem<sup>55</sup>. AHP approaches comprises of three hierarchies that include, goal, criteria and alternatives (Figure 4.2). The goal was to predict whether a protein is extremophilic or non-extremophilic on the basis of selected significant features. The criteria level represented all the statistically significant features. The alternative level comprised of the proteins to be categorized as extremophilic or non-extremophilic proteins based on featured ranks. The aim was to rank or prioritize protein features according to their contribution towards extreme-stability of proteins. The primary purpose of AHP was the construction of a pair-wise comparison matrix for ranking the features in accordance to their contribution towards extreme-stability of proteins. The significant attributes for extreme-stability involves the prediction of priority or Eigenvector from the generated reciprocal pairwise comparison matrix. The generated reciprocal pairwise comparison matrix are shown in Appendix Table A2.15 - A2.20. According to Saaty (2008), the aforementioned judgment to derive at the priority vectors can be accepted as consistent if the consistency ratio (CR) value is less than 0.10<sup>55</sup>. The CR of the generated pair-wise comparison matrix for significant features pertaining to extreme-stability of proteins was calculated and presented in Appendix Table A2.15 - A2.20. Through this method, all of the significant features according to its importance in contributing towards extreme-stability of proteins were prioritized. The results of generated ranking model by AHP (MCDM) that predict ranks are presented in Figure 4.3. The ranking results for each dataset showed that the Glu and GT in thermophilic proteins and Gln in psychrophilic protein have the highest ranks. These results were confirmed from Parthiban et al. who showed that Gln is insignificant amino acids for thermostability as they are thermolabile in nature and more prone to deamidation at higher temperatures<sup>14</sup>. Also, the charge accessible surface area (CASA) and  $\alpha$ -helix (AH) are important for alkaliphilic proteins, as they are making the protein structures more robust in the alkaline pH, as reported by Vila et al.<sup>12</sup>. Beside, AAI and DS were found to be most prior attributes for acidophilic proteins as they are contributing the protein stability<sup>53,54</sup>. Similarly, acidic amino acid, aspartic acids in halophilic proteins were

ranked higher. On the other hand, arginine in barophilic proteins were significant for which Di Giulio referred Arg as “barophilic amino acid”<sup>13</sup>.

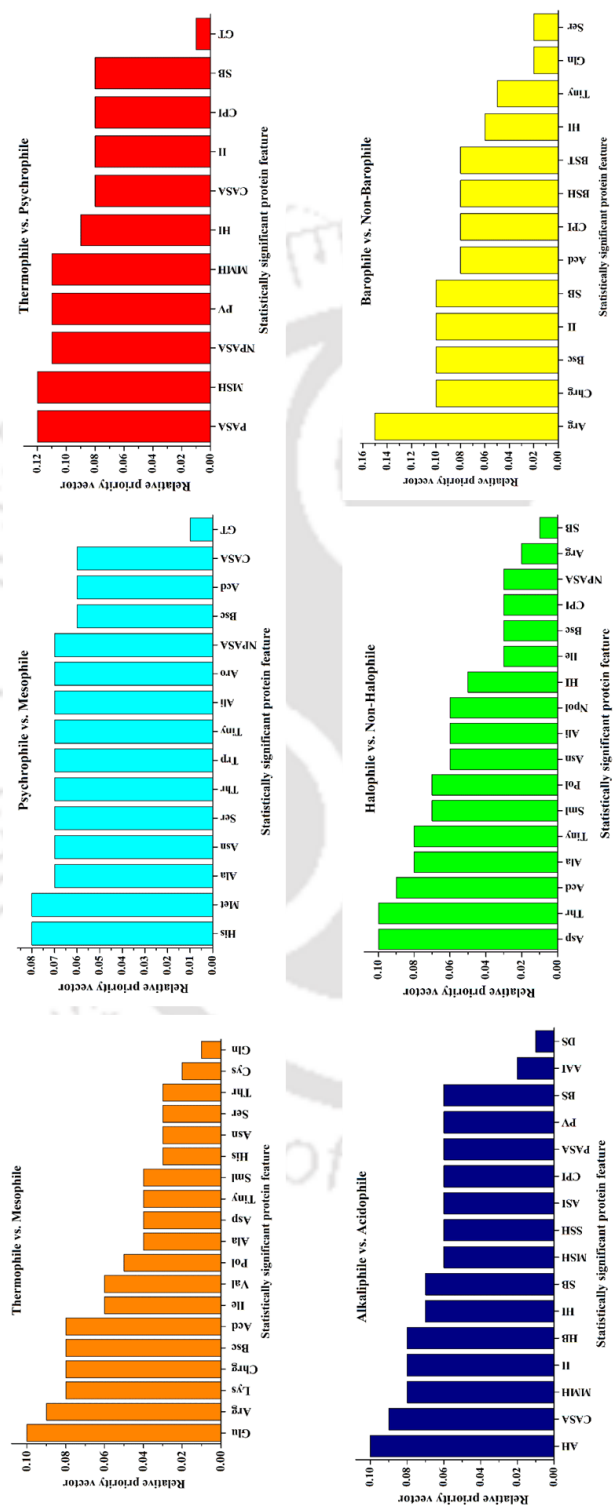


Figure 4.3: Ranks generated by AHP (MCDM) for protein attributes rendering extreme-stability.

#### 4.3.5 Blind test analysis: validation of generated ranking model

In order to rank proteins as extremophilic/non-extremophilic, the protein pairs were chosen randomly as alternatives from the existing dataset. These were processed through generated model by AHP to derive at the ranks. A second blind test was carried out with a set of non-extremophilic proteins with another non-extremophilic counterpart chosen randomly from the existing dataset. The proteins were assigned ranks and the mean rank value. The difference in rank values were calculated from these two types of blind tests i.e. first test compared extremophilic and non-extremophilic proteins whereas the second test compared non-extremophilic and non-extremophilic proteins. The results of blind test validation, accuracy prediction for ranking models and AHP ranks are shown in Table 4.8. The accuracy of prediction for T-M dataset was calculated to be 96 %; P-M was 93%; T-P was 91 %; A-B was 90 %; H-Nh was 90 %; B-Nb was 94 % on the basis of first blind test of randomly chosen 100 thermophilic/mesophilic protein pairs; 100 psychrophilic/mesophilic protein pairs; 100 thermophilic/psychrophilic protein pairs; 100 acidophilic/alkaliphilic protein pairs; 100 halophilic/nonhalophilic protein pairs; and 50 barostable/non-barostable protein pairs, as alternatives from the existing dataset. These accuracy percentage depends upon random frequency of false positive (FP), False Negative (FN), True Positive (TP) and True Negative (TN) values that were obtained in the blind test. Previously, Nath and Subbiah employed the same mathematical formula of accuracy prediction for discriminating piezophilic and non-piezophilic proteins<sup>26</sup>. Further, the second blind test was performed for ranking of 40 non-extremophilic/non-extremophilic protein pairs (except 20 for non-barophilic/non-barophilic protein pairs) from the existing non-extremophilic protein dataset resulted in rank value difference. This is much less than the rank value difference calculated for the first blind test of extremophilic and non-extremophilic protein pairs. The result showed that the final rank differences of the aforementioned two blind test i.e. extremophilic/non-extremophilic proteins and non-extremophilic/non-extremophilic proteins in T-M was obtained at 0.09 and 0.0002, respectively; in P-M was 0.05 and 0.003, respectively; in T-P was 0.01 and 0.0005, respectively; in A-B were 0.14 and 0.004, respectively; in H-Nh were 0.08 and 0.001, respectively; and in B-Nb were 0.2 and 0.01, respectively. Here it can be

conclusively said that this method could correctly differentiate between extremophilic proteins and non-extremophilic proteins. The higher rank value difference for the extremophilic and non-extremophilic protein set also strongly indicates that the priority set to the significant features are correct and this priority vectors can thus be conveniently utilized to distinguish between extreme-stabilizing and non-extreme-stabilizing mutations. Thus, any mutation in a protein which leads to a rank higher than 0.5 compared to its non-extremophilic counterpart through this method can be predicted to be extreme-stabilizing.

Table 4.8: Validation of ranking models by blind test analysis.

| Test Sets   | Blind test set  | No. of proteins             | Accuracy (%) | Mean Rank | Rank Difference |
|-------------|-----------------|-----------------------------|--------------|-----------|-----------------|
| <b>T-M</b>  | 1 <sup>st</sup> | 100 Thermophilic Proteins   | <b>96</b>    | 0.540     | 0.0900          |
|             |                 | 100 Mesophilic Proteins     |              | 0.450     |                 |
|             | 2 <sup>nd</sup> | 40 Mesophilic Proteins      |              | 0.4998    | 0.0002          |
|             |                 | 40 Mesophilic Proteins      |              | 0.5000    |                 |
| <b>P-M</b>  | 1 <sup>st</sup> | 100 Psychrophilic Proteins  | <b>93</b>    | 0.520     | 0.0500          |
|             |                 | 100 Mesophilic Proteins     |              | 0.470     |                 |
|             | 2 <sup>nd</sup> | 40 Mesophilic Proteins      |              | 0.502     | 0.0030          |
|             |                 | 40 Mesophilic Proteins      |              | 0.499     |                 |
| <b>P-T</b>  | 1 <sup>st</sup> | 100 Psychrophilic Proteins  | <b>91</b>    | 0.550     | 0.0100          |
|             |                 | 100 Thermophilic Proteins   |              | 0.450     |                 |
|             | 2 <sup>nd</sup> | 40 Thermophilic Proteins    |              | 0.5003    | 0.0005          |
|             |                 | 40 Thermophilic Proteins    |              | 0.4998    |                 |
| <b>A-B</b>  | 1 <sup>st</sup> | 100 Alkaliphilic Proteins   | <b>90</b>    | 0.570     | 0.1400          |
|             |                 | 100 Acidophilic Proteins    |              | 0.430     |                 |
|             | 2 <sup>nd</sup> | 40 Acidophilic Proteins     |              | 0.502     | 0.0040          |
|             |                 | 40 Acidophilic Proteins     |              | 0.498     |                 |
| <b>H-Nh</b> | 1 <sup>st</sup> | 100 Halophilic Proteins     | <b>90</b>    | 0.540     | 0.0800          |
|             |                 | 100 Non-Halophilic Proteins |              | 0.460     |                 |
|             | 2 <sup>nd</sup> | 40 Non-Halophilic Proteins  |              | 0.500     | 0.0010          |
|             |                 | 40 Non-Halophilic Proteins  |              | 0.499     |                 |
| <b>B-Nb</b> | 1 <sup>st</sup> | 50 Barophilic Proteins      | <b>94</b>    | 0.600     | 0.2000          |
|             |                 | 50 Non-Barophilic Proteins  |              | 0.400     |                 |
|             | 2 <sup>nd</sup> | 20 Non-Barophilic Proteins  |              | 0.505     | 0.0100          |
|             |                 | 20 Non-Barophilic Proteins  |              | 0.495     |                 |

#### 4.3.6 Salient findings

The results of all types of extreme-stability prediction along with their positively and negatively contributing attributes are summarized in Table 4.9. Taking care of these facts,

a model mesophilic protein was chosen and the mutations were designed in such a way to increase the content of these positively contributing attributes.

Table 4.9: Salient findings for attributes contributing different types of extreme-stability.

| Extreme-stability | Dataset | Positively contributing attributes | Negatively contributing attributes |
|-------------------|---------|------------------------------------|------------------------------------|
| Thermostability   | AA      | Glu, Arg, Chrg                     | Gln, Tiny                          |
|                   | ST      | GT, CASA, HI                       | PASA                               |
| Psychrostability  | AA      | Gln, Tiny                          | Glu, Arg, Chrg                     |
|                   | ST      | PASA                               | GT, CASA, HI                       |
| Alkalistability   | AA      | Chrg, Bsc, Glu                     | Aro, Trp, Pol                      |
|                   | ST      | CASA, SB                           | AAI, DS                            |
| Acidostability    | AA      | Aro, Trp, Pol                      | Chrg, Bsc, Glu                     |
|                   | ST      | AAI, DS                            | CASA, SB                           |
| Halostability     | AA+ST   | Asp, Acd, Thr                      | Arg, SB                            |
| Barostability     | AA+ST   | Arg, Chrg, SB, II                  | Ser, Gln                           |

#### 4.4 Conclusion

Present study shows that studying different datasets (such as T-M, P-M, T-P, A-B, H-Nh and B-Nb) at different level (such as AA, ST and AA+ST) can bring about the importance of features in each hierarchy of protein organization which is responsible for protein extreme-stability. The results support the long drawn conclusion that protein amino acid and structural composition is correlated to different types of protein extreme-stability. Further, various machine learning approaches based classification models were generated for prediction of extreme-stability. The best decision tree was induced by various Decision Tree and Random Forest algorithms that showed a combinatorial increase or decrease of significant attributes as decisive features as contributors of extreme-stabilization. With the intention of protein engineering applications, an attempt was made to prioritize the extreme-stability contributors applying AHP. It ranked the attributes according to their calculated priority vectors from reciprocal pairwise comparison matrix for different extreme-stability at sequence level and structural level. The present research converges upon that AHP generated ranking model could successfully predict extreme-stabilizing mutations explaining its rationale basis. The rationale basis of the prediction method reduces time and increases process economy to attain mutants via protein engineering making it instrumental for various industrial applications.

## 4.5 References

1. Chakravorty, D., Khan, M. F. & Patra, S. Thermostability of Proteins Revisited Through Machine Learning Methodologies: From Nucleotide Sequence to Structure. *Current Biotechnology* **6**, 39–49 (2017).
2. Karshikoff, a & Ladenstein, R. Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. *Protein Eng. Des. Sel.* **11**, 867–872 (1998).
3. Szilágyi, A. & Závodszky, P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: Results of a comprehensive survey. *Structure* **8**, 493–504 (2000).
4. Vogt, G., Woell, S. & Argos, P. Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **269**, 631–43 (1997).
5. Jaenicke, R. & Böhm, G. The stability of proteins in extreme environments. *Curr. Opin. Struct. Biol.* **8**, 738–48 (1998).
6. Creighton, T. E. *Protein structure: a practical approach*. (Oxford University Press, 1997).
7. Metpally, R. P. R. & Reddy, B. V. B. Comparative proteome analysis of psychrophilic versus mesophilic bacterial species: Insights into the molecular basis of cold adaptation of proteins. *BMC Genomics* **10**, 11 (2009).
8. Rao, J. K. & Argos, P. Structural stability of halophilic proteins. *Biochemistry* **20**, 6536–43 (1981).
9. Tokunaga, H., Arakawa, T. & Tokunaga, M. Engineering of halophilic enzymes: two acidic amino acid residues at the carboxy-terminal region confer halophilic characteristics to Halomonas and Pseudomonas nucleoside diphosphate kinases. *Protein Sci.* **17**, 1603–10 (2008).
10. Reed, C. J., Lewis, H., Trejo, E., Winston, V. & Evilia, C. Protein adaptations in archaeal extremophiles. *Archaea* **2013**, 373275 (2013).
11. Goodarzi, H., Torabi, N., Najafabadi, H. S. & Archetti, M. Amino acid and codon usage profiles: Adaptive changes in the frequency of amino acids and codons. *Gene* **407**, 30–41 (2008).
12. Vila, J. A., Ripoll, D. R. & Scheraga, H. A. Physical reasons for the unusual alpha -helix stabilization afforded by charged or neutral polar residues in alanine-rich peptides. *Proc. Natl. Acad. Sci.* **97**, 13075–13079 (2000).
13. Di Giulio, M. A comparison of proteins from Pyrococcus furiosus and Pyrococcus abyssi: barophily in the physicochemical properties of amino acids and in the genetic code. *Gene* **346**, 1–6 (2005).
14. Parthiban, V., Gromiha, M. M. & Schomburg, D. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* **34**, W239–42 (2006).
15. Ebrahimi, M., Lakizadeh, A., Agha-Golzadeh, P., Ebrahimie, E. & Ebrahimi, M. Prediction of Thermostability from Amino Acid Attributes by Combination of Clustering with Attribute Weighting: A New Vista in Engineering Enzymes. *PLoS One* **6**, e23146 (2011).
16. Chakravorty, D., Khan, M. F. & Patra, S. Multifactorial level of extremostability of proteins: can they be exploited for protein engineering? *Extremophiles* **21**, 419–444 (2017).
17. Marsaglia, G., Tsang, W. W. & Wang, J. Evaluating Kolmogorov's Distribution. *J. Stat. Softw.* **8**, (2003).
18. Biesiada, J. & Duch, W. Feature selection for high-dimensional data: A kolmogorov-smirnov correlation-based filter. in *Computer Recognition Systems* 95–103 (Springer, 2005).
19. Subrahmanyam, K., Sankar, N. S., Baggam, S. P. & Raghavendra, R. S. A Modified KS-test for Feature Selection. *IOSR J. Comput. Eng.* 73–79 (2013).
20. Ebrahimie, E., Ebrahimi, M., Sarvestani, N. R. & Ebrahimi, M. Protein attributes contribute to halo-stability, bioinformatics approach. *Saline Systems* **7**, 1 (2011).
21. Han, J., Kamber, M. & Tung, K. H. Spatial Clustering Methods in Data Mining: A Survey. Harvey J. Miller and Jiawei Han (eds.), *Geographic Data Mining and Knowledge Discovery*. (2001).
22. Nugent, R. & Meila, M. An overview of clustering applied to molecular biology. *Methods Mol. Biol.* **620**, 369–404 (2010).
23. Rokach, L. A survey of clustering algorithms. in *Data mining and knowledge discovery handbook* 269–298 (Springer, 2009).
24. Saaty, T. L. *Decision making for leaders: the analytic hierarchy process for decisions in a complex world*. (RWS publications, 1990).



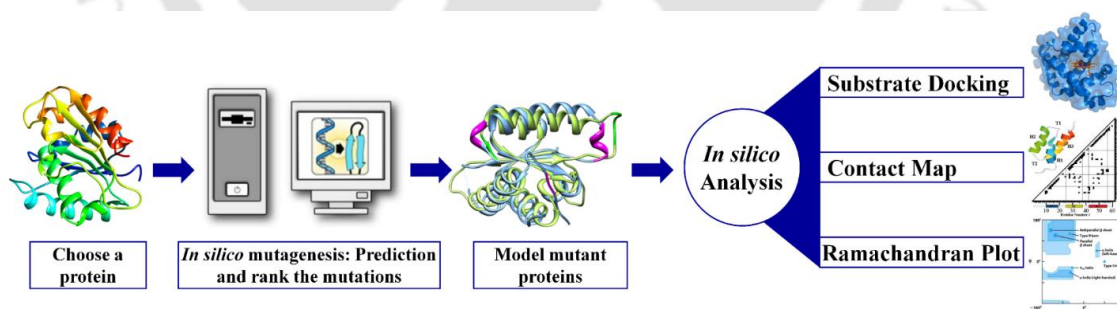
25. ALONSO, J. A. & LAMATA, M. T. CONSISTENCY IN THE ANALYTIC HIERARCHY PROCESS: A NEW APPROACH. *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.* **14**, 445–459 (2006).
26. Nath, A. & Subbiah, K. Insights into the molecular basis of piezophilic adaptation: Extraction of piezophilic signatures. *J. Theor. Biol.* **390**, 117–26 (2016).
27. Fillion, G. J. The signed Kolmogorov-Smirnov test: Why it should not be used. *GigaScience* **4**, (2015).
28. Pasieka, T. J., Woolson, R. F. & Grose, C. Viral induced fusion and syncytium formation: Measurement by the Kolmogorov-Smirnov statistical test. *J. Virol. Methods* **111**, 157–161 (2003).
29. Di Giulio, M. The origin of the genetic code: theories and their relationships, a review. *Biosystems*. **80**, 175–84 (2005).
30. Suplatov, D. *et al.* Computational design of a pH stable enzyme: understanding molecular mechanism of penicillin acylase's adaptation to alkaline conditions. *PLoS One* **9**, e100643 (2014).
31. Reed, J.C., Lewis, H. Trejo, E., Winston, V. and Evilia, C. Protein Adaptations in Archaeal Extremophiles. *Archaea* **2013**, 1–14 (2013).
32. Sadeghi, M., Naderi-Manesh, H., Zarrabi, M. & Ranjbar, B. Effective factors in thermostability of thermophilic proteins. *Biophys. Chem.* **119**, 256–270 (2006).
33. Michael Gromiha, M. & Huang, L.-T. Machine Learning Algorithms for Predicting Protein Folding Rates and Stability of Mutant Proteins: Comparison with Statistical Methods. *Curr. Protein Pept. Sci.* **12**, 490–502 (2011).
34. Kumar, S., Tsai, C.-J. & Nussinov, R. Factors enhancing protein thermostability. *Protein Eng. Des. Sel.* **13**, 179–191 (2000).
35. Kumwenda, B., Litthauer, D., Bishop, O. T. & Reva, O. Analysis of protein thermostability enhancing factors in industrially important thermus bacteria species. *Evol. Bioinform. Online* **9**, 327–42 (2013).
36. Borovskii, G. B., Stupnikova, I. V., Antipina, A. I., Vladimirova, S. V. & Voinikov, V. K. Accumulation of dehydrin-like proteins in the mitochondria of cereals in response to cold, freezing, drought and ABA treatment. *BMC Plant Biol.* **2**, 5 (2002).
37. Yafremava, L. S., Di Giulio, M. & Caetano-Anollés, G. Comparative Analysis of Barophily-Related Amino Acid Content in Protein Domains of *Pyrococcus abyssi* and *Pyrococcus furiosus*. *Archaea* **2013**, 1–9 (2013).
38. Siddiqui, K. S. & Thomas, T. *Protein adaptation in extremophiles*. (Nova Publishers, 2008).
39. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. in *Advances in Knowledge Discovery and Data Mining* 160–172 (2013). doi:10.1007/978-3-642-37456-2\_14
40. Arora, P., Deepali & Varshney, S. Analysis of K-Means and K-Medoids Algorithm For Big Data. *Procedia Comput. Sci.* **78**, 507–512 (2016).
41. McGarry, L. M. *et al.* Quantitative classification of somatostatin-positive neocortical interneurons identifies three interneuron subtypes. *Front. Neural Circuits* **4**, 12 (2010).
42. Tarca, A. L., Carey, V. J., Chen, X., Romero, R. & Drăghici, S. Machine Learning and Its Applications to Biology. *PLoS Comput. Biol.* **3**, e116 (2007).
43. Zhang, Z. Naïve Bayes classification in R. *Ann. Transl. Med.* **4**, 241–241 (2016).
44. Wu, X. & Kumar, V. *The Top Ten Algorithms in Data Mining*. Chapman & Hall/CRC Press (2009). doi:10.1007/s10115-007-0114-2
45. Zhu, J. & Hastie, T. Kernel Logistic Regression and the Import Vector Machine. *J. Comput. Graph. Stat.* **14**, 185–205 (2005).
46. Ben-Hur, A., Horn, D., Siegelmann, H. T. & Vapnik, V. Support vector clustering. *J. Mach. Learn. Res.* **2**, 125–137 (2001).
47. Ambard, M. & Rotter, S. Support vector machines for spike pattern classification with a leaky integrate-and-fire neuron. *Front. Comput. Neurosci.* **6**, 78 (2012).
48. Karamizadeh, S., Abdullah, S. M., Halimi, M., Shayan, J. & Rajabi, M. J. Advantage and drawback of support vector machine functionality. in *2014 International Conference on Computer, Communications, and Control Technology (I4CT)* 63–65 (IEEE, 2014). doi:10.1109/I4CT.2014.6914146
49. Qi, Y. Random forest for bioinformatics. in *Ensemble Machine Learning: Methods and Applications* 307–323 (2012). doi:10.1007/9781441993267\_10
50. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

51. Karshikoff, A., Nilsson, L. & Ladenstein, R. Rigidity versus flexibility: the dilemma of understanding protein thermal stability. *FEBS J.* **282**, 3899–3917 (2015).
52. Chakravorty, D., Parameswaran, S., Dubey, V. K. & Patra, S. In silico characterization of thermostable lipases. *Extremophiles* **15**, 89–103 (2011).
53. Burley, S. & Petsko, G. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science (80-. )*. **229**, 23–28 (1985).
54. Serrano, L., Bycroft, M. & Fersht, A. R. Aromatic-aromatic interactions and protein stability. *J. Mol. Biol.* **218**, 465–475 (1991).
55. Saaty, T. L. & Peniwati, K. *Group decision making: drawing out and reconciling differences*. (RWS publications, 2013).



## 5

## *In silico* prediction of mutations to attain extreme-stability of proteins and their validation



## Abstract

This chapter gives a pertinent account of in silico prediction of plausible mutations and their validation in *Bacillus subtilis* lipase (a chosen mesophilic enzyme) by substituting the preferred protein attributes to attain extreme-stability of proteins by designing in silico homology models. Accordingly, three thermostable mutants (bsl\_the1: V149K, Q150E, bsl\_the2: F41K, W42E, V149K, Q150E and bsl\_the3: F41K, W42E, P119E, Q121K, V149K, Q150E), two psychrostable mutants (bsl\_psy1: G52H, P53T and bsl\_psy2: G52H, P53T, G92M, G93A), two acidostable mutants (bsl\_acd1: G52F, V54Y, G111Y and bsl\_acd2: G52F, V54Y, T110H, G111Y), two alkalistable mutants (bsl\_alk1: T110E, G111Y and bsl\_alk2: W31E, S32H, T45H, T110E, G111Y), two halostable mutants (bsl\_hal1: G52D, P53D, P119D, N120D and bsl\_hal2: G52D, P53D, G93D, P119D, N120D) and one barostable mutant (bsl\_bar: S24R, P119R, Q164R, G175R) were designed. In silico validation of mutant was done by AHP generated ranking models, various mutation prediction servers (HotSpot Wizard, I-Mutant2, Cupsat, iPTREE-STAB, WET-STAB and ERIS webserver), Substrate Docking (Autodock 4.2 version), Ramachandran analysis (PROCHECK) and Contact Map analysis (CMView 1.1.1). Interestingly, results showed that the designed mutations did not affect the catalytic activity of the lipase and thus, mutation did not disturb the catalytic properties.

## 5.1 Introduction

During the past decade, many researchers have attempted developing desired enzymes possessing high activity and stability suitable for various industrial processes. Enzyme stability is an important parameter in biocatalytic applications, and researchers have always been in search for robust enzymes. The alternative approach is identification of extreme-stabilizing features in stable enzymes so that they can be further utilized to engineer extreme-stable enzymes by pre predicted site-directed mutagenesis techniques. A wide variety of thermostable enzymes have been cloned and expressed in mesophilic organisms. Another method have been carrying out random mutations and screening for desirable candidates with extremophilic parameters. However, one drawback is the randomness of the strategies and that such approaches may also lead to compromised stability and flexibility with the functional state of proteins<sup>1</sup>. Directed evolution strategies applied are still partially successful rational concepts. This approach is practically limited as it requires ample amount of time and is also labor intensive. Furthermore, it involves application of selective pressure to achieve mutagenesis followed by colony screening for the desired trait<sup>2</sup>. Additionally, repeated rounds of random mutagenesis are the key to success for this strategy. The lacuna is that it may lead to neutral and deleterious mutations<sup>2</sup>. Correspondingly, Reetz et al. proposed an innovative method for improving protein thermostability of mesostable enzymes without compromising the activity<sup>3</sup>. The method chooses the appropriate sites at which rigidity is to be increased in the wild type mesostable enzyme on the basis of atomic displacement parameters obtained from X-ray data, namely the B-Factors<sup>4,5</sup>. Their strategy constituted a combination of rational design and directed evolution. The method requires the availability of the wild type protein structure and illustrated that iterative cycles of saturation mutagenesis were performed at high B-Factors sites of protein<sup>6</sup>. The B-factor based site-directed mutagenesis has been exploited to enhance the thermostability of *Serratia plymuthica* sucrose isomerase<sup>7</sup>, *Candida antarctica* lipase B<sup>8</sup>, *Rhizomucor miehei* lipase<sup>9</sup>, *Coprinus cinereus* peroxidase<sup>10</sup> and *Bacillus subtilis* lipase<sup>3</sup>. Some of these recipes cease to exist which can render proteins extreme-stable through protein engineering approaches. Moreover, the predictive power of these concepts is rather limited, the targeted mutations are random and require

to be tested individually by site-directed mutagenesis<sup>11</sup>. Thus, to overcome the demerits of such approaches there is a need to develop a guided method which can predict mutations as extremostabilizing so that such mutations can be carried out *in vitro* through site directed mutagenesis.

Directed evolution strategies have been so far utilized to study and design enzymes with specific thermostabilizing properties. This involves using a series of cumbersome experimental techniques for mutation and recombination of the protein and a random selection on the basis of a certain property<sup>2</sup>. The features which lead to thermostabilize the proteins engineered through directed evolution can only be analyzed after obtaining the purified protein and protein crystallization<sup>12</sup>. Through our previous study a two-step protocol to perform site directed mutagenesis to engineer thermostable proteins was devised. Through this approach the rationale behind the stability can be analysed beforehand through *in silico* approaches circumventing the requirement of protein purification and crystallization<sup>13</sup>.

Furthermore, through the present study of ranking of features according to their importance in contributing towards extreme-stability of proteins, it was observed that as glutamic acid, arginine at sequence level and gamma turns, hydrophobic interaction, electrostatic interaction and salt bridges at structure level are abundant in thermophilic, alkaliphilic and barophilic proteins. To contrary, glutamine in psychrophilic protein have the highest ranks. Parthiban et al. also showed glutamine are insignificant amino acids for thermostability as they are thermolabile in nature and more prone to deamidation at higher temperatures<sup>14</sup>. Also, charge accessible surface area and alpha-helix are important for alkaliphilic proteins, as they are making the protein structures more robust in the alkaline pH, as reported by Vila et al.<sup>15</sup>. Similarly, aspartic acids in halophilic proteins were ranked higher as reported by Ebrahimie et al.<sup>16</sup> and arginine in barophilic proteins were significant for which Di Giulio referred arginine as “barophilic amino acid” at structure levels and charged amino acids are sequence levels are the highest ranked features AHP generated model and machine learning approaches<sup>17</sup>. Therefore, *in silico* point mutation(s) that lead to increase in non-covalent interactions in mutant proteins have also been attempted. The mutations were further ranked through AHP ranking



models and those having higher rank than the wild type structures were selected for further *in vitro* mutations.

The criteria for choosing mutations rendering protein extreme-stable was also deduced from Chapter 4. The protocol to mutate structures to render them extreme-stable has been deduced as:

1. Only residues which show high mutability propensity should be chosen. The mutations should not be done on stabilizing centre residues of a protein.
2. The mutated residues should not belong to the active site pocket of the proteins.
3. The mutations should be done on the loop regions and surface exposed region.

To validate the aforementioned methodology, in this chapter, the chosen mesostable enzyme *Bacillus subtilis* lipase A (estA) which is stable at 35 °C and pH 8.0, was chosen as the model enzyme for carrying out mutations<sup>18,19</sup>. As most of the thermostable proteins are lipases and *Bacillus subtilis* lipase is an industrially important enzyme especially in chiral synthesis<sup>12,18,19</sup>, it qualified as an ideal candidate for the application of this protocol. Furthermore it lacks lid and thus has no problem in operating in oil water interphase<sup>20</sup>. Therefore, the aim was to validate AHP ranking models in designing extreme-stabilizing mutations that can be performed through site directed mutagenesis to increase extreme-stability of the lipase. To further study the features which are rendering mutations desired extremostable parameter was validated through Ramachandran plot analysis, structure superimposition analysis, substrate docking and contact map analysis. Finally, the secondary structural change in the mutant lipases were predicted through PSIPRED server.

## 5.2 Methodology

The statistically significant features rendering proteins extreme-stability were ranked using AHP analysis and higher ranked feature were chosen for designing extreme-stabilizing mutations in an exemplary mesophilic enzyme. To validate the predicted

mutations, various computational approaches were carried out that have been illustrated as the scheme of the methodology in Figure 5.1.

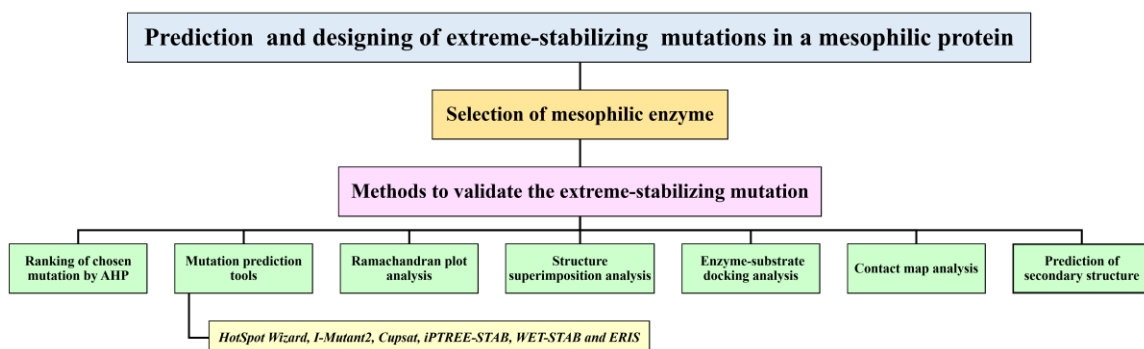


Figure 5.1: Flowchart exhibiting the methods employed for validation of chosen mutations.

### 5.2.1 Selection of mesophilic candidate enzyme for experimentation

*Bacillus subtilis* lipase was chosen as an exemplary mesophilic enzyme since its optimum activity was reported at 35 °C and pH 8.0<sup>18</sup>. The enzyme structure lacks lid, disulphide bonds and metal coordination which have been previously related to thermostability of lipases<sup>21</sup>.

### 5.2.2 *In silico* prediction and validation of mutations

#### *Testing of ranks of chosen mutations for validation of AHP ranking models*

The generated AHP ranking models were used to predict ranks for selected point mutations in *B. subtilis* lipase. The ranks were generated as the cumulative effect of statistically significant attributes after mutagenesis. Therefore, the combinations of mutations were chosen that cumulatively increases AHP ranks of the mutant lipase w.r.t. the wild type lipase. For the same, the chosen mutants were modelled first through I-TASSER server using the wild type structure as template to carried out all these analysis<sup>22</sup>. Further, the contents of statistically significant features of wild type and the predicted mutant lipases were compared and enumerated after *in silico* mutagenesis using PEPSTATS, PIC, VADAR, ESBRI and Promotif webserver<sup>23–26</sup>.

### ***Checking the mutability propensity in terms of $\Delta\Delta G$ for chosen mutations***

The predicted mutations were validated through various online available prediction tools such as HotSpot Wizard, I-Mutant2, Cupsat, iPTREE-STAB, WET-STAB and ERIS. These servers predicted stability of point mutation in terms of  $\Delta\Delta G$  value<sup>14,27–32</sup>.

### ***Structure validation of generated mutations through Ramachandran plot analysis***

The modelled enzyme structure of mutant lipases were validated through Ramachandran plot. It finds the stable conformations of polypeptide chain and illustrates structural stability through steric hindrance between the side-chains of the amino acids present in proteins<sup>33</sup>. PROCHECK program was used to generate Ramachandran plots and the percentage of residues in most favored region were determined for each mutant lipase<sup>34</sup>.

### ***Superimposition analysis of wild-type and mutant lipase structures***

Additionally, the wild-type and mutant lipase structures were superimposed using PyMolV0.99 and their RMSD values were also determined<sup>35</sup>.

### ***Functional efficiency of generated mutations by Enzyme-substrate docking analysis***

The molecular docking is often considered fundamental for understanding the strength of substrate binding with wild type and mutant lipases on the basis of their binding energy. It was performed through Autodock 4.2 version<sup>36</sup> using C8 (*p*-NPO, *p*-nitrophenol octanoate) substrate.

### ***Generation of contact map for enumeration of interactions in lipases***

As the present approach, the chosen mutation enhances stability via increasing the contents of positive attributes that relatively increases the number of interactions in mutant lipases than that of wild-type lipases. Protein contact maps represent residue interaction networks in two dimensions which facilitate the identification of structural features such as interactions within and between secondary structural elements and domains<sup>37</sup>. It helps in enumerating the unique interactions formed in mutant lipase *w.r.t.*

to wild type. Calculation and visualization of protein contact maps were performed using CMView 1.1.1<sup>38</sup>.

### ***Prediction of secondary structure contents in in wild-type and mutant lipases***

Finally, the lipases were validated through secondary structure prediction in terms of content of helices and sheets of wild-type and mutant lipases using PSIPRED server<sup>39</sup>.

## **5.3 Results and discussion**

### **5.3.1 Predicted mutations in *Bacillus subtilis* lipase**

Lipases (E.C 3.1.1.3) are potent industrial biocatalysts. Although their extremophilic natural sources are available, extraction and purification of lipases from such sources, is till date industrially infeasible. *Bacillus subtilis* 168 lipase A (estA) as an exemplary mesostable protein for prediction of mutations. More than 50 stabilizing mutations carried out in the lipase (PDB: 1i6w) were tested for each dataset through their predicted AHP ranking models. Selected stabilizing mutations are presented in Table 5.1. As analyzed in Chapter 4, the lipase sequence and structure were retrieved from UniProtKB (accession number: P37957) and RCSB PDB (PDB id: 1I6W), respectively and subjected to prediction of mutations that stabilize the enzyme under extreme *milieus*. Table 5.1 enlist the chosen point mutations in the lipase for generating extreme-stabilizing mutants.

Table 5.1: Selected mutations in *B. subtilis* lipase for generating various extreme-stabilizing mutants.

| Extreme-stable mutant types | Lipase mutants | Point mutations | Sites of point mutation                |
|-----------------------------|----------------|-----------------|----------------------------------------|
| Thermostablizing mutants    | bsl_the1       | 2               | V149K, Q150E                           |
|                             | bsl_the2       | 4               | F41K, W42E, V149K, Q150E               |
|                             | bsl_the3       | 6               | F41K, W42E, P119E, Q121K, V149K, Q150E |
| Psychrotablizing mutants    | bsl_psy1       | 2               | G52H, P53T                             |
|                             | bsl_psy2       | 4               | G52H, P53T, G92M, G93A                 |
| Acidostablizing mutants     | bsl_acd1       | 3               | G52F, V54Y, G111Y                      |
|                             | bsl_acd2       | 4               | G52F, V54Y, T110H, G111Y               |
| Alkalistablizing mutants    | bsl_alk1       | 2               | T110E, G111Y                           |
|                             | bsl_alk2       | 5               | W31E, S32H, T45H, T110E, G111Y         |
| Halostablizing mutants      | bsl_hal1       | 4               | G52D, P53D, P119D, N120D               |
|                             | bsl_hal2       | 5               | G52D, P53D, G93D, P119D, N120D         |
| Barostablizing mutants      | bsl_bar        | 4               | S24R, P119R, Q164R, G175R              |

### 5.3.2 *In silico* validation of designed mutations through AHP ranking models

More than 50 combinations of mutations were tested out in wild-type *B. subtilis* 168 lipase for each dataset. The mutations were carried out to increase the content of highest priority attributes obtained in the AHP ranking analysis for overall increasing the stability of lipase to sustain different extreme *milieus* (such extremes of temperatures and pH, higher pressure and salt concentrations). Only those mutations were chosen for *in silico* validation that led to an increase in the rank of the mutant w.r.t. the wild type. The 3D structure of wild-type lipase and modelled protein structures of mutant lipases were used to enumerate the statistically significant features. Further, the combinations of mutations were chosen for validation by *in silico* mutagenesis. Therefore, the mutational stability of the changed residues was also analyzed through the developed AHP models for individual extremophile dataset in terms of their ranks. The ranks of the mutated and wild-type lipases were calculated through AHP models for each dataset. Figure 5.2 represents the enumeration of statistically significant attributes and calculation of relative AHP ranks.

| Lipase   | Thermostabilizing mutations            | Change in content (%) of attributes in lipase after mutations |      |       |       |       |       |       |       |       |       |       |       |      |        |       |       | Rank difference ( $R_{mut} - R_{wt}$ ) |       |      |                        |
|----------|----------------------------------------|---------------------------------------------------------------|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|--------|-------|-------|----------------------------------------|-------|------|------------------------|
|          |                                        | Ala                                                           | Cys  | Asp   | Glu   | His   | Ile   | Lys   | Asn   | Gln   | Arg   | Ser   | Thr   | Val  | Tiny   | Sml   | Pol   | Chrg                                   | Bsc   | Acid |                        |
| bsl wt   | -                                      | 6.63                                                          | 0    | 4.97  | 1.66  | 2.76  | 5.52  | 6.08  | 9.39  | 3.31  | 2.76  | 7.18  | 5.52  | 9.39 | 32.60  | 58.56 | 43.65 | 18.23                                  | 11.60 | 6.63 |                        |
| bsl the1 | V149K, Q150E                           | 6.63                                                          | 0    | 4.97  | 2.21  | 2.76  | 5.52  | 6.63  | 9.39  | 2.95  | 2.76  | 7.18  | 5.52  | 8.84 | 32.60  | 58.56 | 44.39 | 19.33                                  | 12.15 | 7.19 | $0.51 - 0.49 = 0.02$   |
| bsl the2 | F41K, W42E, V149K, Q150E               | 6.63                                                          | 0    | 4.97  | 2.76  | 2.76  | 5.52  | 7.18  | 9.39  | 2.95  | 2.76  | 7.18  | 5.52  | 8.84 | 32.60  | 58.56 | 45.11 | 20.44                                  | 12.70 | 7.73 | $0.52 - 0.48 = 0.04$   |
| bsl the3 | F41K, W42E, P119E, Q121K, V149K, Q150E | 6.63                                                          | 0    | 4.97  | 3.31  | 2.76  | 5.52  | 7.73  | 9.39  | 2.21  | 2.76  | 7.18  | 5.52  | 8.84 | 32.60  | 57.46 | 45.86 | 21.55                                  | 13.26 | 8.29 | $0.53 - 0.47 = 0.06$   |
| Lipase   | Psychro stabilizing mutations          | Change in content (%) of attributes in lipase after mutations |      |       |       |       |       |       |       |       |       |       |       |      |        |       |       | Rank difference ( $R_{mut} - R_{wt}$ ) |       |      |                        |
|          |                                        | Ala                                                           | His  | Met   | Asn   | Ser   | Thr   | Trp   | Tiny  | Ala   | Arg   | Pro   | Bsc   | Acid |        | NPASA | CASA  |                                        | GT    |      |                        |
| bsl wt   | -                                      | 6.63                                                          | 2.76 | 2.21  | 9.39  | 7.18  | 5.52  | 1.10  | 32.60 | 30.39 | 11.05 | 11.60 | 6.63  | 6.63 |        | 0.60  | 0.13  | 0.55                                   |       |      |                        |
| bsl psy1 | G52H, P53T                             | 6.63                                                          | 3.31 | 2.21  | 9.39  | 7.18  | 6.08  | 1.10  | 32.32 | 30.94 | 11.60 | 12.15 | 6.63  | 6.63 |        | 0.60  | 0.14  | 1.66                                   |       |      | $0.51 - 0.49 = 0.02$   |
| bsl psy2 | G52H, P53T, G92M, G93A                 | 7.18                                                          | 3.31 | 2.76  | 9.39  | 7.18  | 6.08  | 1.10  | 32.04 | 30.94 | 11.60 | 12.15 | 6.63  | 6.63 |        | 0.60  | 0.15  | 2.76                                   |       |      | $0.52 - 0.48 = 0.04$   |
| Lipase   | Acidostabilizing mutations             | Change in content (%) of attributes in lipase after mutations |      |       |       |       |       |       |       |       |       |       |       |      |        |       |       | Rank difference ( $R_{mut} - R_{wt}$ ) |       |      |                        |
|          |                                        | HI                                                            | DS   | MMH   | MSH   | SSH   | II    | AAI   | ASI   | CPI   | SB    | HB    | PASA  | CASA | PV     | AH    | BS    |                                        |       |      |                        |
| bsl wt   | -                                      | 5.26                                                          | 0    | 13.34 | 4.96  | 1.85  | 0.56  | 0.04  | 0.04  | 0.15  | 0.48  | 79.00 | 0.26  | 0.13 | 858.78 | 5.52  | 3.31  |                                        |       |      |                        |
| bsl acd1 | G52F, V54Y, G111Y                      | 5.48                                                          | 0    | 13.81 | 5.06  | 1.45  | 0.56  | 0.12  | 0.04  | 0.15  | 0.42  | 84.00 | 0.27  | 0.13 | 846.78 | 5.52  | 3.31  |                                        |       |      |                        |
| bsl acd2 | G52F, V54Y, T110H, G111Y               | 5.60                                                          | 0    | 14.19 | 5.15  | 1.27  | 0.65  | 0.18  | 0.04  | 0.22  | 0.40  | 84.00 | 0.28  | 0.14 | 837.12 | 5.52  | 3.31  |                                        |       |      |                        |
| Lipase   | Alkalistabilizing mutations            | Change in content (%) of attributes in lipase after mutations |      |       |       |       |       |       |       |       |       |       |       |      |        |       |       | Rank difference ( $R_{mut} - R_{wt}$ ) |       |      |                        |
|          |                                        | HI                                                            | DS   | MMH   | MSH   | SSH   | II    | AAI   | ASI   | CPI   | SB    | HB    | PASA  | CASA | PV     | AH    | BS    |                                        |       |      |                        |
| bsl wt   | -                                      | 5.26                                                          | 0    | 13.34 | 4.96  | 1.85  | 0.56  | 0.04  | 0.04  | 0.15  | 0.48  | 79.00 | 0.26  | 0.13 | 858.78 | 5.52  | 3.31  |                                        |       |      |                        |
| bsl alk1 | T110E, G111Y                           | 5.25                                                          | 0    | 13.58 | 5.16  | 1.89  | 0.64  | 0.07  | 0.04  | 0.15  | 0.44  | 79.00 | 0.26  | 0.14 | 848.78 | 5.52  | 3.31  |                                        |       |      |                        |
| bsl alk2 | W31E, S32H, T45H, T110E, G111Y         | 5.24                                                          | 0    | 14.01 | 5.67  | 1.96  | 0.91  | 0.07  | 0.04  | 0.18  | 0.44  | 82.00 | 0.24  | 0.17 | 832.34 | 5.52  | 3.31  |                                        |       |      |                        |
| Lipase   | Halostabilizing mutations              | Change in content (%) of attributes in lipase after mutations |      |       |       |       |       |       |       |       |       |       |       |      |        |       |       | Rank difference ( $R_{mut} - R_{wt}$ ) |       |      |                        |
|          |                                        | Ala                                                           | Asp  | Ile   | Asn   | Arg   | Thr   | Tiny  | Sml   | Ala   | Npol  | Pol   | Bsc   | Acid |        | HI    | CPI   | SB                                     | NPASA |      |                        |
| bsl wt   | -                                      | 7.18                                                          | 4.97 | 5.52  | 9.39  | 2.76  | 6.08  | 32.60 | 58.56 | 30.39 | 56.35 | 43.65 | 11.60 | 6.63 |        | 5.48  | 0.15  | 0.48                                   | 0.60  |      |                        |
| bsl hal1 | G52D, P53D, P119D, N120D               | 7.18                                                          | 7.11 | 5.52  | 8.56  | 2.76  | 6.08  | 32.60 | 58.56 | 30.39 | 56.32 | 45.03 | 11.60 | 8.55 |        | 5.34  | 0.12  | 0.43                                   | 0.60  |      | $0.52 - 0.48 = 0.04$   |
| bsl hal2 | G52D, P53D, G93D, P119D, N120D         | 7.18                                                          | 8.29 | 5.52  | 8.29  | 2.76  | 6.08  | 32.60 | 58.56 | 30.39 | 55.25 | 45.86 | 11.60 | 9.94 |        | 5.29  | 0.11  | 0.41                                   | 0.60  |      | $0.525 - 0.475 = 0.05$ |
| Lipase   | Barostabilizing mutations              | Change in content (%) of attributes in lipase after mutations |      |       |       |       |       |       |       |       |       |       |       |      |        |       |       | Rank difference ( $R_{mut} - R_{wt}$ ) |       |      |                        |
|          |                                        | Glu                                                           | Arg  | Ser   | Tiny  | Chrg  | Bsc   | Acid  | HI    | II    | CPI   |       |       |      |        | SB    | BSH   | BST                                    |       |      |                        |
| bsl wt   | -                                      | 3.31                                                          | 2.76 | 7.18  | 32.60 | 18.23 | 11.60 | 6.63  | 5.48  | 0.56  | 0.15  |       |       |      |        | 0.48  | 0.55  | 3.31                                   |       |      |                        |
| bsl bar  | S24R, P119R, Q164R, G175R              | 2.76                                                          | 4.97 | 6.63  | 31.49 | 20.44 | 13.81 | 6.63  | 5.28  | 0.65  | 0.14  | 0.50  |       |      |        | 0.50  | 0.55  | 3.31                                   |       |      | $0.53 - 0.47 = 0.06$   |

Increasing attributes w.r.t. bsl wt      Decreasing attributes w.r.t. bsl wt

Figure 5.2: Prediction of ranks by AHP generated models for different extreme-stabilization in lipase.



### 5.3.3 Validation of mutants through various mutation prediction tools/servers

For validation of chosen mutations and predicting their protein stability changes upon point mutations (single amino acid mutations) was tested by various web servers. They required the modelled protein 3D structure in Protein Data Bank format or single point mutated sequence with location of the residue to be mutated. The generated modelled mutants were subjected to these employed web server to identification of the mutability propensity of the chosen mutations in terms of  $\Delta\Delta G$  value change in the mutant *w.r.t.* wild-type lipase for engineering proteins' stability. The mutations are decided on the basis  $\Delta\Delta G$  value (in kcal/mol) by all of these web servers which is enumerated by the following equation:

$$\Delta\Delta G = \Delta G (\text{Mutant Protein}) - \Delta G (\text{WildType}) \quad (5.1)$$

These web-servers adopt the same input code by identifying the  $\Delta\Delta G$  values, they are of basically two labels: one represents the increased protein stability of mutant ( $\Delta\Delta G > 0$ , label is +), the other is associated with the destabilizing mutation ( $\Delta\Delta G < 0$ , label is -)<sup>28</sup>. The overall applicability of using these web-servers to confirmed the thermodynamic stability ( $\Delta\Delta G$ ) for the chosen point mutations and assists protein engineers with the rational design of site-specific mutations<sup>32</sup>. Table 5.2 summarizes stabilizing mutations (represented by S) and destabilizing mutations (represented by D).

Table 5.2: Validation of predicted mutations by online available tools/servers.

| Lipase mutants | Point Mutations | HOTSPOT | SCIDE | CUPSAT | I-MUTANT 2.0 | I-PTREE STAB | ERIS | WET STAB |
|----------------|-----------------|---------|-------|--------|--------------|--------------|------|----------|
| bsl_the        | F41K            | S       | S     | S      | D            | S            | S    | S        |
|                | W42E            | S       | S     | S      | D            | S            | D    | S        |
|                | P119E           | S       | S     | S      | S            | S            | S    | S        |
|                | Q121K           | S       | S     | S      | S            | S            | S    | S        |
|                | V149K           | S       | S     | S      | S            | S            | S    | S        |
|                | Q150E           | S       | S     | S      | D            | S            | S    | S        |
| bsl_psy        | G52H            | S       | S     | S      | S            | S            | S    | S        |
|                | P53T            | S       | S     | S      | D            | S            | S    | S        |
|                | G92M            | S       | S     | S      | D            | S            | D    | S        |
|                | G93A            | S       | S     | S      | S            | S            | S    | S        |
| bsl_acd        | G52F            | S       | S     | S      | S            | S            | D    | S        |

|         |        |   |   |   |   |   |   |   |
|---------|--------|---|---|---|---|---|---|---|
|         | V54Y   | S | S | S | D | S | D | S |
|         | T110H  | S | S | S | S | S | S | S |
|         | G111Y  | S | S | S | D | S | S | S |
| bsl_alk | W31E   | S | S | S | D | S | D | S |
|         | S32H   | S | S | S | S | S | S | S |
|         | T45H   | S | S | S | D | S | S | S |
|         | G46Y   | S | S | S | S | S | S | S |
|         | T110E  | S | S | S | S | S | S | S |
|         | G111Y  | S | S | S | D | S | D | S |
| bsl_hal | G52D   | S | S | S | S | S | S | S |
|         | P53D   | S | S | S | S | S | S | S |
|         | G93D   | S | S | S | S | S | S | S |
|         | N 94 D | S | S | S | S | S | S | S |
|         | P119D  | S | S | S | S | S | S | S |
|         | N120D  | S | S | S | S | S | S | S |
| bsl_bar | S24R   | S | S | S | D | S | S | S |
|         | P119R  | S | S | S | S | S | S | S |
|         | Q164R  | S | D | S | D | S | D | S |
|         | G175R  | S | S | S | D | S | S | S |

S: stabilizing mutations; D: destabilizing mutations

### 5.3.4 Validation of mutants by Ramachandran plot

The rotations of the polypeptide backbone of wild-type and mutant lipases around their bonds between N-C<sub>α</sub> (called Phi, φ) and C<sub>α</sub>-C (called Psi, ψ) was determined by Ramachandran plot analysis. It provides an easy way to view the distribution of these torsion angles in a protein structure and an overview of excluded regions that show which rotations of the polypeptide are not allowed due to steric hindrance (collisions between atoms). Thus, Ramachandran plots were generated for each extreme-stabilizing mutations obtained from PROCHECK to validate the backbone structure of wild-type and mutant lipases. They have assigned allowed and disallowed regions represented by different color codes such as red indicates low-energy regions, brown allowed regions, yellow the so-called generously-allowed regions and pale-yellow marks disallowed regions. Figure 5.3 represents the generated Ramachandran plots for wild-type and mutant lipases and Table 5.3 summarizes the percentage of residues in the allowed (favored) region.

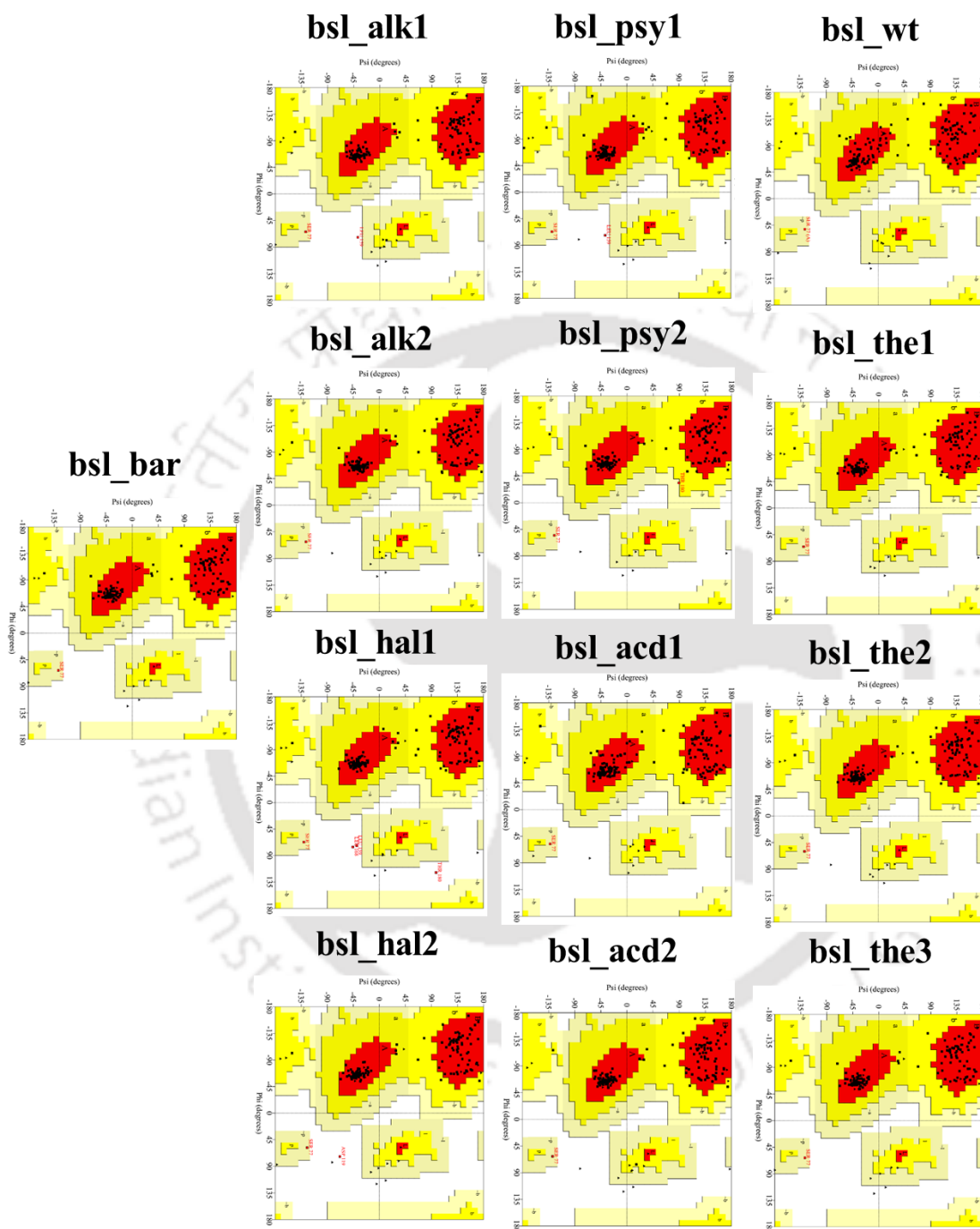


Figure 5.3: Analysis of residues in the most favorable region by Ramachandran plot.

Table 5.3: Ramachandran plot enumeration of generated mutations.

| Lipases  | Total residue | No. of Gly (23) + Pro (3) + End residues (2) | No. of Non-Gly and Non-Pro | Residues in most favoured region | Percentage of residues in most favoured region |
|----------|---------------|----------------------------------------------|----------------------------|----------------------------------|------------------------------------------------|
| bsl_wt   | 181           | 28                                           | 153                        | 135                              | 88.2                                           |
| bsl_the1 | 181           | 28                                           | 153                        | 140                              | 91.5                                           |
| bsl_the2 | 181           | 28                                           | 153                        | 142                              | 92.8                                           |
| bsl_the3 | 181           | 28                                           | 153                        | 143                              | 93.5                                           |
| bsl_psy1 | 181           | 28                                           | 153                        | 140                              | 91.5                                           |
| bsl_psy2 | 181           | 28                                           | 153                        | 142                              | 92.8                                           |
| bsl_acd1 | 181           | 28                                           | 153                        | 141                              | 92.2                                           |
| bsl_acd2 | 181           | 28                                           | 153                        | 143                              | 93.5                                           |
| bsl_alk1 | 181           | 28                                           | 153                        | 139                              | 90.8                                           |
| bsl_alk2 | 181           | 28                                           | 153                        | 142                              | 92.8                                           |
| bsl_hal1 | 181           | 28                                           | 153                        | 140                              | 91.5                                           |
| bsl_hal2 | 181           | 28                                           | 153                        | 145                              | 94.8                                           |
| bsl_bar  | 181           | 28                                           | 153                        | 142                              | 92.8                                           |

### 5.3.5 Structure superimposition validation of generated mutations

The structural superimpositions were analysed in terms of root mean square deviation (RMSD) alignment of two different conformations of the same protein can be skewed by the difference between the mobile regions. Higher the RMSD value, more target structure deviate from template structure. Here, structural superimpositions of generated lipase mutants *w.r.t.* wild-type lipase was performed using PyMol V0.99. The analyses revealed that the wild type and mutated structures were alike without any massive structural changes since the RMSD value of mutants lipases were found to be close to the RMSD value of wild-type lipase. Table 5.4 summarizes RMSD values of mutants obtained *w.r.t.* wild type lipase in the superimposition analysis.

Table 5.4: Obtained RMSD values of mutant lipases in superimposition analysis.

| Mutant   | Superimposition (RMSD) | Mutant   | Superimposition (RMSD) |
|----------|------------------------|----------|------------------------|
| bsl_the1 | 0.225                  | bsl_acd2 | 0.223                  |
| bsl_the2 | 0.234                  | bsl_alk1 | 0.218                  |
| bsl_the3 | 0.242                  | bsl_alk2 | 0.229                  |
| bsl_psy1 | 0.201                  | bsl_hal1 | 0.231                  |
| bsl_psy2 | 0.204                  | bsl_hal2 | 0.243                  |
| bsl_acd1 | 0.215                  | bsl_bar  | 0.223                  |

### 5.3.6 Enzyme-substrate docking analysis for functional efficiency

Comparative molecular docking of generated mutants and wild-type lipase using C8 (*p*-NPO) substrate were independently performed in Autodock 4.2 version. The docking results indicated enhanced stability and better substrate binding of the mutated lipase structures having substitutions residues that increase the content of positively contributing attributes which did not disturb the catalytic properties. Wild type structure of *B. subtilis* 168 lipase with no substitutions showed relatively low stability and substrate binding capacity when compared to the mutant structures. Moreover, the binding pockets of mutated structures were intact as the wild type structures. The binding energy in Autodock suite is the sum of intermolecular energy and torsional energy. The results showed that the mutant lipases had slightly less binding energy than that of the native wild type structure ( $-5.49 \text{ kcal mol}^{-1}$ ) (Table 5.5). However, the binding pocket was intact and thus this change can be attributed negligible due to the structural stability of the lipase (Figure 5.4). Larger Inhibition constant (*k*<sub>I</sub>) value indicates bad docking. There was a massive decrease in *k*<sub>I</sub> value observed in mutated structures as well. The results also showed variation in the docking parameters in lipase mutants although the chosen mutations were far away from the site of docking (catalytic site). It may be that the mutations affected the number of residue and hydrogen bond involved in the interaction with *p*-NPO substrate at the active site. This leads to the change in proximity of substrate at the active site and affects the docking parameter even in the rigid docking as the mutational changes remained in the structure before performing rigid docking.

Table 5.5: Molecular docking of lipase mutants with C8 (*p*-NPO) substrate.

| Lipase   | Binding Energy ( $\text{kcal mol}^{-1}$ ) | Inhibition Constant ( $\mu\text{M}$ ) | Unbound Extended Energy ( $\text{kcal/mol}$ ) | Residues involved in interaction with <i>p</i> -NPO | No. of H-bond formed in lipase- <i>p</i> -NPO interaction |
|----------|-------------------------------------------|---------------------------------------|-----------------------------------------------|-----------------------------------------------------|-----------------------------------------------------------|
| bsl_wt   | -5.99                                     | 40.71                                 | -0.49                                         | G11, N18, H156                                      | 4                                                         |
| bsl_the1 | -6.11                                     | 16.53                                 | -0.18                                         | G11, N18, H76                                       | 4                                                         |
| bsl_the2 | -6.36                                     | 21.69                                 | -0.28                                         | G11, N18, H156                                      | 4                                                         |
| bsl_the3 | -7.61                                     | 2.52                                  | -0.87                                         | G11, N18, H67, H156                                 | 5                                                         |
| bsl_psy1 | -6.17                                     | 25.39                                 | -0.54                                         | G11, N18, H67, H156                                 | 4                                                         |
| bsl_psy2 | -6.27                                     | 18.25                                 | -0.58                                         | G11, N18, H156                                      | 4                                                         |

|          |       |       |       |                     |   |
|----------|-------|-------|-------|---------------------|---|
| bsl_acd1 | -5.93 | 41.98 | -0.50 | G11, N18, H76       | 3 |
| bsl_acd2 | -5.96 | 40.04 | -0.50 | G11, H67, H156      | 4 |
| bsl_alk1 | -7.10 | 9.69  | -0.79 | G11, N18, H67, H156 | 5 |
| bsl_alk2 | -7.60 | 2.69  | -0.86 | G11, N18, H67, H156 | 5 |
| bsl_hal1 | -6.12 | 29.04 | -0.41 | G11, H67, H156      | 4 |
| bsl_hal2 | -6.25 | 26.15 | -0.43 | G11, N18, H76       | 4 |
| bsl_bar  | -7.50 | 5.69  | -0.83 | G11, N18, H67, H156 | 5 |

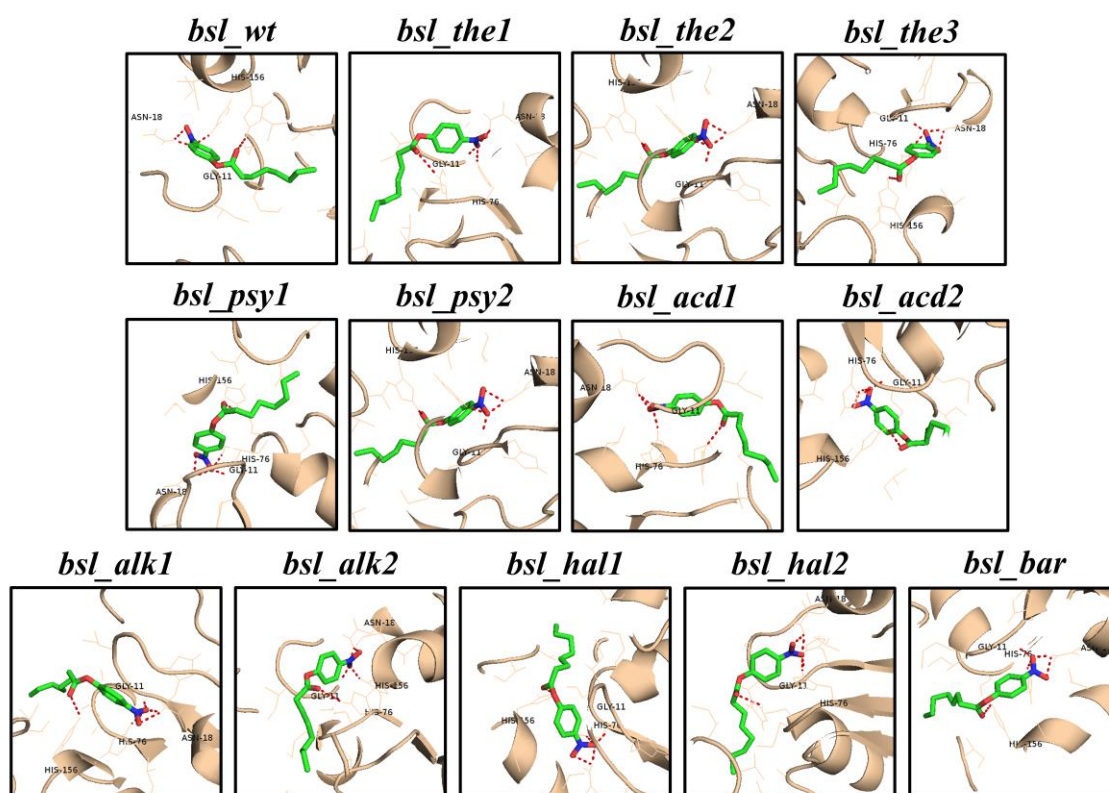


Figure 5.4 Interaction of *p*-NPO with active site of wild type and mutant lipases

### 5.3.7 Contact map analysis for enumeration of interactions in lipases

Contact map represent the residue interaction networks in two dimensions, which facilitate the identification of structural features such as interactions within and between secondary structures and domains. Unique contacts with respect to modelled structures of mutated structures of *Bacillus subtilis* lipases (bsl\_the1, bsl\_the2, bsl\_the3, bsl\_psy1, bsl\_psy2, bsl\_acd1, bsl\_acd2, bsl\_alk1, bsl\_alk2, bsl\_hal1, bsl\_hal2 and bsl\_bar) and their available PDB structure of wild type lipase (bsl\_wt) were analyzed using CMView



1.1.1. The contact maps and their protein model structure showing unique interactions have been presented in Appendix Figure A2.21. In contact map, the black dots represents common contacts in wild-type and mutant lipases, pink dots for contacts unique to the wild-type structure and green for contacts unique to the lipase mutants. It can be clearly observed that the total number of unique contacts for all the mutants was observed to be greater than the wild type. The total number of contacts formed is higher than the number of contacts lost suggesting an increase in the compactness of the protein. The graphical representations unique contacts in all the mutant lipases were compared with wild-type lipases have been illustrated in Figure 5.5. It can be concluded that as the mutants generated through AHP ranking models showed similar trends of unique contacts with those of already reported extreme-stable mutants, they have high probability of being extreme-stable.

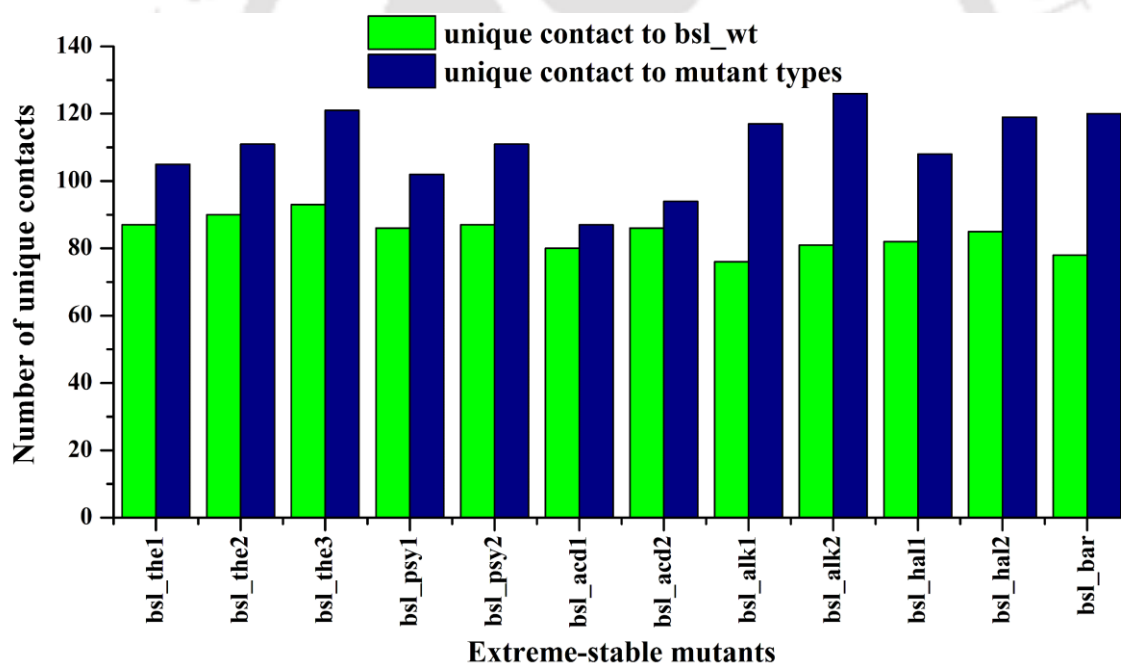


Figure 5.5: Graphical representation of unique contacts formed in wild-type and mutants.

### 5.3.8 *In silico* prediction of secondary structure of mutants

The secondary structure such as  $\alpha$ -helices and  $\beta$ -sheets played a crucial role in making the protein structure stable<sup>22,23</sup>. They stabilize the protein structure under extreme *milieus* by improving the contents of secondary structure such as hydrophobic contacts,

backbone hydrogen bonding, etc.<sup>24</sup>. The  $\alpha$ -helices are largely stabilized by backbone hydrogen bonding. That is, local interactions dominate in a helix, whereas a sheet is stabilized by long range contacts. So, a sheet is slightly inferior in terms of stability. The prediction of secondary structure of wild-type and mutant lipases were done through PSIPRED server<sup>39</sup>. The amino acid contributing to formation of helices and sheets were found to be increased to the wild-type lipases, expect that of psychrophilic mutants (bsl\_psy1 and bsl\_psy2). The increased content of  $\alpha$ -helices and  $\beta$ -sheets was reported which makes the lipase mutants more rigid by reducing the content of random coil makes the protein rigid to sustain extreme-conditions. Thus, chosen mutants were considered as stabilizing mutants as improved secondary structures in mutant lipases. Table 5.6 represents the percentage of secondary structure in the wild-type and mutant lipases.

Table 5.6: Predicted percentage of secondary structure of wild-type and mutant lipases.

| Lipases  | Secondary structures (%) |                 |             |
|----------|--------------------------|-----------------|-------------|
|          | $\alpha$ -helices        | $\beta$ -sheets | Random coil |
| bsl_wt   | 45.2                     | 33.3            | 21.5        |
| bsl_the1 | 47.1                     | 34.2            | 18.7        |
| bsl_the2 | 49.2                     | 33.3            | 17.5        |
| bsl_the3 | 55.2                     | 29.1            | 15.7        |
| bsl_psy1 | 43.6                     | 35              | 21.4        |
| bsl_psy2 | 41.3                     | 35.2            | 23.5        |
| bsl_acd1 | 44.2                     | 34.3            | 21.5        |
| bsl_acd2 | 47.2                     | 30.9            | 21.9        |
| bsl_alk1 | 51.7                     | 34.8            | 13.5        |
| bsl_alk2 | 56.7                     | 37.9            | 5.4         |
| bsl_hal1 | 48.2                     | 30.1            | 21.7        |
| bsl_hal2 | 50.8                     | 31.1            | 18.1        |
| bsl_bar  | 53.2                     | 30.3            | 18.5        |

## 5.4 Conclusion

All the *in silico* validation methods in *Bacillus subtilis* lipase showed enhanced stability of the generated mutants through this rationalized approach. The preferred protein attributes to attain extreme-stability of proteins were substituted by designing *in silico* homology models. Three thermostable mutants, two psychrostable mutants, two acidostable mutants, two alkalistable mutants, two halostable mutants and one barostable

mutant were designed. AHP generated ranking models, various mutation prediction servers (HotSpot Wizard, I-Mutant2, Cupsat, iPTREE-STAB, WET-STAB and ERIS webserver), Substrate Docking (Autodock 4.2 version), Ramachandran analysis (PROCHECK) and Contact Map analysis (CMView 1.1.1) were used for *in silico* validation of mutant. Results showed that the designed mutations did not affect the catalytic activity of the lipase and thus, mutation did not disturb the catalytic properties.



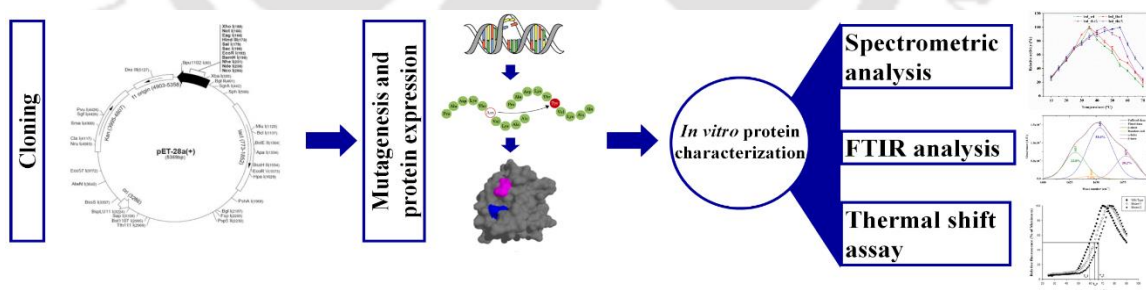
## 5.5 References

1. Jaenicke, R. Do ultrastable proteins from hyperthermophiles have high or low conformational rigidity? *Proc. Natl. Acad. Sci.* **97**, 2962–2964 (2000).
2. Lehmann, M. & Wyss, M. Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr. Opin. Biotechnol.* **12**, 371–5 (2001).
3. Reetz, M. T., Carballeira, J. D. & Vogel, A. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed. Engl.* **45**, 7745–51 (2006).
4. Karplus, P. A. & Schulz, G. E. Prediction of chain flexibility in proteins. *Naturwissenschaften* **72**, 212–213 (1985).
5. Parthasarathy, S. & Murthy, M. R. Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng.* **13**, 9–13 (2000).
6. Schlessinger, A., Yachdav, G. & Rost, B. PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics* **22**, 891–893 (2006).
7. Duan, Y. *et al.* The role of leucine and its metabolites in protein and energy metabolism. *Amino Acids* **48**, 41–51 (2016).
8. Xie, Y. *et al.* Enhanced enzyme kinetic stability by increasing rigidity within the active site. *J. Biol. Chem.* **289**, 7994–8006 (2014).
9. Zhang, J. *et al.* High-throughput screening of B factor saturation mutated *Rhizomucor miehei* lipase thermostability based on synthetic reaction. *Enzyme Microb. Technol.* **50**, 325–30 (2012).
10. Kim, S. J. *et al.* The development of a thermostable CiP (*Coprinus cinereus* peroxidase) through in silico design. *Biotechnol. Prog.* **26**, 1038–46 (2010).
11. Spector, S. *et al.* Rational modification of protein stability by the mutation of charged surface residues. *Biochemistry* **39**, 872–9 (2000).
12. Srivastava, A. & Sinha, S. Thermostability of in vitro evolved *Bacillus subtilis* lipase A: a network and dynamics perspective. *PLoS One* **9**, e102856 (2014).
13. Lutz, S. & Patrick, W. M. Novel methods for directed evolution of enzymes: quality, not quantity. *Curr. Opin. Biotechnol.* **15**, 291–7 (2004).
14. Parthiban, V., Gromiha, M. M. & Schomburg, D. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* **34**, W239–42 (2006).
15. Vila, J. A., Villegas, M. E., Baldoni, H. A. & Scheraga, H. A. Predicting  $^{13}\text{C } \alpha$  chemical shifts for validation of protein structures. *J. Biomol. NMR* **38**, 221–235 (2007).
16. Ebrahimi, M., Lakizadeh, A., Agha-Golzadeh, P., Ebrahimi, E. & Ebrahimi, M. Prediction of Thermostability from Amino Acid Attributes by Combination of Clustering with Attribute Weighting: A New Vista in Engineering Enzymes. *PLoS One* **6**, e23146 (2011).
17. Di Giulio, M. The origin of the genetic code: Theories and their relationships, a review. *BioSystems* **80**, 175–184 (2005).
18. Dartois, V., Baulard, A., Schanck, K. & Colson, C. Cloning, nucleotide sequence and expression in *Escherichia coli* of a lipase gene from *Bacillus subtilis* 168. *Biochim. Biophys. Acta - Gene Struct. Expr.* **1131**, 253–260 (1992).
19. Acharya, P., Rajakumara, E., Sankaranarayanan, R. & Rao, N. M. Structural Basis of Selection and Thermostability of Laboratory Evolved *Bacillus subtilis* Lipase. *J. Mol. Biol.* **341**, 1271–1281 (2004).
20. Westers, H. *et al.* Genes involved in SkfA killing factor production protect a *Bacillus subtilis* lipase against proteolysis. *Appl. Environ. Microbiol.* **71**, 1899–1908 (2005).
21. Van Pouderoyen, G., Eggert, T., Jaeger, K.-E. & Dijkstra, B. W. The crystal structure of *Bacillus subtilis* lipase: a minimal  $\alpha/\beta$  hydrolase fold enzyme. *J. Mol. Biol.* **309**, 215–226 (2001).
22. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–38 (2010).
23. Willard, L. *et al.* VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.* **31**, 3316–9 (2003).
24. Tina, K. G., Bhadra, R. & Srinivasan, N. PIC: Protein Interactions Calculator. *Nucleic Acids Res.* **35**, W473–6 (2007).
25. Costantini, S., Colonna, G. & Facchiano, A. M. ESBRI: a web server for evaluating salt bridges in proteins. *Bioinformatics* **3**, 137–8 (2008).
26. Figueroa-Montiel, A. *et al.* In Silico Identification of Protein Disulfide Isomerase Gene Families in the

- De Novo Assembled Transcriptomes of Four Different Species of the Genus *Conus*. *PLoS One* **11**, e0148390 (2016).
27. Capriotti, E., Fariselli, P. & Casadio, R. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* **20 Suppl 1**, i63-8 (2004).
  28. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **33**, (2005).
  29. Yin, S., Ding, F. & Dokholyan, N. V. Eris: an automated estimator of protein stability. *Nat. Methods* **4**, 466 (2007).
  30. Capriotti, E., Fariselli, P., Rossi, I. & Casadio, R. A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* **9 Suppl 2**, S6 (2008).
  31. Pavelka, A., Chovancova, E. & Damborsky, J. HotSpot Wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Res.* **37**, W376-83 (2009).
  32. Bendl, J. *et al.* HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Res.* **44**, W479–W487 (2016).
  33. Hoof, R. W. W., Sander, C. & Vriend, G. Objectively judging the quality of a protein structure from a Ramachandran plot. *Bioinformatics* **13**, 425–430 (1997).
  34. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
  35. Kelley, L. A. & Sternberg, M. J. E. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363–371 (2009).
  36. Rizvi, S. M. D., Shakil, S. & Haneef, M. A simple click by click protocol to perform docking: Autodock 4.2 made easy for non-bioinformaticians. *EXCLI J.* **12**, 830–857 (2013).
  37. Barah, P. & Sinha, S. Analysis of protein folds using protein contact networks. *Pramana* **71**, 369–378 (2008).
  38. Vehlow, C. *et al.* CMView: interactive contact map visualization and analysis. *Bioinformatics* **27**, 1573–4 (2011).
  39. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405 (2000).

## 6

## *In vitro* characterization and validation of predicted mutations





## Abstract

This chapter deals with experimentation employed for *in vitro* cloning, mutagenesis, expression, purification and characterization of lipases. Results revealed that the optimum activity and stability increased from 35 °C to 55 °C in bsl\_the3, pH 8 to pH 10 in bsl\_alk2 but reduced to pH 6.5 in bsl\_acd2 w.r.t. wild type lipase (bsl\_wt). Further, the bsl\_wt was halo-sensitive (0.5 M NaCl) but in bsl\_hal2 halostability was increased to 1.5 M NaCl. The content of secondary structure enhanced extreme-stability (with increased  $T_m$  value) which was confirmed through FT-IR (ATR mode) and Differential Scanning Fluorimetry (Thermal shift assay).

## 6.1 Introduction

Numerous research efforts have been there to improve extreme-stability of proteins with most efforts focusing on thermostability<sup>1-4</sup>. Most methods pertained to using of directed evolution approaches to attain the desired improved candidates<sup>2,3,5</sup>. The present chapter is a concerted effort towards developing a rationalized novel method to attain extremostable proteins. The method follows two steps: first, identification of extreme-stabilizing mutants with the help of numerical ranks calculated by AHP derived rank (discussed in earlier chapter); second, mutation by site directed mutagenesis and protein characterization for validation. The method is specific, pre predicted and reduces the cumbersome steps of colony screening and directed evolution. The rank provided to the mutant proteins is based on the overall contributory effect of their intra-molecular interactions (that have been earlier enumerated in the chapters of this dissertation) to be involved in extreme-stabilizing proteins<sup>3,5</sup>. Thus, the cumulative effect of stabilizing factors is considered while calculating the ranks. The method predicts that higher the rank of the mutated protein in comparison to the wild type, the more stable it will be rendered with respect to any extremostable parameter. The aforementioned protocol was employed to achieve mutants of *Bacillus subtilis* lipase. The wild type lipase is 181 amino acid residues, 19.4 kDa and can be regarded as a minimal  $\alpha/\beta$ -hydrolase fold enzyme<sup>6-8</sup>. *Bacillus subtilis* lipase was chosen since it is an industrially important enzyme finding use in detergent formulations and chiral synthesis. It does not show interfacial activation<sup>8</sup>. Multiple mutants were predicted to be extreme-stabilizing such as thermostabilizing, psychro stabilizing, acidostabilizing, alkalistabilizing and halostabilizing mutations. The generated AHP ranking models provided the mutants with higher ranks than the wild type lipase. The experimental validation of theoretical approach was initiated with cloning, mutagenesis, expression, purification and characterization of the selected mesostable enzyme *Bacillus subtilis* 168 lipase A (estA) for comparing the extreme-stability of wild-type and produced mutant lipases. Thus, the present work employed various molecular biology techniques for enzymes characterization and validation of attained mutations for further application.

## 6.2 Methodology

### 6.2.1 Strains, plasmids, enzymes, reagents and instruments

*Escherichia coli* DH5 $\alpha$  and BL21 (DE3) strains were procured from Invitrogen (USA) for cloning and protein overexpression, respectively. Both the strains were grown and maintained in Luria-Bertani (LB) medium (Sigma-Aldrich). The cloned pET-21b vector with wild type *Bacillus subtilis* 168 lipase was a kind gift from Dr. N. Madhusudhan Rao of Centre for Cellular & Molecular Biology (CCMB), India. The plasmid pET-28a(+) was procured from Addgene (USA) and used for sub-cloning and expression of recombinant enzyme production. Quick Ligation<sup>TM</sup> kit, BamHI and NdeI enzymes, DNA and protein markers were purchased from New England Biolabs (UK) and StrataPrep Plasmid Miniprep Kit, StrataPrep DNA Gel Extraction Kit, QuikChange Site-Directed Mutagenesis Kit and SYPRO orange dye were purchased from Agilent Technologies (USA). Ni-NTA resin and polypropylene columns were obtained from Qiagen (USA) for His-tagged enzyme purification. All other reagents, solvents used were of molecular biology/analytical grade and obtained from Merck (USA) and Himedia (India). SDS PAGE equipment was from Bio-Rad laboratories, RT-PCR (MX3005P) and Sure Cyclor from Agilent Technologies. Absorbance measurements were done on microtitre plate reader in TECAN Infinite Series M200 microplate spectrophotometer (TECAN, Salzburg, Austria).

### 6.2.2 Cloning and mutagenesis of wild-type and mutant enzymes

*B. subtilis* 168 lipase gene was first sub-cloned in pET-28a(+) plasmid using restrictions sites BamHI and NdeI since, pET-28a(+) plasmid have N-terminal poly-Histidine tag that helps in purification of recombinant enzyme using Ni-NTA affinity chromatography. The digested gene product (639bp) was gel extracted and ligated into pET-28a(+) plasmid using DNA Gel Extraction Kit and Quick Ligation<sup>TM</sup> Kit, respectively. The recombinant plasmid was transformed into competent *E. coli* DH5 $\alpha$  and plated on Luria-Bertani (LB) agar plates with 50  $\mu$ g/ml kanamycin for screening of putative clones. The recombinant vector having wild type lipase was utilized for site directed mutagenesis. The chosen mutations were selected for improving protein extreme-stability and were derived as per the ranking given

by generated ranking model by multi-criteria decision-making approach (mentioned in Chapter 5). The designed primers for multiple point mutations are enlisted in Table 6.1. Cloning and mutagenesis of wild-type and mutant lipase was confirmed through restriction digestion and sequence analysis (1<sup>st</sup> BASE DNA Sequencing Services, Singapore). The site directed mutagenesis was carried out using QuikChange Primer Design (Agilent Genomics) according to their protocol (Figure 6.1)

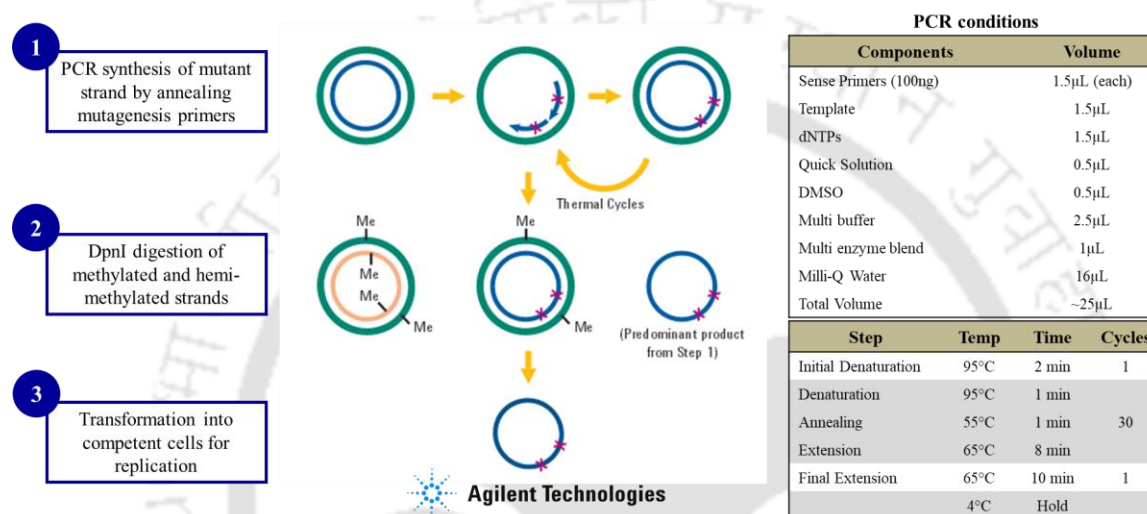


Figure 6.1: Methodology of multi-site mutagenesis using QuikChange Lightning Multi Site-Directed mutagenesis and PCR conditions. (Adapted from Agilent Technologies, QuikChange Lightning Multi Site-Directed Mutagenesis Kit Instruction Manual).

Table 6.1 Primers designed for predicted point mutations in *Bacillus subtilis* 168 lipase.

| S. No.                        | Mutations | Wild-type |       | Mutant |       | Mutagenesis Primer Sequence (for sense strand)                                    |                 | Primer Name |
|-------------------------------|-----------|-----------|-------|--------|-------|-----------------------------------------------------------------------------------|-----------------|-------------|
|                               |           | AA        | Codon | AA     | Codon |                                                                                   |                 |             |
| Thermostabilizing Mutations   |           |           |       |        |       |                                                                                   |                 |             |
| 1                             | F41K      | F         | TTT   | K      | AAA   | 5'-cattgtataattggcctgtctgtcttctttatcaactgcatacagctgtgccgaccag-3'                  | The_F41K_W42E   |             |
| 2                             | W42E      | W         | TGG   | E      | GAA   |                                                                                   |                 |             |
| 3                             | P119E     | P         | CCA   | E      | GAA   | 5'-tggatgtataaaaatctttttatctgttcccggaagcgcttgcc-3'                                | The_P119_Q121K  |             |
| 4                             | Q121K     | Q         | CAA   | K      | AAA   |                                                                                   |                 |             |
| 5                             | V149K     | V         | GTT   | K      | AAA   | 5'-ccgatgtccaacgccatggattctttgtttctagcaccatctaactcttg-3'                          | The_V149K_Q150E |             |
| 6                             | Q150E     | Q         | CAA   | E      | GAA   |                                                                                   |                 |             |
| Psychro stabilizing Mutations |           |           |       |        |       |                                                                                   |                 |             |
| 1                             | G52H      | G         | GGA   | H      | CAT   | 5'-taaaacctttgcacaaatcgtgataaacggatgatgttataattgtgcctgtctgtgccaa-3'               | Psy_G52H_P53T   |             |
| 2                             | P53T      | P         | CCG   | T      | ACC   |                                                                                   |                 |             |
| 3                             | G92M      | G         | GGC   | M      | ATG   | 5'-caacgcgcagcgttgcaactttatcgcctgtccagatttttagtagtaaaagtgtg-3'                    | Psy_G92M_G93A   |             |
| 4                             | G93A      | G         | GGA   | A      | GCG   |                                                                                   |                 |             |
| Acidostabilizing Mutations    |           |           |       |        |       |                                                                                   |                 |             |
| 1                             | G52F      | G         | GGA   | F      | TTT   | 5'-cgtttcatcaaaacctttgcacaaatcgtgataaacggaaaatgttataattgtgcctgtctgtcccaaaatcaa-3' | Acd_G52F_V54Y   |             |
| 2                             | V54Y      | V         | GTA   | Y      | TAT   |                                                                                   |                 |             |
| 3                             | T110H     | T         | ACA   | H      | CAT   | 5'-atctgttcccggaagcgcttataatgctgaacggctcgcgccg-3'                                 | Acd_T110H_G111Y |             |
| 4                             | G111Y     | G         | GGC   | Y      | TAT   |                                                                                   |                 |             |
| Alkalistabilizing Mutations   |           |           |       |        |       |                                                                                   |                 |             |
| 1                             | W31E      | W         | TGG   | E      | GAA   | 5'-caaaaatcaactgcatacagctgttcccgatgttcgcccttgagatacagagatagctttaattc-3'           | Alk_W31E_S32H   |             |
| 2                             | S32H      | S         | TCG   | H      | CAT   |                                                                                   |                 |             |
| 3                             | T45H      | T         | ACA   | H      | CAT   | 5'-tcgtgataataccggtccatgttataattgtataatgctgtcccaaaatcaactgcatacagcttgtc-3'        | Alk_T45H_G46Y   |             |
| 4                             | G46Y      | G         | GGC   | Y      | TAT   |                                                                                   |                 |             |
| 5                             | T110E     | T         | ACA   | E      | GAA   | 5'-ctgttcccggaagcgcttattccgcacacggctcgcgccg-3'                                    | Alk_T110E_G111Y |             |
| 6                             | G111Y     | G         | GGC   | Y      | TAT   |                                                                                   |                 |             |
| Halostabilizing Mutations     |           |           |       |        |       |                                                                                   |                 |             |
| 1                             | G52D      | G         | GGA   | D      | GAT   | 5'-ctaaaacctttgcaaaatcgtgataatacatcatcattgttataattgtcctgtctgtccaaa-3'             | Hal_G52D_P53D   |             |
| 2                             | P53D      | G         | CCG   | D      | GAT   |                                                                                   |                 |             |
| 3                             | G93D      | G         | GGA   | D      | GAT   | 5'-ccgtcacgacgttgcaactttatcatcgccgtccagatttttagtag-3'                             | Hal_G93D_N94D   |             |
| 4                             | N94D      | N         | AAT   | D      | GAT   |                                                                                   |                 |             |
| 5                             | P119D     | P         | CCA   | D      | GAT   | 5'-aatggatgtataaaatcttttgatcatcatctgttccgggaagcgcttgcc-3'                         | Hal_P119D_N120D |             |
| 6                             | N120D     | N         | AAT   | D      | GAT   |                                                                                   |                 |             |

### 6.2.3 Overexpression and purification of wild-type and mutant enzymes

The recombinant plasmids of wild-type and mutant lipases were purified using StrataPrep Plasmid Miniprep Kit<sup>9</sup>. The purified plasmids were transformed into competent *E. coli* BL21 (DE3) and were cultured for optimization of lipase overexpression in 5 ml culture broth after induction in 50  $\mu$ M and 100  $\mu$ M IPTG at 16 °C, 25 °C and 37 °C for 4 h, 8 h and 12 h. The large-scale expression and purification of lipases was performed under optimized conditions in 1 l culture broth. The induced cultured cells were harvested by centrifugation at 8000 rpm for 10 minutes and then re-suspended in 10 ml sodium phosphate (20mM) lysis buffer (pH 8) containing 0.5 mg ml<sup>-1</sup> of lysozyme. The suspended cells were sonicated to recover the soluble enzymes by centrifugation at 14,000  $\times$  g, 4 °C for 20 minutes. Wild-type and mutant lipases were purified from obtained supernatants by affinity chromatography using Ni-NTA columns<sup>10</sup>. Finally, the bound enzymes were eluted using sodium phosphate buffer (pH 8) having 250 mM imidazole in 1 ml fractions. The purification of lipases was confirmed by SDS-PAGE. The free imidazole in the purified lipases was removed by dialysis in the same buffer at 4 °C, followed by determination of concentrations of lipases by Bradford assay using BSA as standard (Appendix Figure A2.23)<sup>11,12</sup>.

### 6.2.4 Kinetic characterization of wild-type and mutant enzymes

Activity of wild-type and mutant lipases were determined spectrophotometrically using *p*-nitrophenyl octanoate (*p*-NPO) as substrate as per Glogauer et al. with slight modifications<sup>13</sup>. The enzyme-substrate reactions were incubated at 35 °C for 10 min with varying *p*-NPO concentrations (0.002 to 0.5 mM) and constant lipase concentration (0.1 mg ml<sup>-1</sup>). For preparation of *p*-NPO solution, *p*-NPO was dissolved in 2-propanol to make the concentration of 3 mg ml<sup>-1</sup>. 0.2 % triton X 100 and 0.43 M tetrahydrofuran were added. Finally, the substrate solution consisted of 0.1 ml of *p*-NPO solution and 2.8 ml phosphate buffer (20 mM, pH8). The temperature and substrate was chosen as they were reported to be optimum for the wild type lipase<sup>2,6</sup>. The reactions were immediately stopped by immersing the reaction mixture in ice and adding 100  $\mu$ l of 0.1 M HCl. Then, 125  $\mu$ l was transferred to a 96-well microtiter plate, completed to 250  $\mu$ l with deionized water and the



amount of released *p*-nitrophenol (*p*-NP) was measured at 410 nm. The molar extinction coefficient of *p*-NP ( $\epsilon = 15.1 \text{ mM}^{-1} \text{ cm}^{-1}$ ) was estimated from the absorbance measured at 410 nm of standard solutions of *p*-NP. One enzyme unit (U) was defined as the lipase activity that liberated  $1 \text{ } \mu\text{mol}$  equivalent of *p*-NP  $\text{ml}^{-1} \text{ min}^{-1}$  under the standard assay conditions. All assays were performed independently and in triplicates for all lipase types (wild-type and mutant). The enzyme kinetic parameter such as activity, specific activity,  $K_m$  value, turnover number ( $k_{\text{cat}}$ ) and catalytic efficiency ( $k_{\text{cat}}/K_m$ ) was calculated for the wild type and mutant enzymes.

### 6.2.5 Effect of temperature, pH and NaCl concentration on lipase activity and stability

The effect of temperature, pH and NaCl concentrations on lipases activity was determined by incubating the enzyme-substrate reaction mixture for 10 min at varying range of temperature (from  $10^\circ\text{C}$  to  $70^\circ\text{C}$ ) and pH (from pH 3 to pH 12) and salt concentrations (from 0 M to 3.5 M NaCl), respectively. Then, the optimum activity under aforementioned extreme environment conditions were determined independently.

On the other hand, wild-type and mutant lipases were used to determine the thermal stability at temperature  $50^\circ\text{C}$  (for thermostabilizing lipase mutants – bsl\_the1, bsl\_the2 and bsl\_the3) and  $25^\circ\text{C}$  (for psychro stabilizing lipase mutants – bsl\_psy1 and bsl\_psy2); pH stability at pH 6.5 (for acidostabilizing lipase mutants – bsl\_acd1 and bsl\_acd2) and pH 9 (for alkalistabilizing lipase mutants – bsl\_alk1 and bsl\_alk2); and halostability at 1M NaCl concentration (for halostabilizing lipase mutants – bsl\_hal1 and bsl\_hal2). The stability studies were carried out by incubating the mutant and wild-type lipases upto 24 h performed under aforementioned stability conditions and then assayed under the standard assay conditions.

### 6.2.6 Differential scanning fluorimetry analysis for determining unfolding ( $T_m$ value)

The differential scanning fluorimetry (DSF) can identify the temperature at which a protein unfolds ( $T_m$ ). This is measured by an increase in the fluorescence of a dye which binds with affinity for hydrophobic parts of the protein as the protein unfolds<sup>14</sup>. Real-time PCR

instrument (Agilent Mx3005p) was used to measure the fluorescence using 200X SYPRO orange dye. The excitation and emission spectra of the dye is at 492 nm and 610 nm respectively. PCR tubes were prepared by aliquoting 0.1 mg ml<sup>-1</sup> lipase in 20 mM phosphate buffer (pH 8.0) and 5X final concentration of dye to a final volume of 50 µl, mixed well by pipetting, and then centrifuged at 2000 × *g* to remove any air bubbles. PCR tubes were placed in triplicates into the PCR instrument and temperature scan was run from 25 to 95 °C, at 1 °C min<sup>-1</sup> ramp rate. No protein controls having the dye and 20 mM phosphate buffer, pH 8 were used for each experiment. This assay was performed for all the lipases types to measure the melting temperature (*T<sub>m</sub>*) of the wild type and the mutant lipases.

The melting temperatures (*T<sub>m</sub>*) of the transitions for melting curves were calculated from the midpoint of transition of the percentage relative fluorescence intensity of the maximum as a function of temperature. The midpoint of transition is the temperature at which 50 % of the protein has denatured, and is a measure of the protein's inherent thermal stability<sup>14</sup>.

#### 6.2.7 ATR-FTIR for determination of secondary changes after mutagenesis in lipase

The change in secondary structures of the wild-type and mutant lipases was also investigated using attenuated total reflection infrared (ATR-FTIR) particularly in the amide I range 1600-1700 cm<sup>-1</sup> was performed with constant nitrogen purging using a Spectrum Two™ FTIR Spectrometers (Perkin-Elmer). All lipase preparations were desalted by dialysis and briefly, these lipases were dried at 37 °C before recording spectra by FTIR-ATR instrument. A drop of 20 µl lipase solution on one side of the ATR was placed and measure the peak positions in amide I range and the whole experiment was carried out at room temperature. Peak positions in amide I range were resolved and fitted with a Gaussian shape using second derivative using OriginPro 8.5 software.

### 6.3 Results and Discussion

This study aimed to induce wild-type and mutants of *Bacillus subtilis* lipases with the nature of lipase activity at extreme conditions such as at extremes temperature, extremes

pH and high salt (NaCl) concentrations. The mutations were chosen as per the respected AHP ranking model. The lipase gene was cloned, mutated through site directed mutagenesis and purified. The kinetic stability, melting temperature and secondary structure analysis confirmed the stability of the mutants at extreme conditions compared to the wild type.

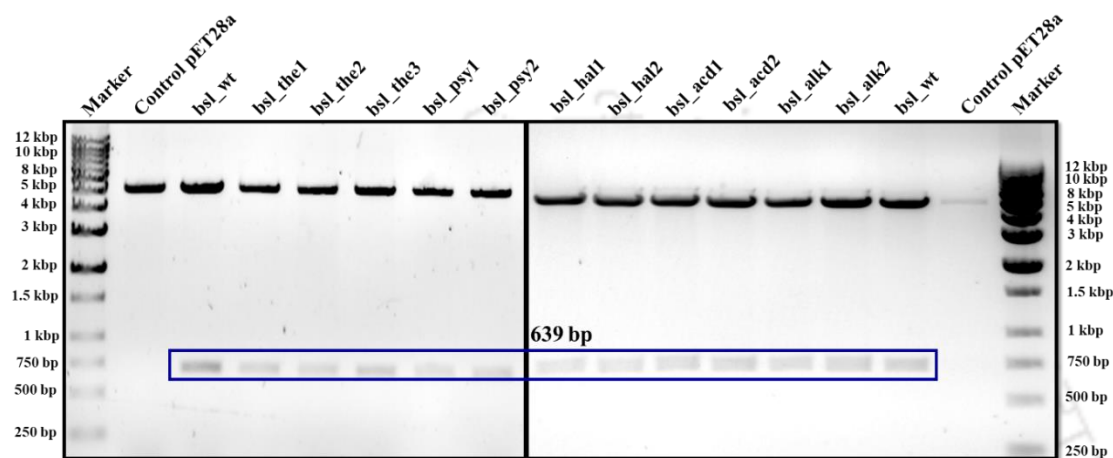
### 6.3.1 Cloning and *in vitro* mutagenesis of *Bacillus subtilis* lipase

The wild type lipase was successfully cloned into pET28a vector with N-terminal His-tag. Multiple primers were designed for each mutation using Quick Change Primer Design (Agilent Genomics) and then synthesized and used for mutagenesis according to their protocol (Figure 6.1). The designed primers for mutagenesis have been listed in Table 6.1. The mutagenesis was carried out at V149K, Q150E in *bsl\_the1*; F41K, W42E, V149K, Q150E in *bsl\_the2* and F41K, W42E, P119E, Q121K, V149K, Q150E in *bsl\_the3*; G52H, P53T in *bsl\_psy1* and G52H, P53T, G92M, G93A in *bsl\_psy2*; G52F, V54Y, G111Y in *bsl\_acd1* and G52F, V54Y, T110H, G111Y in *bsl\_acd2*; T110E, G111Y in *bsl\_alk1* and W31E, S32H, T45H, T110E, G111Y in *bsl\_alk2*; G52D, P53D, P119D, N120D in *bsl\_hal1* and G52D, P53D, G93D, P119D, N120D in *bsl\_hal2* and S24R, P119R, Q164R, G175R in *bsl\_bar* performed through site directed mutagenesis and was successful as observed through sequencing results. Figure 6.2 is the gel electrophoresis picture after restriction digestion of the wild type and mutants through NdeI and BamHI. The clones were confirmed by sequencing from 1<sup>st</sup> BASE, Singapore and the results are presented in Figure Appendix A2.22.

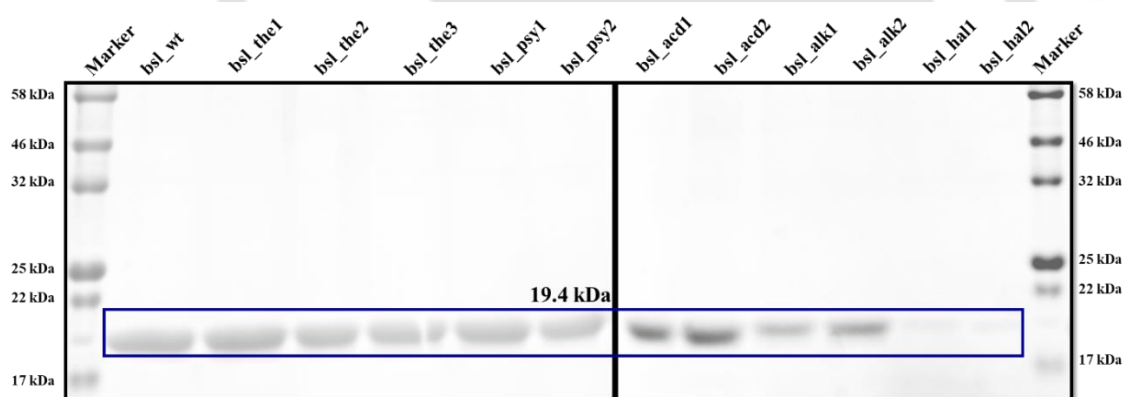
### 6.3.2 Enzyme expression and purification

The wild-type and mutant lipases were expressed and purified in 4 steps - (i) induction of primary culture was done by 100  $\mu$ M IPTG, after initial growth of 2 h at 37 °C; (ii) harvest the bacterial pellet and cell lysis; (iii) purify lipase using Ni-NTA affinity chromatography (iv) check protein expression and purification of lipases and was confirmed through SDS PAGE analysis. The purified wild type and mutant proteins gave bands on a stained SDS PAGE gel which is corresponding to 19.4 kDa (Figure 6.3). It is clear that the wild type

and mutant lipases were successfully purified and the concentrations of purified lipases are enlisted in Table 6.2. Coomassie Brilliant Blue staining revealed a single protein band with a molecular weight of about 19.4 kDa. The molecular weight obtained is same as that reported earlier<sup>2,6</sup>.



**Figure 6.2:** Restriction digestion of *Bacillus subtilis* 168 lipase wild type and mutants in pET28a (5.4 kb) was digested using BamHI and NdeI and released wild-type and mutant lipase genes was obtained at 639 bp.



**Figure 6.3:** SDS PAGE (12 %) analysis of purified wild type and mutant enzymes.

**Table 6.2:** Concentration of lipases obtained after purification ( $\text{g l}^{-1}$ ).

| Lipase   | Enzyme yield ( $\text{g l}^{-1}$ ) | Lipase   | Enzyme yield ( $\text{g l}^{-1}$ ) |
|----------|------------------------------------|----------|------------------------------------|
| bsl_wt   | 5.5                                | bsl_acd1 | 7.0                                |
| bsl_the1 | 4.3                                | bsl_acd2 | 8.8                                |
| bsl_the2 | 5.2                                | bsl_alk1 | 4.6                                |
| bsl_the3 | 7.6                                | bsl_alk2 | 9.2                                |
| bsl_psy1 | 5.1                                | bsl_hal1 | 2.2                                |
| bsl_psy2 | 5.6                                | bsl_hal2 | 2.1                                |

### 6.3.3 Enzyme kinetics studies

Kinetic stability was measured for the wild type and mutant lipases using *p*-nitrophenyl octanoate (*p*-NPO) as the substrate as earlier reports have shown that *Bacillus subtilis* lipase showed maximum substrate specificity for C8 triacylglycerol<sup>8</sup>. The kinetic parameters such as enzyme activity ( $\text{U ml}^{-1}$ ), specific activity ( $\text{U mg}^{-1}$ ), Michaelis-Menten constant ( $K_m$ ), turnover number ( $k_{cat}$ ), catalytic efficiency ( $k_{cat}/K_m$ ) on hydrolysis of *p*-NPO by the purified wild-type and mutant lipases was determined from Michaelis-Menten and Lineweaver-Burk plots. These parameter were estimated using the formulae present in equation (6.1) to (6.5). The Michaelis-Menten and Lineweaver-Burk plots has been illustrated in Figure 6.4 and 6.5 that shows kinetics of extreme-stable mutant and wild-type lipases. A constant enzyme concentration i.e.  $0.1 \text{ mg ml}^{-1}$  with varying *p*-NPO concentrations from  $0.002 \text{ mM}$  to  $0.5 \text{ mM}$  was used for kinetics analysis of different wild-type and mutant lipases. The kinetics analysis revealed extreme-stable mutant lipases enhanced activity with decreased  $K_m$  value and improved catalytic efficiency (higher  $k_{cat}/K_m$ ) than that of wild-type lipase.

$$\text{Specific activity} = \text{Conc. of product } (\mu\text{mol L}^{-1}) \times \frac{\text{Total rxn vol. (L)}}{\text{Time (min)} \times \text{Enzyme (mg)}} \quad (6.1)$$

$$\text{Conc. of product } (\mu\text{mol L}^{-1}) = \frac{OD \text{ (at 410 nm)}}{\epsilon \text{ (p-NPO=15.1 mM}^{-1}\text{cm}^{-1}) \times l \text{ (=0.5 cm)}} \quad (6.2)$$

$$\text{Enzyme (mg)} = \text{Conc. of enzyme (mg ml}^{-1}) \times \frac{\text{Vol. of enzyme in rxn } (\mu\text{l})}{1000} \quad (6.3)$$

$$\text{Turnover number or } k_{cat} \text{ (}\mu\text{M)} = \frac{V_{max}}{\text{Total enzyme conc. in rxn}} \quad (6.4)$$

$$\text{Catalytic efficiency } (\mu\text{M}^{-1}\text{sec}^{-1}) = \frac{k_{cat}}{K_m} \quad (6.5)$$

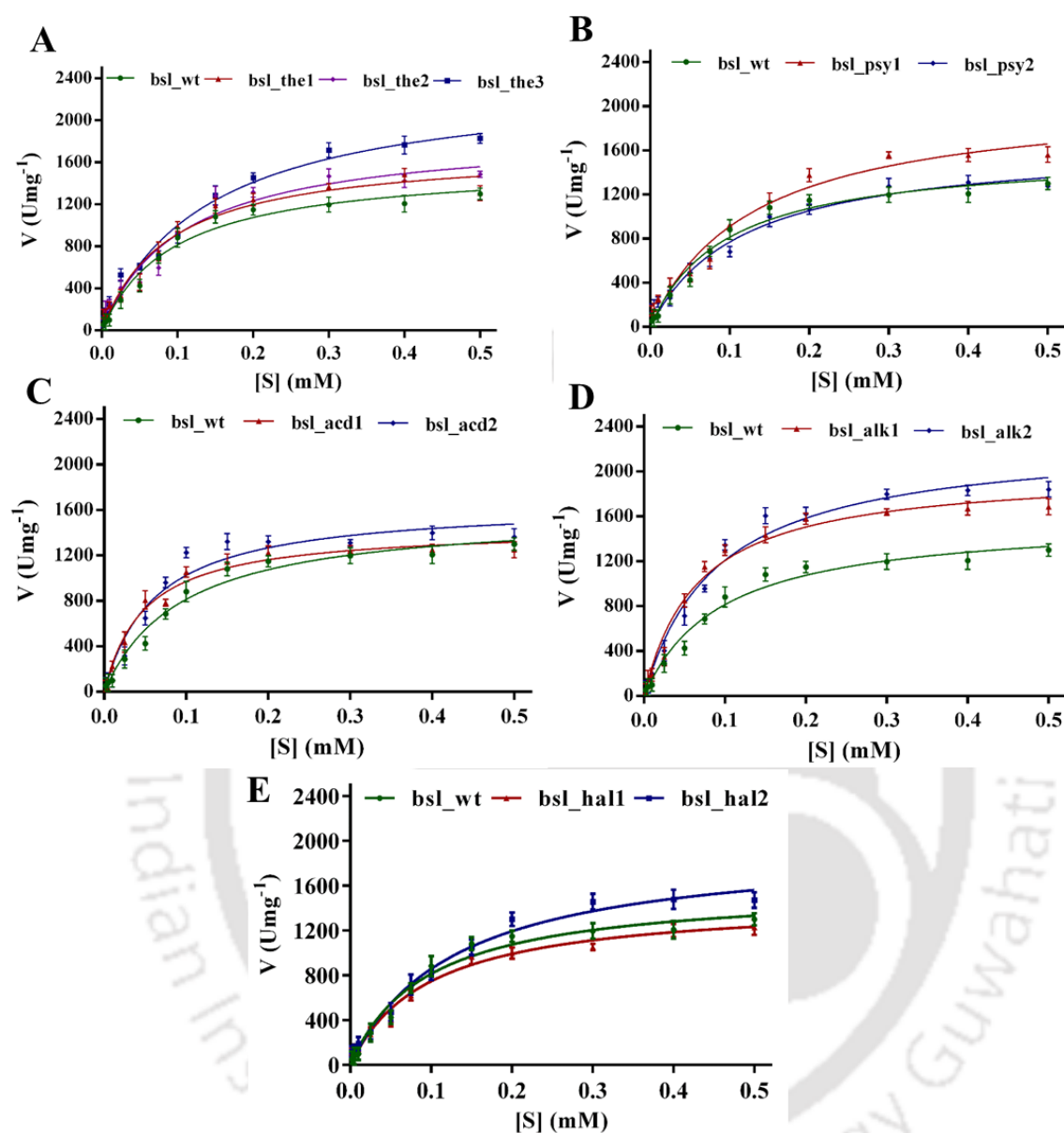


Figure 6.4: Comparative analysis of Michaelis-Menten kinetics studies of wild-type with (A) thermostable mutant, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases.



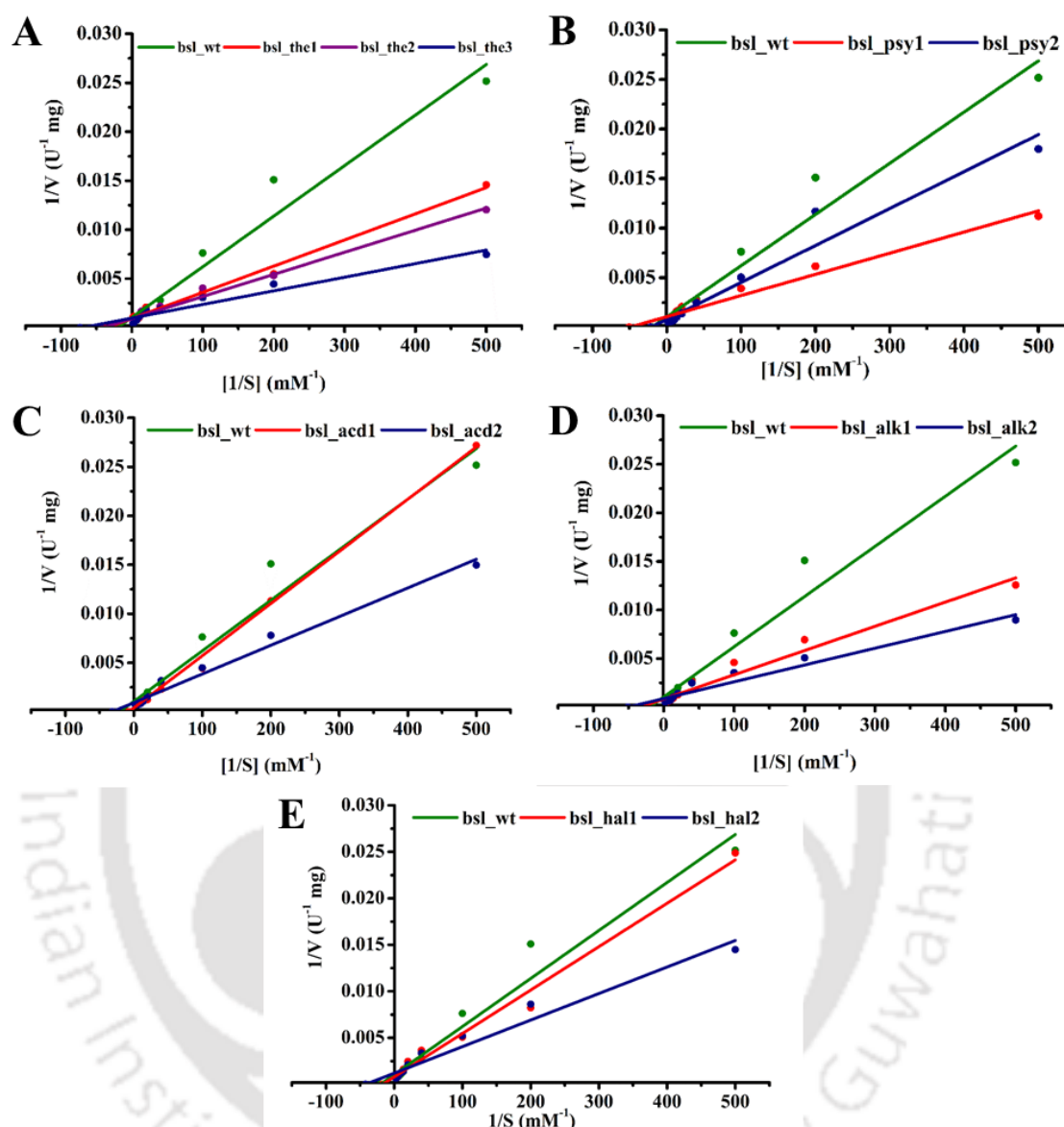


Figure 6.5: Comparative analysis of Lineweaver-Burk kinetics studies of wild-type with (A) thermostable mutant, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipase.

### *Thermostable mutant of lipase*

The comparative analysis of three thermostable mutants of lipase with wild-type lipase (bsl\_wt) showed 1.5 fold increase in the activity of bsl\_the3, 1.2 fold in bsl\_the2 and 1.1 fold in bsl\_the1. The chosen mutations in these thermostabilizing mutants showed increase in  $K_m$  value when compared with wild-type lipase which can cause reduction in enzyme-substrate affinity. However, the thermostabilizing mutants showed increased turnover

number and somehow constant catalytic efficiency. Thus, the overall analysis of lipase performance showed an increasing trend  $bsl\_wt < bsl\_the1 < bsl\_the2 < bsl\_the3$ .

### ***Psychrostable mutant of lipase***

The comparative analysis of two psychrostable mutants of lipase with wild-type lipase ( $bsl\_wt$ ) showed 1.32 fold increase in the activity of  $bsl\_psy1$  but not much change in the activity of  $bsl\_psy2$ . The chosen mutations caused slight increase in the  $K_m$  value than the wild-type lipase which can lead to reduction in enzyme-substrate affinity. These psychro stabilizing mutants also showed a little increase in turnover number. The kinetics study revealed that mutation in both the  $bsl\_psy1$  and  $bsl\_psy2$  affected overall catalytic efficiency.

### ***Acidostable mutant of lipase***

The mutant  $bsl\_acd1$  showed reduction in enzyme activity whereas,  $bsl\_acd2$  showed a little increase in enzyme activity when compared to  $bsl\_wt$ . But both the acidostable mutants showed increased  $K_m$  value which is bad for enzyme-substrate affinity. A significant increase in catalytic efficiency was noticed in both the mutants since reduced  $K_m$  value.

### ***Alkalistable mutant of lipase***

The comparative analysis of two alkalistable mutants of lipase with wild-type lipase ( $bsl\_wt$ ) showed 1.45 fold increase in the activity of  $bsl\_alk2$  and 1.26 fold in  $bsl\_alk1$ . Interestingly, both the alkalistable mutants showed decreased  $K_m$  value that was essential for better enzyme-substrate binding and thus, leading to a significant increase in the increased turnover number and somehow constant catalytic efficiency. The highest catalytic efficiency was obtained with  $bsl\_alk1$  followed by  $bsl\_alk2$ .

### ***Halostable mutant of lipase***

The mutant  $bsl\_hal1$  showed reduction in enzyme activity whereas,  $bsl\_hal2$  showed a little increase in enzyme activity when compared to  $bsl\_wt$ . But both the halostable mutants showed decreased  $K_m$  value which is good for enzyme-substrate binding affinity. Because of reduced  $K_m$  value, a significant increase in catalytic efficiency was noticed in both the mutants.

The overall analysis of all generated extreme-stable mutants showed that most of the lipase mutants have more or less equal catalytic efficiency ( $k_{\text{cat}}/K_m$ ) than wild-type lipase. Out of these mutants, bsl\_the3 was the best mutant in terms of activity and turnover number ( $k_{\text{cat}}$ ) but its  $K_m$  value was increased leading to overall loss in the catalytic efficiency ( $k_{\text{cat}}/K_m$ ). On the other hand, alkalistable mutants (bsl\_alk1 and bsl\_alk2) and halostable mutants (bsl\_hal1 and bsl\_hal2) also showed a significant increase in catalytic efficiency ( $k_{\text{cat}}/K_m$ ). The order of performance of lipases in terms of catalytic efficiency revealed: bsl\_hal1 > bsl\_alk1 > bsl\_alk2 > bsl\_hal2 > bsl\_the1 > bsl\_the2, bsl\_the3 > bsl\_wt > bsl\_psy1 > bsl\_acd2 > bsl\_acd1 > bsl\_psy2. All the enumerated kinetic parameters for each mutant have been enlisted in Table 6.3.

Table 6.3: Enumerated enzyme kinetics of the wild type and mutant lipases.

| Lipase and mutants | Kinetic parameters (Best fitted values)     |                                         |                                       |            |                                       |                                                              |
|--------------------|---------------------------------------------|-----------------------------------------|---------------------------------------|------------|---------------------------------------|--------------------------------------------------------------|
|                    | Enzyme concentration (mg ml <sup>-1</sup> ) | Specific activity (U mg <sup>-1</sup> ) | Enzyme activity (U ml <sup>-1</sup> ) | $K_m$ (μM) | $K_{\text{cat}}$ (sec <sup>-1</sup> ) | $k_{\text{cat}} / K_m$ (μM <sup>-1</sup> sec <sup>-1</sup> ) |
| bsl_wt             | 0.1                                         | 1583                                    | 15.83                                 | 94         | 511                                   | 5.43                                                         |
| bsl_the1           | 0.1                                         | 1725                                    | 17.25                                 | 88         | 557                                   | 6.33                                                         |
| bsl_the2           | 0.1                                         | 1891                                    | 18.91                                 | 106        | 610                                   | 5.76                                                         |
| bsl_the3           | 0.1                                         | 2396                                    | 23.96                                 | 140        | 773                                   | 5.52                                                         |
| bsl_psy1           | 0.1                                         | 2087                                    | 20.87                                 | 129        | 673                                   | 5.22                                                         |
| bsl_psy2           | 0.1                                         | 1668                                    | 16.68                                 | 116        | 538                                   | 4.64                                                         |
| bsl_acd1           | 0.1                                         | 1473                                    | 14.73                                 | 97         | 475                                   | 4.90                                                         |
| bsl_acd2           | 0.1                                         | 1963                                    | 19.63                                 | 129        | 633                                   | 4.91                                                         |
| bsl_alk1           | 0.1                                         | 2002                                    | 20.02                                 | 66         | 646                                   | 9.79                                                         |
| bsl_alk2           | 0.1                                         | 2297                                    | 22.97                                 | 90         | 741                                   | 8.24                                                         |
| bsl_hal1           | 0.1                                         | 1444                                    | 14.44                                 | 49         | 466                                   | 9.51                                                         |
| bsl_hal2           | 0.1                                         | 1659                                    | 16.59                                 | 62         | 535                                   | 8.63                                                         |

### 6.3.4 Temperature stability analysis of wild type and mutated lipases

The temperature effects on catalysis of wild-type and mutant lipases from *Bacillus subtilis* in terms of enzyme activity and stability has been illustrated in Figure 6.6 and 6.7, respectively. The optimum temperature of lipases was analyzed by incubating the enzyme-substrate mixture at different temperature ranges from 20 °C to 70 °C (at pH 8) and further

the percentage relative activity was determined<sup>2,6</sup>. The optimum enzyme activity of wild type lipase (bsl\_wt) was obtained maximum at 35 °C. In case of thermostable mutants, the maximum thermostability was obtained at 55 °C for bsl\_the3, 50 °C for bsl\_the2 and 40 °C in bsl\_the1. Interesting conclusions have been drawn by comparison of the achieved temperature activity of the bsl\_the3 mutant (at 55 °C) through 6 point mutations, with previously reported thermostable mutants of *Bacillus subtilis* lipase (PDB ID: 3D2C with 9 mutations) obtained by directed evolution strategies reported at 55 °C<sup>15</sup>. To the best of our knowledge, this represents less number of predicted mutations that resulted in such enhancement in thermostability. In contrary, the psychrostable mutant bsl\_psy2 showed lowest optimum temperature at 25 °C and bsl\_psy1 showed at 30 °C. The other mutants also showed improved optimum thermal activity such as acidostable mutant bsl\_acd2 showed optimum temperature at 40 °C and bsl\_acd1 at 35 °C; alkalistable mutant, bsl\_alk2 at 50 °C and bsl\_alk1 at 40 °C; halostable mutant, bsl\_hal2 at 45 °C and bsl\_hal1 at 40 °C. The wild-type and all extreme-stable mutant lipases showed a sharp decrease in the activity at temperatures beyond 60 °C. The order of thermal activity revealed: bsl\_psy2 (at 25 °C) > bsl\_psy1 (at 30 °C) > bsl\_wt = bsl\_acd1 (at 35 °C) > bsl\_the1 = bsl\_acd2 = bsl\_alk1 = bsl\_hal1 (at 40 °C) > bsl\_hal2 (at 45 °C) > bsl\_the2 = bsl\_alk2 (at 50 °C) > bsl\_the3 (at 55 °C). The thermal analysis of wild-type and mutant lipases have been presented in Figure 6.6A – 6.6E.

Further, the mutant and wild-type lipases were exploited for thermal stability analysis. They were incubated at 50 °C for 24 h except, psychrostable mutants (bsl\_psy1 and bsl\_psy2). Result showed that thermostable mutant bsl\_the3 showed significant thermal tolerance i.e. retained 70% of its activity after an incubation period of 24 h. The mutant bsl\_the2 and bsl\_alk2 retained upto 50% of its activity after an incubation of 24 h. The thermostability also revealed that the bsl\_the1, bsl\_acd1, bsl\_acd2, bsl\_alk1, bsl\_hal1 and bsl\_hal2 were found to be thermosensitive since they have retained < 30 % of its activity after an incubation of 24 h. In contrary, the psychrostable mutant, bsl\_psy2 showed 82 % of activity was retained after an incubation at 25 °C for 24 h. Similarly, bsl\_psy1 retained 68 % activity after same incubation. The results of temperature stability have been presented in Figure 6.7.

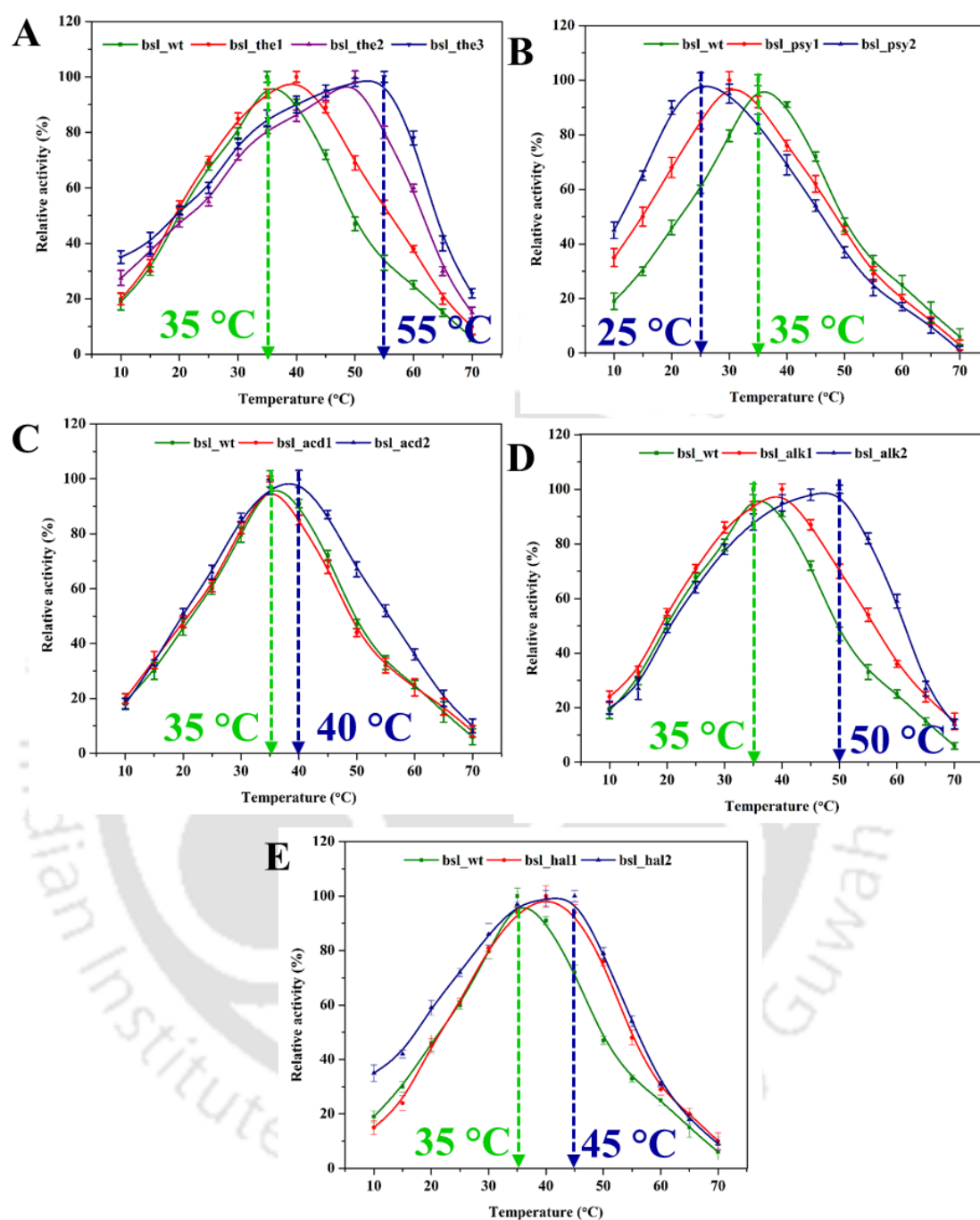


Figure 6.6: Graph showing percentage relative activity for identifying the optimum temperature (from temperatures 10 °C - 70 °C) of wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Green arrow represent optimum activity of wild-type lipase and blue arrow represents optimum activity of most stable mutant. Assays conducted at pH 8 using *p*-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.

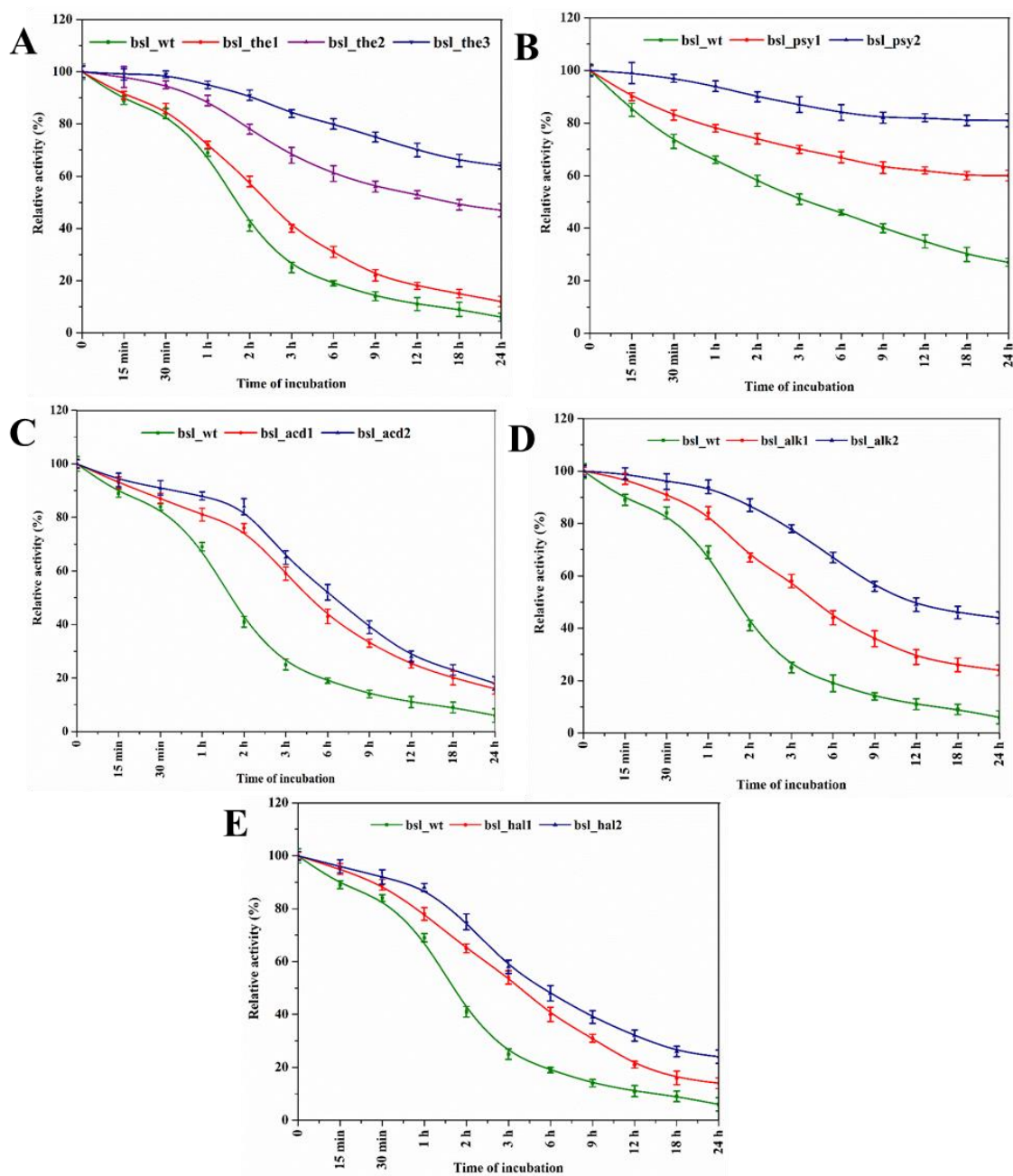


Figure 6.7: Graph showing percentage relative activity for identifying the temperature stability at 50 °C (except psychrophilic mutants, temperature stability at 25 °C) of wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Assays conducted at pH 8 using *p*-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.



### 6.3.5 pH stability analysis of wild type and mutated lipases

It is known that the changes in pH affects the enzyme structure, stability as well as activity<sup>16</sup>. The effect of pH on *Bacillus subtilis* wild type and extreme-stability mutant was performed for determination of acidostability and alkalistability of the enzyme. Figure 6.8 and Figure 6.9, illustrated effect of different pH at constant temperature 35 °C on activity and stability of wild-type and mutant lipases. The results showed optimum pH of wild type was obtained at pH 8 which maximally increased by 2 unit to pH 10 in case of alkalistable mutants *i.e.* bsl\_alk2 whereas it was decreased to pH 6.5 in case of acidostable mutants *i.e.* bsl\_acd2. The pH optimum of wild-type lipase corroborated with the previous studies which was reported to be pH 8.0<sup>2</sup>. For other lipases, the optimum pH was obtained at pH 7.0 for bsl\_acd1; pH 8.0 for bsl\_the1, bsl\_psy1 and bsl\_psy2; pH 8.5 for bsl\_hal1; pH 9.0 for bsl\_the2, bsl\_the3, bsl\_alk1 and bsl\_hal2 (Figure 6.8). The activity dropped significantly after pH 10.0 for all the lipase types.

The chosen pH was 9.0 for alkalistability studies for all lipases by incubating them upto 24 h. The maximum relative activity was retained upto 78 % for bsl\_alk2 followed by 72 % for bsl\_the3 mutant after 24 h incubation at pH 9.0 (Figure 6.9). For acidostability analysis, only bsl\_acd1 and bsl\_acd2 mutants were incubated at pH 6.5 for an incubation of 24 h since the lowest pH optimum was obtained pH 6.5 for bsl\_acd2. The mutant bsl\_acd2 showed maximum activity retained to 46 % at pH 6.5. Thus, the overall analysis of high or low pH values generally result in complete loss of activity for all the lipase types with time.

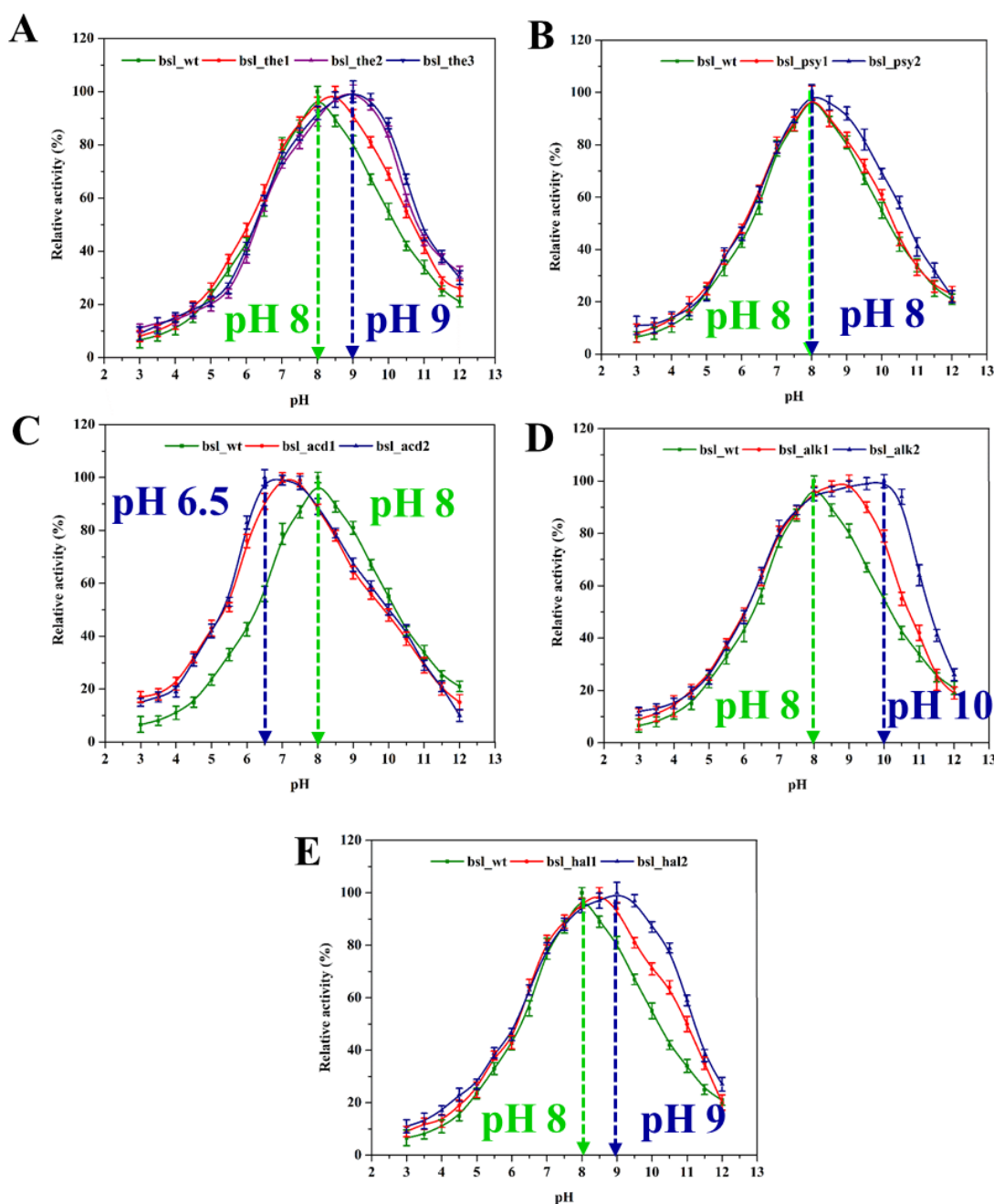


Figure 6.8: Graph showing percentage relative activity for identifying the optimum pH (from pH 2 – pH 12) of wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Green arrow represent optimum activity of wild-type lipase and blue arrow represents optimum activity of most stable mutant. Assays conducted at 35 °C using *p*-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.

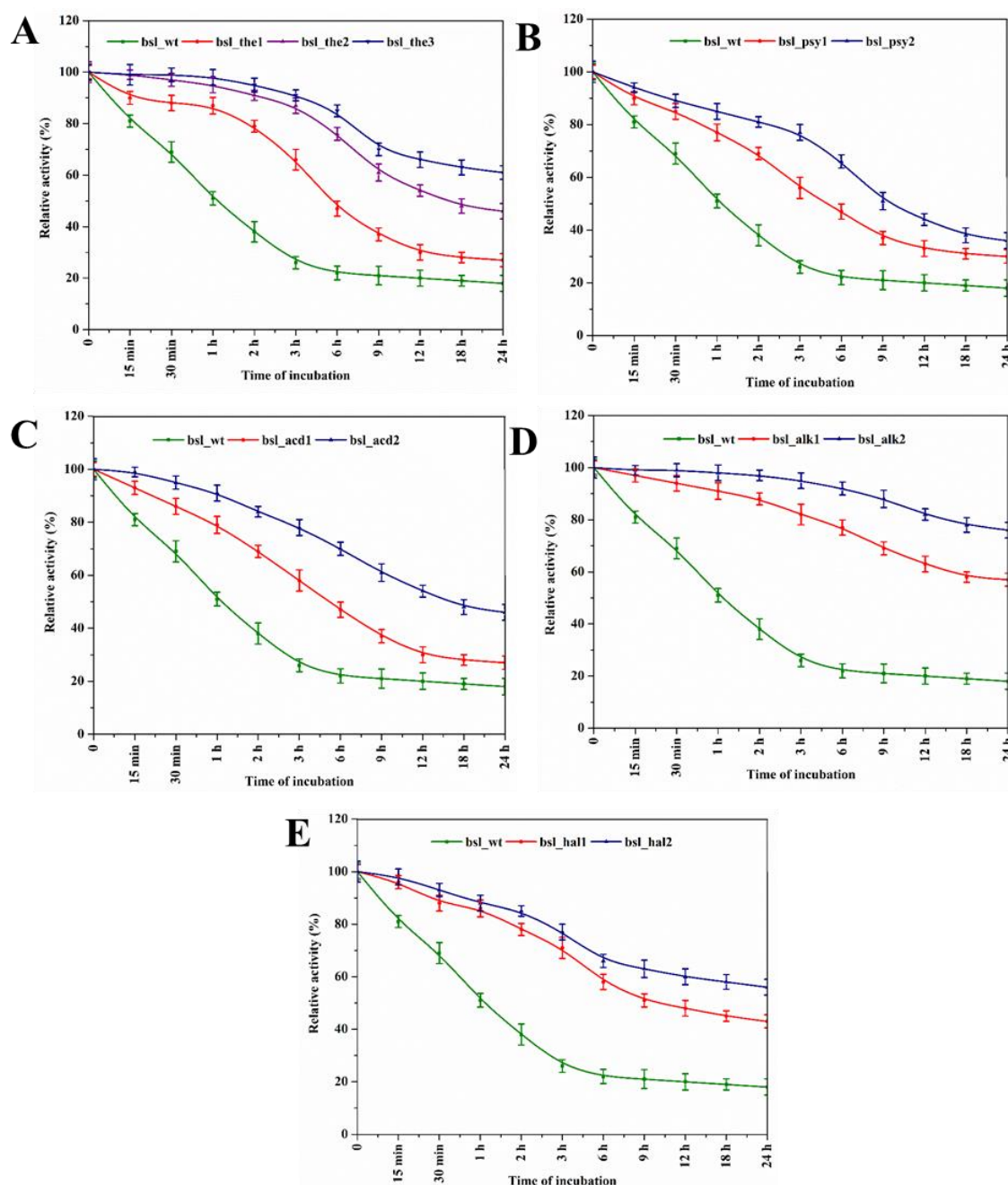


Figure 6.9: Graph showing percentage relative activity for identifying the pH stability at pH 9 (except acidophilic mutants, pH stability at pH 6.5) of wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutan, (D) alkalistable mutant and (E) halostable mutant lipases. Assays conducted at 35 °C using *p*-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.

### 6.3.6 Halostability analysis of wild type and mutated lipases

Enzyme halostability and activity of wild type and mutant lipases was carried out in terms of adaptability of enzyme performance in the presence of increasing concentrations of NaCl<sup>17</sup>. The effect of NaCl on *Bacillus subtilis* wild type and extreme-stability mutant was performed for determination of halostability and activity of the lipases. Figure 6.10 and Figure 6.11, illustrated effect of NaCl on activity and stability of wild-type and mutant lipases. The results showed that the wild-type lipase (bsl\_wt) was found to be halosensitive, since their activity and stability was gradually decreases as the concentration of NaCl increase in the enzyme-substrate reaction mix. The best halostable mutant i.e. bsl\_hal2 showed a tremendous halotolerance upto 1.5 M NaCl (Figure 6.10). The other mutants revealed a continuous decrease in the activity and stability of lipases. The chosen mutations by AHP models were found to be significant for generating lipase mutants.

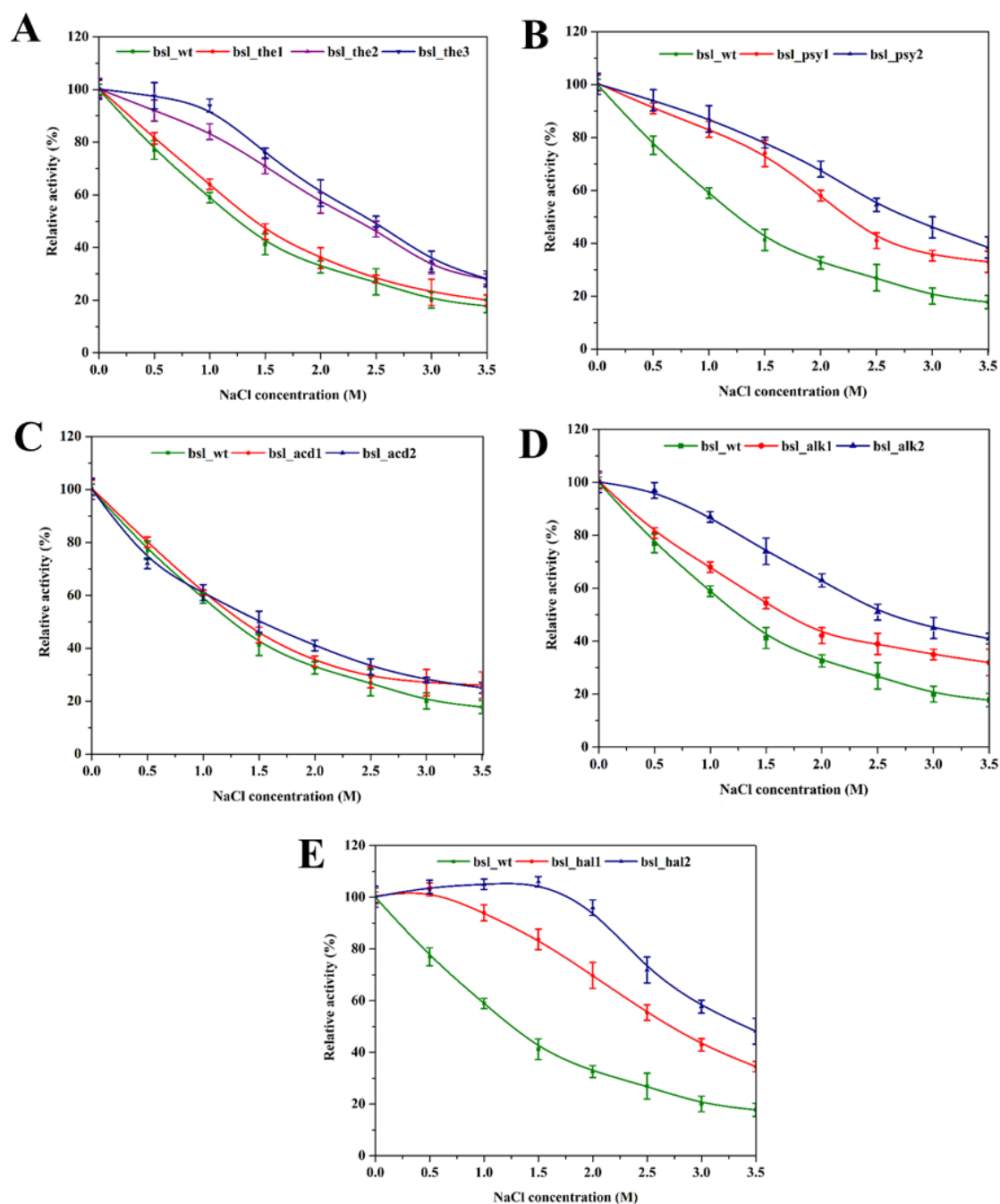


Figure 6.10: Graph showing percentage relative activity for identifying the optimum NaCl concentration (from 0 M - 3.5 M) of wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Assays conducted at 35 °C and pH 8 using *p*-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.

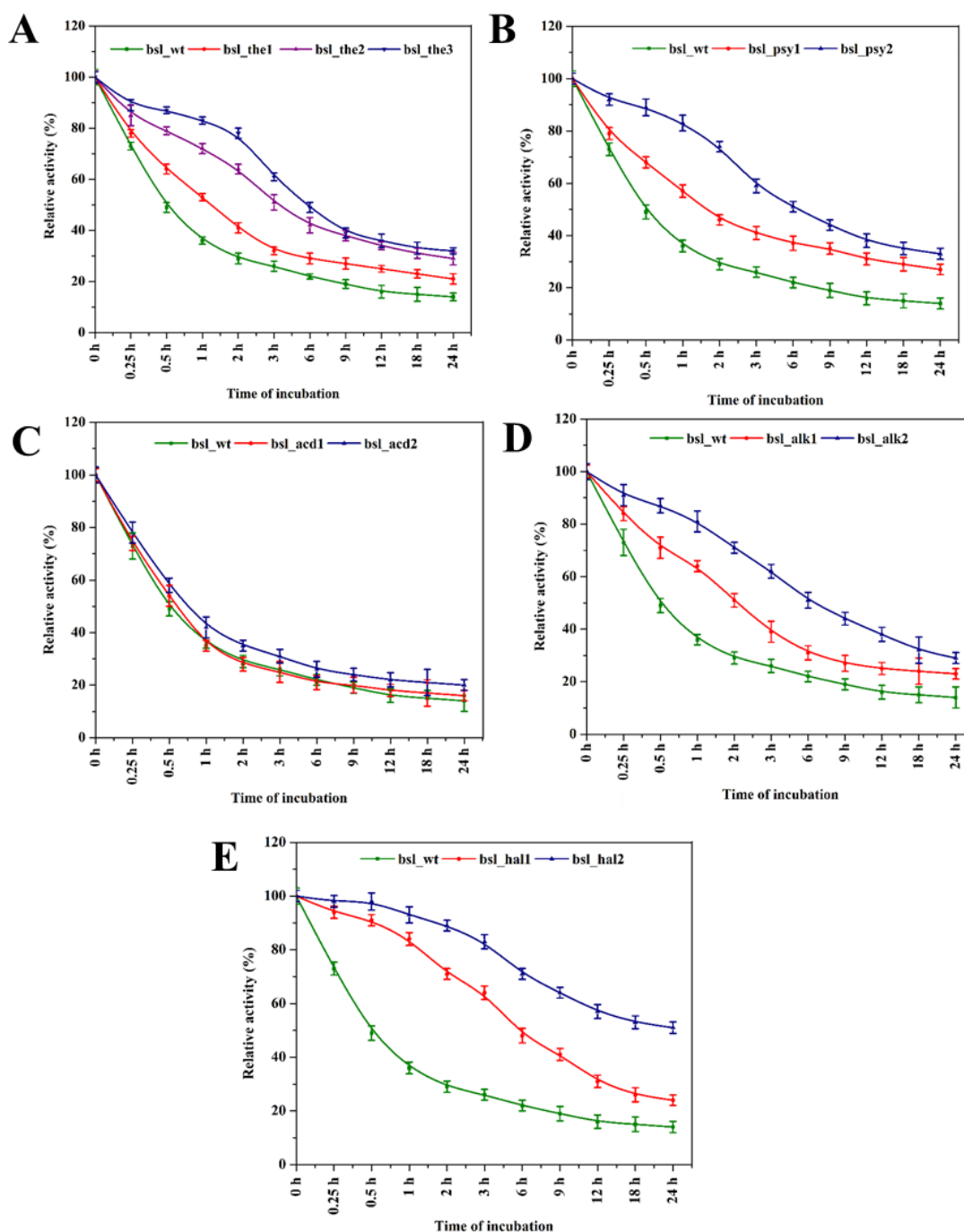


Figure 6.11: Graph showing percentage relative activity for identifying the halostability at 1 M NaCl for wild-type lipase and compared with (A) thermostable mutants, (B) psychrostable mutant, (C) acidostabilizing mutant, (D) alkalistable mutant and (E) halostable mutant lipases. Assays conducted at 35 °C and pH 8 using *p*-NPO as substrate. Error bars are the standard deviations of experiments done independently and in triplicates.



The results of optimum temperature, pH and salt [NaCl] concentration and their stabilities are summarized in Table 6.4.

Table 6.4: Analysis comparing the optimum activity and stability of wild-type and mutant lipases.

| Lipase   | Opt. Temp. (°C) | Temp. Stability (% activity retained after 24 h at 50 °C) | Opt. pH | pH Stability (% activity retained after 24 h at pH 9) | Opt. [NaCl] conc. | Temp. Stability (% activity retained after 24 h in 1M [NaCl]) |
|----------|-----------------|-----------------------------------------------------------|---------|-------------------------------------------------------|-------------------|---------------------------------------------------------------|
| bsl_wt   | 35              | 6.0±1.5                                                   | 8.0     | 18.0±3.1                                              | 0.5               | 14.0±1.5                                                      |
| bsl_the1 | 40              | 12.0±2.0                                                  | 8.5     | 27.0±2.5                                              | 0.5               | 21.0±2.0                                                      |
| bsl_the2 | 50              | 47.0±2.5                                                  | 9.0     | 46.0±3.0                                              | 0.5               | 29.0±2.5                                                      |
| bsl_the3 | 55              | 64.0±1.2                                                  | 9.0     | 61.0±2.6                                              | 1.0               | 32.0±1.5                                                      |
| bsl_psy1 | 30              | 60.0±2.0 (incubated @ 25°C; bsl_wt = 27.0±1.5)            | 8.0     | 30.0±2.5                                              | 0.5               | 27.0±2.0                                                      |
| bsl_psy2 | 25              | 81.2±2.5 (-do-)                                           | 8.0     | 36.0±3.0                                              | 0.5               | 33.0±2.1                                                      |
| bsl_acd1 | 35              | 16.0±2.0                                                  | 7.0     | 26.0±2.0 (incubated @ pH 6.5; bsl_wt = 17.0±1.6)      | 0.5               | 16.0±2.0                                                      |
| bsl_acd2 | 40              | 18.5±2.5                                                  | 6.5     | 45.0±3.0 (-do-)                                       | 0.5               | 20.0±2.2                                                      |
| bsl_alk1 | 40              | 24.0±2.0                                                  | 9.0     | 57.0±2.3                                              | 0.5               | 23.0±2.0                                                      |
| bsl_alk2 | 50              | 44.0±2.3                                                  | 10.0    | 76.0±3.1                                              | 1.0               | 31.0±2.2                                                      |
| bsl_hal1 | 40              | 14.0±2.0                                                  | 8.5     | 43.0±2.5                                              | 0.5               | 24.0±2.0                                                      |
| bsl_hal2 | 45              | 25.0±2.5                                                  | 9.0     | 56.0±3.0                                              | 1.5               | 51.0±1.9                                                      |

### 6.3.7 Differential Scanning Fluorimetry of wild type and mutated lipases

Differential scanning fluorimetry or thermal shift assay was employed as an alternative to Circular dichroism and Differential scanning calorimetry (DSC) to monitor the melting temperature of wild type and mutants of *Bacillus subtilis* lipase. This method was chosen as it was reported that thermal profiles of proteins obtained by CD or DSC is similar to that obtained from DSF and they also produce similar data profiles<sup>18,19</sup>. The advantages of the method over circular dichroism (CD) and differential scanning calorimetry (DSC) are manifold. DSF is superior to DSC with respect to throughput or the assay time is lower for DSF than DSC. The SYPRO orange dye used provides high signal to noise ratio. The excitation for SYPRO orange is near 500 nm which decreases the likelihood that any small molecule would interfere with the optical properties of the dye and cause quenching of the fluorescence intensity<sup>14</sup>.

The measurement of fluorescence was carried out using Agilent Technologies, Mx3005p Real Time PCR. The theory behind thermal shift assay is based on monitoring protein

thermal curves (Figure 6.12). With increase in temperature unfolding of protein occur which results in exposure of hydrophobic regions and eventual protein aggregation<sup>20</sup>. The assay involves monitoring this changes folding by incorporation of fluorescence signal of SYPRO orange dye. The dye's fluorescence signal is quenched in the aqueous environment of a properly folded protein in solution, but becomes unquenched when exposed to the protein's hydrophobic core upon unfolding. The temperature at which concentrations of folded and unfolded protein are equal is considered as melting temperature ( $T_m$ )<sup>14</sup>. The midpoint ( $T_m$  value) of the protein melting curve will increase in proteins that are more thermostable<sup>14</sup>.

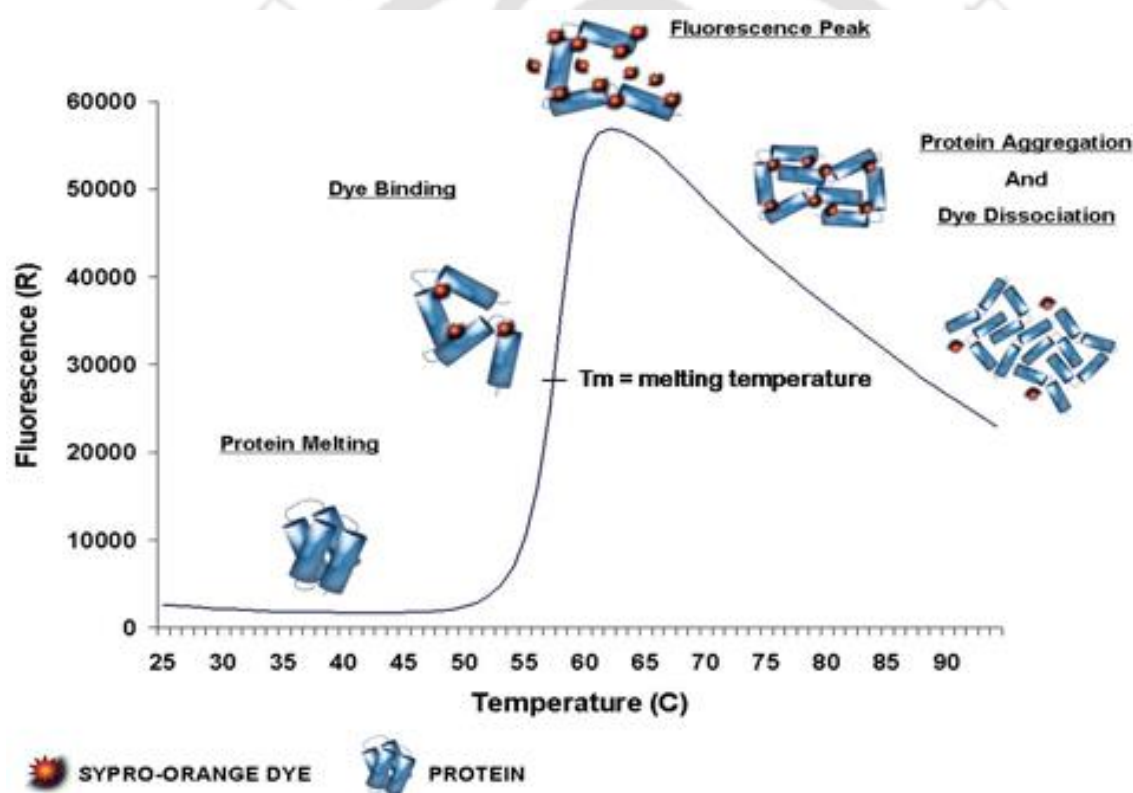


Figure 6.12: The schematic of the rationale behind thermal shift assay.  
(Adopted from, [http://www.wikiwand.com/en/Thermal\\_shift\\_assay](http://www.wikiwand.com/en/Thermal_shift_assay))

The thermal profiles of wild type and mutant lipases under their extreme condition in addition to temperature, have been illustrated in Figure 6.13. The mid-point of fluorescence transition or where 50% relative fluorescence is achieved, is the  $T_m$  of the enzymes. The results of melting temperature ( $T_m$  values) and their thermal profiles of wild-type and mutant lipases are represented in the Figure 6.13. It can be concluded here that predicted

mutations in *Bacillus subtilis* lipase resulted in site directed mutagenesis possessed increased temperature stability than the wild-type lipase. The  $T_m$  of the mutant lipases are relative higher than that of wild-type lipase indicating structural rigidity. The bsl\_the3 mutant showed highest increase in the  $T_m$  value to 66.0 °C from wild-type lipase (bsl\_wt)  $T_m$  value i.e. 59.0 °C. Except that of psychrostable mutants (the bsl\_psy1 have  $T_m$  =57.5 °C and bsl\_psy2 have  $T_m$  =55.0 °C) with decrease in the  $T_m$  value than wild-type lipase (bsl\_wt) are relatively less rigid or more flexible in nature. The other mutant like bsl\_the1 ( $T_m$  = 60.5 °C), bsl\_the2 ( $T_m$  = 63.0 °C), bsl\_acd1 ( $T_m$  = 60.5 °C), bsl\_acd2 ( $T_m$  = 61.5 °C), bsl\_alk1 ( $T_m$  = 60.5 °C), bsl\_alk2 ( $T_m$  = 65.0 °C), bsl\_hal1 ( $T_m$  = 61.0 °C) and bsl\_hal2 ( $T_m$  = 62.5 °C) also showed increase in their  $T_m$  value when compared with wild-type lipase (bsl\_wt). The different types of substitutions were found in these mutants, and these substitutions resulted in an increase in positively contributing attributes via substituting amino acids in the wild-type enzyme. The previous reports also showed that enzymes' thermal stability and catalytic activity were improved through evolutionary molecular engineering by a significant increase in the  $T_m$  values of mutant enzyme<sup>21,22</sup>. Table 6.5 enlists the obtained  $T_m$  of wild-type and mutant lipases and applied experimental conditions.

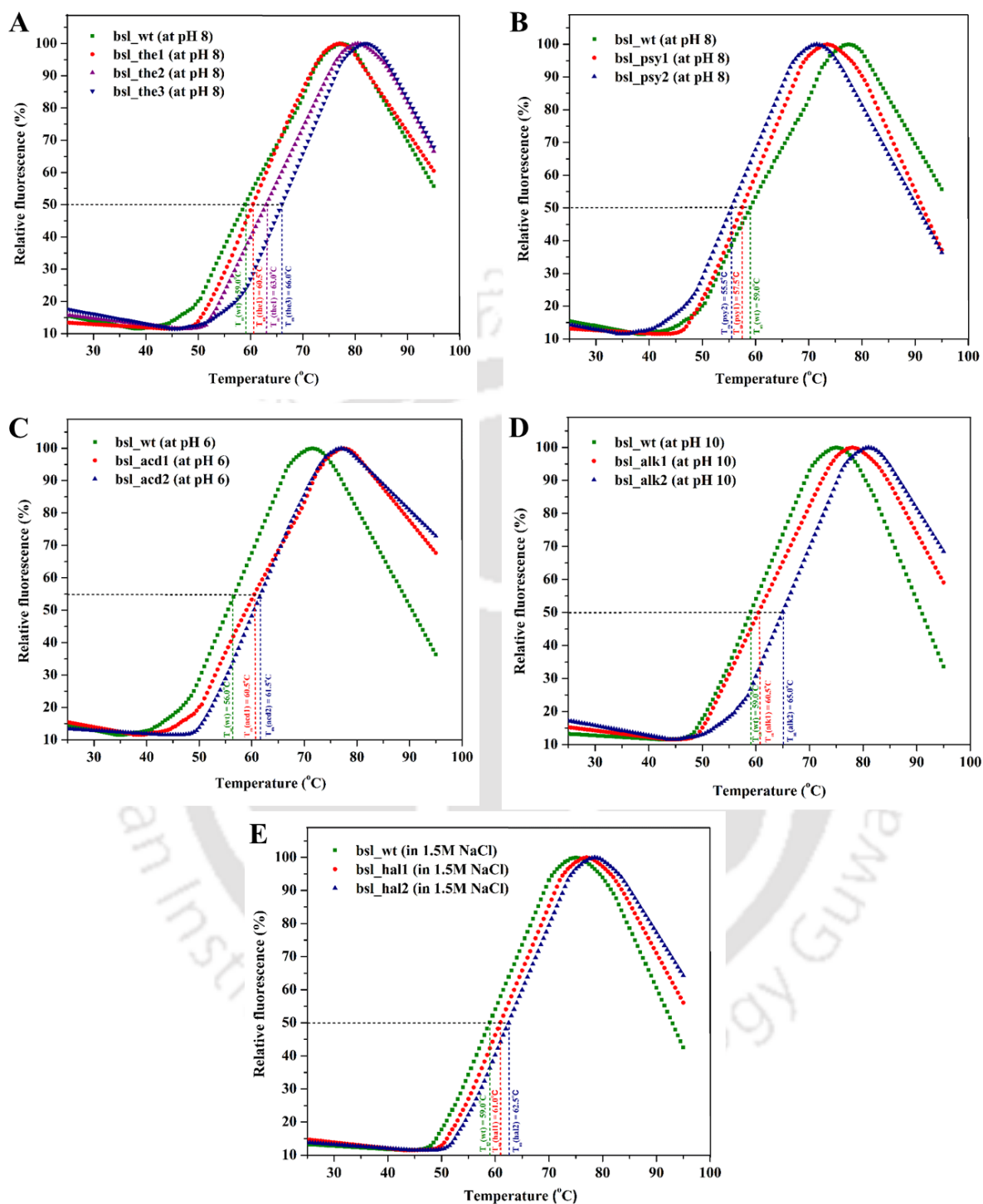


Figure 6.13: Thermograph of wild-type and mutant lipases.

Table 6.5: Comparing  $T_m$  values of wild-type and mutant lipases at different experimental conditions.

| Lipases  | $T_m$ (°C) | Experimental Condition |
|----------|------------|------------------------|
| bsl_wt   | 59.0       | pH 8 or 10; 1.5M NaCl  |
| bsl_wt   | 56.0       | pH 6                   |
| bsl_the1 | 60.5       | pH 8                   |
| bsl_the2 | 63.0       | pH 8                   |
| bsl_the3 | 66.0       | pH 8                   |
| bsl_psy1 | 57.5       | pH 8                   |
| bsl_psy2 | 55.5       | pH 8                   |
| bsl_acd1 | 60.5       | pH 6                   |
| bsl_acd2 | 61.5       | pH 6                   |
| bsl_alk1 | 60.5       | pH 10                  |
| bsl_alk2 | 65.0       | pH 10                  |
| bsl_hal1 | 61.0       | pH 8; 1.5M NaCl        |
| bsl_hal2 | 62.5       | pH 8; 1.5M NaCl        |

### 6.3.8 Secondary structure prediction of wild-type and mutant lipase by FTIR (ATR)

The amide vibrations determined the peptide group present in protein. The structural repeat unit of proteins, gives up to 9 characteristic bands named amide A, B, I, II, ... VII. The amide A band (about  $3500\text{ cm}^{-1}$ ) and amide B (about  $3100\text{ cm}^{-1}$ ) originate from a Fermi resonance between the first overtone of amide II and the N-H stretching vibration. Amide I and amide II bands are two major bands of the protein infrared spectrum. The amide I band (between  $1600$  and  $1700\text{ cm}^{-1}$ ) is mainly associated with the C=O stretching vibration (70-85 %) and is directly related to the backbone conformation (Figure 6.14). Amide II results from the N-H bending vibration (40-60 %) and from the C-N stretching vibration (18-40 %). This band is conformationally sensitive. Amide III and IV are very complex bands resulting from a mixture of several coordinate displacements. The out-of-plane motions are found in amide V, VI and VIII.

The amide I range ( $1600$ - $1700\text{ cm}^{-1}$ ) of the spectra were fitted with Gaussian curves using Origin 8.5 software. Relative content of various secondary structures ( $\alpha$ -helix,  $\beta$ -sheet,  $\beta$ -turn, random and side chain) were determined and are shown in Figure 6.14. Briefly, peak

for  $\alpha$ -helix is present at around  $1655\text{ cm}^{-1}$ , peaks for  $\beta$ -sheet are found between  $1620$  to  $1636\text{ cm}^{-1}$ , peaks for  $\beta$ -turn are mostly be in the range  $1662$  to  $1688\text{ cm}^{-1}$  while random peak is found close at  $1645\text{ cm}^{-1}$ .

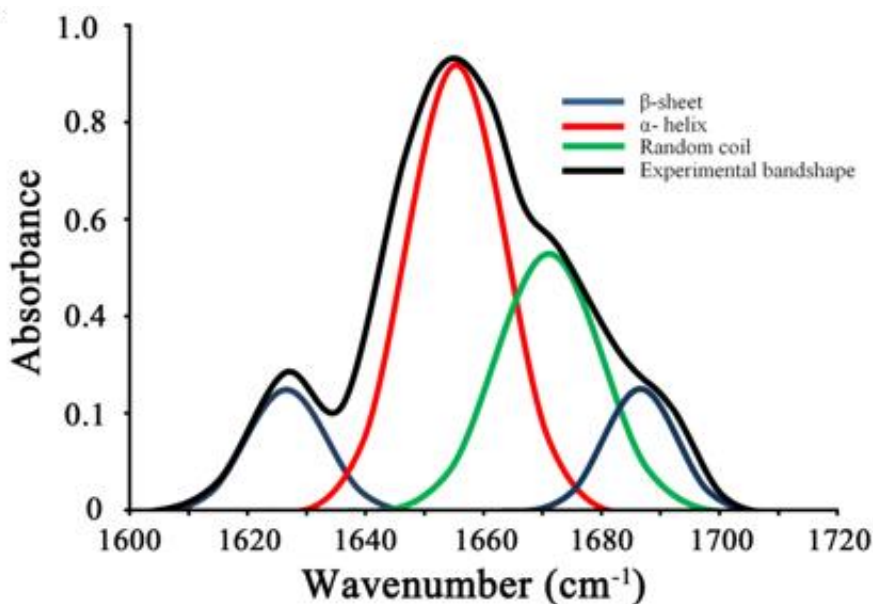


Figure 6.14: A graph representing amide I region from  $1600 - 1700\text{ cm}^{-1}$  fitted with peaks of secondary structures in protein of the rationale behind thermal shift assay. (Adopted from, <http://nptel.ac.in/courses/102103044/module2/lec10/8.html>).

The change in secondary structures of the wild-type and mutant lipases were observed in attenuated total reflectance (ATR) FTIR spectra investigated in the amide I range ( $1600-1700\text{ cm}^{-1}$ ). The amide I range of the spectra was fitted with Gaussian curves using Origin 8.5 software. The relative contents of various secondary structures ( $\alpha$ -helix,  $\beta$ -sheet,  $\beta$ -turn and random coil) of wild-type and mutant lipases were determined and are shown in Figure 6.15 and summarized in Table 6.6. It reflects wild type and mutant lipases were analyzed in solution to observe the structural changes after mutagenesis on secondary structure. It was found that the increase in  $\alpha$ -helix and  $\beta$ -sheet contents of mutant lipases after mutagenesis, which is attributed to the increased hydrogen bonding between amine and carboxylic groups of the proteins. From, RCSB PDB and UniProtKB database, it was revealed that the wild-type *B. subtilis* lipase have 46.93%  $\alpha$ -helix, 21.79%  $\beta$ -sheets, 19.55% turns and 11.73% random coil. The increase in secondary structure contents by reducing the content of random coil makes the protein rigid and improve stability. The best results was obtained with a significant increase in the  $\alpha$ -helix of bsl\_the2, bsl\_the3,



bsl\_alk2 and bsl\_hal2 and it was around ~55% and  $\beta$ -sheets contents remains the same. The results are corroborated with Abrusán and Marsh,  $\alpha$ -helices are more susceptible to make the protein structure robust than  $\beta$ -sheets after mutagenesis<sup>23</sup>.

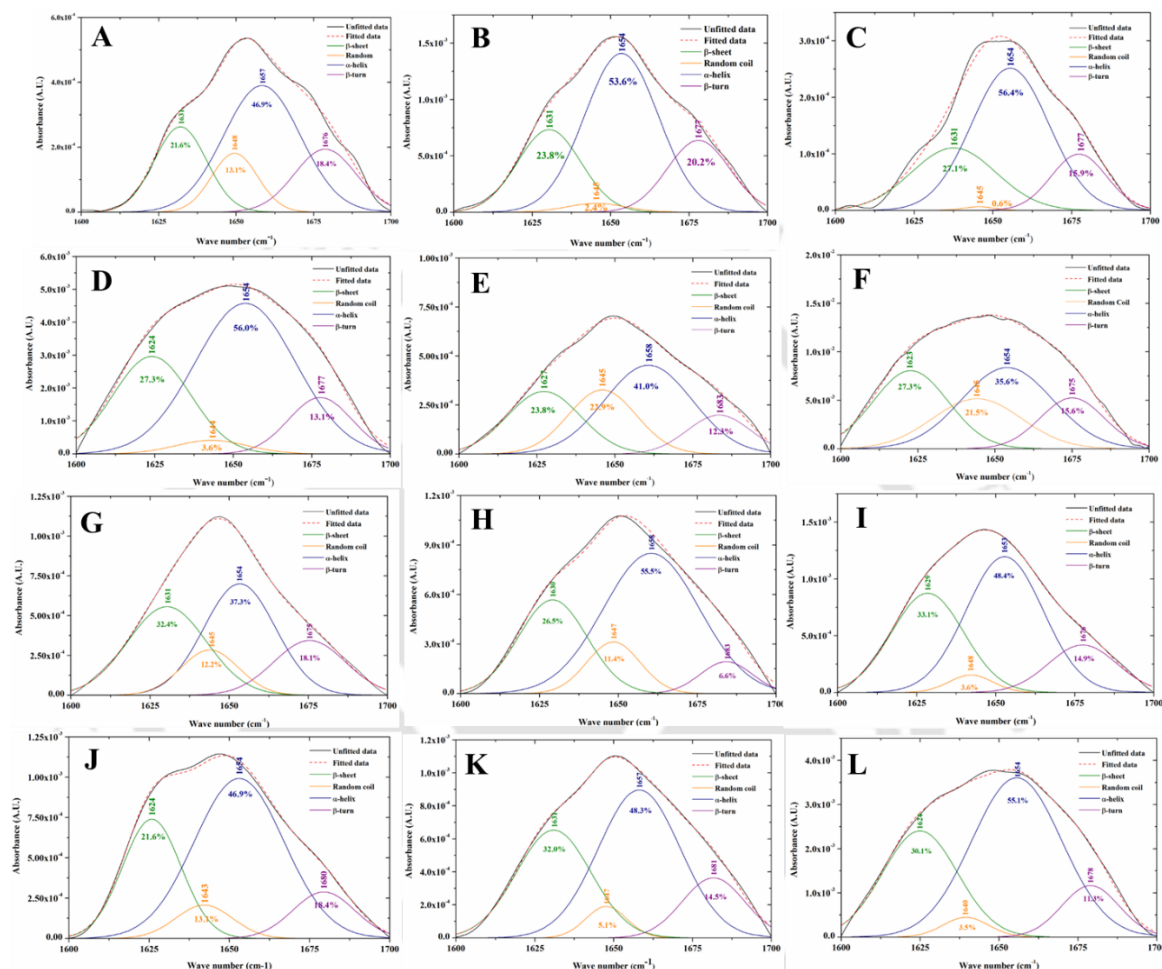


Figure 6.15: FTIR-ATR analysis for structural change in wild-type (A. bsl\_wt) and mutant lipases (B. bsl\_the1, C. bsl\_the2, D. bsl\_the3, E. bsl\_psy1, F. bsl\_psy2, G. bsl\_acd1, H. bsl\_acd2, I. bsl\_alk1, J. bsl\_alk2, K. bsl\_hal1 and L. bsl\_hal2).

Table 6.6: Determination of secondary structures by FTIR-ATR.

| Lipase and mutants | Secondary structures (%) (Best fitted values) |                 |       |              |
|--------------------|-----------------------------------------------|-----------------|-------|--------------|
|                    | $\alpha$ -helices                             | $\beta$ -sheets | Turns | Random coils |
| bsl_wt (database)  | 46.93                                         | 21.79           | 19.55 | 11.73        |
| bsl_wt             | 46.87                                         | 21.56           | 18.47 | 13.09        |
| bsl_the1           | 53.56                                         | 23.86           | 20.19 | 2.37         |
| bsl_the2           | 56.43                                         | 27.05           | 15.91 | 0.58         |
| bsl_the3           | 55.99                                         | 27.30           | 13.11 | 3.58         |
| bsl_psy1           | 40.85                                         | 23.86           | 12.32 | 22.95        |
| bsl_psy2           | 35.57                                         | 27.42           | 15.59 | 21.40        |
| bsl_acd1           | 37.28                                         | 32.42           | 18.11 | 12.18        |
| bsl_acd2           | 48.33                                         | 32.01           | 14.50 | 5.14         |
| bsl_alk1           | 48.48                                         | 33.09           | 14.86 | 3.56         |
| bsl_alk2           | 55.16                                         | 26.70           | 10.86 | 7.26         |
| bsl_hal1           | 48.33                                         | 32.01           | 14.50 | 5.14         |
| bsl_hal2           | 55.15                                         | 30.13           | 11.26 | 3.44         |

## 6.4 Conclusion

The *in vitro* cloning, mutagenesis, expression, purification and characterization of lipases was achieved. Results showed that in bsl\_the3, the optimum activity and stability increased from 35 °C to 55 °C, pH 8 to pH 10 in bsl\_alk2 but reduced to pH 6.5 in bsl\_acd2 when compared with wild type lipase (bsl\_wt) whereas, bsl\_wt was halo-sensitive (0.5M NaCl) but bsl\_hal2 halostability was increased to 1.5 M NaCl. Differential scanning fluorimetry analysis (Thermal shift assay) showed improved  $T_m$  values of mutant lipases and subsequently, FT-IR-ATR confirmed that the content of secondary structure enhanced the extreme-stability.

## 6.5 Reference

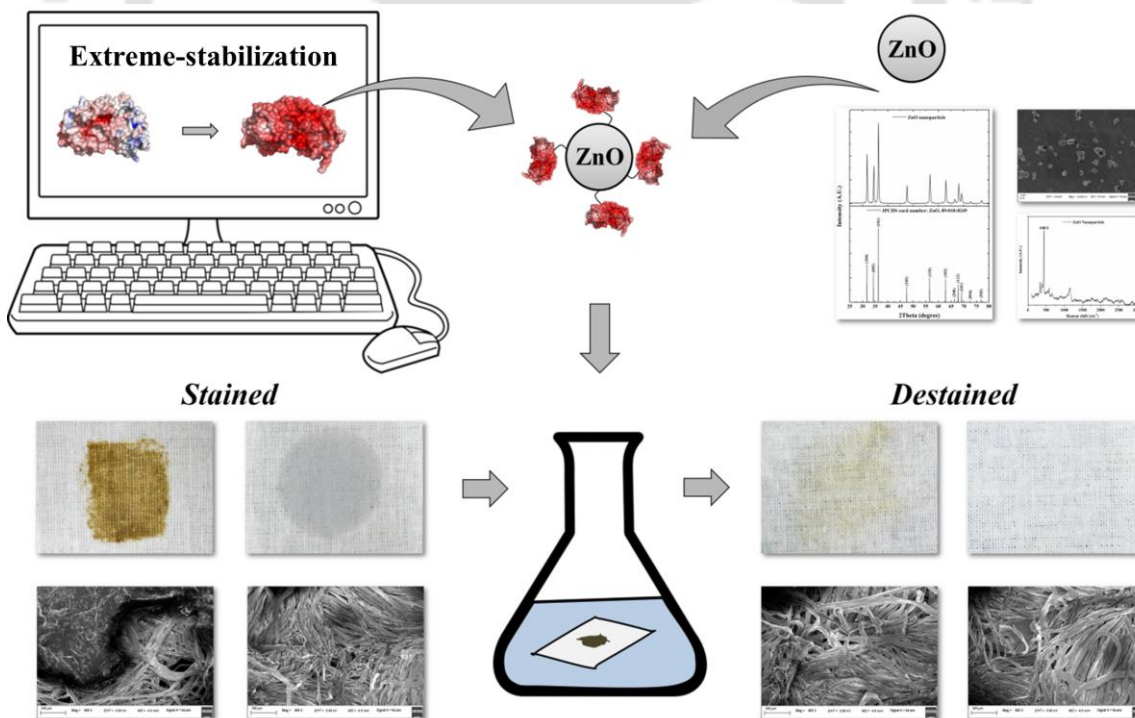
1. Kumar, S., Tsai, C.-J. & Nussinov, R. Factors enhancing protein thermostability. *Protein Eng. Des. Sel.* **13**, 179–191 (2000).
2. Acharya, P., Rajakumara, E., Sankaranarayanan, R. & Rao, N. M. Structural Basis of Selection and Thermostability of Laboratory Evolved *Bacillus subtilis* Lipase. *J. Mol. Biol.* **341**, 1271–1281 (2004).
3. Vogt, G., Woell, S. & Argos, P. Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **269**, 631–43 (1997).
4. Huang, L.-T., Gromiha, M. M. & Ho, S.-Y. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics* **23**, 1292–1293 (2007).
5. Trivedi, S., Gehlot, H. S. & Rao, S. R. Protein thermostability in Archaea and Eubacteria. *Genet. Mol. Res.* **5**, 816–27 (2006).
6. Dartois, V., Baulard, A., Schanck, K. & Colson, C. Cloning, nucleotide sequence and expression in *Escherichia coli* of a lipase gene from *Bacillus subtilis* 168. *Biochim. Biophys. Acta - Gene Struct. Expr.* **1131**, 253–260 (1992).
7. LESUISSE, E., SCHANCK, K. & COLSON, C. Purification and preliminary characterization of the extracellular lipase of *Bacillus subtilis* 168, an extremely basic pH-tolerant enzyme. *Eur. J. Biochem.* **216**, 155–160 (1993).
8. Van Pouderoyen, G., Eggert, T., Jaeger, K.-E. & Dijkstra, B. W. The crystal structure of *Bacillus subtilis* lipase: a minimal  $\alpha/\beta$  hydrolase fold enzyme<sup>1</sup>. *J. Mol. Biol.* **309**, 215–226 (2001).
9. Aranishi, F. Flexprep: scale-flexible rapid plasmid preparation for analysis of recombinant clones. *Mol. Biotechnol.* **21**, 39–41 (2002).
10. Nthangeni, M. B., Patterson, H.-G., van Tonder, A., Vergeer, W. P. & Litthauer, D. Over-expression and properties of a purified recombinant *Bacillus licheniformis* lipase: a comparative report on *Bacillus* lipases. *Enzyme Microb. Technol.* **28**, 705–712 (2001).
11. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–54 (1976).
12. Kruger, N. J. The Bradford Method for Protein Quantitation. in *Basic Protein and Peptide Protocols* **32**, 9–16 (Humana Press, 2002).
13. Glogauer, A. *et al.* Identification and characterization of a new true lipase isolated through metagenomic approach. *Microb. Cell Fact.* **10**, 54 (2011).
14. Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat. Protoc.* **2**, 2212–2221 (2007).
15. Ahmad, S., Kamal, M. Z., Sankaranarayanan, R. & Rao, N. M. Thermostable *Bacillus subtilis* lipases: in vitro evolution and structural insight. *J. Mol. Biol.* **381**, 324–40 (2008).
16. OHNISHI, K.; YOSHIDA, Y. TOITA, J.; SEKIGUCHI, J. Purification and characterization of a novel Lipolytic Enzyme from *Aspergillus oryzae*. *J. Fermentation Bioeng.* **78**, 413–419 (1994).
17. Wang, C.-Y. *et al.* Purification and characterization of a novel halostable cellulase from *Salinivibrio* sp. strain NTU-05. *Enzyme Microb. Technol.* **44**, 373–379 (2009).
18. Allali-Hassani, A. *et al.* A survey of proteins encoded by non-synonymous single nucleotide polymorphisms reveals a significant fraction with altered stability and activity. *Biochem. J.* **424**, 15–26 (2009).
19. Vedadi, M. *et al.* Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 15835–40 (2006).
20. Poklar, N., Lah, J., Salobir, M., Maček, P. & Vesnaver, G. pH and Temperature-Induced Molten Globule-Like Denatured States of Equinatoxin II: A Study by UV-Melting, DSC, Far- and Near-UV CD Spectroscopy, and ANS Fluorescence †. *Biochemistry* **36**, 14345–14352 (1997).
21. Song, J. K. & Rhee, J. S. Simultaneous enhancement of thermostability and catalytic activity of phospholipase A1 by evolutionary molecular engineering. *Appl. Environ. Microbiol.* **66**, 890–894

- (2000).
22. Kuo, T.-Y., Tsai, C.-C. & Fu, H.-W. Enhanced Mutant Screening in One-step PCR-based Multiple Site-directed Plasmid Mutagenesis by Introduction of Silent Restriction Sites for Structural and Functional Study of Proteins. *Biol. Proced. Online* **19**, 12 (2017).
  23. Abrusán, G. & Marsh, J. A. Alpha Helices Are More Robust to Mutations than Beta Strands. *PLOS Comput. Biol.* **12**, e1005242 (2016).



## 7

## Employing predicted extreme-stable mutants for applications



## Abstract

This chapter employs generated extreme-stable mutants for applications. Detergent formulation was the chosen application here. To further increase the enzyme stability, bsl\_wt and engineered lipases (bsl\_the3, bsl\_psy2, bsl\_acd2, bsl\_alk2 and bsl\_hal2 have better extreme-stability) were immobilized on synthesized ZnO nanoparticles as it is cheap, biocompatible and environmental friendly. The synthesized ZnO nanoparticles and lipase immobilization was confirmed using UV-Vis, FT-IR, Raman spectroscopy, DLS (for particle size), powder XRD, TEM and FESEM analysis. Interestingly, ZnO immobilized engineered lipases further enhanced the storage stability, reusability and their compatibility with various oxidizing agents, surfactants and commercial detergents and was applied for oil and grease stain removal on cotton fabrics. For selection of better washing conditions, the washing process was done at different temperatures and time intervals using different concentrations of detergent and lipase activities. The order of performance in oil and grease stain removal from the stained cotton fabric pieces by immobilized lipases on ZnO nanoparticle- bsl\_alk2 > bsl\_the3 > bsl\_hal2 > bsl\_psy2 > bsl\_acd2 > bsl\_wt. Results also showed that the maximum removal of oil and grease stains were up to 90% and 82%, respectively.



## 7.1 Introduction

Recent biotechnological advancement, particularly in areas of protein engineering have provided important tools for the efficient development of new enzymes. Thus, enzymes with improved properties are employed for various tailor-made applications. However, in spite of their excellent catalytic properties, enzymes have many other characteristics that are not very suitable for their use in industries such as low stability, inhibition by high concentrations of substrates and products, low activity and selectivity towards unnatural substrates under non-conventional conditions. To overcome such conditions, several immobilization techniques are used. The term ‘immobilized enzymes’ refers to enzymes physically confined or localized in a certain defined region of space with retention of their catalytic activities that can be used repeatedly and continuously<sup>1</sup>.

Immobilizing enzymes can be categorized into irreversible and reversible methods<sup>2</sup>. Irreversible immobilization means that once the biocatalyst is attached to the support it cannot be detached without destroying either the biological activity of the enzyme or the support. The most common procedures of irreversible enzyme immobilization are covalent coupling, entrapment or micro-encapsulation, and cross-linking. Immobilization of proteins by methods based on the formation of covalent bonds are among the most widely used<sup>3</sup>. An advantage of this method is that, because of the stable nature of the bonds formed between enzyme and matrix, the enzyme is not released into the solution upon use. Further, the covalent reactions commonly employed give rise to enzymes linked to the support through either amide, ether, thio-ether, or carbamate bonds. Therefore, the enzyme is strongly bound to the matrix. The entrapment method is based on the occlusion of an enzyme within a polymeric network that allows the substrate and products to pass through but retains the enzyme<sup>3</sup>. There are different approaches for entrapping enzymes such as gel<sup>4</sup> or fiber entrapping<sup>5</sup> and micro-encapsulation<sup>6</sup>. Moreover, cross-linking is performed by formation of intermolecular cross-linkages between the enzyme molecules by means of bi- or multifunctional reagents. The most commonly used cross-linking reagent is glutaraldehyde as it is economical and easily obtainable in large quantities. It has been used for cross-linking proteins via reaction of the free amino groups on the surface of enzyme molecules with oligomers or polymers of glutaraldehyde. The use of reversible methods

for enzyme immobilization is highly attractive and economic. The simplest immobilization method is nonspecific adsorption, which is mainly based on physical adsorption or ionic binding<sup>7,8</sup>. Another is the ‘metal link immobilization’ method where transition metal salts or hydroxides deposited on the surface of organic carriers become bound by coordination with nucleophilic groups on the matrix<sup>9–11</sup>.

Recently, nanoparticle-based immobilization of enzymes has gained immense interest. Nanostructure materials exhibits remarkable properties such as large surface to volume ratio, high surface reaction activity, high catalytic efficiency and strong adsorption ability that makes them potential candidate materials to play an important role in enzyme immobilization<sup>12</sup>. Enzymes immobilized on nanoparticles shows a broader working pH and temperature range with high thermal stability than the native enzymes. Moreover, enzyme immobilized nanoparticles are easy to synthesize without using toxic reagents, are homogeneous, and particle size can be conveniently tailored within utility limits. Recently, immobilization of industrially important enzymes onto nanomaterials with improved performance has paved the way to myriad of application-based commercialization<sup>13</sup>.

This chapter pertains to the synthesis of zinc oxide nanoparticles to enhance the stability of the enzyme for application in detergent formulation. The synthesized nanoparticles were characterized and the lipases were subjected to covalent immobilization using glutaraldehyde as the cross linking agent. This gives a promising perspective of using low-cost nanoparticle conjugated biocatalysts for detergent formulation to overcome the drawbacks of the chemical counterparts.

## 7.2 Methodology

### 7.2.1 Chemicals and instrumentation

Reagents such as 4-nitrophenyl octanoate (p-NPO), branched polyethyleneimine (PEI), (3-aminopropyl) triethoxysilane (APTES) and glutaraldehyde (25%, aqueous solution) were purchased from Sigma-Aldrich (Germany). Zinc Sulphate ( $\text{ZnSO}_4$ ), sodium hydroxide (NaOH), surfactants (Tween 20; Tween 80; Triton X-100; NP-40 and sodium dodecyl sulfate, SDS), oxidizing agents (hydrogen peroxide,  $\text{H}_2\text{O}_2$ ; sodium perborate,  $\text{NaBO}_3$  and

sodium hypochlorite, NaOCl) were procured from Hi-Media (India). Commercial detergents like Ariel, Surf Excel, Henko, Tide and Vanish were used for destaining of grease and oil stained from cotton fabrics. All other reagents, solvents used were of molecular biology/analytical grade and obtained from Merck (USA) and Himedia (India).

The instruments like Field Emission Scanning Electron Microscope (FESEM) (Make: Zeiss, Model: Sigma), Transmission Electron Microscope (TEM) (JEM – 2100 HRTEM, Make – JEOL, Japan), Dynamic Light Scattering (DLS) instrument Delsa Nano C particle analyzer (Beckmen Coulter Inc.), UV-Vis Spectrophotometer UV-2600 (Shimadzu), Spectrum Two™ FTIR Spectrometers (Perkin-Elmer), Laser Micro Raman System (Horiba Jobin Vyon) and Powder X-ray diffractometer (Rigaku TTRAX, Seifert XRD 3003 T/T) were used for the present work.

### **7.2.2 Testing lipase compatibility with surfactants, oxidizing agents and commercial detergents before and after immobilization**

To investigate the compatibility of wild-type and extreme-stable mutants of *Bacillus subtilis* lipase with various surfactants, oxidizing agents and commercial detergents, the activity of lipase was tested in their presence at a concentration of 10 % v/v or w/v and assayed under standard conditions and expressed in terms of percent relative activity<sup>14</sup>. In case of commercial detergent, the endogenous lipases present were inactivated by heating the diluted detergents at 90 °C for 1 h prior to the addition of enzyme preparation. To determine the stability, an aliquot of wild-type and mutant lipases (50 U ml<sup>-1</sup>) was incubated with detergent components (10% v/v or w/v of respective detergent) in phosphate buffer (20 mM, pH 8.0) for 1 h at 35 °C. The percentage activity (%) of each sample was determined and compared with the control of wild-type lipase (bsl\_wt) without detergent. The relative activity of control was defined as the enzyme activity without detergent, incubated under the similar conditions, and was taken as 100 %. The detergent compatible wild-type and mutant lipases were chosen and then subjected for immobilization on ZnO nanoparticle to further improve stability and reusability of enzymes. Furthermore, these most compatible immobilized mutants of lipases were again tested for surfactants, oxidizing agents and commercial detergents.

### 7.2.3 Synthesis of functionalized ZnO nanoparticles

ZnO nanoparticles were synthesized by direct precipitation method using zinc sulphate and NaOH as precursors. In this work, the aqueous solution (0.45 M) of zinc sulphate ( $\text{ZnSO}_4$ ) and the solution (0.9 M) of NaOH were prepared. The pre-heated NaOH solution (at 55 °C) was slowly added into  $\text{ZnSO}_4$  solution at 55 °C under vigorous stirring (at 300 rpm), which resulted in the formation of a white suspension. The white was washed repeatedly with distilled water to reach the suspension  $\sim$  pH 7. After washing, the nanoparticles were dried at 80 °C overnight. During drying, complete conversion of  $\text{Zn}(\text{OH})_2$  into ZnO takes place. The obtained product was calcined at 500 °C in air atmosphere for 2 h<sup>15</sup>. The grafting of PEI onto the surface of naked ZnO nanoparticles was carried. Here, PEI (1.5 ml) was suspended in a solution containing 10 % APTES (in 10ml ethanol). To this mixture, 10 mg of ZnO nanoparticles were added and stirred vigorously to react for 24 h. The composite obtained was isolated by centrifugation (10,000×g for 15 min) and repeatedly washed with ethanol and milliQ in order to remove excess PEI.

### 7.2.4 Covalent immobilization of lipase on synthesized ZnO nanoparticles

The covalent immobilization procedure was developed using mild glutaraldehyde (GLU) as cross-linking agents between the lipase and the amino groups located on the surface of the ZnO-PEI/APTES. In a typical procedure, 10 mg of functionalized ZnO-PEI/APTES was incubated in 10 ml phosphate buffer solution (20 mM, pH 8.0) containing 0.25 % (w/v) GLU and allowed to react for 2 h at 100 rpm at room temperature. Then, 1 ml of lipase solution (20 mg ml<sup>-1</sup>) was incorporated into the mixture and stirred for 2 h. The stepwise protocol of synthesis, coating and immobilization is also shown in Figure 7.1.

The resultant nanobioconjugant obtained were dried under vacuum (37 °C for 1 h) and further preceded for characterization through FESEM and TEM analysis, DLS (for particle size), UV-Vis, FT-IR, Raman spectroscopy, powder XRD analysis. The crystalline structure of synthesized ZnO nanoparticles and their chemical modification on the surface of nanoparticles before and after functionalization and lipase immobilization was studied

by FESEM, TEM, DLS, UV-Vis spectroscopy, FT-IR spectroscopy, Raman spectroscopy and powder XRD analysis.

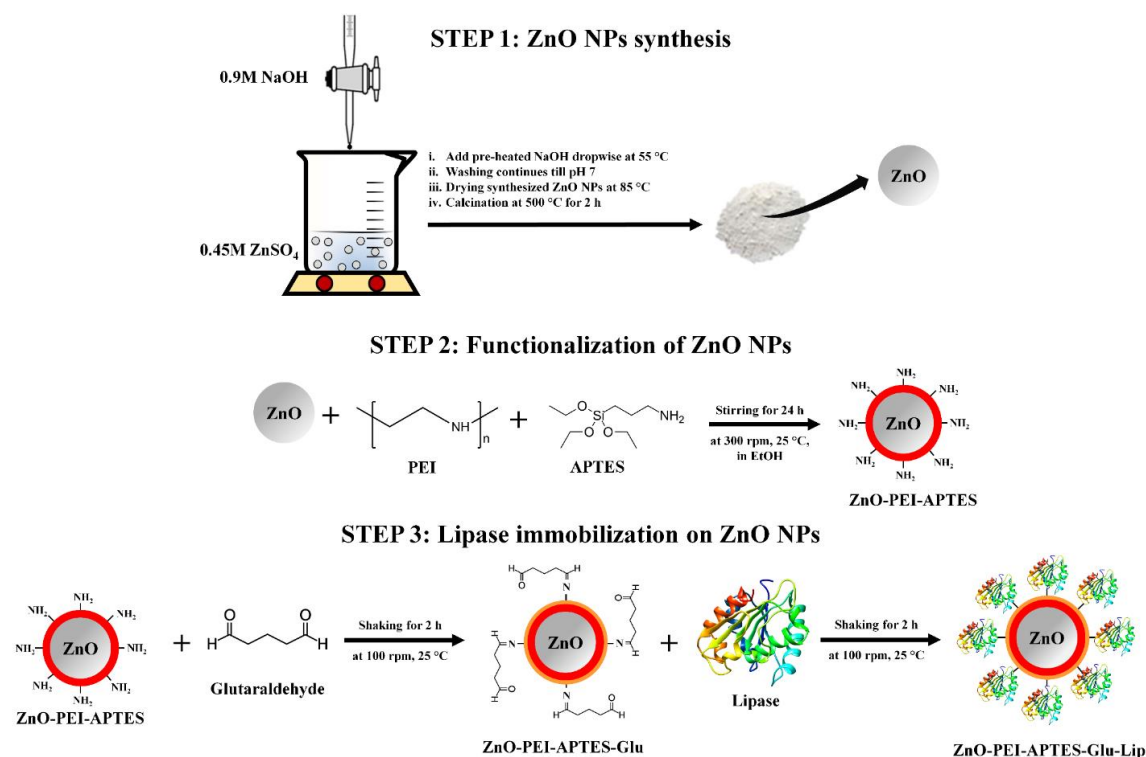


Figure 7.1: Scheme of ZnO nanoparticle synthesis and lipase immobilization.

## 7.2.5 Studying morphology and size of naïve and bioconjugated ZnO nanoparticle

### FESEM analysis

The morphology of naïve and bioconjugated ZnO nanoparticles was studied by Field Emission Scanning Electron Microscope (FESEM) (Make: Zeiss, Model: Sigma) and Transmission Electron Microscope (FEI TECNAI G2 20S-TWIN TEM). The sample preparation for FESEM analysis involved making solution of dry nanoparticle powder in water having concentration of 0.1 mg ml<sup>-1</sup> and then sonicate the suspension for 15 min at room temperature. The prepared sample was placed onto carbon tape (aluminum foil pasted on the one side of the two sided carbon tape) and examined under a FESEM.

### ***TEM analysis***

The morphology of naïve and bioconjugated ZnO nanoparticles was also studied by Transmission Electron Microscope (FEI TECNAI G2 20S-TWIN TEM) as per earlier reports<sup>16–18</sup>. An accelerating voltage of 120 kV was used to image and study the morphology (size and shape) of each sample solution of naïve ZnO and bioconjugated ZnO. The TEM samples were prepared by drop casting samples (10 µl of 1% ZnO nanoparticles dispersed in aqueous suspension) in the carbon coated copper (200 mesh) grid.

### ***DLS analysis***

The dynamic light scattering is a widely used technique for the determination of particle size in colloidal solution. Size distributions of the naïve and bioconjugated ZnO nanoparticles were obtained using a dynamic light scattering instrument Delsa Nano C particle analyzer (Beckmen Coulter Inc.) equipped with a green laser at 532 nm. For DLS measurements suspensions were prepared at 0.2 g l<sup>-1</sup> solid loading of ZnO at pH 8.0. Initially, the solution at pH 8.0 was passed through a 0.2 µm syringe driven filter to minimize the effect of unwanted dust particles and samples were sonicated for 5 min.

## **7.2.6 Characterization of naïve and bioconjugated ZnO nanoparticle using spectroscopic analysis**

### ***UV-Vis spectroscopic analysis***

UV-Vis absorption spectroscopy is widely being used technique to examine the optical properties of nanosized particles. ZnO nanoparticles were characterized in a UV-Vis Spectrophotometer UV-2600 (Shimadzu). The scanning range for the samples was 200–700 nm. The spectrophotometer was equipped with “UVWinlab” software to record and analyze data. Base correction of the spectrophotometer was carried out by using a blank reference. The UV-Vis absorption spectra of the samples were recorded and numerical data were plotted.



### ***FTIR spectroscopic analysis***

Fourier-transform infrared spectroscopy (FTIR) technique was used to determine the confirmation of synthesized ZnO nanoparticle, binding of PEI/APTES to ZnO nanoparticle, binding of APTES, glutaraldehyde and lipase to ZnO nanoparticle. The wild-type and mutant lipases (20 mg ml<sup>-1</sup>) were dissolved into 10 ml functionalized ZnO-PEI/APTES nanoparticles (10 mg). All of these samples were used as in the solid products after lyophilization in a freeze dryer. FTIR scanning was performed with constant nitrogen purging using a Spectrum Two™ FTIR Spectrometers (Perkin-Elmer) equipped with a DTGS KBr detector and a KBr beam splitter with constant nitrogen purging. Then, 1 – 2 mg of fine sample powder and 90 – 100 mg of KBr powder were mixed and dried completely then transferred that mixture to make a nearly transparent and homogeneous pellet. All spectra were taken at 4 cm<sup>-1</sup> resolution, averaged over 20 scans in the range of 400 cm<sup>-1</sup> to 4000 cm<sup>-1</sup>. The final spectra were collected after subtracting the background spectra of D<sub>2</sub>O.

### ***Raman spectroscopic analysis***

Laser Micro Raman System (Horiba Jobin Vyon) was used to record spectra with a 514 nm laser source. It was also applied to study the interaction of ZnO nanoparticle with coating material (PEI and APTES) and lipases. All wild-type and selected mutant lipases (20 mg ml<sup>-1</sup>) were independently dissolved into 10 ml functionalized ZnO-PEI/APTES nanoparticles and subjected for lyophilization in a freeze dryer. It was scanned in the range of 100–2000 cm<sup>-1</sup> with a 2 cm<sup>-1</sup> resolution.

### **7.2.7 Analysis of naïve and bioconjugated ZnO nanoparticles using powder X-ray diffractometers**

The crystallinity of naïve and bioconjugated ZnO nanoparticles were determined using powder X-ray diffractometer (Rigaku TTRAX, Seifert XRD 3003 T/T). Analysis was performed by using an XRD equipped with a Cu K $\alpha$  ( $\lambda$ =1.54 Å) source, maintaining applied voltage of 40 kV and current at 30 mA. About 0.3 g of dried ZnO samples were

deposited as a randomly oriented powder into a plexiglass sample container, and the XRD patterns were recorded between 25° and 70° angles, with speed of 5.0 degrees.

### 7.2.8 Enzyme activity assay and biochemical characterization of immobilized lipases

The lipase bound ZnO nanoparticles were used for further measurements. The amount of enzyme immobilized on nanoparticles was estimated by Bradford assay at 595 nm using bovine serum albumin as a standard<sup>19</sup>. Immobilization efficiency was calculated using equation (7.1) and immobilization yield using equation (7.2):

$$\text{Efficiency of immobilization (\%)} = \frac{E_0 - E_1}{E_0} \times 100 \quad (7.1)$$

$$\text{Immobilization yield (\%)} = \frac{Y_1}{Y_0} \times 100 \quad (7.2)$$

In equation (7.1) of immobilization efficiency, the  $E_0$  and  $E_1$  indicate the amounts of total protein introduced for immobilization and the amount of protein present in the filtrate solutions after immobilization, respectively. On the other hand, the yield of immobilization in equation (7.2), where,  $Y_1$  is the enzymatic activity used for immobilization and  $Y_0$  the activity immobilized on the support.

The activity assay was conducted in order to examine the binding lipase residual activity. The activities of ZnO immobilized lipases were determined spectrophotometrically using *p*-nitrophenyl octanoate (*p*-NPO) as substrate<sup>20</sup>. The substrate used was 0.2 mM *p*-NPO dissolved in phosphate buffer (20 mM, pH 8); the resulting solution was uniformly mixed with lipase and incubated at 35°C for 10 min. Subsequently, the *p*-NPO was added into the reaction solutions and the absorbance was measured at 410 nm. One unit of ZnO immobilized lipase activity is defined as the amount of enzyme that produced 1 μmol of *p*-NP in 10 min at 35 °C and pH 8. Further, the ZnO immobilized wild-type and mutant lipases were used to determine optimum temperature, pH and salt concentration.

After activity analysis, the ZnO immobilized lipases were further used for determination of optimum temperature (in the range of 20 °C to 70 °C), optimum pH (in the range of pH 2 to pH 12) and optimum salt concentration (in the range of 0 M to 3.5 M NaCl). Finally,

utilizing the optimized biochemical parameter and the immobilized enzymes were used for detergent application of lipases.

### 7.2.9 Reuse and storage stability of ZnO immobilized lipases

The storage stability of the free and immobilized wild-type and selected mutant lipases was investigated by calculating the residual activity till 60 days of experiment. The free and immobilized lipases were stored in phosphate buffer (20 mM, pH 8.0) at room temperature (25 °C) and their activities were assayed at regular intervals of 4 days for 60 days. In order to investigate the operational stability, the immobilized lipase on ZnO nanoparticle was separated by centrifugation, washed with phosphate buffer (20 mM, pH 7.0) and eventually reused with fresh substrates for upto 20 cycles of experiment.

### 7.2.10 Application of ZnO immobilized lipases for removal of oil and grease stains from cotton fabrics

#### *Preparation of olive oil and grease soiled cotton fabric*

Defatting of cotton fabric (4 cm × 4 cm) supposed to be soiled was carried out by boiling in chloroform for 4 h and was repeated thrice. Soiling of defatted cotton fabric was done twice with micropipette by spotting olive oil in benzene solution (0.5 ml, 100 mg ml<sup>-1</sup> concentration) and grease coating was also done on defatted cotton.

#### *Destaining of oil and grease stains from cotton fabrics*

There are three kind of washing solutions namely, B-L, B-D and B-D-L and the composition for preparing 100 ml of respective washing solution are given in Table 7.1. The solution B-D-L, consists of buffer, detergent (any commercial detergent at which the best activity was obtained by the ZnO immobilized lipase) and lipase solutions, was prepared by taking the measured buffer and surfactant solutions in Erlenmeyer flask with ground stopper. Preheat (35 °C for 20 min) the contents of the flask before adding the lipases. Similarly, B-L and B-D solutions were also prepared in the same manner. The final volume of the respective washing solutions was adjusted to 100 mL with distilled water. Washing of soiled fabric dipped in respective washing solutions was done by shaking the

flask contents at 100 rpm at 35 °C for 15 min. After 10 min, the washed soiled fabric was rinsed thrice with 100 mL distilled water for 2 min at 37 °C and air dried.

For selection of better or optimized washing conditions, the washing process was done at different washing temperatures and time intervals using different concentrations of detergents and lipases<sup>14</sup>. The optimization was done for the destaining of oil and grease stained cotton fabrics with varying concentration of detergent (upto 9 % w/v), lipase (upto 70 U ml<sup>-1</sup>), with varying temperature range (30 – 70 °C) and with varying interval of washing time (10 – 70 min). The percentage of oil and grease removal was calculated by using the following formula in equation (7.3)

$$\% \text{ Oil or Grease Removal} = \frac{W_3 - W_1}{W_2 - W_1} \times 100 \quad (7.3)$$

where,  $W_1$  is weight of cloth before applying oil/grease (mg);  $W_2$  is weight of cloth after applying oil/grease (mg) and  $W_3$  is weight of cloth after total oil/grease removal (mg).

Table 7.1: Compositions of washing solution for making 100 ml solution.

| Components                       | Total volume (100 ml)           |                                 |                                                  |
|----------------------------------|---------------------------------|---------------------------------|--------------------------------------------------|
|                                  | B <sup>1</sup> + L <sup>2</sup> | B <sup>1</sup> + D <sup>3</sup> | B <sup>1</sup> + D <sup>3</sup> + L <sup>2</sup> |
| Phosphate buffer (20 mM, pH 8.0) | 40                              | 40                              | 40                                               |
| Detergent (0.5%)                 | —                               | 50                              | 50                                               |
| Lipase (50 U/mL)                 | 10                              | —                               | 10                                               |
| Distilled water                  | 50                              | 10                              | —                                                |

<sup>1</sup>Buffer; <sup>2</sup>Lipase (ZnO immobilized lipase); <sup>3</sup>Detergent.

### 7.2.11 Statistical analysis

All the experiments were carried out in triplicate, and results are expressed as mean ± standard deviation for at least n = 3. GraphPad Prism version 6.0 was used to determine the statistically significant differences between the means of different groups using one-way analysis of variance (ANOVA) with the Tukey test. Differences at p < 0.05 were considered significant.

## 7.3 Results and Discussion

The produced wild-type and mutant lipases have examined high stability toward extremes of temperature, pH and salt concentrations were used for applications. To further increase the enzyme stability, activity and reusability, the wild-type and mutant lipases were immobilized on synthesized Zinc oxide (ZnO) nanoparticles as they are cheap, non-toxic, and eco-friendly. Finally, the immobilized enzymes were used for application in detergent formulation.

### 7.3.1 Compatibility of lipase with surfactants, oxidizing agents and commercial detergents

Lipases is effectively used for remove oily and greasy stains from the fabrics and thus, have applications in detergent formulations. The present approach attempts to first engineer the lipase to improve its compatibility, robustness and stability under such harsh processes of detergent application. To test the compatibility of lipase, the enzymes were incubated with various surfactant, oxidizing agents and commercial detergents and then assayed under standard optimum experimental conditions. The compatibility of wild-type and mutant lipases with various surfactant, oxidizing agents and commercial detergents have been summarized in Table 7.2 and Figure 7.2. The analysis revealed that the most compatible mutants with various detergents were found to be bsl\_the3, bsl\_psy2, bsl\_acd2, bsl\_alk2 and bsl\_hal2 in the individual class of extreme-stability. These mutant lipases were selected for immobilization on ZnO nanoparticle to further improve stability and reusability of enzymes.

Table 7.2: Effect of surfactants, oxidizing agents and commercial detergents on wild-type and mutant lipase activity.

| Detergent component | Relative activity of lipase in presence of detergent component |            |            |             |             |            |             |            |             |            |             |            |             |
|---------------------|----------------------------------------------------------------|------------|------------|-------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|
|                     | Control                                                        | bsl_wt     | bsl_the1   | bsl_the2    | bsl_the3    | bsl_psy1   | bsl_psy2    | bsl_acd1   | bsl_acd2    | bsl_alk1   | bsl_alk2    | bsl_hal1   | bsl_hal2    |
| Tween 20            | 100 ± 4.0                                                      | 72.1 ± 3.0 | 73.4 ± 2.0 | 74.5 ± 3.0  | 89.5 ± 3.4  | 74.5 ± 2.0 | 78.4 ± 3.0  | 73.5 ± 2.5 | 78.4 ± 3.0  | 87.5 ± 3.6 | 105.5 ± 2.5 | 73.5 ± 2.0 | 106.5 ± 3.4 |
| Tween 80            | 100 ± 4.0                                                      | 79.0 ± 2.5 | 84.0 ± 2.5 | 77.2 ± 2.3  | 93.2 ± 1.2  | 77.2 ± 2.5 | 84.0 ± 2.3  | 77.2 ± 3.6 | 84.0 ± 2.3  | 88.2 ± 2.3 | 108.2 ± 3.6 | 82.2 ± 2.5 | 106.2 ± 1.2 |
| Triton X-100        | 100 ± 4.0                                                      | 80.2 ± 3.6 | 78.8 ± 1.5 | 91.9 ± 2.5  | 100.9 ± 2.7 | 83.9 ± 1.5 | 88.8 ± 2.5  | 81.9 ± 3.6 | 88.8 ± 2.5  | 89.9 ± 3.2 | 108.9 ± 2.5 | 81.9 ± 3.2 | 106.9 ± 3.0 |
| NP-40               | 100 ± 4.0                                                      | 65.4 ± 2.3 | 67.2 ± 2.0 | 71.2 ± 4.0  | 83.2 ± 1.2  | 74.2 ± 2.0 | 91.2 ± 4.0  | 65.2 ± 2.3 | 91.2 ± 4.0  | 85.2 ± 2.0 | 103.2 ± 3.6 | 84.2 ± 2.0 | 100.2 ± 2.5 |
| SDS                 | 100 ± 4.0                                                      | 62.4 ± 3.2 | 60.2 ± 3.4 | 60.8 ± 2.6  | 79.2 ± 1.6  | 70.2 ± 3.4 | 80.8 ± 2.6  | 65.2 ± 3.2 | 80.8 ± 2.6  | 81.2 ± 3.4 | 99.2 ± 2.3  | 70.2 ± 3.4 | 82.2 ± 3.6  |
| H2O2                | 100 ± 4.0                                                      | 95.5 ± 2.9 | 95.0 ± 2.5 | 93.5 ± 2.5  | 104.0 ± 3.0 | 85.0 ± 3.0 | 92.5 ± 2.5  | 83.0 ± 2.0 | 92.5 ± 2.5  | 86.0 ± 3.5 | 102.0 ± 2.0 | 88.0 ± 3.2 | 107.0 ± 2.0 |
| NaBO3               | 100 ± 4.0                                                      | 98.7 ± 3.2 | 99.0 ± 3.0 | 95.9 ± 1.8  | 109.0 ± 2.0 | 91.6 ± 2.0 | 105.9 ± 1.8 | 89.0 ± 3.0 | 105.9 ± 1.8 | 87.0 ± 2.0 | 110.0 ± 3.0 | 91.6 ± 2.0 | 104.0 ± 4.0 |
| NaClO               | 100 ± 4.0                                                      | 96.3 ± 2.1 | 90.2 ± 3.5 | 100.3 ± 2.7 | 113.0 ± 2.0 | 92.1 ± 2.0 | 102.3 ± 2.7 | 91.0 ± 4.0 | 102.3 ± 2.7 | 91.0 ± 3.0 | 106.0 ± 4.0 | 92.1 ± 2.5 | 109.0 ± 4.0 |
| Tide                | 100 ± 4.0                                                      | 85.3 ± 3.0 | 84.0 ± 2.0 | 83.4 ± 2.9  | 105.0 ± 2.6 | 60.2 ± 2.6 | 83.4 ± 2.9  | 55.0 ± 2.3 | 83.4 ± 2.9  | 70.2 ± 3.0 | 99.0 ± 2.5  | 55.2 ± 2.6 | 99.0 ± 3.0  |
| Vanish              | 100 ± 4.0                                                      | 70.5 ± 3.0 | 69.0 ± 1.3 | 76.8 ± 2.6  | 97.0 ± 1.9  | 66.0 ± 1.9 | 86.8 ± 2.6  | 62.0 ± 3.2 | 86.8 ± 2.6  | 71.0 ± 2.0 | 93.0 ± 1.5  | 66.0 ± 1.9 | 103.5 ± 4.0 |
| Ariel               | 100 ± 4.0                                                      | 84.0 ± 3.5 | 84.0 ± 1.5 | 80.4 ± 2.3  | 101.0 ± 1.6 | 74.0 ± 1.6 | 90.4 ± 2.3  | 63.5 ± 4.0 | 90.4 ± 2.3  | 74.5 ± 2.5 | 108.0 ± 4.0 | 73.5 ± 1.6 | 105.0 ± 2.9 |
| Henko               | 100 ± 4.0                                                      | 89.3 ± 2.6 | 87.0 ± 3.2 | 89.9 ± 3.0  | 94.0 ± 2.6  | 70.0 ± 2.6 | 89.9 ± 3.0  | 65.0 ± 2.0 | 89.9 ± 3.0  | 71.0 ± 3.0 | 102.0 ± 2.0 | 73.0 ± 2.6 | 109.0 ± 2.0 |
| Surf Excel          | 100 ± 4.0                                                      | 81.9 ± 3.1 | 78.0 ± 2.1 | 82.8 ± 3.5  | 99.0 ± 2.4  | 67.0 ± 2.4 | 92.8 ± 3.5  | 64.0 ± 2.4 | 92.8 ± 3.5  | 77.0 ± 4.0 | 111.0 ± 2.4 | 69.0 ± 3.4 | 102.0 ± 2.4 |



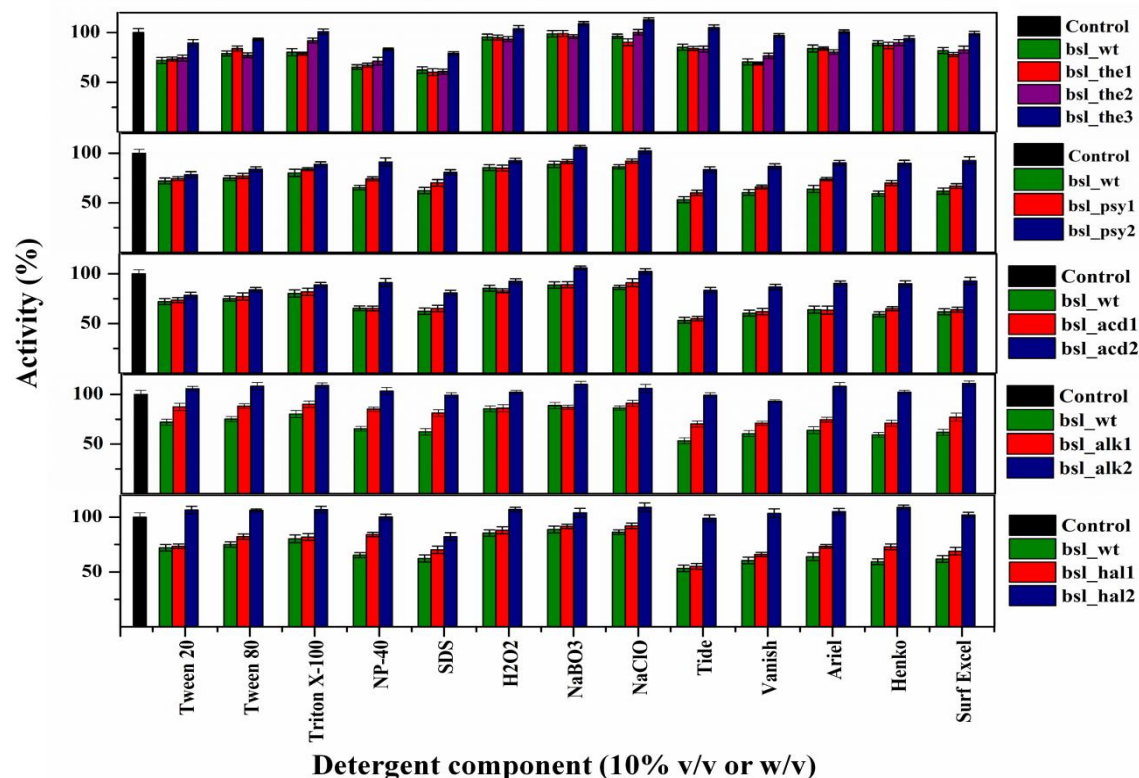


Figure 7.2: Effect of various surfactant, detergents and oxidizing agents on lipase (before immobilization). All experiments were conducted in triplicate and represented as mean  $\pm$  SD (error bars);  $p$ -value  $< 0.05$ .

### 7.3.2 Morphological characterization of naïve and bioconjugated ZnO nanoparticles using FESEM, TEM and DLS analysis

There are varieties of ZnO nanostructures that had been discovered, which are in the form of nanorods, nanotubes, nanobelts, nanostar, nanoflower, nanosprings, nanospirals, nanorings and many more<sup>21</sup>. Morphology study has found out synthesized ZnO are in form of flower-like structure as zinc sulphate were used as precursor. It was supported by other findings, which claimed that the morphology of ZnO nano-powder using zinc sulphate is smoother than in zinc nitrate<sup>22</sup>. Moreover, precursor concentration plays a great role on morphological features of ZnO nanostructure. According to Pourrahimi et al., ZnO can be form in different structures due to the type of precursors that has been used, such as zinc acetate nanorod like structure, zinc chloride and sulphate-nano-prism or nanostar or nanoflower structure and zinc nitrate-prism flower shape structure<sup>23</sup>. Kashif et al. claimed that increase in the concentration of precursor would increase the growth and density of ZnO nanoparticles<sup>24</sup>.

From Figure 7.3 and 7.5, it is observed that ZnO nanoparticles are flower/star shaped (multi-pod structure) with the leaf like parts at the end of the branches. The ordered regular-shaped particle nature of ZnO nanoparticle are shown in Figure 7.3A - 7.3D. The FESEM and TEM micrographs of as-synthesized (naïve) ZnO nanoparticle are represented in Figure 7.3B and 7.3D (at higher magnifications; 100 nm scale for FESEM and 20 nm scale for TEM). It showed star-like or flower-like morphology of ZnO nanoparticles which was previously reported in many literatures<sup>25</sup>. As clearly seen from Figure 7.3B and 7.3D, most of the particle's shape consists of 4-6 petal-like structures attached together in the center where ZnO crystal seed is located. The chemical reaction is expected to be:



Moreover, it can evaluate the content of  $\text{Zn(OH)}_2$  in the final product by analyzing the weight loss of the sample after heat treatment<sup>26</sup>. The final product was first dried in air at 85 °C for overnight to remove the water absorbed in air and weighed immediately to get the sample weight before annealing. Then it was annealed in air at 500 °C for 2 h to decompose  $\text{Zn(OH)}_2$  into ZnO and  $\text{H}_2\text{O}$ . After cooling down to 80 °C, the sample was weighed immediately again to get the weight after annealing. In this process, weight loss is expected to be due to  $\text{H}_2\text{O}$  coming from the decomposition of  $\text{Zn(OH)}_2$ . The decomposition equation of  $\text{Zn(OH)}_2$  showed



The suggesting mechanism is that elongated structure around the zinc hydroxide crystals lead to the formation of petal-like nucleation sites of ZnO. Therefore, during the reaction time, with increasing the quantity of the growth unit of reactant, the ZnO nuclei will be subsequently increased in a way that conglomeration of the ZnO nuclei structure can lead to three-dimensional nanostar-like particles (Figure 7.3A - 7.3D). As showed in Figure 7.3E, for particle size analysis, it has confirmed that the synthesized ZnO is in nanosize form. After analyzing data, it was found that ZnO nanoparticles size were in the range of 100 – 225 nm and is in close agreement with earlier literature. The average particle size of

synthesized ZnO powder was around 157 nm and mono-dispersed particles could be obtained with polydispersity index (PDI) of 0.176. This result is based on their length<sup>27,28</sup>.

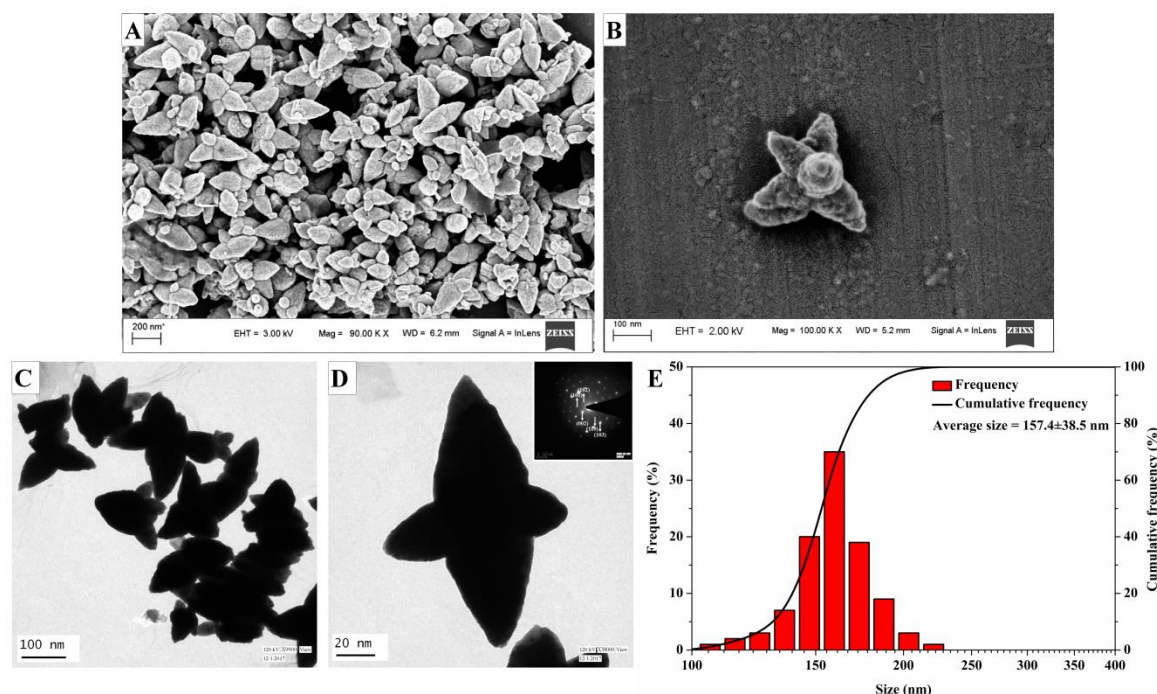


Figure 7.3: FESEM, TEM and DLS analysis of synthesized ZnO nanoparticles.

The morphology and size distribution of lipase immobilized ZnO nanoparticles were also studied using FESEM, TEM and DLS, respectively. The coating of nanoparticles contains free  $\text{NH}_2$ -group at the surface of the functionalized ZnO nanoparticle that can help in the covalent immobilization of lipase via  $\text{NH}_2$ -group of proteins (Lys and Arg) using glutaraldehyde as cross-linking agent. As Figure 7.4B and 7.4C clearly depicted that the successful coated layer of PEI and APTES for functionalization of ZnO nanoparticles as well as lipase immobilization on to the as-synthesized nanoparticles. In this figure the scale bar is 100 nm. In order to establish the coating/immobilization, surface edges were zoomed at a scale bar of 10 nm in Figure 7.4. Here one can clearly observe the coating layers/thin films of ~ 2 to 5 nm formed on the surface (in case of ZnO NPs coated) and visible traces of protein deposits formed on the surface (in case of ZnO NPs immobilized lipase). Further resolution can be achieved using HR-TEM.

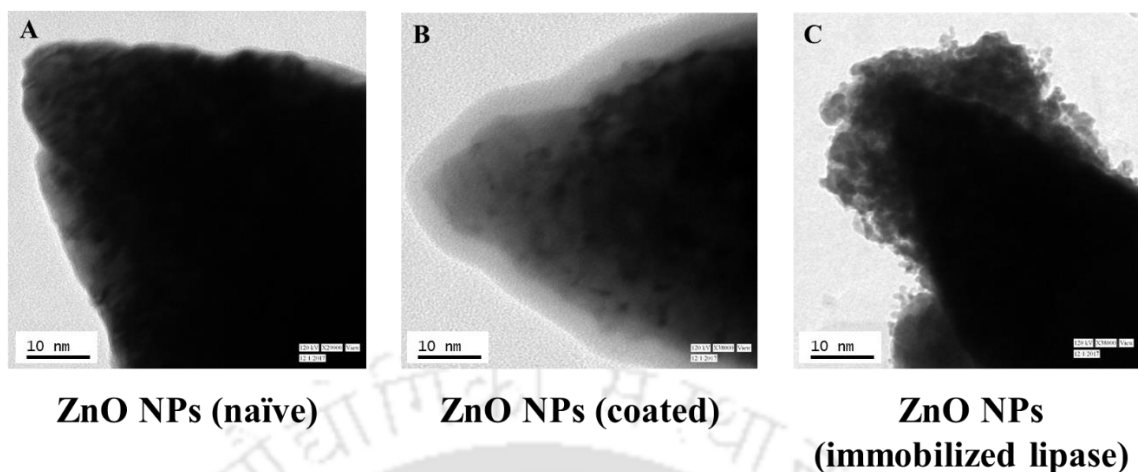


Figure 7.4: TEM micrographs of naïve (as-synthesized), coated and lipase immobilized ZnO nanoparticles at higher resolution (scale of 10 nm).

After analyzing data, it was found that ZnO nanoparticles size were in the range of 100 – 300 nm after immobilizations. The DLS analysis confirmed average size of bsl\_wt was around 208.0 nm (PdI = 0.221); bsl\_the3 was 238.6 nm (PdI = 0.252); bsl\_psy2 was 183.1 nm (PdI = 0.203); bsl\_acd2 was 187.1 nm (PdI = 0.208); bsl\_alk2 was 253.8 nm (PdI = 0.287); and bsl\_hal2 was 220.1 nm (PdI = 0.238). The increase in the average size of naïve ZnO from 157 nm revealed that lipases were successfully immobilized on to the nanoparticles (Figure 7.5).



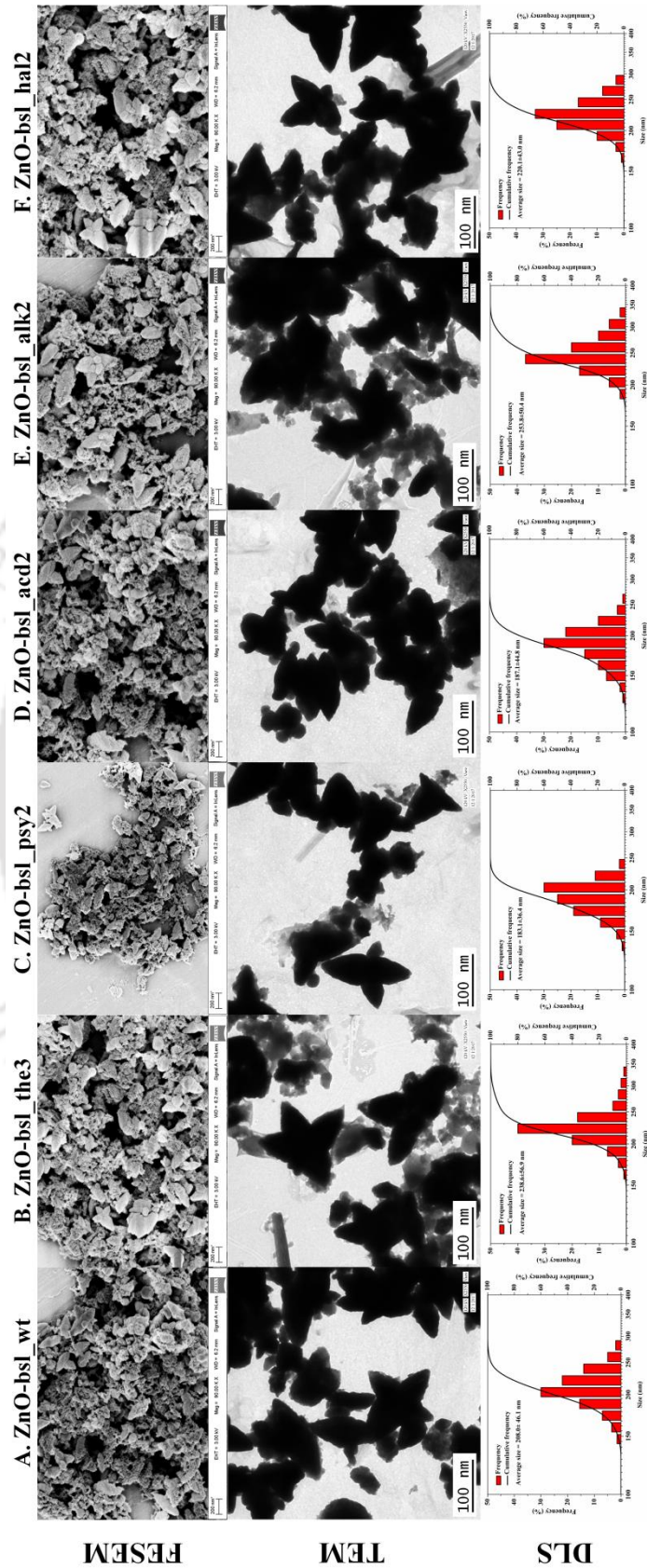


Figure 7.5: FESEM, TEM and DLS analysis of bsl\_wt, bsl\_the3, bsl\_psy2, bsl\_acd2, bsl\_alk2 and bsl\_hal2 on ZnO nanoparticles.

### 7.3.3 UV-Vis spectroscopic analysis for characterization of naïve and bioconjugated ZnO nanoparticles

UV-Vis absorption spectroscopy is widely being used technique to examine the optical properties of nanosized particles<sup>29</sup>. Generally, the optical properties of synthesized ZnO nanoparticles are measured by UV-vis spectrometry that shows an intense absorption band in the range of 340–390 nm<sup>30</sup>. The UV visible spectrum of the naïve ZnO and lipase immobilized ZnO samples are shown in Figure 7.6. The samples maximally absorbed the UV radiations at 375 nm and almost all the visible spectrum radiations were transmitted by the ZnO nanoparticles. The absorption spectrum of ZnO nano-powder is shown in Figure 7.5. The lipase immobilized ZnO also exhibited a strong absorption band at about 375 nm<sup>31</sup>. In addition, 280 nm of proteins' characteristic absorbance peak were also seen that confirmed the immobilization of lipases on ZnO.

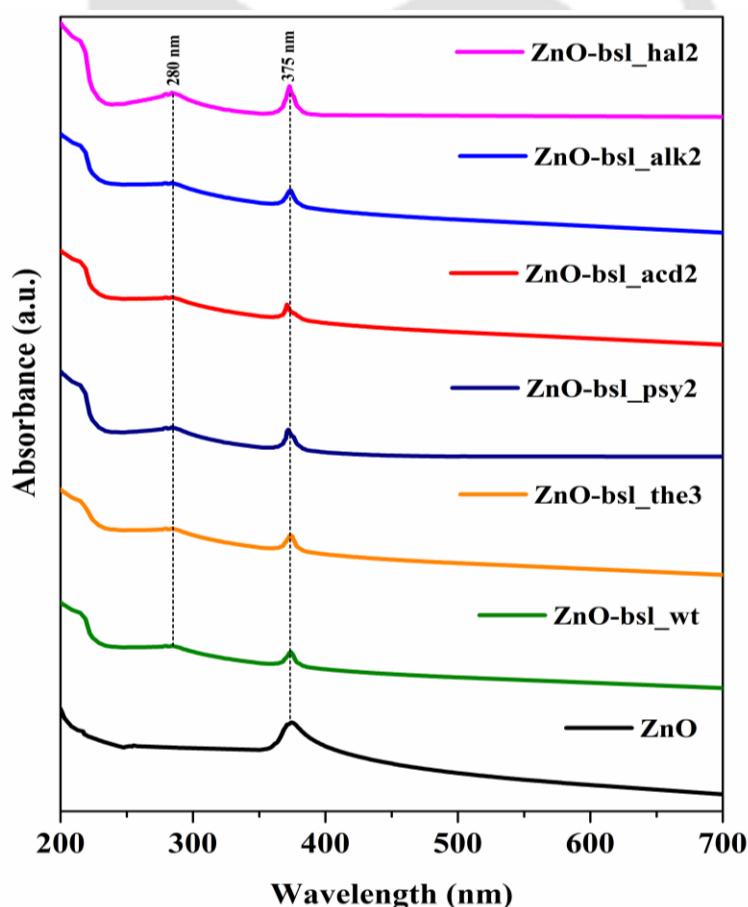


Figure 7.6: UV-Vis spectroscopic analysis of wild-type and mutant lipases.



### 7.3.4 FTIR spectroscopic analysis for characterization of naïve and bioconjugated ZnO nanoparticles

The FTIR spectrum of ZnO, shown in Figure 7.5A, indicates the typical metal oxide (Zn–O) peak beyond  $600\text{ cm}^{-1}$ , suggesting the formation of ZnO material. Spectra were taken in the range from  $1450$  to  $1750\text{ cm}^{-1}$  particularly for the characterization of the amide bond in protein. The peaks lying from  $1500$  to  $1700$ ,  $2300$  to  $2700$  and at  $3000\text{ cm}^{-1}$  represent the functional groups corresponding to C–O symmetric and anti-symmetric stretching mode, C–H stretching mode and O–H stretching mode, respectively. Presence of O–H group represents the presence of water molecules on the surface of ZnO nanoparticles. The ZnO-PEI shows strong twin peaks at  $1636$  and  $1044\text{ cm}^{-1}$  and a broad peak in the  $3200 - 3450\text{ cm}^{-1}$  region. The strong signature at  $1636\text{ cm}^{-1}$  is assigned to the N–H bending mode of the amine group overlapped with the C–H bending mode of the methylene ( $-\text{CH}_2$ ) group. The peak at  $1044\text{ cm}^{-1}$  is due to the C–H bending vibration of the methyl ( $-\text{CH}_3$ ) group. The symmetric and asymmetric stretching modes of N–H appear at higher frequency located around  $3389$  and  $3250\text{ cm}^{-1}$ . The peaks in the  $2830 - 3000\text{ cm}^{-1}$  ( $2926\text{ cm}^{-1}$ ) region are attributed to the symmetric and asymmetric C–H stretching modes of the  $-\text{CH}_2$  and  $-\text{CH}_3$  groups. The C–N stretch appears in the  $1000 - 1250\text{ cm}^{-1}$  region; the peak at  $645\text{ cm}^{-1}$  is attributed to the open long chain ( $-\text{CH}_2$ )<sub>n</sub> bending. The amide bond linking the PEI generally comes in the near region of  $1600\text{ cm}^{-1}$ . In the present spectrum, it appears to be overlapped with the N–H bending mode as the number of amines is comparatively higher in the compound. The most sensitive spectral region to the protein secondary structural components is the amide I band ( $1700 - 1600\text{ cm}^{-1}$ ), which is due almost entirely to the C=O stretch vibrations of the peptide linkages. These additional peaks are specific to nanoparticle coating material and immobilized proteins that confirms the coating and immobilization of lipases (Figure 7.7). In  $\text{H}_2\text{O}$  solution, the bands between  $1654$  and  $1658\text{ cm}^{-1}$  are assigned to  $\alpha$ -helix. Water absorbs strongly in the most important spectral region at approximately  $1640\text{ cm}^{-1}$ . The spectrum related to glutaraldehyde-activated aminopropyl ZnO nanoparticle having the peak in the region  $1036\text{ cm}^{-1}$  could be assigned to the stretching vibration of Si–O<sup>17</sup>.

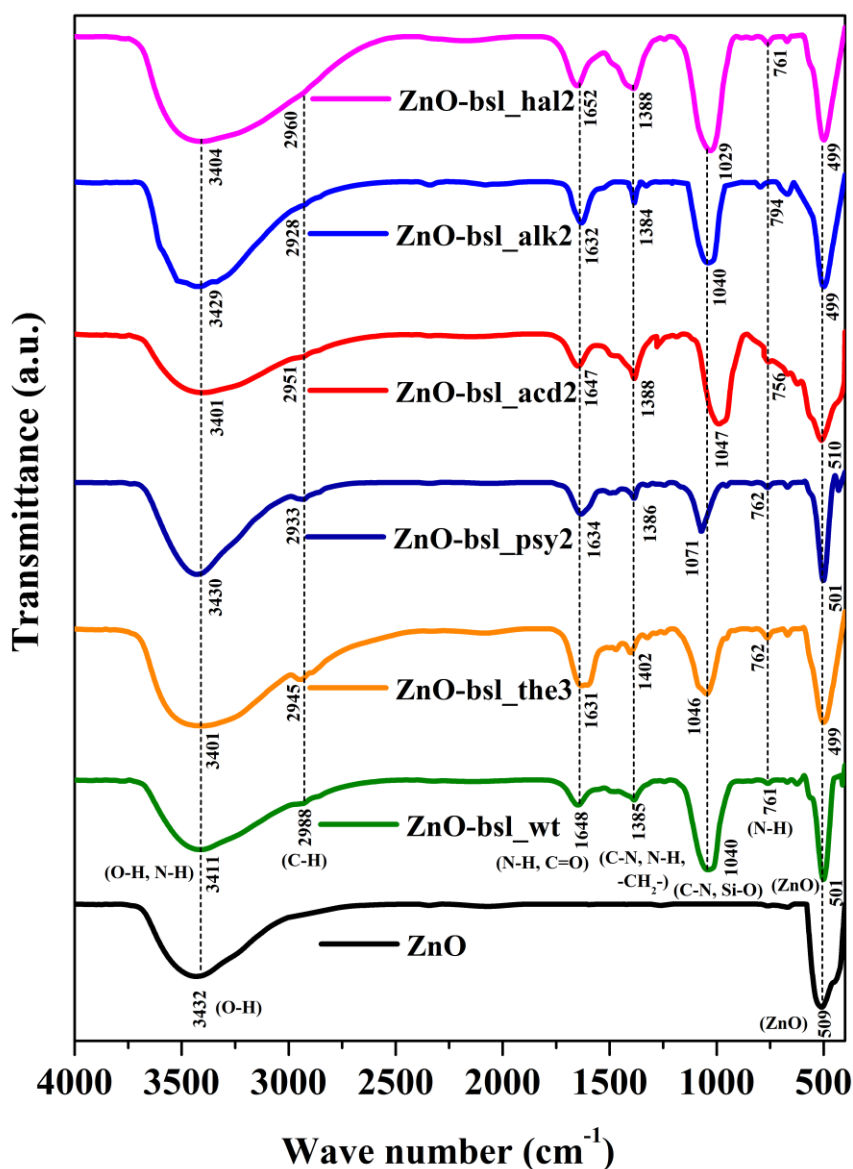


Figure 7.7: FT-IR spectroscopic analysis of wild-type and mutant lipases.

### 7.3.5 Raman spectroscopic analysis for characterization of naïve and bioconjugated ZnO nanoparticles

In order to investigate how surface coating and protein (enzyme) immobilization influence the structural properties of ZnO NPs, Raman spectroscopy was employed to observe the vibrational modes of ZnO NPs. Raman spectroscopy is a nondestructive characterization method for many recent studies on the vibrational properties of ZnO nanoparticles. The changes (shift and/or peak broadening) in Raman spectra can arise due to structural

disorder, defects (oxygen deficiency, surface impurities), and phonon confinement. Raman spectroscopy has been extensively used in protein structural characterization, especially for determining changes in the secondary and tertiary structure of proteins in different matrices.

In the present study, Raman spectrum was recorded with a green laser of 514 nm (from an Ar<sup>+</sup> laser source). This laser is good for inorganic materials and resonance Raman experiments. ZnO nanoparticles showed two characteristic peak at 338 and 440 cm<sup>-1</sup> (Figure 7.8). Notably from the Raman spectra samples, naïve ZnO showed a characteristic Raman vibration mode of the dominant E<sub>2</sub> (high) of Wurtzite ZnO at ~440 cm<sup>-1</sup> commonly observed for bulk ZnO samples at 437 cm<sup>-1</sup> which was slightly blue shifted to higher frequencies for the lipase immobilized ZnO samples as illustrated here<sup>32</sup>. This blue shift in this Raman spectra might be due to particle size increase (Naïve ZnO=157 nm; lipase immobilized ZnO=183-253 nm). Nevertheless, the blue shift is not a significant one. The other peak at ~338 cm<sup>-1</sup> is the second-order Raman spectrum, originating from the zone boundary phonons 3E<sub>2H</sub>-E<sub>2L</sub><sup>33</sup>. The blue shift is not considerable here as well and can be due to insufficient optical phonon confinement by nanostructures, laser induced heating and/or the phonon localization by defects or impurities in the nanostructures or due to anisotropic internal strains corresponding to different growth directions.

Several vibrational modes of Raman spectroscopy can be used for the analysis of protein structure. Most characteristic bands are associated with the -CONH group, amide I to VII (I: 1621 cm<sup>-1</sup>, II: 1492.5 cm<sup>-1</sup>, III: 1289.6 cm<sup>-1</sup>, IV: 958.4 cm<sup>-1</sup>, V: 642.5 and 648.7 cm<sup>-1</sup>). The immobilized lipase onto ZnO NPs is clearly established by analyzing amide I (stretching vibration of carbonyl, C=O of the peptide group) and amide III bands (both associated with coupled C–N stretching and N–H bending vibrations of the peptide group) shown in Figure 7.8. Amide I and amide III bands are expected in the spectral intervals of 1640 to 1680 cm<sup>-1</sup> and 1230 to 1310 cm<sup>-1</sup>, respectively. Both the bands are the main signal for the identification of protein and the band shapes and peak positions are sensitive to the protein secondary structure. Accordingly, these Raman amide bands are often used as indicators of protein secondary structure<sup>34,35</sup>. These were not present in naïve ZnO NPs. Nevertheless, the additional peaks were further specific to nanoparticle coating material

and immobilized proteins that confirms the coating and immobilization of lipases. However, the band maxima shift of amide I-III bands *w.r.t.* wild type and mutant lipases and its correlation with  $\alpha$ -helix content, effect of unstructured folding and  $\beta$ -sheet structures, and microenvironment of aromatic amino acid residues warrants further study.

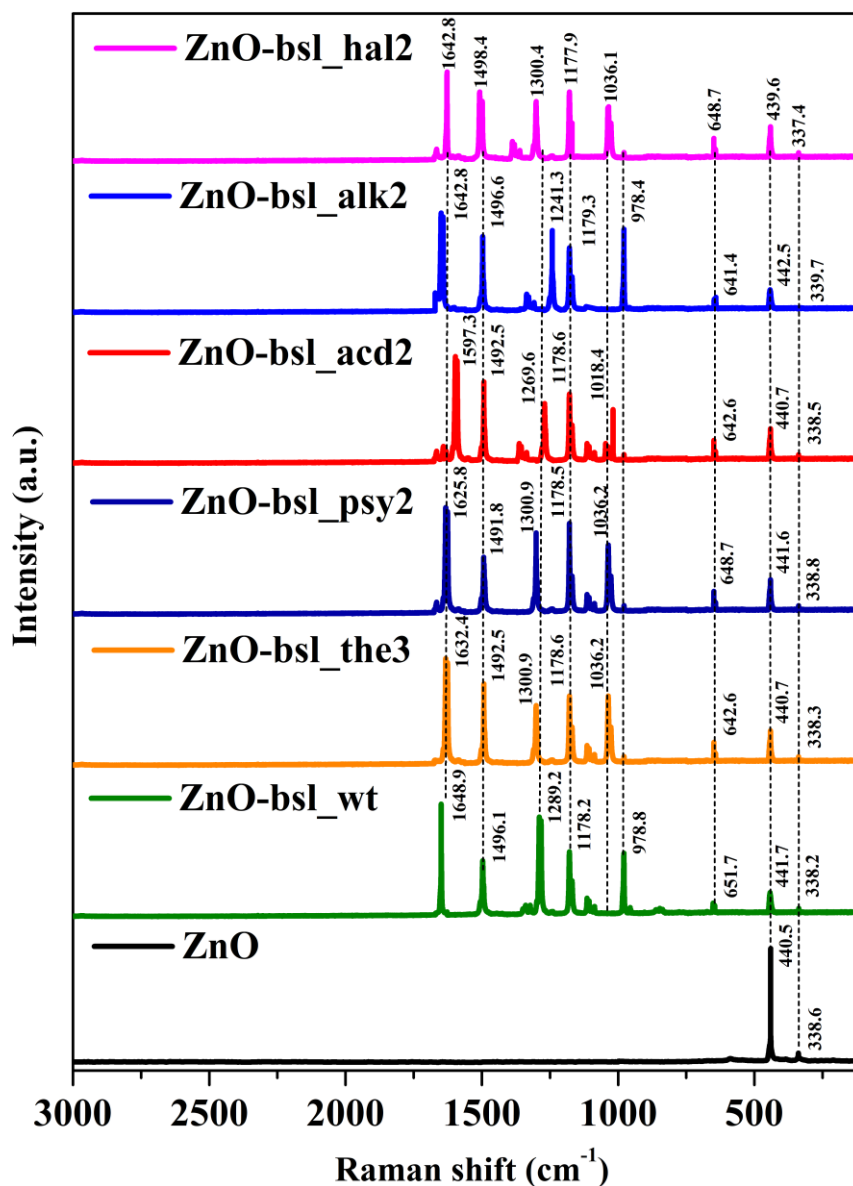


Figure 7.8: Raman spectroscopic analysis of wild-type and mutant lipases.

### 7.3.6 XRD analysis for characterization of naïve and bioconjugated ZnO nanoparticles

According to the spectrogram of the crystal structure of naïve ZnO, the well-defined peaks typical of ZnO in the crystal structure are clearly noticed. Results of XRD analysis of naïve ZnO showed the diffraction peaks located at  $31.84^\circ$ ,  $34.52^\circ$ ,  $36.33^\circ$ ,  $47.63^\circ$ ,  $56.71^\circ$ ,  $62.96^\circ$ ,  $68.13^\circ$ , and  $69.18^\circ$  have been keenly indexed as hexagonal wurtzite phase of ZnO (Figure 7.9). This is in compliance with the reports of the norms of the joint committee on powder diffraction standard ICDD card number 98-010-8249<sup>36</sup>. X-ray diffraction pattern for the ZnO nanoparticles was recorded using an X-ray diffractometer using Cu K $\alpha$  radiation of wavelength  $\lambda=0.1541$  nm in the scan range  $2\theta=20-90^\circ$ .

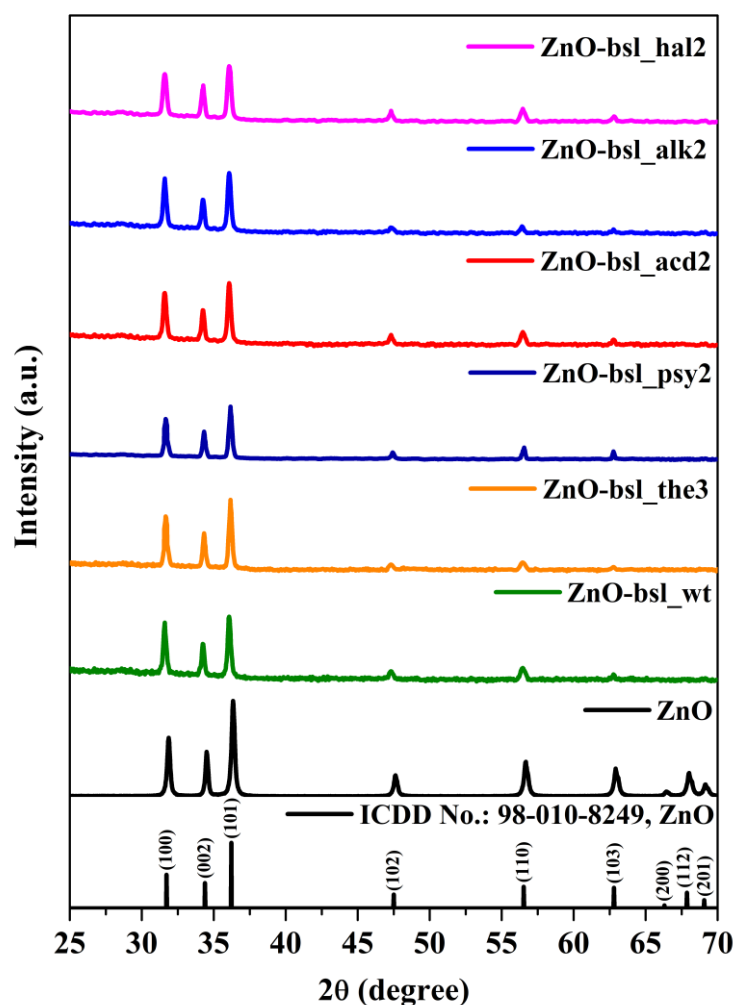


Figure 7.9: XRD spectrogram of wild-type and mutant lipases.

Interestingly, in case of lipase immobilized ZnO nanoparticles showed peaks at same located but their intensity is significantly decrease which is due loss of crystalline structure of ZnO nanoparticles after immobilization of organic coating materials and proteins. Also, The XRD pattern for naïve ZnO nanoparticles obtained shows much sharper peaks. This means that the crystallinity is more ZnO nano-particles than bioconjugated ZnO nanoparticle (Figure 7.9).

### 7.3.7 Biochemical characterization of immobilized lipase: Immobilization efficiency and yield

The amounts of total attached wild-type and mutant lipases were estimated by the Bradford protein assay method<sup>19</sup>, and activity assay was carried out to calculate the activity retention of the immobilized lipase<sup>37</sup>. Various factors can affect lipase immobilization efficiency as well as the activity of bound enzyme<sup>38</sup>. The kinetic parameters, immobilization efficiency and yield of enzyme were determined after immobilization of wild-type and mutant lipases and are presented in Table 7.3. The results revealed that ZnO-bsl\_the3 and ZnO-bsl\_alk2 lipases mutants showed most efficient immobilization of lipase out of other lipase i.e. 88.00% and 88.50%, respectively. The other mutant such as ZnO-bsl\_wt, ZnO-bsl\_psy2, ZnO-bsl\_acd2 and ZnO-bsl\_hal2 showed immobilization efficiency of 82.50%, 82.00%, 82.60% and 82.95%, respectively. The immobilized lipase activity was determined in terms of immobilization yield. The result of immobilization yield showed that almost all the immobilized lipase mutants gave activity more or less 80% of immobilized enzymes. The lowest yield was obtained in case of immobilized wild-type lipase (ZnO-bsl\_wt) i.e. 71.79%. The possible reason for decrease in the activity after immobilization is due to covalent binding of enzyme between the aldehyde group (provided by glutaraldehyde) of nanoparticles and the amino group of lipases<sup>39</sup>.

To investigate thermal, pH and halostability of immobilized lipases, the enzyme activity has been determined at different temperature, pH and NaCl concentration on enzyme activity and the results are summarized in Table 7.4. The analysis revealed that the improved activity of ZnO-bsl\_alk2 i.e. maximally at 55 °C, pH 10.5 and 1.5 M NaCl, it was followed by ZnO-bsl\_the3 i.e. at 55 °C, pH 9.5 and 1.5 M NaCl. The maximum



halostability was obtained in ZnO-bsl\_hal2 *i.e.* at 2.0 M NaCl and maximum psychrostability was obtained in ZnO-bsl\_psy2 *i.e.* at 20 °C. The main reason for improving thermal, pH and halostability was that the covalent bond formation between aldehyde and amino groups of immobilized lipase might also reduce conformational flexibility that leads to higher activation energy for the molecule to reorganize proper conformation of substrate binding site to the substrate<sup>40</sup>.

Table 7.3: Enumeration of immobilization parameters and formula for immobilization efficiency and yield.

| Lipase       | Lipase used for immobilization (mg/ml) = E0 | Lipase in filtrate after immobilization (mg/ml) = E1 | % Efficiency | Specific activity of lipase (U/mg) | Activity of immobilized lipase (U/ml) = Y1 | Activity of free lipase (U/ml) = Y0 | % Yield |
|--------------|---------------------------------------------|------------------------------------------------------|--------------|------------------------------------|--------------------------------------------|-------------------------------------|---------|
| ZnO-bsl_wt   | 20                                          | 3.5                                                  | 82.50        | 1381                               | 113.93                                     | 158.3                               | 71.97   |
| ZnO-bsl_the3 | 20                                          | 2.3                                                  | 88.50        | 2151                               | 190.36                                     | 239.6                               | 79.45   |
| ZnO-bsl_psy2 | 20                                          | 3.6                                                  | 82.00        | 1660                               | 136.12                                     | 166.8                               | 81.61   |
| ZnO-bsl_acd2 | 20                                          | 3.48                                                 | 82.60        | 1778                               | 146.86                                     | 165.9                               | 88.52   |
| ZnO-bsl_alk2 | 20                                          | 2.4                                                  | 88.00        | 2162                               | 190.26                                     | 229.7                               | 82.83   |
| ZnO-bsl_hal2 | 20                                          | 3.41                                                 | 82.95        | 1509                               | 125.17                                     | 147.3                               | 84.98   |

Table 7.4: Biochemical characterization of lipase after and before immobilization

| Lipase       | Optimum temperature (°C) |                    | Optimum pH  |                    | Optimum NaCl conc. (M) |                    |
|--------------|--------------------------|--------------------|-------------|--------------------|------------------------|--------------------|
|              | Free lipase              | Immobilized lipase | Free lipase | Immobilized lipase | Free lipase            | Immobilized lipase |
| ZnO-bsl_wt   | 35                       | 35                 | 8.0         | 8.5                | Halo-sensitive         | 0.5                |
| ZnO-bsl_the3 | 55                       | 55                 | 9.0         | 9.5                | 1.0                    | 1.5                |
| ZnO-bsl_psy2 | 25                       | 20                 | 8.0         | 9.0                | 0.5                    | 1.0                |
| ZnO-bsl_acd2 | 40                       | 45                 | 6.5         | 6.5                | Halo-sensitive         | 0.5                |
| ZnO-bsl_alk2 | 50                       | 55                 | 10.0        | 10.5               | 1.0                    | 1.5                |
| ZnO-bsl_hal2 | 45                       | 45                 | 9.0         | 9.0                | 1.5                    | 2.0                |

### 7.3.8 Reusability and storage stability of immobilized lipases

The major limitations of enzymes in industrial process are that reusing them in continuous reactions. The immobilized wild-type and extreme-stable lipase onto functionalized ZnO-PEI/APTES can be easily separated from the reaction mixture with the help of centrifugation and consequently can be reused for hydrolysis of oils, fats and lipids. The reusability of immobilized wild-type and mutant lipases (such as ZnO-bsl\_wt, ZnO-bsl\_the3, ZnO-bsl\_psy2, ZnO-bsl\_acd2, ZnO-bsl\_alk2 and ZnO-bsl\_hal2) were assayed over 20 reaction cycles at their optimum pH and temperature. The relative activity of the

first run was set to 100% and the results are presented in Figure 7.10. As it shown in this figure, the ZnO-bsl\_wt could still have a high level of about 38% of its initial activity after 20 reaction cycles whereas, 80 % for ZnO-bsl\_the3, 72 % for ZnO-bsl\_psy2, 60 % for ZnO-bsl\_acd2, 81 % for ZnO-bsl\_alk2 and 62 % for ZnO-bsl\_hal2 activity retained after upto 20 cycles in the reaction mixture under the optimum reaction conditions. The possible reason for decreases in enzymatic activity is the repeated use of substrate to the active site of immobilized lipases that weakens the binding strength between the ZnO matrix and the immobilized lipases causing leaching of enzyme in recycling experiments and consequently loss of activity. Furthermore, the recurrent encountering of substrate with the active site of immobilized enzyme causes its distortion and leads to loss of activity.

For determination of storage stability, the free and immobilized lipases (in 20 mM phosphate buffer, pH 8.0) were stored at room temperature and their activities were evaluated for a period of 60 days. The catalytic activity was examined at every 4 days interval. The results revealed that the storage stability was improved due to immobilization on ZnO nanoparticle (Figure 7.10). This figure showed that retained activity of immobilized wild-type lipase i.e. ZnO-bsl\_wt was obtained around 63 % whereas, for immobilized mutant lipases such as ZnO-bsl\_the3 of 82%, ZnO-bsl\_psy2 of 88%, ZnO-bsl\_acd2 of 71%, ZnO-bsl\_alk2 of 87% and ZnO-bsl\_hal2 of 72% of its initial activity after 60 days of experiment. On the other hand, the free wild-type and mutant lipases could retain less than 50% of the initial activity at the same condition. According to these results, the activity of lipase usually declined after a long time. Thus, immobilization significantly prevented enzyme deactivation and subsequently improved the storage stability as well as reusability of the enzymes<sup>41</sup>.

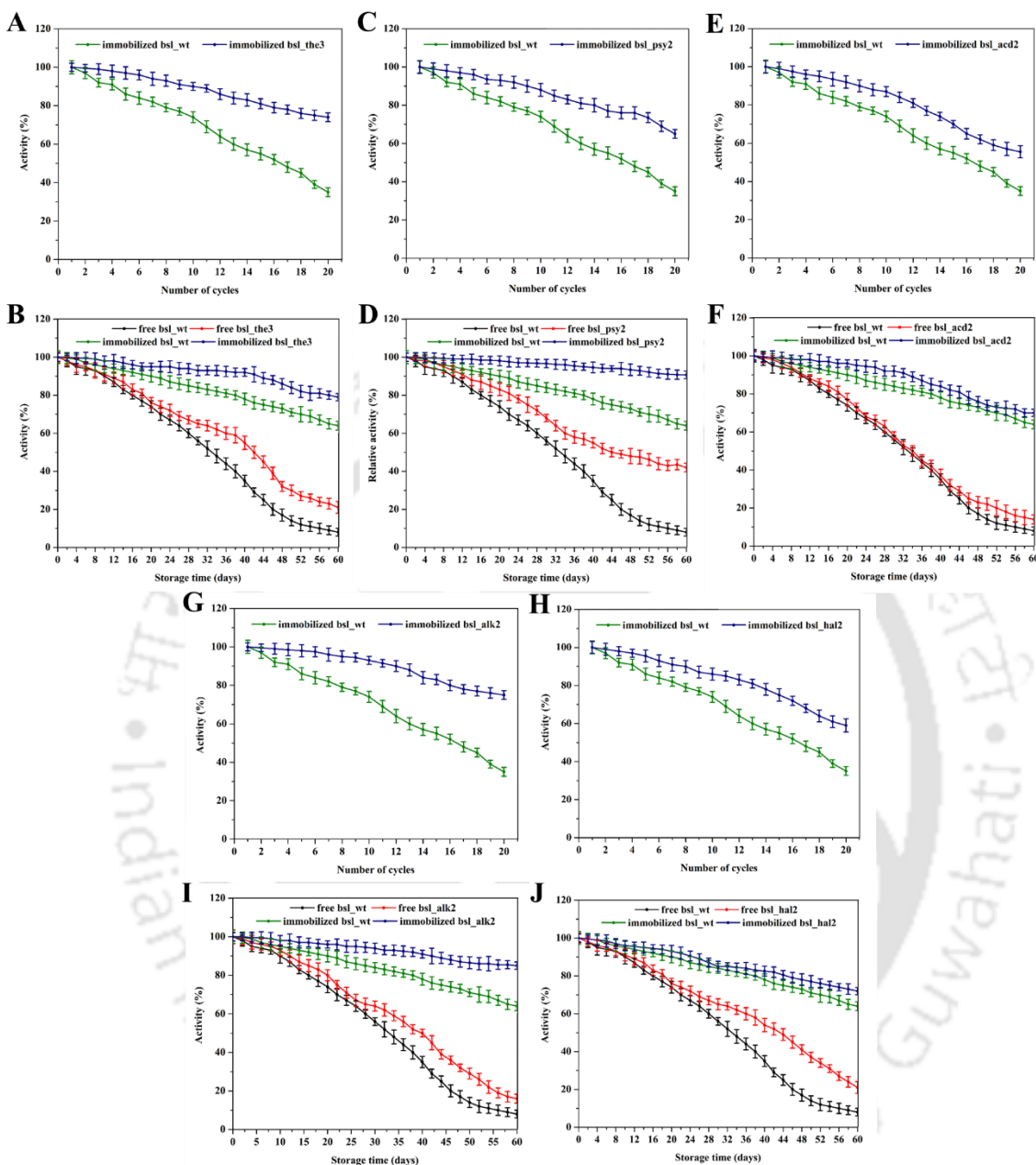


Figure 7.10: Comparative analysis of reusability and storage stability of immobilized wild-type lipases and (A and B) thermostable mutant lipase, ZnO-bsl\_the3; (C and D) psychrostable mutant lipase, ZnO-bsl\_psy2 (E and F) acidostable mutant lipase, ZnO-bsl\_acd2; (G and H) alkalistable mutant, Zn-bsl\_alk2; and (I and J) halostable mutant lipase, ZnO-bsl\_hal2, respectively. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).

### 7.3.9 Lipase compatibility with detergents and oxidizing agents after immobilization

All the selected immobilized lipases were again examined for their compatibility with same surfactants, oxidizing agents and commercial detergents. The activity of immobilized wild-type lipase (ZnO-bsl\_wt) tested without any detergent components was considered as 100 % and then the relative activity in presence of detergent components were enumerated for both type of immobilized wild-type and mutant lipases. Figure 7.11 is the graphical representation of comparative analysis of effect of various surfactant, detergents and oxidizing agents on lipase after immobilization. It clearly depicted that out of all mutant lipases, the ZnO-bsl\_the3 and ZnO-bsl\_alk2 gave best results and are represented by orange and red bars respectively.

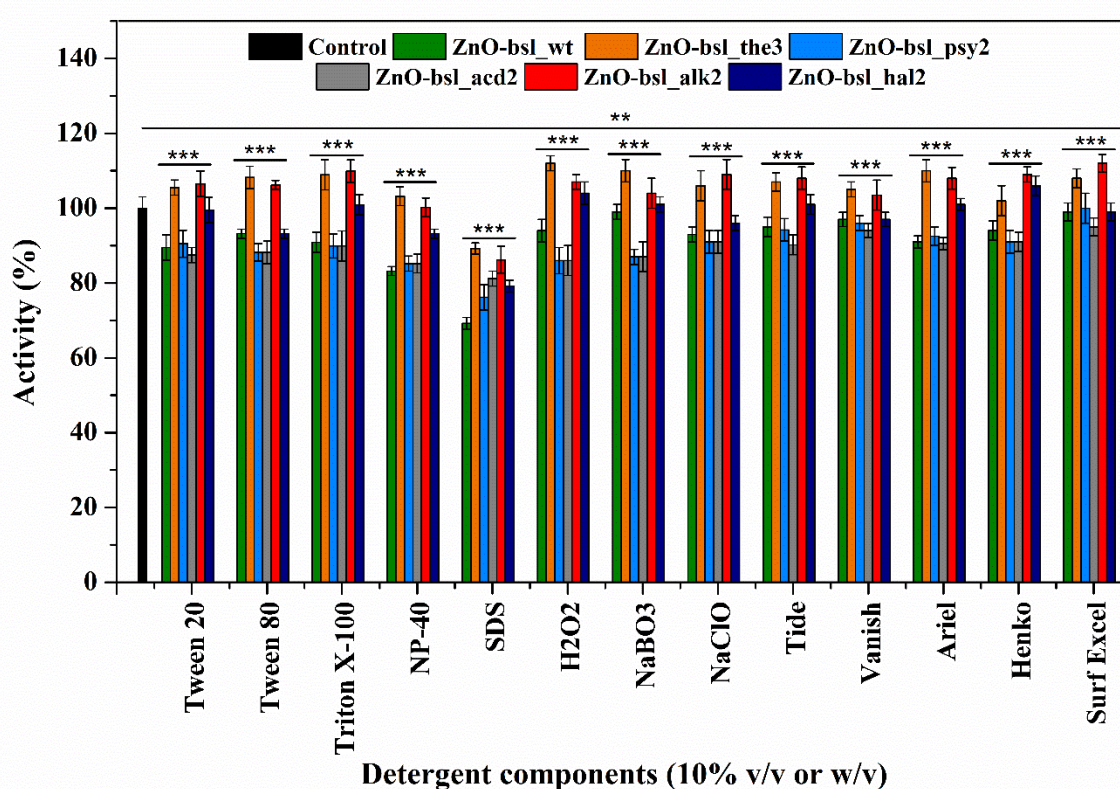


Figure 7.11: Effect of various surfactant, detergents and oxidizing agents on lipase after immobilization. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars); \*\* denotes  $p$ -value  $< 0.05$  and \*\*\* denotes  $p$ -value  $< 0.005$ .

The comparative analysis showed that after immobilization, ZnO-bsl\_the3 is most compatible with Ariel, ZnO-bsl\_psy2 is with Surf Excel, ZnO-bsl\_acd2 is with Vanish, ZnO-bsl\_alk2 is with Surf Excel and ZnO-bsl\_hal2 is with Henko. Furthermore, these most

compatible immobilized mutants of lipases with detergents were exploited for oil and grease destaining from cotton fabrics.

### 7.3.10 Destaining of oil and grease stains from cotton fabrics

Optimization results of detergent concentration, lipase concentration, washing temperature and time for oil and grease removal of thermostable mutant (ZnO-bsl\_the3), psychrostable mutant (ZnO-bsl\_psy2), acidostable mutant (ZnO-bsl\_acd2), alkalistable mutant (ZnO-bsl\_alk2), and halostable mutant (ZnO-bsl\_hal2) have been presented in Figure 7.12 to 7.21. The photographs and their respective FESEM micrographs of stained and destained fabric by ZnO-bsl\_the3, ZnO-bsl\_psy2, ZnO-bsl\_acd2, ZnO-bsl\_alk2 and ZnO-bsl\_hal2 have been presented in the same figures.

#### *Performance of immobilized thermostable lipase mutant on destaining*

The optimization of immobilized thermostable lipase mutant (ZnO-bsl\_the3) concentration was obtained at 50 U ml<sup>-1</sup> for both oil and grease removal. The immobilized mutant showed best detergent compatibility with Ariel at 5% for oil removal and 6% Ariel for grease removal from the fabric. The optimum washing temperature for efficient oil and grease removal showed highest temperature stability at 55 °C with ZnO-bsl\_the3 whereas, 45 °C was suitable for ZnO-bsl\_wt. The optimal washing time of ZnO-bsl\_the3 for oil removal from the fabric was 30 min and grease removal was 40 min. In control, only detergent was less efficient and attained highest oil removal of 21 % and grease removal of 17 % with any lipase. But, the optimized parameters showed efficient and maximum oil removal upto 90 % and grease removal upto 82 % with ZnO-bsl\_the3 lipase. In case of ZnO-bsl\_wt, the maximum oil removal was obtained only upto 62 % and grease removal upto 57% using same optimized conditions. The results of oil and grease removal from the fabric for comparative analysis of ZnO-bsl\_the3 and ZnO-bsl\_wt have been presented in Figure 7.12 and 7.13, respectively.



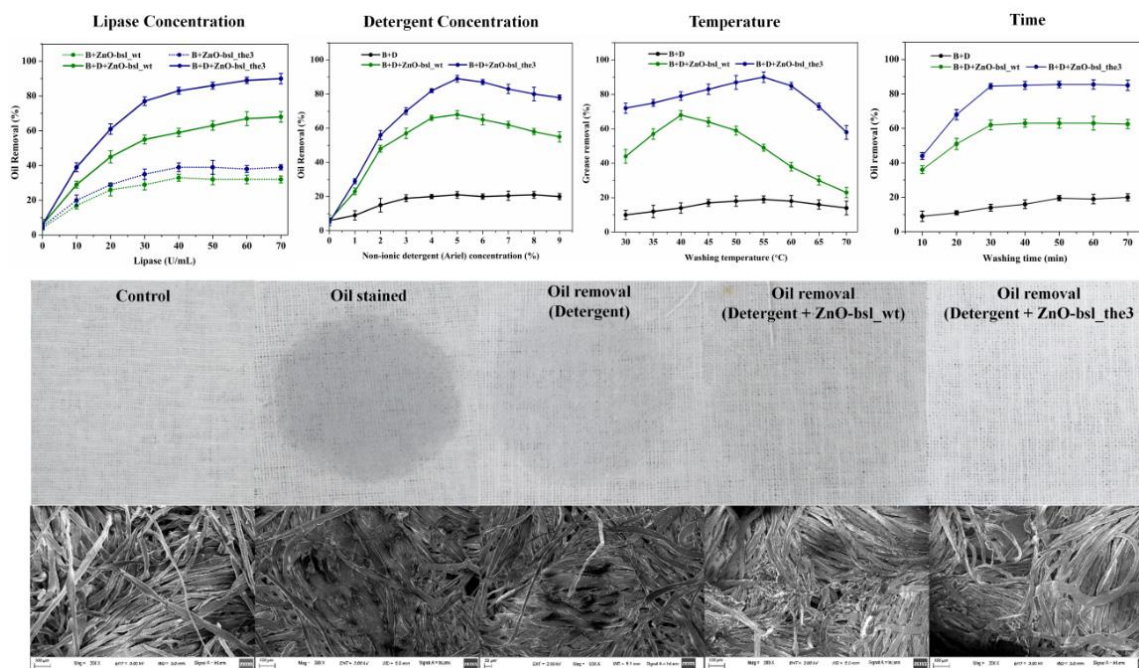


Figure 7.12: Optimization of oil removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and thermostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).

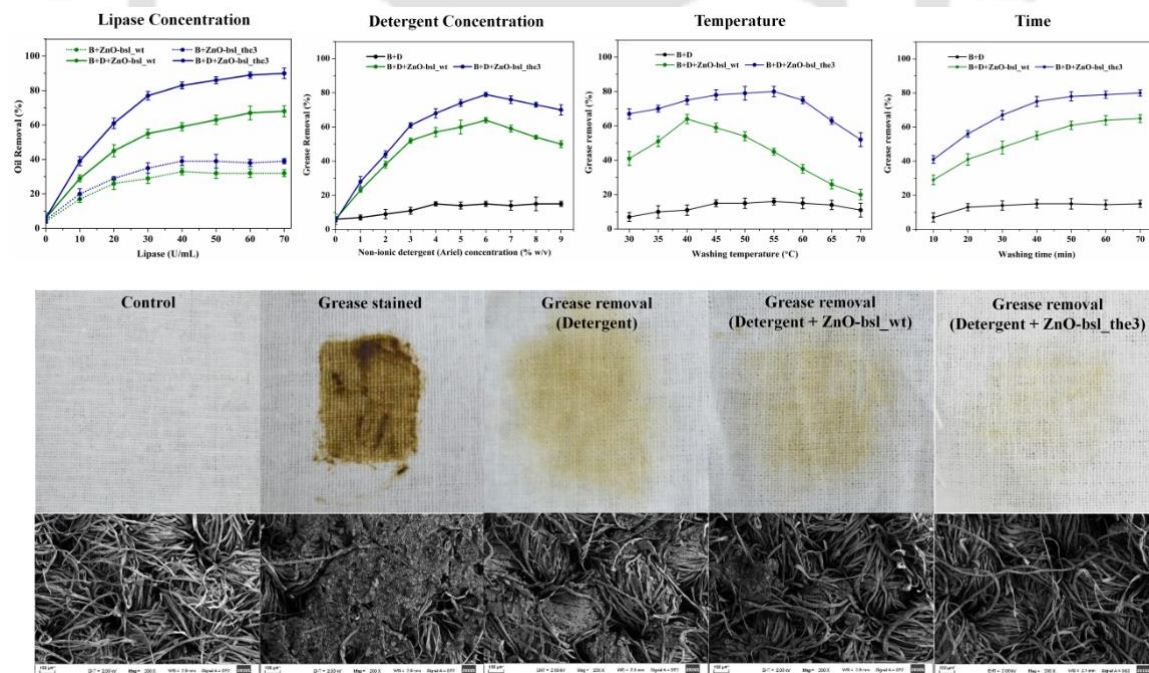


Figure 7.13: Optimization of grease removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and thermostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).



### Performance of immobilized psychrostable lipase mutant on destaining

The optimization of immobilized psychrostable lipase mutant (ZnO-bsl\_psy2) concentration for oil removal from the fabric was obtained at 40 U ml<sup>-1</sup> and for grease removal was obtained at 50 U ml<sup>-1</sup>. The immobilized mutant showed best detergent compatibility with Surf Excel at 5 % for oil removal from the fabric and 6 % Surf Excel for grease removal from the fabric. ZnO-bsl\_psy2 showed optimum washing temperature for efficient and maximum oil and grease removal was obtained at 25 °C. The optimal washing time of ZnO-bsl\_psy2 for oil removal from the fabric was 30 min and grease removal was 40 min. In control, only detergent was less efficient and attained highest oil removal of 21 % and grease removal of 17 % with any lipase. But, the optimized parameters showed efficient and maximum oil removal upto 80 % and grease removal upto 76 % with ZnO-bsl\_psy2 lipase. In case of ZnO-bsl\_wt, the maximum oil removal was obtained only upto 62 % and grease removal upto 57 % using same optimized conditions. The results of optimization of oil and grease removal from the fabric for comparative analysis of ZnO-bsl\_psy2 and ZnO-bsl\_wt have been presented in Figure 7.14 and 7.15.

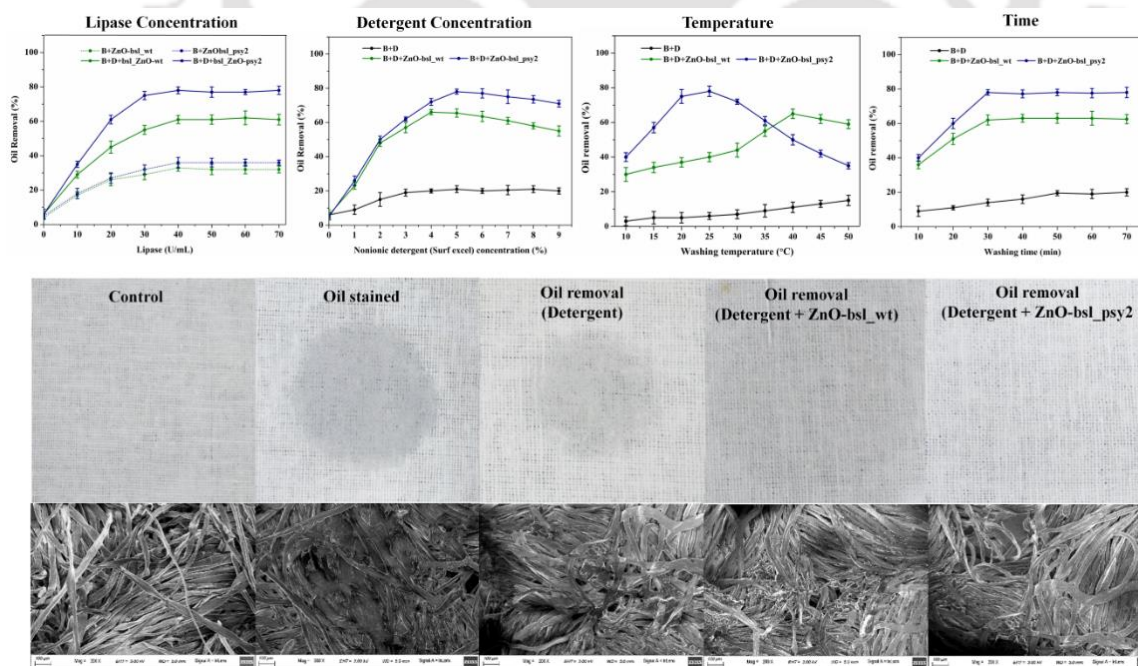


Figure 7.14: Optimization of oil removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and psychrostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).

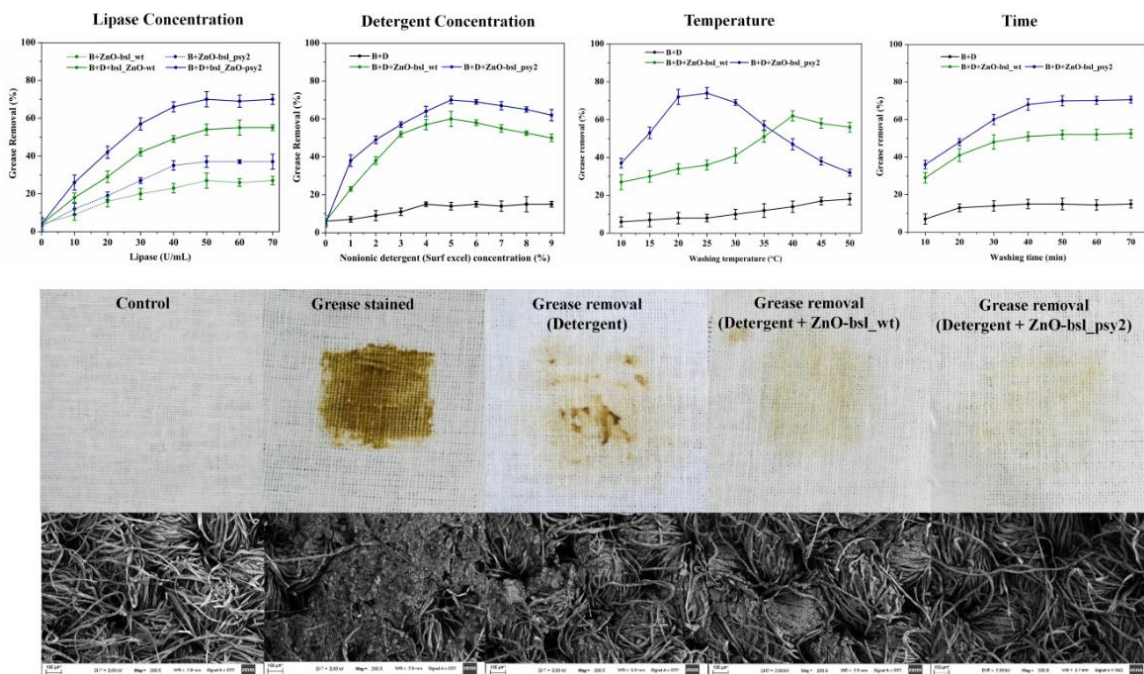


Figure 7.15: Optimization of grease removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and psychrostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).

### Performance of immobilized acidostable lipase mutant on destaining

The optimization of immobilized acidostable lipase mutant (ZnO-bsl\_acd2) concentration for oil removal from the fabric was obtained at  $40 \text{ U ml}^{-1}$  and for grease removal was obtained at  $50 \text{ U ml}^{-1}$ . The immobilized mutant showed best detergent compatibility with Vanish at 5 % for oil removal from the fabric and 4 % Vanish for grease removal from the fabric. ZnO-bsl\_acd2 showed optimum washing temperature for efficient and maximum oil and grease removal was obtained at  $40^\circ\text{C}$ . The optimal washing time of ZnO-bsl\_acd2 for oil removal from the fabric was 30 min and grease removal was 50 min. In control, only detergent was less efficient and attained highest oil removal of 21 % and grease removal of 17 % with any lipase. But, the optimized parameters showed efficient and maximum oil removal upto 68 % and grease removal upto 66 % with ZnO-bsl\_acd2 lipase. In case of ZnO-bsl\_wt, the maximum oil removal was obtained only upto 62 % and grease removal upto 57 % using same optimized conditions. The results of optimization of oil and grease removal from the fabric for comparative analysis of ZnO-bsl\_acd2 and ZnO-bsl\_wt have been presented in Figure 7.16 and 7.17, respectively.



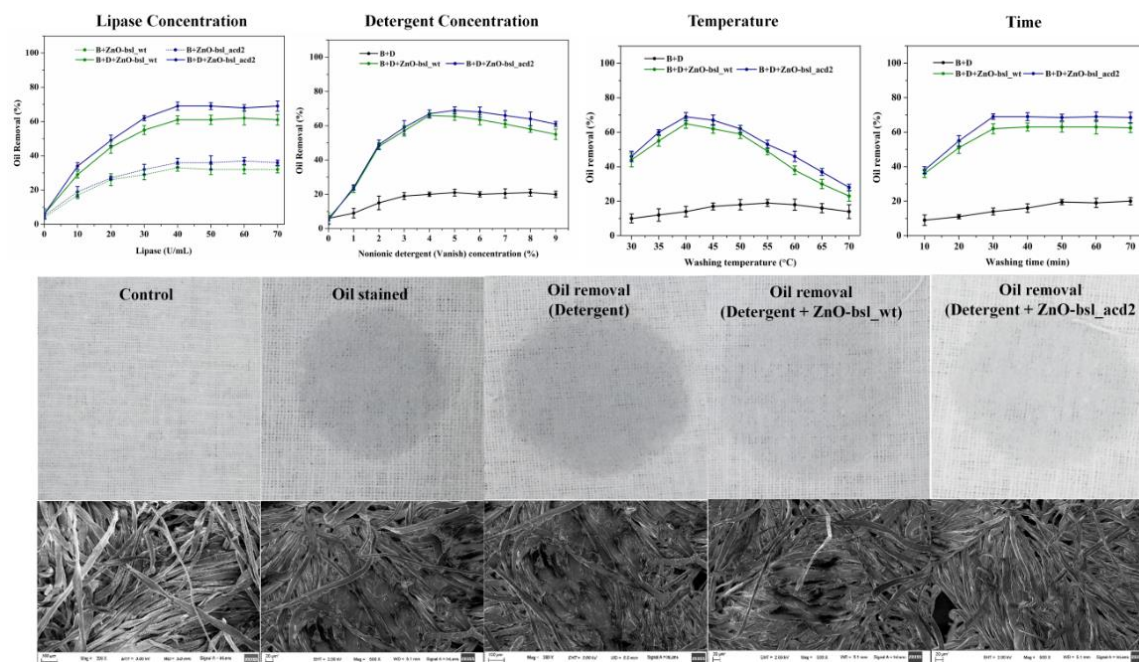


Figure 7.16: Optimization of oil removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and acidostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).

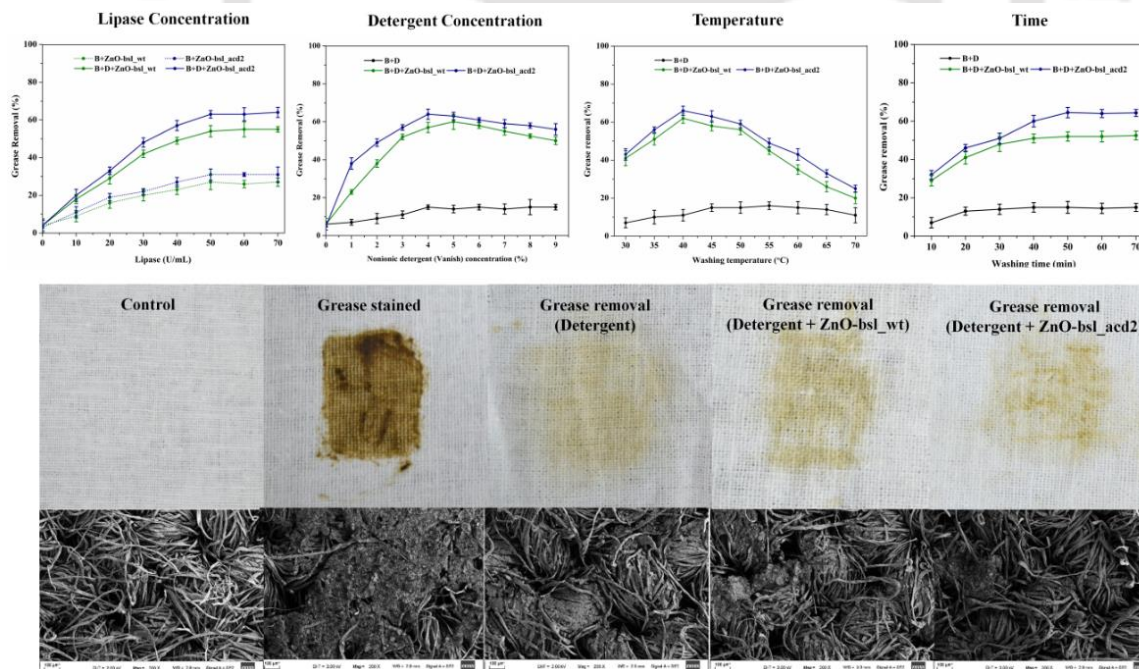


Figure 7.17: Optimization of grease removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and acidostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).

### Performance of immobilized alkalistable lipase mutant on destaining

The optimization of immobilized alkalistable lipase mutant (ZnO-bsl\_alk2) concentration for oil removal from the fabric was obtained at 40 U ml<sup>-1</sup> and for grease removal was obtained at 50 U ml<sup>-1</sup>. The immobilized mutant showed best detergent compatibility with Surf Excel at 5% for oil and grease removal from the fabric. ZnO-bsl\_alk2 showed optimum washing temperature for efficient and maximum oil and grease removal was obtained at 50 °C. The optimal washing time of ZnO-bsl\_alk2 for oil removal from the fabric was 30 min and grease removal was 50 min. In control, only detergent was less efficient and attained highest oil removal of 21 % and grease removal of 17 % with any lipase. But, the optimized parameters showed efficient and maximum oil removal upto 90 % and grease removal upto 82 % with ZnO-bsl\_alk2 lipase. In case of ZnO-bsl\_wt, the maximum oil removal was obtained only upto 62 % and grease removal upto 57 % using same optimized conditions. The results of optimization of oil and grease removal from the fabric for comparative analysis of ZnO-bsl\_alk2 and ZnO-bsl\_wt have been presented in Figure 7.18 and 7.19, respectively.

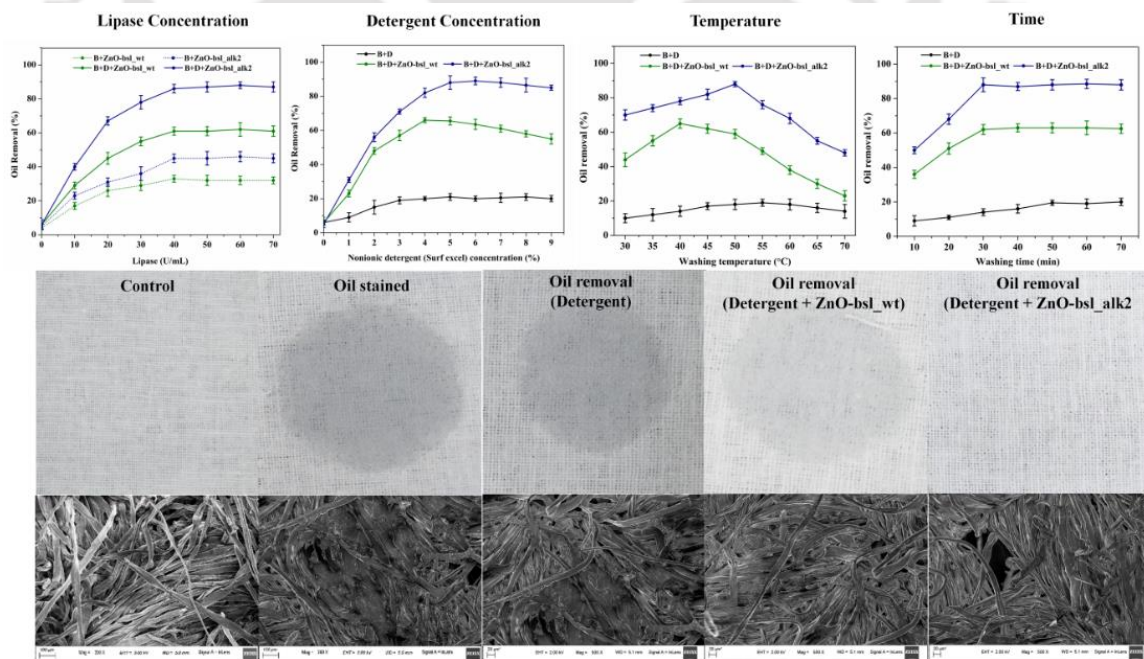


Figure 7.18: Optimization of oil removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and alkalistable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).



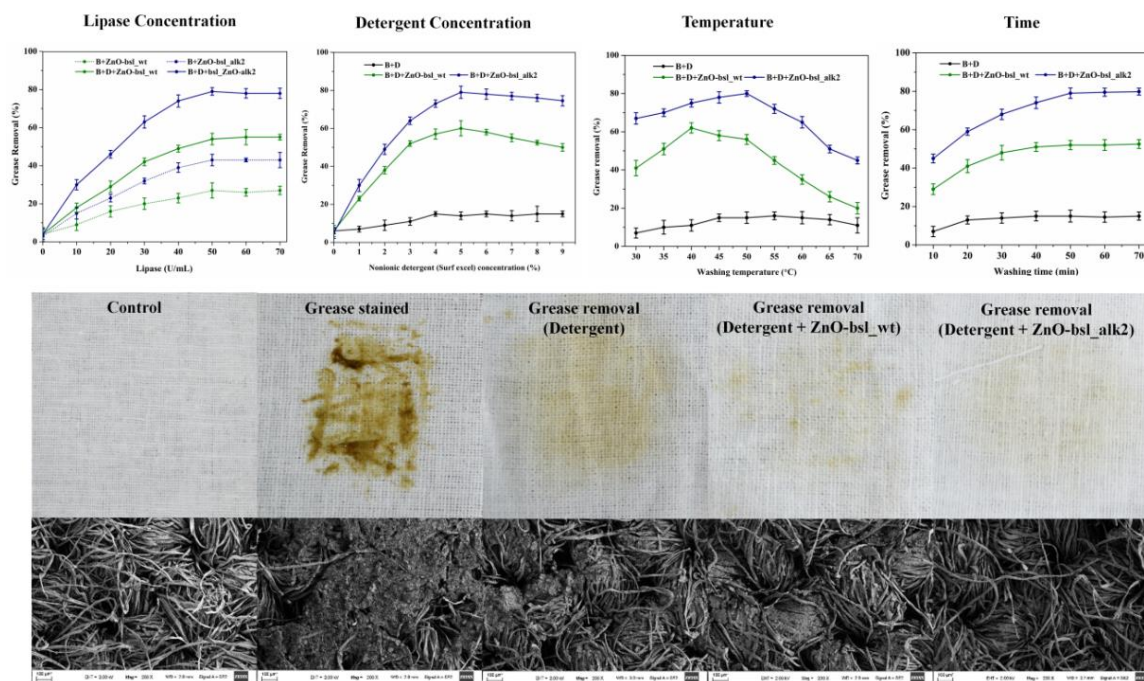


Figure 7.19: Optimization of grease removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and alkalistable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).

### Performance of immobilized halostable lipase mutant on destaining

The optimization of immobilized halostable lipase mutant (ZnO-bsl\_hal2) concentration for oil removal from the fabric was obtained at 40 U ml<sup>-1</sup> and for grease removal was obtained at 50 U ml<sup>-1</sup>. The immobilized mutant showed best detergent compatibility with Surf Excel at 5% for oil removal from the fabric and 4% Surf Excel for grease removal from the fabric. ZnO-bsl\_hal2 showed optimum washing temperature for efficient and maximum oil and grease removal was obtained at 45 °C. The optimal washing time of ZnO-bsl\_hal2 for oil removal from the fabric was 30 min and grease removal was 50 min. In control, only detergent was less efficient and attained highest oil removal of 21 % and grease removal of 17 % with any lipase. But, the optimized parameters showed efficient and maximum oil removal upto 79 % and grease removal upto 75 % with ZnO-bsl\_hal2 lipase. In case of ZnO-bsl\_wt, the maximum oil removal was obtained only upto 62 % and grease removal upto 57 % using same optimized conditions. The results of optimization of oil and grease removal from the fabric for comparative analysis of ZnO-bsl\_hal2 and ZnO-bsl\_wt have been presented in Figure 7.20 and 7.21, respectively.

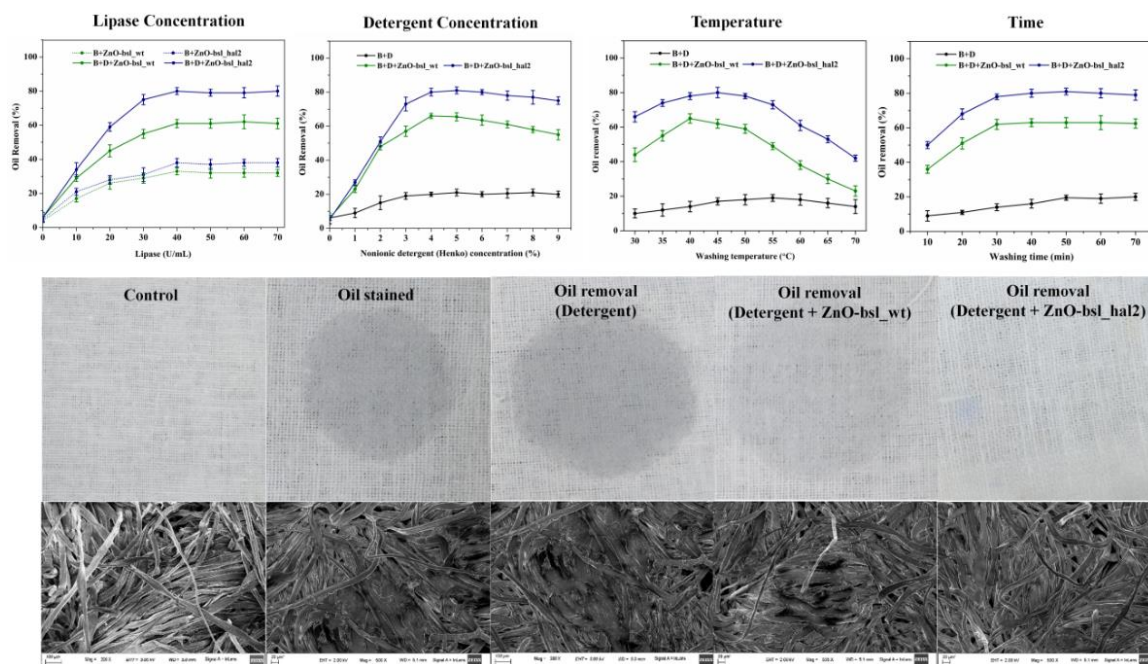


Figure 7.20: Optimization of oil removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and halostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).

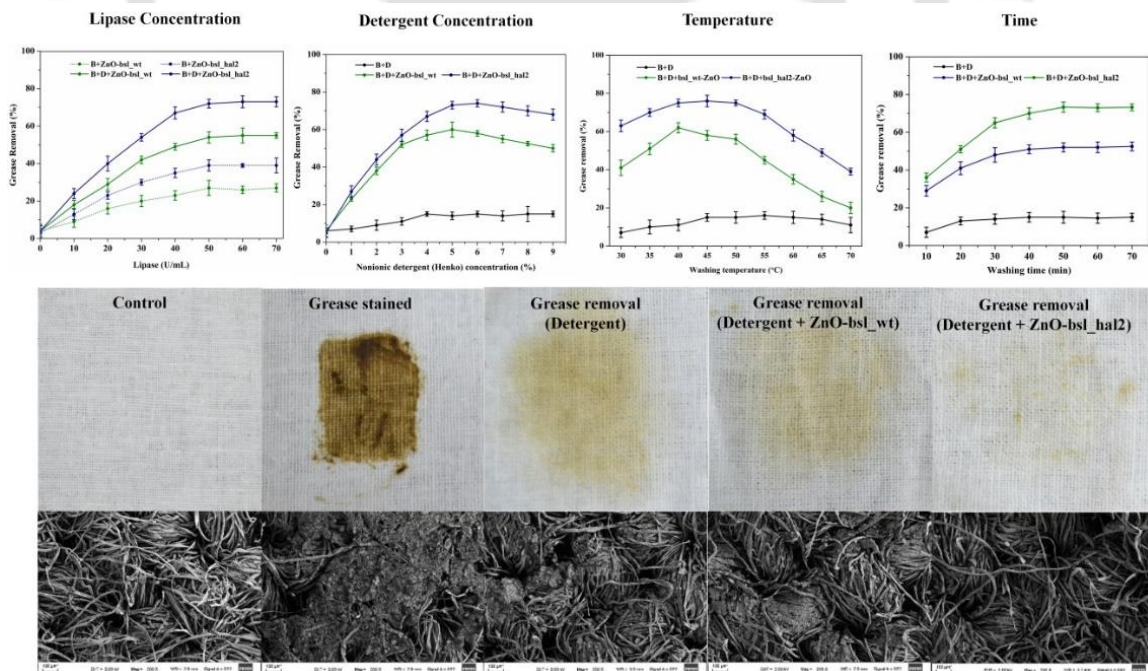


Figure 7.21: Optimization of grease removal for lipase concentrations, detergent concentrations, washing temperature and washing time by wild-type and halostable lipase mutant. Photographs and FESEM micrographs of stained and destained fabric. All experiments were conducted in triplicate and the data have been represented as mean  $\pm$  SD (error bars).



Previously, enhanced removal of oily and greasy stains was found using commercial detergent when it was indigenously added with lipase of *Pseudozyma* sp. NII 08165 or lipase of *Pseudomonas aeruginosa* lipases<sup>42,43</sup>. Similarly, Flipsen et al. (1998) and Sajna et al. (2013) revealed that the lipases as additive in detergent formulation might be more effective than that of detergent formulation with lipases<sup>42,44</sup>. They also revealed that non-ionic detergent performed better than anionic detergents since non-ionic detergent did not inhibit the enzyme activity<sup>42,45</sup>. Higher temperature is the prerequisite for better oil and grease removal from the fabrics<sup>46,47</sup>. In the present study, the optimum washing temperature for oil and grease removal was obtained in the range of 40 – 55 °C using different immobilized lipase mutants. However, the thermostable lipase mutant, ZnO-bsl\_the3 showed highest temperature stability and maximum oil and grease removal at 55 °C. ZnO-bsl\_alk2 also showed better compatibility and improved thermostability in destaining of fabric. The present study also optimized washing time for all the immobilized lipases mutant for oil from the fabric and was 30 min for all other immobilized lipase. But, the efficient and optimized grease removal was noted in either 40 min for ZnO-bsl\_the3 and ZnO-bsl\_psy2 or 50 min for rest of the immobilized lipase. These results corroborated with previous reports that showed stain removal is more in presence of both lipases and biosurfactants which usually increases the wash time<sup>42,43</sup>. Table 7.5 summarizes washing process optimization for detergent concentration, lipase concentration, washing temperature and washing time for oil and grease removal from the cotton fabrics.

Table 7.5: Summary of washing process optimization of stain removal from the cotton fabrics.

| Ingredients*         | Stain removal | Optimized condition for maximum oil and grease removal |                                 |                          |                                 | Stain removal percentage |
|----------------------|---------------|--------------------------------------------------------|---------------------------------|--------------------------|---------------------------------|--------------------------|
|                      |               | Lipase concentration (U/ml)                            | Detergent concentration (% w/v) | Washing temperature (°C) | Washing time optimization (min) |                          |
| B+D                  | Oil           | 40                                                     | Remain constant after 4         | 45                       | 30                              | 21                       |
|                      | Grease        | 50                                                     | Remain constant after 5         | 45                       | 50                              | 17                       |
| B+D+ZnO-<br>bsl_wt   | Oil           | 40                                                     | 4                               | 45                       | 30                              | 62                       |
|                      | Grease        | 50                                                     | 5                               | 45                       | 50                              | 57                       |
| B+D+ZnO-<br>bsl_the3 | Oil           | 50                                                     | 5                               | 55                       | 30                              | 90                       |
|                      | Grease        | 50                                                     | 6                               | 55                       | 40                              | 82                       |
| B+D+ZnO-<br>bsl_psy2 | Oil           | 40                                                     | 5                               | 25                       | 30                              | 80                       |
|                      | Grease        | 50                                                     | 6                               | 25                       | 40                              | 76                       |
| B+D+ZnO-<br>bsl_acd2 | Oil           | 40                                                     | 5                               | 40                       | 30                              | 68                       |
|                      | Grease        | 50                                                     | 4                               | 40                       | 50                              | 66                       |
| B+D+ZnO-<br>bsl_alk2 | Oil           | 40                                                     | 5                               | 50                       | 30                              | 90                       |
|                      | Grease        | 50                                                     | 5                               | 50                       | 50                              | 82                       |
| B+D+ZnO-<br>bsl_hal2 | Oil           | 40                                                     | 4                               | 45                       | 30                              | 79                       |
|                      | Grease        | 50                                                     | 5                               | 45                       | 50                              | 75                       |

\*B: buffer, D: detergent

## 7.4 Conclusion

The generated extreme-stable mutants were employed for application in detergent formulations. The engineered lipases were immobilized on synthesized ZnO nanoparticles. The synthesized ZnO nanoparticles and lipase immobilization was confirmed using UV-Vis, FT-IR, Raman spectroscopy, DLS (for particle size), powder XRD, TEM and FESEM analysis. Moreover, the ZnO immobilized engineered lipases enhanced the storage stability, and compatibility with various oxidizing agents, surfactants and commercial detergents. They were subjected to oil and grease stain removal on cotton fabrics. The order of performance in oil and grease stain removal from the stained cotton fabric pieces by immobilized lipases on ZnO nanoparticle were bsl\_alk2 > bsl\_the3 > bsl\_hal2 > bsl\_psy2 > bsl\_acd2 > bsl\_wt. A total of 90% and 82% of oil and grease stains were removed from stained fabrics, respectively.

## 7.5 References

1. Katchalski-Katzir, E. Immobilized enzymes--learning from past successes and failures. *Trends Biotechnol.* **11**, 471–8 (1993).
2. Gupta, M. N. & Mattiasson, B. Unique applications of immobilized proteins in bioanalytical systems. *Methods Biochem. Anal.* **36**, 1–34 (1992).
3. O'Driscoll, K. F. [12] Techniques of enzyme entrapment in gels. in *Methods in Enzymology* **44**, 169–183 (1976).
4. Bernfeld, P. & Wan, J. Antigens and Enzymes Made Insoluble by Entrapping Them into Lattices of Synthetic Polymers. *Science (80-. )*. **142**, 678–679 (1963).
5. Dinelli, D., Marconi, W. & Morisi, F. [18] Fiber-entrapped enzymes. in *Methods in Enzymology* **44**, 227–243 (1976).
6. Wadiak, D. T. & Carbonell, R. G. Kinetic Behavior of Microencapsulated  $\beta$ -Galactosidase. *Biotechnol. Bioeng.* **17**, 1157–1181 (1975).
7. Messing, R. A. Adsorption and inorganic bridge formations. *Methods Enzymol.* **44**, 148–69 (1976).
8. Woodward, J. Immobilized enzymes: adsorption and covalent coupling. *Immobil. Cells Enzym. A Pract. Approach* 3–17 (1985).
9. Kennedy, J. F. & Cabral, J. M. S. Immobilisation of biocatalysts by metal-link/chelation processes. *Artif. Cells, Blood Substitutes, Biotechnol.* **23**, 231–252 (1995).
10. Cabral, J. M. S., Novais, J. M. & Kennedy, J. F. Immobilization studies of whole microbial cells on transition metal activated inorganic supports. *Appl. Microbiol. Biotechnol.* **23**, 157–162 (1986).
11. Cabral, J. M. & Kennedy, J. F. Covalent and coordination immobilization of proteins. *Bioprocess Technol.* **14**, 73–138 (1991).
12. Yoon, H. C., Hong, M.-Y. & Kim, H.-S. Functionalization of a poly (amidoamine) dendrimer with ferrocenyls and its application to the construction of a reagentless enzyme electrode. *Anal. Chem.* **72**, 4420–4427 (2000).
13. Wolf, M., Wirth, M., Pittner, F. & Gabor, F. Stabilisation and determination of the biological activity of L-asparaginase in poly (D, L-lactide-co-glycolide) nanospheres. *Int. J. Pharm.* **256**, 141–152 (2003).
14. Chauhan, M., Chauhan, R. S. & Garlapati, V. K. Evaluation of a new lipase from *Staphylococcus* sp. for detergent additive capability. *Biomed Res. Int.* **2013**, 374967 (2013).
15. Ghorbani, H., Mehr, F., Pazoki, H. & Rahmani, B. Synthesis of ZnO Nanoparticles by Precipitation

- Method. *Orient. J. Chem.* **31**, 1219–1221 (2015).
16. Kundu, D., Hazra, C., Chatterjee, A., Chaudhari, A. & Mishra, S. Extracellular biosynthesis of zinc oxide nanoparticles using *Rhodococcus pyridinivorans* NT2: Multifunctional textile finishing, biosafety evaluation and in vitro drug delivery in colon carcinoma. *J. Photochem. Photobiol. B Biol.* (2014). doi:10.1016/j.jphotobiol.2014.08.001
  17. Shah, E., Mahapatra, P., Bedekar, A. V. & Soni, H. P. Immobilization of *Thermomyces lanuginosus* lipase on ZnO nanoparticles: mimicking the interfacial environment. *RSC Adv.* **5**, 26291–26300 (2015).
  18. Patel, V., Shah, C., Deshpande, M. & Madamwar, D. Zinc Oxide Nanoparticles Supported Lipase Immobilization for Biotransformation in Organic Solvents: A Facile Synthesis of Geranyl Acetate, Effect of Operative Variables and Kinetic Study. *Appl. Biochem. Biotechnol.* **178**, 1630–1651 (2016).
  19. Bradford, M. M. & Marion, M. B. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254 (1976).
  20. Glogauer, A. *et al.* Identification and characterization of a new true lipase isolated through metagenomic approach. *Microb. Cell Fact.* **10**, 54 (2011).
  21. Bahadur, V. & Garimella, S. V. Electrowetting-based control of droplet transition and morphology on artificially microstructured surfaces. *Langmuir* **24**, 8338–45 (2008).
  22. Bahadur, H., Srivastava, A. K., Sharma, R. K. & Chandra, S. Morphologies of sol–gel derived thin films of ZnO using different precursor materials and their nanostructures. *Nanoscale Res. Lett.* **2**, 469 (2007).
  23. Pourrahimi, A. M. *et al.* Water-based synthesis and cleaning methods for high purity ZnO nanoparticles – comparing acetate, chloride, sulphate and nitrate zinc salt precursors. *RSC Adv.* **4**, 35568–35577 (2014).
  24. Kashif, M. *et al.* Effect of Different Seed Solutions on the Morphology and Electrooptical Properties of ZnO Nanorods. *J. Nanomater.* **2012**, 1–6 (2012).
  25. Ramimoghadam, D., Bin Hussein, M. & Taufiq-Yap, Y. Hydrothermal synthesis of zinc oxide nanoparticles using rice as soft biotemplate. *Chem. Cent. J.* **7**, 136 (2013).
  26. Zhu, Y. Z. *et al.* Electronic structure and phase stability of MgO, ZnO, CdO, and related ternary alloys. *Phys. Rev. B* **77**, 245209 (2008).
  27. Hasnidawani, J. N. *et al.* Synthesis of ZnO Nanostructures Using Sol-Gel Method. *Procedia Chem.* **19**, 211–216 (2016).
  28. Farley, N. R. S. *et al.* Magnetic properties of sol-gel-derived doped ZnO as a potential ferromagnetic semiconductor: a synchrotron-based study. *New J. Phys.* **10**, 055012 (2008).
  29. Kulkarni, S. S. & Shirsa, M. D. Optical and structural properties of zinc oxide nanoparticles. *IJARPS* **2**, 14–18 (2015).
  30. Hassan, S. S. M., El Azab, W. I. M., Ali, H. R. & Mansour, M. S. M. Green synthesis and characterization of ZnO nanoparticles for photocatalytic degradation of anthracene. *Adv. Nat. Sci. Nanosci. Nanotechnol.* **6**, 45012 (2015).
  31. Jin, B. ., Im, S. & Lee, S. . Violet and UV luminescence emitted from ZnO thin films grown on sapphire by pulsed laser deposition. *Thin Solid Films* **366**, 107–110 (2000).
  32. Manjón, F. J., Marí, B., Serrano, J. & Romero, A. H. Silent Raman modes in zinc oxide and related nitrides. *J. Appl. Phys.* **97**, 053516 (2005).
  33. Khan, A. Raman Spectroscopic Study of the ZnO Nanostructures. *J. Pakistan Mater. Soc.* **4**, 5–9 (2010).
  34. Nemecek, D., Stepanek, J. & Thomas, G. J. Raman Spectroscopy of Proteins and Nucleoproteins. *Curr. Protoc. Protein Sci.* **71**, 17.8.1–17.8.52 (2013).
  35. Žukienė, R. & Snitka, V. Zinc oxide nanoparticle and bovine serum albumin interaction and nanoparticles influence on cytotoxicity in vitro. *Colloids Surfaces B Biointerfaces* **135**, 316–323 (2015).
  36. Kisi, E. H. & Elcombe, M. M. u parameters for the wurtzite structure of ZnS and ZnO using powder neutron diffraction. *Acta Crystallogr. Sect. C Cryst. Struct. Commun.* **45**, 1867–1870 (1989).
  37. Ghasemi, S., Heidary, M., Faramarzi, M. A. & Habibi, Z. Immobilization of lipase on Fe<sub>3</sub>O<sub>4</sub>/ZnO core/shell magnetic nanoparticles and catalysis of Michael-type addition to chalcone derivatives. *J. Mol. Catal. B Enzym.* **100**, 121–128 (2014).
  38. Chen, Y. *et al.* Synthesis of biodiesel from waste cooking oil using immobilized lipase in fixed bed reactor. *Energy Convers. Manag.* **50**, 668–673 (2009).
  39. Rodrigues, R. C. *et al.* Positive effects of the multipoint covalent immobilization in the reactivation of partially inactivated derivatives of lipase from *Thermomyces lanuginosus*. *Enzyme Microb. Technol.* **44**, 386–393 (2009).

40. Mazlan, S. Z. & Hanifah, S. A. Effects of Temperature and pH on Immobilized Laccase Activity in Conjugated Methacrylate-Acrylate Microspheres. *Int. J. Polym. Sci.* **2017**, (2017).
41. Defaei, M., Taheri-Kafrani, A., Miroliaei, M. & Yaghmaei, P. Improvement of stability and reusability of  $\alpha$ -amylase immobilized on naringin functionalized magnetic nanoparticles: A robust nanobiocatalyst. *Int. J. Biol. Macromol.* (2018). doi:10.1016/j.ijbiomac.2018.02.147
42. Sajna, K. V. *et al.* Studies on biosurfactants from *Pseudozyma* sp. NII 08165 and their potential application as laundry detergent additives. *Biochem. Eng. J.* **78**, 85–92 (2013).
43. Grbavčić, S. *et al.* Production of lipase and protease from an indigenous *Pseudomonas aeruginosa* strain and their evaluation as detergent additives: Compatibility study with detergent ingredients and washing performance. *Bioresour. Technol.* **102**, 11226–11233 (2011).
44. Flipsen, J. A. C., Appel, A. C. M., Van Der Hijden, H. T. W. M. & Verrips, C. T. Mechanism of removal of immobilized triacylglycerol by lipolytic enzymes in a sequential laundry wash process. *Enzyme Microb. Technol.* (1998). doi:10.1016/S0141-0229(98)00050-7
45. Flipsen, J. A. C., Appel, A. C. M., Van der Hijden, H. & Verrips, C. T. Mechanism of removal of immobilized triacylglycerol by lipolytic enzymes in a sequential laundry wash process. *Enzyme Microb. Technol.* **23**, 274–280 (1998).
46. Horchani, H., Mosbah, H., Salem, N. Ben, Gargouri, Y. & Sayari, A. Biochemical and molecular characterisation of a thermoactive, alkaline and detergent-stable lipase from a newly isolated *Staphylococcus aureus* strain. *J. Mol. Catal. B Enzym.* **56**, 237–245 (2009).
47. Romdhane, I. B.-B., Fendri, A., Gargouri, Y., Gargouri, A. & Belghith, H. A novel thermoactive and alkaline lipase from *Talaromyces thermophilus* fungus for use in laundry detergents. *Biochem. Eng. J.* **53**, 112–120 (2010).

# 8

## Conclusion and future perspectives

---





Extremostable behavior is relative to normal. The present thesis entitled “Exploring molecular adaptations of extremophilic proteins: A platform for protein engineering” is an effort in this direction. The molecular adaptations which the extreme-stable proteins undergo to be robust as compared to their normal counterparts divulges the attributes contributing to extreme-stability of proteins. It tries to address questions as:

- Which is the most important contributing factor among them?
- Can all such factors be ranked according to their role in contributing towards extreme-stability?
- Can these factors be exploited for classifying extremostable proteins from their normal counterparts?
- Can the divulged factors be exploited to arrive at pre predicted mutations leading to extremostability?

**Chapter one** provides an insight into till date research pertaining to the multifactorial level of extreme-stability of proteins. Protein stability can be affected at three hierarchies of protein organization – DNA level, protein sequence and structure level (including primary sequence features, various intra-protein interactions, secondary and tertiary structural features).

**Chapter two** is on creating a curated database on Extreme-stable Protein Database (ExProtDB). The database architecture is on a platform based on the Linux-Apache-MySQL-PHP (LAMP). It contains information of 536 organisms (175 archaea, 255 bacteria, 100 eukaryotes and 6 viruses) and 795 protein entries (375 thermophilic proteins, 115 psychrophilic proteins, 67 acidophilic proteins, 94 alkaliphilic proteins, 100 halophilic proteins and 43 barophilic proteins) and 277 engineered extreme-stable proteins. It is integrated with a user friendly search and browse interface where search can be performed by user entered organism name; protein name; engineered extreme-stable proteins; literature and patents; E.C. No.; physicochemical properties (temperature, pH, pressure, salt concentrations, etc.) Additionally, the amino acid compositions and intra-protein interaction data have been integrated with each protein entry in the database. The tools for calculation of amino acid compositions and intra-protein interaction were

also incorporated as the download option for users. Therefore, it is a knowledgebase about all the information available in the World Wide Web about extremophiles and their extreme-stable proteins. The novelty of the database lies in the fact that it gives consolidated and comprehensive information of all types of extremophiles and extremophilic proteins.

**Chapter three** attempts at understanding the specific codon usage pattern in extremophiles. The collected genomic and proteomic data were then utilized for comparative analyses of extremophiles and non-extremophiles in the proceeding chapters. This chapter throws light on primary level i.e. genomic adaptations of extremophilic proteins. It deciphers the rationale behind the specific codon usage patterns in extremophiles. Relative abundance analysis, 1-9 scale ranking, nucleotide compositions, attribute weighting and machine learning algorithms were employed to arrive at findings. Analyses revealed AGG (Arg) in thermophiles and barophiles, CAA (Gln) in mesophiles and psychrophiles, TGG (Trp) in acidophiles, GAG (Glu) in alkaliphiles and GAC (Asp) in halophiles had highest preference. The preference of GC-rich and G/C-ending codons were observed in halophiles and barophiles whereas a decreasing trend was reflected in psychrophiles and alkaliphiles. GC-rich codons were found to decrease and G/C-ending codons increased in thermophiles whereas, acidophiles showed equal contents of GC-rich and G/C-ending codons. Codon usage patterns exhibited harmony among different extremophiles and has been detailed. However, the codon attribute preferences and their selectivity of extremophiles varied in comparison to non-extremophiles. Although codon study is being partially deciphered in thermophiles, this is probably the first report of codon preferences pertaining to all extremophile classes.

**Chapter four** is on analysis of attributes contributing to extreme-stability of proteins, further classification of extreme-stable proteins from their normal counterparts and generation of ranking models. It encompassed on prediction and ranking models for analyzing protein sequence and structure attributes contributing towards extreme-stability through machine learning and multi-criteria decision making (also known as Analytic Hierarchy Process, AHP) approaches. The analyses revealed thermophilic, alkaliphilic

and barophilic proteins preferred charged amino acids (Glu, Arg, Lys), ionic interactions and charge accessible surface area whereas, psychrophilic and halophilic proteins preferred small amino acids (Ser, Gln, Asp) and polar accessible surface area and acidophilic proteins preferred aromatic amino acids (Trp) and aromatic-aromatic interactions. These positively contributing attributes have been further exploited to design mutations in a mesophilic (non-extremophilic) protein in such a way to increase the content of these attributes to endorse the prediction of generated models. This chapter reports novel contributing factors for extreme-protein stabilization, which were not reported earlier.

**Chapter five** comprises of *in silico* prediction of possible mutations to attain extreme-stability of proteins and their validation. This chapter gives a pertinent account of *in silico* prediction of plausible mutations and their validation in *Bacillus subtilis* lipase (a chosen mesophilic enzyme) by substituting the preferred protein attributes to attain extreme-stability of proteins by designing *in silico* homology models. Accordingly, three thermostable mutants (bsl\_the1: V149K, Q150E, bsl\_the2: F41K, W42E, V149K, Q150E and bsl\_the3: F41K, W42E, P119E, Q121K, V149K, Q150E), two psychrostable mutants (bsl\_psy1: G52H, P53T and bsl\_psy2: G52H, P53T, G92M, G93A), two acidostable mutants (bsl\_acd1: G52F, V54Y, G111Y and bsl\_acd2: G52F, V54Y, T110H, G111Y), two alkalistable mutants (bsl\_alk1: T110E, G111Y and bsl\_alk2: W31E, S32H, T45H, T110E, G111Y), two halostable mutants (bsl\_hal1: G52D, P53D, P119D, N120D and bsl\_hal2: G52D, P53D, G93D, P119D, N120D) and one barostable mutant (bsl\_bar: S24R, P119R, Q164R, G175R) were designed. *In silico* validation of mutant was done by AHP generated ranking models, various mutation prediction servers (HotSpot Wizard, I-Mutant2, Cupsat, iPTREE-STAB, WET-STAB and ERIS web servers), Substrate Docking (Autodock 4.2 version), Ramachandran analysis (PROCHECK) and Contact Map analysis (CMView 1.1.1). Interestingly, results showed that the designed mutations did not affect the catalytic activity of the lipase and thus, mutation did not disturb the catalytic properties.

**Chapter six** pertains to the *in vitro* validation of predicted mutations. It deals with experimentation employed for *in vitro* cloning, mutagenesis, expression, purification and

characterization of lipases. Results revealed that the optimum activity and stability increased from 35 °C to 55 °C in bsl\_the3, pH 8 to pH 10 in bsl\_alk2 but reduced to pH 6.5 in bsl\_acd2 *w.r.t.* wild type lipase (bsl\_wt). Further, the bsl\_wt was halo-sensitive (0.5M NaCl) but in bsl\_hal2 halostability was increased to 1.5M NaCl. The content of secondary structure enhanced extreme-stability (with increased  $T_m$  value) which was confirmed through FT-IR (ATR mode) and Differential Scanning Fluorimetry (Thermal shift assay).

**Chapter seven** is about the applications of the generated mutations. Detergent formulations was the chosen application here. To further increase the enzyme stability, bsl\_wt and engineered lipases (bsl\_the3, bsl\_psy2, bsl\_acd2, bsl\_alk2 and bsl\_hal2 have better extreme-stability) were immobilized on synthesized ZnO nanoparticles as it is cheap, biocompatible and environmental friendly. The synthesized ZnO nanoparticles and lipase immobilization was confirmed using UV-Vis, FT-IR, Raman spectroscopy, DLS (for particle size), powder XRD, TEM and FESEM analysis. Interestingly, ZnO immobilized engineered lipases further enhanced the storage stability, reusability and their compatibility with various oxidizing agents, surfactants and commercial detergents and was applied for oil and grease stain removal on cotton fabrics. For selection of better washing conditions, the washing process was done at different temperatures and time intervals using different concentrations of detergent and lipase activities. The order of performance in oil and grease stain removal from the stained cotton fabric pieces by immobilized lipases on ZnO nanoparticle- bsl\_alk2 > bsl\_the3 > bsl\_hal2 > bsl\_psy2 > bsl\_acd2 > bsl\_wt. Results also showed that the maximum removal of oil and grease stains were up to 87% and 82%, respectively.

**Chapter eight** concludes the thesis summarizing the findings of the present investigations and states the future scopes and upcoming challenges. It can be concluded that thesis work has been successful to develop a single design strategy to mutate amino acid residues which can render proteins stable at different extremophilic range. Conclusively, this work reveals a promising perspective of using low-cost nanoparticle conjugated biocatalysts which was produced through rationalized protein engineering

approach and was instrumented for detergent formulation to overcome the drawbacks of the chemical counterparts.

In a nutshell,

- A curated database of extremophilic organisms and extremophilic proteins was created which will be accessible to all researchers.

#### ***In molecular adaptations of extremophilic proteins;***

- *Codon attributes showed* AGG in thermophiles and barophiles; CAA in psychrophiles; TGG in acidophiles; GAG in alkaliphiles; and GAC in halophiles are preferred.
- Codons usage patterns showed harmony among different extremophiles.
- *Amino acid attributes showed* Charged amino acid (Glu, Arg, Lys) in thermophilic, alkaliphilic and barophilic proteins; Small, tiny amino acids and Gln in psychrophilic proteins; aromatic and polar amino acids in acidophilic proteins; and acidic amino acids (Asp) in halophilic protein are preferred.
- *At structural attributes, the* Gamma turns, charge accessible surface area, electrostatic interactions, salt bridge and hydrophobic interactions in thermophilic, alkaliphilic and barophilic proteins; polar accessible surface area in psychrophilic proteins; and aromatic-aromatic interactions in acidophilic protein are preferred.

#### ***Further, for applications in protein engineering;***

- The positively contributing attributes were ranked and used for rationalized protein engineering approach that changes the regime from random mutagenesis to pre-predicted approach.

The ***novelty of the thesis*** lies in moving from random mutagenesis to a pre-predicted approach by ranking the contributing protein attributes of extreme-stability for a rationalized protein engineering platform.

Thus, the ***future perspectives*** stemming from the present work are

- Codon optimization for heterologous expression of extremophilic proteins.

- Studying the regulatory elements of extremostable proteins and designing of suitable expression vector.

Though protein extreme-stability has been extensively studied for many decades, universal rules to render proteins extremostable do not exist. Random protein engineering approaches are still utilized for extremostabilizing proteins with limited success, hence consuming time and capital. Though a wide range of *in silico* approaches have been utilized to study the difference among extremophilic and non-extremophilic proteins and to design stabilizing mutations, the approaches are rather individualistic with the lack of commonality between them. Thus, further newer methods need to surface with one goal of designing a universal protocol to extreme-stabilize proteins through site directed mutagenesis approach with investment of much lesser time and money.



# Research output

## **Thesis publications**

### ***Published***

1. **Khan, M. F.** and Patra, S. Deciphering the rationale behind specific codon usage pattern in extremophiles. *Sci. Rep.* **8**, 15548 (2018).
2. **Khan, M. F.\***, Chakravorty, D.\* and Patra, S. Multifactorial level of extremostability of proteins: can they be exploited for protein engineering? *Extremophiles* **21**, 419–444 (2017). \*contributed equally
3. **Khan, M. F.\***, Chakravorty, D.\* and Patra, S. Thermostability of Proteins Revisited Through Machine Learning Methodologies: From Nucleotide Sequence to Structure. *Current Biotechnology* **6**, 39–49 (2017). \*contributed equally

### ***Under Revision / Communicated***

4. **Khan, M. F.** and Patra, S. Molecular adaptations of barophilic proteins: the rationale basis and possible applications in protein engineering thereafter. (Minor revision submitted – *Extremophiles*, ID: EXT-18-Mar-0042).
5. **Khan, M. F.**, Kundu, D., Hazra, C. and Patra, S. Enhancing thermostability of *Bacillus subtilis* lipase through rationalized engineering and immobilization on zinc oxide nanoparticles for enzymatic detergent formulation. (Submitted in ACS – *Synthetic Biology*)
6. **Khan, M. F.**, Chakravorty, D. and Patra, S. ExProtDB: Extremophile Protein Database, a comprehensive web resource of extremophilic genomes and proteomes. (Submitted in Oxford's – *Database*)

### ***Manuscript under preparation***

7. **Khan, M. F.** and Patra, S. Molecular characterization of psychrophilic proteins: A study of attributes pertaining to cold adaptability
8. **Khan, M. F.** and Patra, S. An insight into the mechanisms of pH stability of proteins, rational design of alkaline and acidic stable mutants.
9. **Khan, M. F.** and Patra, S. Learning from evolution to design halophilic mutants by a predictive learning approach and their exploitation for industrial application.
10. **Khan, M. F.** and Patra, S. Designing of alkali-halo-thermostable lipase mutants using rationalized protein engineering method for detergent applications.

## **Allied publications**

1. **Khan, M. F.\***, Saha, P.\* and Patra, S. Truncated  $\alpha$ -amylase: an improved candidate for textile processing. *Prep. Biochem. Biotechnol.* 1–11 (2018). \*contributed equally

2. **Khan, M. F.\***, Saha, P.\* and Patra, S. Application of immobilized truncated  $\alpha$ -amylase in desizing of fabrics in textile industries. (Manuscript under preparation)
3. Kedare, M. M., **Khan, M. F.**, Singh K. K. and Patra, S. Challenges encountered in metagenome purification from manure compost – Reaching an acceptable purity level for PCR amplification. (Submitted in Elsevier's – Journal of Microbiological Methods)
4. Kedare, M. M., **Khan, M. F.** and Patra, S. Uncultured and non culturable microbes: Novel insights from metagenomics. (Manuscript under preparation)

### **Book Chapters**

1. **Khan, M. F.\***, Kundu, D.\* and Patra, S. Enzyme responsive and enzyme immobilized nanoplateforms for therapeutic delivery: an overview of research innovations and biomedical applications. \*contributed equally (Under review – Springer's Nanopharmaceuticals: Principles and Applications)
2. **Khan, M. F.\***, Kundu, D.\* and Patra, S. Environmental impact and econanotoxicity of engineered nanomaterials. \*contributed equally (Under review – Springer's Nanopharmaceuticals: Principles and Applications)

### **Patents**

1. **Khan, M. F.**, Kundu, D. and Patra, S. A rationalized protein engineering approach for enhanced thermostability, alkalistability and halostability of industrially important enzymes. (Ready to file)

### **Conferences**

1. **Mohd Faheem Khan**, Debamitra Chakravorty, Sanjukta Patra. In silico prediction of protein stability at extreme temperature through machine learning methodologies. Exploring Mechanisms in Biology: Theory and Experiment. 25 – 27 November, 2015. Bioinformatics Institute (BII) A\*STAR, Singapore. (Poster Presentaion)
2. **Mohd Faheem Khan** and Sanjukta Patra. A protein engineering platform to improve stability of proteins for industrial applications. Research Conclave'18. 8<sup>th</sup> – 11<sup>th</sup> March, 2018, IIT Guwahati, India. (Awarded Best Post Presentation)
3. Mayur Mahindra Kedare, **Mohd Faheem Khan**, Sanjukta Patra. Prospecting for novel thermostable lipase and other industrially important enzymes by exploiting metagenome. Research Conclave'18. 8<sup>th</sup> – 11<sup>th</sup> March, 2018, IIT Guwahati, India.
4. Poulomi Saha, **Mohd Faheem Khan** and Sanjukta Patra. Potential application of  $\alpha$ -amylase for desizing of fabrics in textile industries. Research Conclave'18. 8<sup>th</sup> – 11<sup>th</sup> March, 2018, IIT Guwahati, India.
5. **Mohd Faheem Khan** and Sanjukta Patra. Immobilization of engineered thermostable *Bacillus subtilis* lipase on ZnO nanoparticles for application in detergent formulation. 5<sup>th</sup> International Conference on Advanced Nanomaterial and Nanotechnology

- (ICANN-2017). 18 – 21 December, 2017, Centre for Nanotechnology, IIT Guwahati, Guwahati, India. (Poster Presentation)
6. Poulomi Saha, **Mohd Faheem Khan** and Sanjukta Patra. Exploring the potential of ZnO-NP immobilized *Bacillus subtilis*  $\alpha$ -amylases for desizing of fabrics in textile industry. 5<sup>th</sup> International Conference on Advanced Nanomaterial and Nanotechnology (ICANN-2017). 18 – 21 December, 2017, Centre for Nanotechnology, IIT Guwahati, Guwahati India. (Poster Presentation)
  7. **Mohd Faheem Khan** and Sanjukta Patra. Protein engineering of *Bacillus subtilis* lipase to improve alkalistability and thermostability for detergent application. 5<sup>th</sup> Bioprocessing India. 9 – 11 December, 2017, IIT Guwahati, Guwahati, India. (Poster Presentation)
  8. Mayur Mahindra Kedare, **Mohd Faheem Khan** and Sanjukta Patra. Metagenomic approach for mining industrially relevant thermophilic enzymes. 5<sup>th</sup> Bioprocessing India. 9 – 11 December, 2017, IIT Guwahati, Guwahati, India. (Poster Presentation)
  9. Poulomi Saha, **Mohd Faheem Khan** and Sanjukta Patra. Potential applications of *Bacillus subtilis*  $\alpha$ -amylase immobilized ZnO-NP for desizing of fabrics in textile industry. 5<sup>th</sup> Bioprocessing India. 9 – 11 December, 2017, IIT Guwahati, Guwahati. (Poster Presentation)
  10. **Mohd Faheem Khan** and Sanjukta Patra. Understanding molecular and physicochemical adaptations in barophilic proteins: from codon usage to protein structure 57<sup>th</sup> Annual Conference of Association of Microbiologists of India & International Symposium on “Microbes and Biosphere: What’s New What’s Next”. 24 – 27 November, 2016. Department of Botany, Guwahati University and IASST, Guwahati, India. (Poster Presentation)
  11. Mayur Mahindra Kedare, **Mohd Faheem Khan** and Sanjukta Patra. Exploiting metagenome for industrially important enzymes: Prospecting for novel thermostable lipase. 57<sup>th</sup> Annual Conference of Association of Microbiologists of India & International Symposium on “Microbes and Biosphere: What’s New What’s Next”. 24 – 27 November, 2016. Department of Botany, Guwahati University and IASST, Guwahati, India. (Poster Presentation)
  12. **Mohd Faheem Khan**, Sanjukta Patra. Design Design of halophilic mutants by predictive approaches. 3<sup>rd</sup> International Symposium on Protein Folding and Dynamics. 8 – 11 November 2016, NCBS-TIFR, Bangalore, India. (Poster Presentaion)
  13. **Mohd Faheem Khan**, Debamitra Chakravorty, Sanjukta Patra. An insight into the behavior of psychrophilic and thermophilic proteins. MCB75: from Molecules to Organisms. 11 – 14 December, 2015. Department of Microbiology and Cell Biology, IISc Bangalore, India. (Poster Presentaion)
  14. **Mohd Faheem Khan**, Debamitra Chakravorty, Sanjukta Patra. Predicting thermostability of proteins revisited through machine learning methodology: from

- nucleotide sequences to structures. 5<sup>th</sup> Annual International Conference Bio Tech 2015 Advances in Biotechnology 13 – 15 March 2015, IIT Kanpur, India. (Poster Presentaion)
15. **Mohd Faheem Khan**, Debamitra Chakravorty, Sanjukta Patra. Attaining psychrostaibility of proteins: Insights into the molecular basis. Research Conclave 17th-20th Mar 2016, IIT Guwahati, India. (Poster Presentaion)
  16. **Mohd Faheem Khan**. Predicting protein thermostability from sequence to structure. Advances in Computational Biology and Computer Aided Drug Design 24th-26th Jun 2015, IIT Guwahati, India. (Young Investigators' Oral Talk)
  17. **Mohd Faheem Khan**, Debamitra Chakravorty, Sanjukta Patra. Predicting thermostability of proteins revisited through machine learning methodology: from nucleotide sequences to structures. Research Conclave 23 – 26 March 2015, IIT Guwahati, India. (Poster Presentaion)
  18. **Mohd Faheem Khan**, Recent Advances in Cancer Biology and Therapeutics. 5 December 2014, IIT Guwahati, Guwahati, India. (Conference Attended)

### **Workshops**

1. Symposium cum Workshop on Advances in Computational Biology and Computer Aided Drug Design, 24 – 26 June 2015, IIT Guwahati, India.
2. Workshop on Flow Cytometry Data Analysis organized by BIO-RAD, 23 – 24 Janaury 2015, IIT Guwahati, India.
3. Workshop on Intellectual Property Rights - Invent, Protect & Prosper, 30 November - 1 December 2016, IIT Guwahati IPR Cell.
4. Workshop on “From Basic Invention to Advanced Biomedical Devices” (ICANN-2017) 18 December 2017, Centre for Nanotechnology, IIT Guwahati, Guwahati.
5. Workshop on Advanced Excel, Advanced Origin, ImageJ, Patent writing and Copyright at Research Conclave'18. 8<sup>th</sup>-11<sup>th</sup> March, 2018, SAB, IIT Guwahati, Guwahati, India.

# Appendices



## Appendix I

Table A1.1: Thermophile-Mesophile (T-M) dataset.

| Thermophile |              |        |                                                                 |                                                                                                                                                                  | Mesophile  |              |        |
|-------------|--------------|--------|-----------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|--------------|--------|
| ENA ID      | UniProtKB ID | PDB ID | Protein name                                                    | Organism name                                                                                                                                                    | ENA ID     | UniProtKB ID | PDB ID |
| AAD36062.1  | Q9X078       | 1JDQ   | UPF0033 protein TM_0983                                         | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAC21758.1 | P43933       | 1NO5   |
| AAD35268.1  | Q9WY19       | 1VKU   | Acyl carrier protein (ACP)                                      | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAC74997.1 | P0AA31       | 1JE3   |
| AAB85436.1  | O27021       | 1IHN   | Uncharacterized protein                                         | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | AAC44308.1 | P80643       | 1F80   |
| CAA40902.1  | P15452       | 1AYG   | Cytochrome c-552 (Cytochrome c552)                              | <i>Hydrogenobacter thermophilus</i> (strain DSM 6534 / IAM 12695 / TK-6)                                                                                         | CAA35703.1 | P00099       | 1DVV   |
| AAC07338.1  | O67367       | 1OZ9   | Endoribonuclease YbeY (EC 3.1.-.-)                              | <i>Aquifex aeolicus</i> (strain VF5)                                                                                                                             | AAC21683.1 | P71335       | 1XAX   |
| AAB07763.1  | P71143       | 2BM3   | Scaffolding dockerin binding protein A                          | <i>Clostridium thermocellum</i> ( <i>Ruminiclostridium thermocellum</i> )                                                                                        | AAP48995.1 | Q7WYN3       | 1QZN   |
| AAB03727.1  | Q56313       | 1IN4   | Holliday junction ATP-dependent DNA helicase RuvB (EC 3.6.4.12) | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | BAA11175.1 | P15412       | 1RDV   |
| AAD35484.1  | Q9WYN0       | 1KGS   | Response regulator                                              | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAP59551.1 | O75351       | 1WR0   |
| AAK42515.1  | Q97W73       | 1O7I   | Single-stranded DNA binding protein Ssb (SSB)                   | <i>Sulfolobus solfataricus</i> (strain ATCC 35092 / DSM 1617 / JCM 11322 / P2)                                                                                   | AAC00348.1 | P13792       | 1MVO   |
| CAA27987.1  | P04766       | 1D1N   | Translation initiation factor IF-2                              | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | AAL08288.1 | O49453       | 1WJJ   |
| AAB86164.1  | O27727       | 1JCU   | Conserved protein                                               | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | BAE22940.1 | Q91YJ5       | 2CRV   |
| CAB65651.1  | Q9RHZ6       | 1URD   | Maltose binding protein                                         | <i>Alicyclobacillus acidocaldarius</i> ( <i>Bacillus acidocaldarius</i> )                                                                                        | AAC74349.2 | P0AFR4       | 1K7J   |
| AAD36829.1  | Q9X286       | 1TZZ   | N utilization substance protein B homolog (Protein NusB)        | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAC21773.1 | P35755       | 1D9V   |

|            |        |      |                                                                                              |                                                                                                                                                                  |            |        |      |
|------------|--------|------|----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
| AAB99207.1 | P54066 | 1RA4 | 50S ribosomal protein L7Ae (Ribosomal protein L8e)                                           | <i>Methanocaldococcus jannaschii</i> (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440) ( <i>Methanococcus jannaschii</i> )                        | CAA45737.1 | P0A780 | 1EY1 |
| AAC44889.2 | P77994 | 1TTY | RNA polymerase sigma factor SigA (Sigma-A)                                                   | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | CAD98427.1 | Q7YYQ3 | 2AIF |
| BAD71652.1 | Q5SHA1 | 1IV0 | Putative Holliday junction resolvase (EC 3.1.-.-)                                            | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                                                                                  | AAC76103.1 | P00579 | 1SIG |
| BAD71661.1 | Q5SH92 | 2D2E | SufC protein (ATP-binding protein)                                                           | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                                                                                  | AAC75986.1 | P0A8I1 | 1NMN |
| AAB89725.1 | O28751 | 1HJZ | [Protein ADP-ribosylglutamate] hydrolase AF_1521 (EC 3.2.2.-)                                | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)                                                                    | AAA98234.1 | P08716 | 1MT0 |
| AAC07832.1 | O67859 | 1T6T | Uncharacterized protein                                                                      | <i>Aquifex aeolicus</i> (strain VF5)                                                                                                                             | AAA41560.1 | Q02874 | 1YD9 |
| BAA29671.1 | O58335 | 1UZ5 | 402aa long hypothetical molybdopterin biosynthesis moea protein                              | <i>Pyrococcus horikoshii</i> (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)                                                                    | AAB02999.1 | P0A6P7 | 1PUI |
| AAB89817.1 | O28835 | 1Y8A | Uncharacterized protein                                                                      | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)                                                                    | BAA24003.1 | P12915 | 1BEV |
| AAB91265.1 | O30273 | 1YTL | Acetyl-CoA decarbonylase/synthase complex subunit epsilon 2 (ACDS complex subunit epsilon 2) | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)                                                                    | BAE76284.1 | P0ACN4 | 1TF1 |
| AAB86213.1 | O27775 | 1RYJ | Uncharacterized protein                                                                      | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | CAA24361.1 | P03306 | 1ZBA |
| AAD35304.1 | Q9WY55 | 1ZKO | Glycine cleavage system H protein                                                            | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | CAJ24843.1 | Q3BQX0 | 1XS3 |
| AAC07491.1 | O67517 | 1LFP | Probable transcriptional regulatory protein aq_1575                                          | <i>Aquifex aeolicus</i> (strain VF5)                                                                                                                             | AAA25674.1 | P02904 | 1DCZ |
| AAB90746.1 | O29759 | 1P9Q | Ribosome maturation protein SDO1 homolog                                                     | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)                                                                    | EAW73148.1 | P32455 | 1DG3 |
| AAB86171.1 | O27734 | 1GH8 | Elongation factor 1-beta (EF-1-beta) (aEF-1beta)                                             | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | AAH73755.1 | P13928 | 1W3W |
| AAD36689.1 | Q9X1V7 | 1VR8 | Uncharacterized protein                                                                      | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAN32614.1 | Q8TD84 | 1VA9 |

|            |        |      |                                                                                                                                                    |                                                                                                                                                                  |            |        |      |
|------------|--------|------|----------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
| BAK54261.1 | Q975N2 | 1UD9 | DNA polymerase sliding clamp 1 (Proliferating cell nuclear antigen homolog 1) (PCNA1)                                                              | <i>Sulfolobus tokodaii</i> (strain DSM 16993 / JCM 10545 / NBRC 100140 / 7)                                                                                      | AAC73854.1 | P52697 | 1RI6 |
| AAM23944.1 | Q8RBX6 | 1U4H | Methyl-accepting chemotaxis protein                                                                                                                | <i>Caldanaerobacter subterraneus</i> subsp. <i>tengcongensis</i> (strain DSM 15242 / JCM 11007 / NBRC 100824 / MB4) ( <i>Thermoanaerobacter tengcongensis</i> )  | AAH47750.1 | P55055 | 1P8D |
| AAD36040.1 | Q9X056 | 2ETD | LemA protein                                                                                                                                       | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAG33028.1 | P24817 | 1CF5 |
| AAB85143.1 | O26734 | 1JRM | UPF0235 protein MTH_637                                                                                                                            | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | AAG58084.1 | Q8XCU6 | 1N91 |
| CAA06924.1 | Q9ZF13 | 1BQC | Beta-mannanase (EC 3.2.1.78)                                                                                                                       | <i>Thermobifida fusca</i> ( <i>Thermomonospora fusca</i> )                                                                                                       | CAA83942.1 | Q59232 | 1LF1 |
| AAC06388.1 | P26222 | 1TML | Endoglucanase E-2 (EC 3.2.1.4) (Cellulase E-2) (Cellulase E2) (Endo-1,4-beta-glucanase E-2)                                                        | <i>Thermobifida fusca</i> ( <i>Thermomonospora fusca</i> )                                                                                                       | CAA10867.1 | O34816 | 1Y7M |
| AAF24127.1 | P23360 | 1GOK | Endo-1,4-beta-xylanase (Xylanase) (EC 3.2.1.8) (1,4-beta-D-xylan xylanohydrolase) (TAXI)                                                           | <i>Thermoascus aurantiacus</i>                                                                                                                                   | AAA74578.1 | P49863 | 1MZA |
| AAM77714.2 | Q8NJY3 | 1OLR | Endoglucanase                                                                                                                                      | <i>Humicola grisea</i>                                                                                                                                           | AIX97359.1 | O74705 | 1KS4 |
| CAA29568.1 | P09177 | 1MPP | Mucorpepsin (EC 3.4.23.23) (Mucor rennin)                                                                                                          | <i>Rhizomucor pusillus</i>                                                                                                                                       | AAB63975.1 | P07267 | 1DP5 |
| AAD45520.2 | Q9XBQ3 | 1PZ2 | Intracellular exo-alpha-(1->5)-L-arabinofuranosidase (ABF) (EC 3.2.1.55) (Intracellular arabinan exo-alpha-(1->5)-L-arabinosidase) (Arabinosidase) | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | AAA17747.1 | P45700 | 1NXC |
| AAC12257.1 | O66015 | 1KU0 | Lipase (EC 3.1.1.3)                                                                                                                                | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | AAA50466.1 | P22088 | 1HQD |
| CAC11959.1 | P96084 | 1MT3 | Proline iminopeptidase (PIP) (EC 3.4.11.5) (Prolyl aminopeptidase) (PAP) (Tricorn protease-interacting factor F1)                                  | <i>Thermoplasma acidophilum</i> (strain ATCC 25905 / DSM 1728 / JCM 9062 / NBRC 15155 / AMRC-C165)                                                               | BAA23336.1 | O32449 | 1QTR |
| AAC45089.1 | O05209 | 1CZ4 | VCP-like ATPase                                                                                                                                    | <i>Thermoplasma acidophilum</i> (strain ATCC 25905 / DSM 1728 / JCM 9062 / NBRC 15155 / AMRC-C165)                                                               | BAA16561.2 | P0A7G6 | 1AA3 |
| BAB64339.1 | Q93HT9 | 1WL7 | Intracellular endo-alpha-(1->5)-L-arabinanase (ABN) (EC 3.2.1.99) (Endo-1,5-alpha-L-arabinanase)                                                   | <i>Geobacillus thermodenitrificans</i>                                                                                                                           | AAA16327.1 | Q07009 | 1DF0 |

|            |        |      |                                                                                                                                                                            |                                                                                               |            |        |      |
|------------|--------|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|------------|--------|------|
| BAA12070.1 | P71140 | 2C24 | Endoglucanase J                                                                                                                                                            | <i>Clostridium thermocellum</i> ( <i>Ruminiclostridium thermocellum</i> )                     | CAA59198.1 | Q50245 | 1L0Q |
| ABK52387.1 | P54583 | 1ECE | Endoglucanase E1 (EC 3.2.1.4) (Cellulase E1) (Endo-1,4-beta-glucanase E1) (Endocellulase E1)                                                                               | <i>Acidothermus cellulolyticus</i> (strain ATCC 43068 / 11B)                                  | AAK97759.2 | Q8L5J1 | 1RH9 |
| CAA54291.1 | P00800 | 1FJ3 | Thermolysin (EC 3.4.24.27) (Thermostable neutral proteinase)                                                                                                               | <i>Bacillus thermoproteolyticus</i>                                                           | BAB20416.1 | Q9LBQ9 | 1UG9 |
| AAC44232.1 | P77847 | 1PMH | Beta-1,4-mannanase (EC 3.2.1.78)                                                                                                                                           | <i>Caldicellulosiruptor saccharolyticus</i> ( <i>Caldocellum saccharolyticum</i> )            | AAA25811.1 | P14756 | 1EZM |
| BAD18071.1 | Q75UV1 | 1VC8 | Diadenosine hexaphosphate hydrolase (Ap6A hydrolase) (EC 3.6.1.61) (ATP-generating (di)nucleotide polyphosphate hydrolase) (ATP-generating Ap6A hydrolase) (Nudix protein) | <i>Thermus thermophilus</i>                                                                   | BAE77699.1 | P0ADK6 | 1OYZ |
| AAA24565.1 | P0A7Y4 | 1F21 | Ribonuclease HI (RNase HI) (EC 3.1.26.4) (Ribonuclease H) (RNase H)                                                                                                        | <i>Escherichia coli</i> (strain K12)                                                          | CAB63351.1 | Q9U2M7 | 1KT9 |
| CAA43026.1 | P29253 | 1JL2 | Ribonuclease H (RNase H) (EC 3.1.26.4)                                                                                                                                     | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                               | AAA24565.1 | P0A7Y4 | 1F21 |
| BAD69943.1 | Q5SM23 | 1WF3 | GTPase Era                                                                                                                                                                 | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                               | BAE77587.1 | P25522 | 1RFL |
| CAA55695.1 | P43522 | 1V3Y | Peptide deformylase (PDF) (EC 3.5.1.88) (Polypeptide deformylase)                                                                                                          | <i>Thermus thermophilus</i>                                                                   | AAG02249.1 | P68826 | 1LM4 |
| AGL50123.1 | O33835 | 1ZY9 | Alpha-galactosidase (EC 3.2.1.22)                                                                                                                                          | <i>Thermotoga maritima</i>                                                                    | AAP54412.1 | Q9FXT4 | 1UAS |
| AAD36230.1 | Q9X0N8 | 1PBT | 6-phosphogluconolactonase (6PGL) (EC 3.1.1.31)                                                                                                                             | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | AAF96794.1 | Q9KL51 | 1Y89 |
| AAD36544.1 | Q9X1I7 | 1O0X | Methionine aminopeptidase (MAP) (MetAP) (EC 3.4.11.18) (Peptidase M)                                                                                                       | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | BAG65108.1 | P53582 | 2B3H |
| AGL50674.1 | Q9X264 | 1PW5 | NagD protein, putative (Putative NagD-like phosphatase)                                                                                                                    | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | BAE77597.1 | P0A8Y5 | 1RKQ |
| AGL49592.1 | Q9WZD5 | 1J6O | Putative deoxyribonuclease YcfH (Uncharacterized protein)                                                                                                                  | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | CAA28212.1 | P00720 | 102L |
| AAD36531.1 | Q9X1H4 | 1NZ0 | Ribonuclease P protein component (RNase P protein) (RNaseP protein) (EC 3.1.26.5) (Protein C5)                                                                             | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | AAF61418.1 | P0A0H5 | 1D6T |
| AAB89058.1 | O28085 | 1OI0 | Uncharacterized protein                                                                                                                                                    | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126) | AAZ38713.1 | P20701 | 1CQP |
| AAB90868.1 | O29883 | 1Z0B | Archaeal Lon protease (EC 3.4.21.-) (ATP-dependent protease La homolog)                                                                                                    | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC         | ABD72193.1 | P13569 | 1NBD |

|            |        |      |                                                                                                                                                                                                                             |                                                                                                                                           |            |        |      |
|------------|--------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
|            |        |      |                                                                                                                                                                                                                             | 100126)                                                                                                                                   |            |        |      |
| AAC06997.1 | O67040 | 1T6C | Exopolyphosphatase                                                                                                                                                                                                          | <i>Aquifex aeolicus</i> (strain VF5)                                                                                                      | DAA06783.1 | P38804 | 1NYN |
| AAB99612.1 | Q58989 | 1F5S | Phosphoserine phosphatase (PSP) (PSPase) (EC 3.1.3.3) (O-phosphoserine phosphohydrolase)                                                                                                                                    | <i>Methanocaldococcus jannaschii</i> (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440) ( <i>Methanococcus jannaschii</i> ) | AAT48221.1 | P15043 | 1OYW |
| BAA30713.1 | O59248 | 1X0T | Ribonuclease P protein component 4 (RNase P component 4) (EC 3.1.26.5) (Rpp21)                                                                                                                                              | <i>Pyrococcus horikoshii</i> (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)                                             | CAD97681.1 | P78330 | 1L8L |
| AAD54338.1 | Q7LYT7 | 1MWO | Alpha amylase (EC 3.2.1.1)                                                                                                                                                                                                  | <i>Pyrococcus woesei</i>                                                                                                                  | BAA06696.1 | Q62132 | 1JLN |
| AAL80661.1 | Q8U3D2 | 1U04 | Uncharacterized protein                                                                                                                                                                                                     | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                | ABH07387.1 | P20646 | 1RPA |
| AAL80564.1 | E7FHX6 | 1DQ3 | Vitamin B12-dependent ribonucleoside-diphosphate reductase (B12-dependent RNR) (EC 1.17.4.1) (Ribonucleotide reductase) [Cleaved into: Endonuclease PI-PfuI (EC 3.1.-.-) (Pfu rnr-1 intein); Pfu rnr-2 intein (EC 3.1.-.-)] | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                | AAK28401.1 | O14727 | 1C15 |
| AAD26328.1 | Q9X519 | 1EJJ | 2,3-bisphosphoglycerate-independent phosphoglycerate mutase (23PGA-independent) (BPG-independent PGAM) (Phosphoglyceromutase) (iPGM) (EC 5.4.2.12)                                                                          | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                              | AAK07481.1 | Q9ALZ1 | 1V5C |
| BAA24446.1 | O52980 | 1IX5 | FKBP-type peptidyl-prolyl cis-trans isomerase (PPIase) (EC 5.2.1.8) (MtFK) (Rotamase)                                                                                                                                       | <i>Methanothermococcus thermolithotrophicus</i> ( <i>Methanococcus thermolithotrophicus</i> )                                             | AAA52373.1 | P05198 | 1KL9 |
| AAO40746.1 | Q84FH6 | 1ODE | Chorismate mutase AroH (EC 5.4.99.5)                                                                                                                                                                                        | <i>Thermus thermophilus</i>                                                                                                               | AAL57682.1 | Q9SCY2 | 1U79 |
| BAD71845.1 | Q53WB3 | 1V37 | Alpha-ribazole-5'-phosphate phosphatase                                                                                                                                                                                     | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                                                           | EAK90238.1 | Q5CYJ7 | 2A22 |
| AAD36114.1 | Q9X0C7 | 1QO2 | 1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino]imidazole-4-carboxamide isomerase (EC 5.3.1.16) (Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase)                                   | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                              | CAA97363.1 | P36623 | 1FZT |
| BAA31042.1 | O59580 | 1V7R | Non-canonical purine NTP pyrophosphatase (EC 3.6.1.19) (Non-standard purine NTP pyrophosphatase)                                                                                                                            | <i>Pyrococcus horikoshii</i> (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)                                             | CAB51442.1 | P16250 | 1VZW |



|            |        |      |                                                                                                                                                            |                                                                                                    |            |        |      |
|------------|--------|------|------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|------------|--------|------|
|            |        |      | (Nucleoside-triphosphate diphosphatase)<br>(Nucleoside-triphosphate pyrophosphatase) (NTPase)                                                              |                                                                                                    |            |        |      |
| CAC11654.1 | Q9HKT1 | 2ARS | Lipoate-protein ligase A subunit 1 (EC 2.7.7.63) (Lipoate--protein ligase subunit 1)                                                                       | <i>Thermoplasma acidophilum</i> (strain ATCC 25905 / DSM 1728 / JCM 9062 / NBRC 15155 / AMRC-C165) | AAX93766.1 | P47244 | 1KFI |
| BAD70261.1 | P27000 | 1G59 | Glutamate--tRNA ligase (EC 6.1.1.17) (Glutamyl-tRNA synthetase) (GluRS)                                                                                    | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                    | AAC73774.1 | P00962 | 1EUQ |
| BAD69921.1 | Q5SM45 | 1IQ0 | Arginine--tRNA ligase (EC 6.1.1.19) (Arginyl-tRNA synthetase) (ArgRS)                                                                                      | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                    | AAC73255.3 | P27305 | 1NZJ |
| CAC39622.1 | Q93RP5 | 1IQ0 | Arginine--tRNA ligase (EC 6.1.1.19) (Arginyl-tRNA synthetase)                                                                                              | <i>Thermus thermophilus</i>                                                                        | AAC43103.1 | P15640 | 1GSO |
| AAD36325.1 | Q9X0X7 | 1VKZ | Phosphoribosylamine--glycine ligase (EC 6.3.4.13) (GARS) (Glycinamide ribonucleotide synthetase) (Phosphoribosylglycinamide synthetase)                    | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                       | AAA88210.1 | P15925 | 1FGS |
| AAD35259.1 | Q9WY13 | 1O5Z | Folylpolyglutamate synthase/dihydrofolate synthase                                                                                                         | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                       | AAD03608.1 | P81294 | 1PXZ |
| BAB40336.1 | Q9AJM4 | 1VBL | Pectate lyase 47                                                                                                                                           | <i>Bacillus</i> sp. TS-47                                                                          | BAA14792.1 | P0A877 | 1V7Y |
| AAS81077.1 | P16608 | 1UJP | Tryptophan synthase alpha chain (EC 4.2.1.20)                                                                                                              | <i>Thermus thermophilus</i> (strain HB27 / ATCC BAA-163 / DSM 7039)                                | AAB09590.1 | Q94734 | 1D2U |
| AAC07100.1 | O67135 | 1C3P | Acetoin utilization protein                                                                                                                                | <i>Aquifex aeolicus</i> (strain VF5)                                                               | BAB21059.1 | Q9AQS0 | 1J0M |
| BAA30907.1 | O59452 | 1WR2 | Putative uncharacterized protein PH1788                                                                                                                    | <i>Pyrococcus horikoshii</i> (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)      | AAB38270.1 | P31016 | 1BE9 |
| AAL64050.1 | Q8ZVK6 | 1XQO | N-glycosylase/DNA lyase (8-oxoguanine DNA glycosylase) (EC 3.2.2.-) (AGOG) (DNA-(apurinic or apyrimidinic site) lyase) (AP lyase) (EC 4.2.99.18) (Pa-AGOG) | <i>Pyrobaculum aerophilum</i> (strain ATCC 51768 / IM2 / DSM 7523 / JCM 9630 / NBRC 100827)        | AAA86659.1 | Q12718 | 1GYC |
| CAE00180.1 | Q70KY3 | 1GW0 | Laccase-1 (EC 1.10.3.2) (Benzenediol:oxygen oxidoreductase 1) (Diphenol oxidase 1) (Ligninolytic phenoloxidase) (Urishiol oxidase 1)                       | <i>Melanocarpus albomyces</i>                                                                      | CAA25409.1 | P00374 | 1BOZ |
| BAD76060.1 | Q5KZ26 | 1ZDR | Dihydrofolate reductase (EC 1.5.1.3)                                                                                                                       | <i>Geobacillus kaustophilus</i> (strain HTA426)                                                    | AAC76786.2 | P0AA25 | 1F6M |
| BAD71304.1 | Q5SI93 | 1V98 | Thioredoxin                                                                                                                                                | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                    | AAL90879.1 | Q8RN03 | 1UED |



|            |        |      |                                                                                                                                                                                               |                                                                                                                                                                  |            |        |      |
|------------|--------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
| BAD71988.1 | Q53WG9 | 1WJ9 | CRISPR-associated endoribonuclease Cse3 (EC 3.1.-.-) (Cse3 endoRNase) (Cse3 endoribonuclease)                                                                                                 | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                                                                                  | AAA83534.1 | P06632 | 1A80 |
| AAD36074.1 | Q9X0A2 | 1VP5 | Oxidoreductase, aldo/keto reductase family                                                                                                                                                    | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAA50163.1 | Q02127 | 1D3G |
| AAD35190.1 | Q9WXV1 | 1VHN | tRNA-dihydrouridine synthase (EC 1.-.-.-)                                                                                                                                                     | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAF28842.1 | Q9CQM9 | 1WIK |
| AGL49847.1 | Q9X022 | 1O2D | Alcohol dehydrogenase, iron-containing                                                                                                                                                        | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAA03078.1 | Q59490 | 1L1L |
| CAA73377.1 | O33839 | 1XJE | Ribonucleotide reductase                                                                                                                                                                      | <i>Thermotoga maritima</i>                                                                                                                                       | AAA23381.1 | P00269 | 1RB9 |
| AAL81406.1 | P24297 | 1BQ8 | Rubredoxin (Rd)                                                                                                                                                                               | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                                       | AAA29633.1 | Q27743 | 1CEQ |
| AAA22568.1 | P00344 | 1LDB | L-lactate dehydrogenase (L-LDH) (EC 1.1.1.27)                                                                                                                                                 | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | AAC35399.1 | O87899 | 1YX3 |
| AAB89026.1 | O28055 | 1SAU | Sulfite reductase, desulfovibrin-type subunit gamma (DsvC)                                                                                                                                    | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)                                                                    | CAA42574.1 | P29476 | 1B8Q |
| AAY81375.1 | Q55080 | 1F4T | Cytochrome P450 119 (EC 1.14.-.-) (Peroxidase) (EC 1.11.1.7)                                                                                                                                  | <i>Sulfolobus acidocaldarius</i> (strain ATCC 33909 / DSM 639 / JCM 8929 / NBRC 15157 / NCIMB 11770)                                                             | CAI42779.1 | O95831 | 1M6I |
| AAL81321.1 | Q8U1K9 | 1XHC | NADH oxidase /nitrite reductase                                                                                                                                                               | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                                       | AAA64512.1 | P11182 | 1K8M |
| CAA37630.1 | P11961 | 1B5S | Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex (EC 2.3.1.12) (Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex) (E2) | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | AAB84388.1 | O14792 | 1ZRH |
| AAD20315.1 | Q9X4D0 | 1D0Q | DNA primase (EC 2.7.7.-)                                                                                                                                                                      | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | AAA23610.1 | P69783 | 1F3G |
| ABB58144.1 | Q06904 | 1T4Y | Adaptive-response sensory-kinase SasA (EC 2.7.13.3) ( <i>Synechococcus</i> adaptive sensor protein A)                                                                                         | <i>Synechococcus elongatus</i> (strain PCC 7942) ( <i>Anacystis nidulans</i> R2)                                                                                 | AAC74185.1 | P69786 | 1IBA |
| AAB85393.1 | O26981 | 1ILO | Thioredoxin                                                                                                                                                                                   | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | CCP44148.1 | P9WKE9 | 1S4Q |
| BAD70899.1 | Q5SJD8 | 1WG8 | Ribosomal RNA small subunit methyltransferase H (EC 2.1.1.199) (16S)                                                                                                                          | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                                                                                  | AAC97955.1 | Q9UNQ2 | 1ZQ9 |

|            |        |      |                                                                                                                                                                                                              |                                                                                                     |            |        |      |
|------------|--------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|------------|--------|------|
|            |        |      | rRNA m(4)C1402 methyltransferase)<br>(rRNA (cytosine-N(4)-<br>methyltransferase RsmH)                                                                                                                        |                                                                                                     |            |        |      |
| BAC67244.1 | Q84BQ9 | 1UFK | Ribosomal protein L11 methyltransferase<br>(L11 Mtase) (EC 2.1.1.-)                                                                                                                                          | <i>Thermus thermophilus</i> (strain HB8 / ATCC<br>27634 / DSM 579)                                  | BAA36069.1 | P0A7I0 | 2B3T |
| AAD36367.1 | Q9X119 | 1VLM | Diguanylate cyclase with PAS/PAC<br>sensor (Uncharacterized protein)                                                                                                                                         | <i>Thermotoga maritima</i> (strain ATCC 43589 /<br>MSB8 / DSM 3109 / JCM 10099)                     | AAC74296.1 | P0ACC1 | 1T43 |
| AAD35573.1 | Q9WYV8 | 1NV8 | Release factor glutamine<br>methyltransferase (RF MTase) (EC<br>2.1.1.297) (N5-glutamine<br>methyltransferase PrmC) (Protein-<br>(glutamine-N5) MTase PrmC) (Protein-<br>glutamine N-methyltransferase PrmC) | <i>Thermotoga maritima</i> (strain ATCC 43589 /<br>MSB8 / DSM 3109 / JCM 10099)                     | CAB12110.2 | O31474 | 1XXL |
| AAD35939.1 | Q9WZW1 | 1MRZ | Riboflavin biosynthesis protein (EC<br>2.7.1.26) (EC 2.7.7.2)                                                                                                                                                | <i>Thermotoga maritima</i> (strain ATCC 43589 /<br>MSB8 / DSM 3109 / JCM 10099)                     | AAH07069.1 | Q969G6 | 1NB0 |
| AAD35910.1 | Q9WZT5 | 2AJR | Sugar kinase, pfkB family (Tagatose-6-<br>phosphate kinase / 1-<br>phosphofructokinase) (EC 2.7.1.144) (EC<br>2.7.1.56)                                                                                      | <i>Thermotoga maritima</i> (strain ATCC 43589 /<br>MSB8 / DSM 3109 / JCM 10099)                     | BAH13786.1 | P55263 | 1BX4 |
| AAD36360.1 | Q9X112 | 1T7L | 5-methyltetrahydropteroyltriglutamate--<br>homocysteine methyltransferase (EC<br>2.1.1.14) (Cobalamin-independent<br>methionine synthase) (Methionine<br>synthase, vitamin-B12 independent<br>isozyme)       | <i>Thermotoga maritima</i> (strain ATCC 43589 /<br>MSB8 / DSM 3109 / JCM 10099)                     | AAK82464.1 | O50008 | 1U1H |
| AAD35358.1 | Q9WYA6 | 1J6R | Methionine synthase activation domain<br>protein (EC 2.1.1.13) (Uncharacterized<br>protein)                                                                                                                  | <i>Thermotoga maritima</i> (strain ATCC 43589 /<br>MSB8 / DSM 3109 / JCM 10099)                     | BAA00518.1 | P21404 | 1D4M |
| AAB89084.1 | O28126 | 1R89 | CCA-adding enzyme (EC 2.7.7.72)<br>(CCA tRNA nucleotidyltransferase)<br>(tRNA CCA-pyrophosphorylase) (tRNA<br>adenylyl-/cytidylyl- transferase) (tRNA<br>nucleotidyltransferase) (tRNA-NT)                   | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 /<br>VC-16 / DSM 4304 / JCM 9628 / NBRC<br>100126) | CAA50551.1 | P10538 | 1BFN |
| AAB91236.1 | O30245 | 1TQI | RIO-type serine/threonine-protein kinase<br>Rio2 (AfRio2) (EC 2.7.11.1)                                                                                                                                      | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 /<br>VC-16 / DSM 4304 / JCM 9628 / NBRC<br>100126) | ABD96827.1 | P53355 | 1IG1 |
| AAC07294.1 | O67322 | 1YE8 | Nucleoside-triphosphatase THEP1<br>(NTPase THEP1) (EC 3.6.1.15)<br>(Nucleoside triphosphate                                                                                                                  | <i>Aquifex aeolicus</i> (strain VF5)                                                                | EAW51291.1 | Q9Y3D8 | 1RKB |

|            |        |      |                                                                                                                                                                                                         |                                                                                                                                           |            |        |      |
|------------|--------|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
|            |        |      | phosphohydrolase)                                                                                                                                                                                       |                                                                                                                                           |            |        |      |
| AAB99088.1 | Q58487 | 1KKH | Mevalonate kinase (MK) (MVK) (EC 2.7.1.36)                                                                                                                                                              | <i>Methanocaldococcus jannaschii</i> (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440) ( <i>Methanococcus jannaschii</i> ) | AAA58612.1 | Q01415 | 2A2C |
| BAD86160.1 | O74023 | 1MGT | Methylated-DNA--protein-cysteine methyltransferase (EC 2.1.1.63) (6-O-methylguanine-DNA methyltransferase) (MGMT) (O-6-methylguanine-DNA-alkyltransferase) (Pk-MGMT)                                    | <i>Thermococcus kodakarensis</i> (strain ATCC BAA-918 / JCM 12380 / KOD1) ( <i>Pyrococcus kodakaraensis</i> (strain KOD1))                | AAA59596.1 | P16455 | 1EH6 |
| BAA30480.1 | O50082 | 1IXK | 315aa long hypothetical proliferating-cell nucleolar protein p120                                                                                                                                       | <i>Pyrococcus horikoshii</i> (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)                                             | AAA97507.1 | P0C0R7 | 1EIZ |
| AAL82046.1 | Q8TZR3 | 1JG1 | Protein-L-isoaspartate O-methyltransferase (EC 2.1.1.77) (L-isoaspartyl protein carboxyl methyltransferase) (Protein L-isoaspartyl methyltransferase) (Protein-beta-aspartate methyltransferase) (PIMT) | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                | AAB38386.1 | P22061 | 1I1N |
| AAL80152.1 | Q8U4Q2 | 1VKC | Putative acetyl transferase                                                                                                                                                                             | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                | AAO80753.1 | Q836Z8 | 1U6M |
| AAL80379.1 | Q8U440 | 1RYQ | Transcription elongation factor Spt4                                                                                                                                                                    | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                | AAN33016.1 | P50135 | 1ICZ |

Table A1.2: Psychrophile-Mesophile (P-M) dataset.

| Psychrophile |               |        |                                                                                 |                                                                                                                         | Mesophile  |               |        |
|--------------|---------------|--------|---------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|------------|---------------|--------|
| ENA ID       | UniProtK B ID | PDB ID | Protein name                                                                    | Organism name                                                                                                           | ENA ID     | UniProtK B ID | PDB ID |
| CAD00476.1   | Q8Y4N7        | 1Y9I   | Low temperature requirement C protein, also similar to B. subtilis YutG protein | <i>Listeria monocytogenes</i>                                                                                           | AAC21758.1 | P43933        | 1NO5   |
| AAP78483.1   | Q7X0F0        | 1Y7Y   | C.AhdI                                                                          | <i>Aeromonas hydrophila</i>                                                                                             | AAC74997.1 | P0AA31        | 1JE3   |
| ABK39236.1   | A0KKT0        | 2LGH   | Uncharacterized protein                                                         | <i>Aeromonas hydrophila</i> subsp. <i>hydrophila</i> (strain ATCC 7966 / DSM 30187 / JCM 1027 / KCTC 2358 / NCIMB 9240) | AAC44308.1 | P80643        | 1F80   |
| AAA21938.1   | P09167        | 1PRE   | Aerolysin                                                                       | <i>Aeromonas hydrophila</i>                                                                                             | CAA35703.1 | P00099        | 1DVV   |
| ABK37325.1   | A0KPA6        | 3L6E   | Oxidoreductase, short-chain dehydrogenase/reductase family                      | <i>Aeromonas hydrophila</i> subsp. <i>hydrophila</i> (strain ATCC 7966 / DSM 30187 / JCM 1027 / KCTC 2358 / NCIMB 9240) | AAC21683.1 | P71335        | 1XAX   |
| ABF70178.1   | Q1EHA2        | 3PH0   | AscG (Preprotein translocase G)                                                 | <i>Aeromonas hydrophila</i>                                                                                             | AAP48995.1 | Q7WYN3        | 1QZN   |
| ABF70180.1   | Q1EHA4        | 2Q1K   | AscE                                                                            | <i>Aeromonas hydrophila</i>                                                                                             | BAA11175.1 | P15412        | 1RDV   |
| ABK38882.1   | A0KF03        | 3PSS   | Qnr                                                                             | <i>Aeromonas hydrophila</i> subsp. <i>hydrophila</i> (strain ATCC 7966 / DSM 30187 / JCM 1027 / KCTC 2358 / NCIMB 9240) | AAP59551.1 | O75351        | 1WR0   |
| CAA40386.1   | P26918        | 1X8G   | Beta-lactamase (EC 3.5.2.6)                                                     | <i>Aeromonas hydrophila</i>                                                                                             | AAC00348.1 | P13792        | 1MV O  |
| ABK37415.1   | A0KLG5        | 4BF5   | Alanine racemase (EC 5.1.1.1)                                                   | <i>Aeromonas hydrophila</i> subsp. <i>hydrophila</i> (strain ATCC 7966 / DSM 30187 / JCM 1027 / KCTC 2358 / NCIMB 9240) | AAL08288.1 | O49453        | 1WJJ   |
| ABK36160.1   | A0KH11        | 4BHY   | Alanine racemase (EC 5.1.1.1)                                                   | <i>Aeromonas hydrophila</i> subsp. <i>hydrophila</i> (strain ATCC 7966 / DSM 30187 / JCM 1027 / KCTC 2358 / NCIMB 9240) | BAE22940.1 | Q91YJ5        | 2CRV   |
| ABK35905.1   | A0KEJ7        | 4DZG   | Uncharacterized protein                                                         | <i>Aeromonas hydrophila</i> subsp. <i>hydrophila</i> (strain ATCC 7966 / DSM 30187 / JCM 1027 / KCTC 2358 / NCIMB 9240) | AAC74349.2 | P0AFR4        | 1K7J   |
| AAX19187.1   | Q49TP5        | 3NTS   | VsdC                                                                            | <i>Aeromonas hydrophila</i>                                                                                             | AAC21773.1 | P35755        | 1D9V   |
| ABK35821.1   | A0KHJ5        | 3OD9   | Putative exported protein                                                       | <i>Aeromonas hydrophila</i> subsp. <i>hydrophila</i> (strain ATCC 7966 / DSM 30187 / JCM 1027 / KCTC 2358 /             | CAA45737.1 | P0A780        | 1EY1   |

|            |        |      |                                                                                                             |                                                                                     |            |        |      |
|------------|--------|------|-------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|------------|--------|------|
|            |        |      |                                                                                                             | NCIMB 9240)                                                                         |            |        |      |
| CAL10074.1 | P31490 | 1N5B | YopE regulator                                                                                              | <i>Yersinia enterocolitica</i> serotype O:8 / biotype 1B (strain NCTC 13174 / 8081) | CAD98427.1 | Q7YYQ3 | 2AIF |
| CAL11333.1 | A1JLH5 | 1XVX | Iron(III)-binding periplasmic protein                                                                       | <i>Yersinia enterocolitica</i> serotype O:8 / biotype 1B (strain NCTC 13174 / 8081) | AAC76103.1 | P00579 | 1SIG |
| AAN37556.1 | P0C2V9 | 2BHO | Chaperone protein SycT                                                                                      | <i>Yersinia enterocolitica</i>                                                      | AAC75986.1 | P0A8I1 | 1NMN |
| CAB46570.1 | Q9X9I8 | 2FN0 | Salicylate synthetase, Irp9                                                                                 | <i>Yersinia enterocolitica</i>                                                      | AAA98234.1 | P08716 | 1MT0 |
| CAA54865.1 | P31517 | 2J0P | Hemin transport protein HemS                                                                                | <i>Yersinia enterocolitica</i>                                                      | AAA41560.1 | Q02874 | 1YD9 |
| CAL14084.1 | A1JSS7 | 2JD9 | Putative exported protein                                                                                   | <i>Yersinia enterocolitica</i> serotype O:8 / biotype 1B (strain NCTC 13174 / 8081) | AAB02999.1 | P0A6P7 | 1PUI |
| CBX70720.1 | F4MY12 | -    | Uncharacterized protein ybaJ                                                                                | <i>Yersinia enterocolitica</i> W22703                                               | BAA24003.1 | P12915 | 1BEV |
| AAC15064.1 | O68975 | 2UVE | Exopolygalacturonase                                                                                        | <i>Yersinia enterocolitica</i>                                                      | BAE76284.1 | P0ACN4 | 1TF1 |
| CAL14085.1 | A1JSS8 | 2V8I | Periplasmic pectate lyase                                                                                   | <i>Yersinia enterocolitica</i> serotype O:8 / biotype 1B (strain NCTC 13174 / 8081) | CAA24361.1 | P03306 | 1ZBA |
| CAL10784.1 | A1JJJ6 | 2VH2 | Cell division protein FtsQ                                                                                  | <i>Yersinia enterocolitica</i> serotype O:8 / biotype 1B (strain NCTC 13174 / 8081) | CAJ24843.1 | Q3BQX0 | 1XS3 |
| AAD16831.1 | Q56844 | 2V5G | YscU                                                                                                        | <i>Yersinia enterocolitica</i>                                                      | AAA25674.1 | P02904 | 1DCZ |
| AAN37524.1 | P0C2W0 | 3H7X | Adhesin YadA                                                                                                | <i>Yersinia enterocolitica</i>                                                      | EAW73148.1 | P32455 | 1DG3 |
| CAA32086.1 | P31489 | 1P9H | Adhesin YadA                                                                                                | <i>Yersinia enterocolitica</i>                                                      | AAH73755.1 | P13928 | 1W3W |
| CAL10787.1 | A1JJJ9 | 3NZK | UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase (EC 3.5.1.-) (UDP-3-O-acyl-GlcNAc deacetylase) | <i>Yersinia enterocolitica</i> serotype O:8 / biotype 1B (strain NCTC 13174 / 8081) | AAN32614.1 | Q8TD84 | 1VA9 |
| AAA19860.1 | P15273 | 1LYV | Tyrosine-protein phosphatase YopH (EC 3.1.3.48) (Virulence protein)                                         | <i>Yersinia enterocolitica</i>                                                      | AAC73854.1 | P52697 | 1RI6 |
| AAC37021.1 | Q01245 | 4ALZ | Yop proteins translocation protein D                                                                        | <i>Yersinia enterocolitica</i>                                                      | AAH47750.1 | P55055 | 1P8D |
| AAD16814.1 | O87496 | 2VGX | Chaperone SycD (Type III secretion low calcium response chaperone LcrH/SycD)                                | <i>Yersinia enterocolitica</i>                                                      | AAG33028.1 | P24817 | 1CF5 |
| AAD16812.  | P37132 | 4AM9 | Protein YopD                                                                                                | <i>Yersinia enterocolitica</i>                                                      | AAG58084.  | Q8XCU6 | 1N91 |



|            |        |      |                                                                                                                                                                  |                                                                |            |        |      |
|------------|--------|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|------------|--------|------|
| 1          |        |      |                                                                                                                                                                  |                                                                | 1          |        |      |
| CAF25367.1 | Q663P0 | 1JYA | SycE, yerA putative yopE chaperone                                                                                                                               | <i>Yersinia pseudotuberculosis</i> serotype I (strain IP32953) | CAA83942.1 | Q59232 | 1LF1 |
| CAA68609.1 | P08008 | 1L2W | Outer membrane virulence protein YopE                                                                                                                            | <i>Yersinia pseudotuberculosis</i> serotype I (strain IP32953) | CAA10867.1 | O34816 | 1Y7M |
| BAA07528.1 | Q57221 | 1PM4 | -derived mitogen (Superantigen YPMa) (YPM)                                                                                                                       | <i>Yersinia pseudotuberculosis</i>                             | AAA74578.1 | P49863 | 1MZA |
| CNC69406.1 | Q57329 | 1RKX | CDP-D-glucose-4,6-dehydratase (CDP-glucose 4%2C6-dehydratase) (EC 4.2.1.45)                                                                                      | <i>Yersinia pseudotuberculosis</i>                             | AIX97359.1 | O74705 | 1KS4 |
| CAF25344.1 | Q05608 | 2H7O | Protein kinase YpkA (Protein kinase A) (EC 2.7.11.1) (Targeted effector protein kinase)                                                                          | <i>Yersinia pseudotuberculosis</i> serotype I (strain IP32953) | AAB63975.1 | P07267 | 1DP5 |
| CAF28566.1 | Q6EVP2 | 2IJR | Putative uncharacterized protein api92                                                                                                                           | <i>Yersinia pseudotuberculosis</i>                             | AAA17747.1 | P45700 | 1NXC |
| CAH19996.1 | Q66ED7 | 2WWN | Superoxide dismutase [Cu-Zn] (EC 1.15.1.1)                                                                                                                       | <i>Yersinia pseudotuberculosis</i> serotype I (strain IP32953) | AAA50466.1 | P22088 | 1HQD |
| CAH21499.1 | Q66A71 | 2YJH | Probable thiol peroxidase (EC 1.11.1.-)                                                                                                                          | <i>Yersinia pseudotuberculosis</i> serotype I (strain IP32953) | BAA23336.1 | O32449 | 1QTR |
| AAN23066.1 | Q57323 | 3BB8 | CDP-4-keto-6-deoxy-D-glucose-3-dehydrase (CDP-4-keto-6-deoxy-D-glucose-3-dehydratase) (DdhC protein) (Lipopolysaccharide biosynthesis protein RfbH) (EC 2.6.1.-) | <i>Yersinia pseudotuberculosis</i>                             | BAA16561.2 | P0A7G6 | 1AA3 |
| CAH22830.1 | Q665C6 | 3GG4 | Putative carbohydrate kinase                                                                                                                                     | <i>Yersinia pseudotuberculosis</i> serotype I (strain IP32953) | AAA16327.1 | Q07009 | 1DF0 |
| CAH20441.1 | Q66D48 | 3ODG | Purine nucleoside phosphorylase (EC 2.4.2.1) (Inosine-guanosine phosphorylase)                                                                                   | <i>Yersinia pseudotuberculosis</i> serotype I (strain IP32953) | CAA59198.1 | Q50245 | 1L0Q |
| CAF25413.1 | P40296 | 3UEP | Yop proteins translocation protein Q                                                                                                                             | <i>Yersinia pseudotuberculosis</i> serotype I (strain IP32953) | AAK97759.2 | Q8L5J1 | 1RH9 |
| ACA68167.1 | B1JJ73 | 4AIH | Transcriptional regulator SlyA                                                                                                                                   | <i>Yersinia pseudotuberculosis</i> serotype O:3 (strain YPIII) | BAB20416.1 | Q9LBQ9 | 1UG9 |
| CAF25423.1 | Q663I6 | 4D9V | YscD putative type III secretion protein                                                                                                                         | <i>Yersinia pseudotuberculosis</i> serotype I (strain IP32953) | AAA25811.1 | P14756 | 1EZM |
| AAA27632.1 | P11922 | 1CWV | Invasin                                                                                                                                                          | <i>Yersinia pseudotuberculosis</i> serotype I (strain IP32953) | BAE77699.1 | P0ADK6 | 1OYZ |



|            |        |      |                                                                                     |                                                                                               |            |        |      |
|------------|--------|------|-------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|------------|--------|------|
| CAB94934.1 | Q7AZX0 | 4I2V | Uridine phosphorylase (EC 2.4.2.3)                                                  | <i>Yersinia pseudotuberculosis</i>                                                            | CAB63351.1 | Q9U2M7 | 1KT9 |
| CAG34980.1 | Q6ARP5 | 2PGE | Related to N-acylamino acid racemase (MenC)                                         | <i>Desulfotalea psychrophila</i> (strain LSv54 / DSM 12343)                                   | AAA24565.1 | P0A7Y4 | 1F21 |
| CAG36506.1 | Q6AMB9 | 2PW9 | Protein FdhD homolog                                                                | <i>Desulfotalea psychrophila</i> (strain LSv54 / DSM 12343)                                   | BAE77587.1 | P25522 | 1RFL |
| CAG35507.1 | Q6AQ66 | 2UXQ | Isocitrate dehydrogenase [NADP] (EC 1.1.1.42)                                       | <i>Desulfotalea psychrophila</i> (strain LSv54 / DSM 12343)                                   | AAG02249.1 | P68826 | 1LM4 |
| AAZ25891.1 | Q484T9 | 2M89 | Aha1 domain protein                                                                 | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | AAP54412.1 | Q9FXT4 | 1UAS |
| AAZ26870.1 | Q481E4 | 2JR2 | UPF0352 protein CPS_2611                                                            | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | AAF96794.1 | Q9KL51 | 1Y89 |
| AAZ25333.1 | Q47XN7 | 2V27 | Phenylalanine-4-hydroxylase (EC 1.14.16.1)                                          | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | BAG65108.1 | P53582 | 2B3H |
| AAP94017.1 | Q7WVY1 | 3CIA | Cold-active aminopeptidase (Neutral zinc metallopeptidase, M1 family) (EC 3.4.11.-) | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | BAE77597.1 | P0A8Y5 | 1RKQ |
| AAZ24010.1 | Q47UF8 | 3CNB | DNA-binding response regulator, merR family                                         | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | CAA28212.1 | P00720 | 102L |
| AAZ26474.1 | Q483U6 | 3EQZ | Response regulator                                                                  | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | AAF61418.1 | P0A0H5 | 1D6T |
| AAZ24299.1 | Q488E0 | 3IC9 | Putative dihydrolipoamide dehydrogenase                                             | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | AAZ38713.1 | P20701 | 1CQP |
| AAZ25366.1 | Q47UI6 | 3IWG | Acetyltransferase, GNAT family                                                      | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | ABD72193.1 | P13569 | 1NBD |
| AAZ28225.1 | Q484V4 | 3LO3 | Uncharacterized protein                                                             | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | DAA06783.1 | P38804 | 1NYN |
| AAZ25958.1 | Q482G7 | 3LPX | DNA gyrase subunit A (EC 5.99.1.3)                                                  | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | AAT48221.1 | P15043 | 1OYW |
| AAZ24221.1 | Q47X65 | 3LYG | Uncharacterized protein                                                             | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | CAD97681.1 | P78330 | 1L8L |
| AAZ24666.1 | Q486I1 | 3LYX | Sensory box/GGDEF domain protein                                                    | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | BAA06696.1 | Q62132 | 1JLN |
| AAZ25151.1 | Q47Z72 | 3O0Y | Putative lipoprotein                                                                | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | ABH07387.1 | P20646 | 1RPA |
| AAZ24954.1 | Q47ZP9 | 3QTH | Uncharacterized protein                                                             | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | AAK28401.1 | O14727 | 1C15 |
| AAZ27345.1 | Q48A34 | 4O1W | Cytochrome c552                                                                     | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | AAK07481.1 | Q9ALZ1 | 1V5C |

|            |        |      |                                                                                     |                                                                                               |            |        |      |
|------------|--------|------|-------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|------------|--------|------|
| AAZ28655.1 | Q48AL6 | 4PET | Extracellular solute-binding protein, family 7                                      | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | AAA52373.1 | P05198 | 1KL9 |
| AAZ25171.1 | Q489U8 | 4RHE | Aromatic acid decarboxylase (EC 4.1.1.-)                                            | <i>Colwellia psychrerythraea</i> (strain 34H / ATCC BAA-681) ( <i>Vibrio psychroerythus</i> ) | AAL57682.1 | Q9SCY2 | 1U79 |
| CAI86552.1 | Q3IGF7 | 2ZOO | Probable nitrite reductase (EC 1.7.99.3) (EC 1.7.2.1)                               | <i>Pseudoalteromonas haloplanktis</i> (strain TAC 125)                                        | EAK90238.1 | Q5CYJ7 | 2A22 |
| CAI86290.1 | P84612 | 3LIO | Superoxide dismutase [Fe] (EC 1.15.1.1)                                             | <i>Pseudoalteromonas haloplanktis</i> (strain TAC 125)                                        | CAA97363.1 | P36623 | 1FZT |
| CAI86460.1 | Q3IL66 | 3LS2 | Putative esterase (EC 3.1.1.-)                                                      | <i>Pseudoalteromonas haloplanktis</i> (strain TAC 125)                                        | CAB51442.1 | P16250 | 1VZW |
| ABE52840.1 | Q12UN6 | 3AQ1 | Thermosome subunit                                                                  | <i>Methanococcoides burtonii</i> (strain DSM 6242 / NBRC 107633 / OCM 468 / ACE-M)            | AAX93766.1 | P47244 | 1KFI |
| AAA67362.1 | Q44048 | 1F1R | 3,4-dihydroxyphenylacetate 2,3-dioxygenase                                          | <i>Arthrobacter globiformis</i>                                                               | AAC73774.1 | P00962 | 1EUQ |
| BAB20416.1 | Q9LBQ9 | 1UG9 | Glucodextranase                                                                     | <i>Arthrobacter globiformis</i>                                                               | AAC73255.3 | P27305 | 1NZJ |
| AAK16482.1 | Q9AGP8 | 1PJ5 | Dimethylglycine oxidase (DMGO) (EC 1.5.3.10)                                        | <i>Arthrobacter globiformis</i>                                                               | AAC43103.1 | P15640 | 1GSO |
| AAA18114.1 | P46881 | 1AV4 | Phenylethylamine oxidase (EC 1.4.3.21) (Primary amine oxidase)                      | <i>Arthrobacter globiformis</i>                                                               | AAA88210.1 | P15925 | 1FGS |
| AAS99880.1 | Q7X2H8 | 2JBV | Choline oxidase (EC 1.1.3.17)                                                       | <i>Arthrobacter globiformis</i>                                                               | AAD03608.1 | P81294 | 1PXZ |
| ABM01916.1 | A1SR01 | 4HS5 | Protein CyaY                                                                        | <i>Psychromonas ingrahamii</i> (strain 37)                                                    | BAA14792.1 | P0A877 | 1V7Y |
| ABM03255.1 | A1SUU0 | 4P3M | Serine hydroxymethyltransferase (SHMT) (Serine methylase) (EC 2.1.2.1)              | <i>Psychromonas ingrahamii</i> (strain 37)                                                    | AAB09590.1 | Q94734 | 1D2U |
| ABL74378.1 | A1YIY3 | 3P4G | Antifreeze protein                                                                  | <i>Marinomonas primoryensis</i>                                                               | BAB21059.1 | Q9AQS0 | 1J0M |
| AAZ19911.1 | Q4FPZ7 | 2I9W | Uncharacterized protein                                                             | <i>Psychrobacter arcticus</i> (strain DSM 17307 / 273-4)                                      | AAB38270.1 | P31016 | 1BE9 |
| AAZ18427.1 | Q4FU81 | 2PN2 | Uncharacterized protein                                                             | <i>Psychrobacter arcticus</i> (strain DSM 17307 / 273-4)                                      | AAA86659.1 | Q12718 | 1GYC |
| AAZ18059.1 | Q4FV99 | 2RE7 | Uncharacterized protein                                                             | <i>Psychrobacter arcticus</i> (strain DSM 17307 / 273-4)                                      | CAA25409.1 | P00374 | 1BOZ |
| AAZ17906.1 | Q4FVQ2 | 3HN7 | Putative UDP-N-acetylmuramate:L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase | <i>Psychrobacter arcticus</i> (strain DSM 17307 / 273-4)                                      | AAC76786.2 | P0AA25 | 1F6M |
| AAZ19965.1 | Q4FPU3 | 3IF2 | Aminotransferase (EC 2.6.1.-)                                                       | <i>Psychrobacter arcticus</i> (strain DSM 17307 / 273-4)                                      | AAL90879.1 | Q8RN03 | 1UED |
| AAZ18050.1 | Q4FVA8 | 3KE3 | Possible serine-pyruvate aminotransferase                                           | <i>Psychrobacter arcticus</i> (strain DSM 17307 / 273-4)                                      | AAA83534.1 | P06632 | 1A80 |

|            |        |      |                                                                                                                                        |                                                          |            |        |      |
|------------|--------|------|----------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------|------------|--------|------|
| AAZ19650.1 | Q4FQQ8 | 3PF0 | Uncharacterized protein                                                                                                                | <i>Psychrobacter arcticus</i> (strain DSM 17307 / 273-4) | AAA50163.1 | Q02127 | 1D3G |
| ABE73807.1 | Q1QEU6 | 4NS4 | Alpha/beta hydrolase fold protein                                                                                                      | <i>Psychrobacter cryohalolentis</i> (strain K5)          | AAF28842.1 | Q9CQM9 | 1WIK |
| ABQ94276.1 | A5WF35 | 4QBL | VRR-NUC                                                                                                                                | <i>Psychrobacter</i> sp. (strain PRwf-1)                 | AAA03078.1 | Q59490 | 1L1L |
| ACT35272.1 | C7E5X4 | 4RKC | Aromatic amino acid aminotransferase (EC 2.6.1.57)                                                                                     | <i>Psychrobacter</i> sp. B6                              | AAA23381.1 | P00269 | 1RB9 |
| AAD13225.1 | Q9ZF99 | 1B8P | Malate dehydrogenase (EC 1.1.1.37)                                                                                                     | <i>Aquaspirillum arcticum</i>                            | AAA29633.1 | Q27743 | 1CEQ |
| AAA88910.1 | P50921 | 1AW1 | Triosephosphate isomerase (TIM) (EC 5.3.1.1) (Triose-phosphate isomerase)                                                              | <i>Moritella marina</i> ( <i>Vibrio marinus</i> )        | AAC35399.1 | O87899 | 1YX3 |
| CAM88673.1 | B1VBB0 | 4HMC | Chitinase 60 (EC 3.2.1.14)                                                                                                             | <i>Moritella marina</i> ( <i>Vibrio marinus</i> )        | CAA42574.1 | P29476 | 1B8Q |
| AAC62692.1 | O74049 | 2M1I | Parvulin-like peptidyl-prolyl isomerase (EC 5.2.1.8) (Peptidyl-prolyl cis/trans isomerase)                                             | <i>Cenarchaeum symbiosum</i> (strain A)                  | CAI42779.1 | O95831 | 1M6I |
| CAB37062.1 | Q9Z4P0 | 1QO8 | Fumarate reductase flavoprotein subunit (EC 1.3.5.4) (Fe(3+)-induced flavocytochrome C3) (Ifc3) (Iron(III)-induced flavocytochrome C3) | <i>Shewanella frigidimarina</i> (strain NCIMB 400)       | AAA64512.1 | P11182 | 1K8M |
| CAB38558.1 | P0C278 | 1E39 | Fumarate reductase flavoprotein subunit (EC 1.3.5.4) (Flavocytochrome c) (Flavocytochrome c3) (Fcc3)                                   | <i>Shewanella frigidimarina</i>                          | AAB84388.1 | O14792 | 1ZRH |
| ABI71355.1 | O33731 | 2K3V | Tetraheme cytochrome c-type (Cytochrome c3)                                                                                            | <i>Shewanella frigidimarina</i> (strain NCIMB 400)       | AAA23610.1 | P69783 | 1F3G |
| ABI70437.1 | Q087X8 | 2OOK | Uncharacterized protein                                                                                                                | <i>Shewanella frigidimarina</i> (strain NCIMB 400)       | AAC74185.1 | P69786 | 1IBA |
| ABI72942.1 | Q07YH2 | 2PFW | Cupin 2, conserved barrel domain protein                                                                                               | <i>Shewanella frigidimarina</i> (strain NCIMB 400)       | CCP44148.1 | P9WKE9 | 1S4Q |
| ABI71165.1 | Q085A0 | 2PYX | Tryptophan halogenase                                                                                                                  | <i>Shewanella frigidimarina</i> (strain NCIMB 400)       | AAC97955.1 | Q9UNQ2 | 1ZQ9 |
| ABI71819.1 | Q082J6 | 3BB9 | Putative orphan protein                                                                                                                | <i>Shewanella frigidimarina</i> (strain NCIMB 400)       | BAA36069.1 | P0A7I0 | 2B3T |
| ABI73370.1 | Q07X94 | 3D82 | Cupin 2, conserved barrel domain protein                                                                                               | <i>Shewanella frigidimarina</i> (strain NCIMB 400)       | AAC74296.1 | P0ACC1 | 1T43 |
| ABI71883.1 | Q082D2 | 3FGE | Uncharacterized protein                                                                                                                | <i>Shewanella frigidimarina</i> (strain NCIMB 400)       | CAB12110.2 | O31474 | 1XXL |

|            |        |      |                                                                                                                               |                                                                           |            |        |      |
|------------|--------|------|-------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------|------------|--------|------|
| ABI73132.1 | Q07XY2 | 3LHO | Uncharacterized protein                                                                                                       | <i>Shewanella frigidimarina</i> (strain NCIMB 400)                        | AAH07069.1 | Q969G6 | 1NB0 |
| ABI70397.1 | Q088B8 | 3LWU | Succinylglutamate desuccinylase/aspartoacylase                                                                                | <i>Shewanella frigidimarina</i> (strain NCIMB 400)                        | BAH13786.1 | P55263 | 1BX4 |
| BAL45195.1 | H1AFK5 | 3VND | Tryptophan synthase alpha chain (EC 4.2.1.20)                                                                                 | <i>Shewanella frigidimarina</i>                                           | AAK82464.1 | O50008 | 1U1H |
| ABI72749.1 | Q07Z15 | 4CX9 | Cytochrome c, class II                                                                                                        | <i>Shewanella frigidimarina</i> (strain NCIMB 400)                        | BAA00518.1 | P21404 | 1D4M |
| AAL01059.1 | Q93CG9 | 3R87 | Putative uncharacterized protein                                                                                              | <i>Photobacterium profundum</i> ( <i>Photobacterium</i> sp. (strain SS9)) | CAA50551.1 | P10538 | 1BFN |
| CAG19435.1 | Q6LTE1 | 4K46 | Adenylate kinase (AK) (EC 2.7.4.3) (ATP-AMP transphosphorylase) (ATP:AMP phosphotransferase) (Adenylate monophosphate kinase) | <i>Photobacterium profundum</i> ( <i>Photobacterium</i> sp. (strain SS9)) | ABD96827.1 | P53355 | 1IG1 |
| CAA83122.1 | P41365 | 1LBS | Lipase B (EC 3.1.1.3) (CALB)                                                                                                  | <i>Candida antarctica</i> (Yeast) ( <i>Trichosporon oryzae</i> )          | EAW51291.1 | Q9Y3D8 | 1RKB |

Table A1.3: Thermophile-Psychrophile (T-P) dataset.

| Thermophilic CDS |               |        |                                                                 |                                                                                                                                                                  | Psychrophilic CDS |              |        |
|------------------|---------------|--------|-----------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|--------------|--------|
| ENA ID           | UniProt KB ID | PDB ID | Protein name                                                    | Organism name                                                                                                                                                    | ENA ID            | UniProtKB ID | PDB ID |
| AAD36062.1       | Q9X078        | 1JDQ   | UPF0033 protein TM_0983                                         | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | CAD00476.1        | Q8Y4N7       | 1Y9I   |
| AAD35268.1       | Q9WY19        | 1VKU   | Acyl carrier protein (ACP)                                      | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAP78483.1        | Q7X0F0       | 1Y7Y   |
| AAB85436.1       | O27021        | 1IHN   | Uncharacterized protein                                         | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | ABK39236.1        | A0KKT0       | 2LGH   |
| CAA40902.1       | P15452        | 1AYG   | Cytochrome c-552 (Cytochrome c552)                              | <i>Hydrogenobacter thermophilus</i> (strain DSM 6534 / IAM 12695 / TK-6)                                                                                         | AAA21938.1        | P09167       | 1PRE   |
| AAC07338.1       | O67367        | 1OZ9   | Endoribonuclease YbeY (EC 3.1.-.-)                              | <i>Aquifex aeolicus</i> (strain VF5)                                                                                                                             | ABK37325.1        | A0KPA6       | 3L6E   |
| AAB07763.1       | P71143        | 2BM3   | Scaffolding dockerin binding protein A                          | <i>Clostridium thermocellum</i> ( <i>Ruminiclostridium thermocellum</i> )                                                                                        | ABF70178.1        | Q1EHA2       | 3PH0   |
| AAB03727.1       | Q56313        | 1IN4   | Holliday junction ATP-dependent DNA helicase RuvB (EC 3.6.4.12) | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | ABF70180.1        | Q1EHA4       | 2Q1K   |
| AAD35484.1       | Q9WYN0        | 1KGS   | Response regulator                                              | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | ABK38882.1        | A0KF03       | 3PSS   |
| AAK42515.1       | Q97W73        | 1O7I   | Single-stranded DNA binding protein Ssb (SSB)                   | <i>Sulfolobus solfataricus</i> (strain ATCC 35092 / DSM 1617 / JCM 11322 / P2)                                                                                   | CAA40386.1        | P26918       | 1X8G   |
| CAA27987.1       | P04766        | 1D1N   | Translation initiation factor IF-2                              | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | ABK37415.1        | A0KLG5       | 4BF5   |
| AAB86164.1       | O27727        | 1JCU   | Conserved protein                                               | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | ABK36160.1        | A0KH11       | 4BHY   |
| CAB65651.1       | Q9RHZ6        | 1URD   | Maltose binding protein                                         | <i>Alicyclobacillus acidocaldarius</i> ( <i>Bacillus acidocaldarius</i> )                                                                                        | ABK35905.1        | A0KEJ7       | 4DZG   |
| AAD36829.1       | Q9X286        | 1TZT   | N utilization substance protein B homolog (Protein NusB)        | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAX19187.1        | Q49TP5       | 3NTS   |
| AAB99207.1       | P54066        | 1RA4   | 50S ribosomal protein L7Ae (Ribosomal protein L8e)              | <i>Methanocaldococcus jannaschii</i> (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440) ( <i>Methanococcus jannaschii</i> )                        | ABK35821.1        | A0KHJ5       | 3OD9   |



|            |        |      |                                                                                              |                                                                                                                                                                  |            |        |      |
|------------|--------|------|----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
| AAC44889.2 | P77994 | 1TTY | RNA polymerase sigma factor SigA (Sigma-A)                                                   | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | CAL10074.1 | P31490 | 1N5B |
| BAD71652.1 | Q5SHA1 | 1IV0 | Putative Holliday junction resolvase (EC 3.1.-.-)                                            | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                                                                                  | CAL11333.1 | A1JLH5 | 1XVX |
| BAD71661.1 | Q5SH92 | 2D2E | SufC protein (ATP-binding protein)                                                           | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                                                                                  | AAN37556.1 | P0C2V9 | 2BHO |
| AAB89725.1 | O28751 | 1HJZ | [Protein ADP-ribosylglutamate] hydrolase AF_1521 (EC 3.2.2.-)                                | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)                                                                    | CAB46570.1 | Q9X9I8 | 2FN0 |
| AAC07832.1 | O67859 | 1T6T | Uncharacterized protein                                                                      | <i>Aquifex aeolicus</i> (strain VF5)                                                                                                                             | CAA54865.1 | P31517 | 2J0P |
| BAA29671.1 | O58335 | 1UZ5 | 402aa long hypothetical molybdopterin biosynthesis moea protein                              | <i>Pyrococcus horikoshii</i> (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)                                                                    | CAL14084.1 | A1JSS7 | 2JD9 |
| AAB89817.1 | O28835 | 1Y8A | Uncharacterized protein                                                                      | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)                                                                    | CBX70720.1 | F4MY12 | -    |
| AAB91265.1 | O30273 | 1YTL | Acetyl-CoA decarboxylase/synthase complex subunit epsilon 2 (ACDS complex subunit epsilon 2) | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)                                                                    | AAC15064.1 | O68975 | 2UVE |
| AAB86213.1 | O27775 | 1RYJ | Uncharacterized protein                                                                      | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | CAL14085.1 | A1JSS8 | 2V8I |
| AAD35304.1 | Q9WY55 | 1ZKO | Glycine cleavage system H protein                                                            | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | CAL10784.1 | A1JJJ6 | 2VH2 |
| AAC07491.1 | O67517 | 1LFP | Probable transcriptional regulatory protein aq_1575                                          | <i>Aquifex aeolicus</i> (strain VF5)                                                                                                                             | AAD16831.1 | Q56844 | 2V5G |
| AAB90746.1 | O29759 | 1P9Q | Ribosome maturation protein SDO1 homolog                                                     | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)                                                                    | AAN37524.1 | P0C2W0 | 3H7X |
| AAB86171.1 | O27734 | 1GH8 | Elongation factor 1-beta (EF-1-beta) (aEF-1beta)                                             | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | CAA32086.1 | P31489 | 1P9H |
| AAD36689.1 | Q9X1V7 | 1VR8 | Uncharacterized protein                                                                      | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | CAL10787.1 | A1JJJ9 | 3NZK |
| BAK54261.1 | Q975N2 | 1UD9 | DNA polymerase sliding clamp 1 (Proliferating cell nuclear antigen homolog 1) (PCNA1)        | <i>Sulfolobus tokodaii</i> (strain DSM 16993 / JCM 10545 / NBRC 100140 / 7)                                                                                      | AAA19860.1 | P15273 | 1LYV |



|            |        |      |                                                                                                                                                    |                                                                                                                                                                  |            |        |      |
|------------|--------|------|----------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
| AAM23944.1 | Q8RBX6 | 1U4H | Methyl-accepting chemotaxis protein                                                                                                                | <i>Caldanaerobacter subterraneus</i> subsp. <i>tengcongensis</i> (strain DSM 15242 / JCM 11007 / NBRC 100824 / MB4) ( <i>Thermoanaerobacter tengcongensis</i> )  | AAC37021.1 | Q01245 | 4ALZ |
| AAD36040.1 | Q9X056 | 2ETD | LemA protein                                                                                                                                       | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAD16814.1 | O87496 | 2VGX |
| AAB85143.1 | O26734 | 1JRM | UPF0235 protein MTH_637                                                                                                                            | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | AAD16812.1 | P37132 | 4AM9 |
| CAA06924.1 | Q9ZF13 | 1BQC | Beta-mannanase (EC 3.2.1.78)                                                                                                                       | <i>Thermobifida fusca</i> ( <i>Thermomonospora fusca</i> )                                                                                                       | CAF25367.1 | Q663P0 | 1JYA |
| AAC06388.1 | P26222 | 1TML | Endoglucanase E-2 (EC 3.2.1.4) (Cellulase E-2) (Cellulase E2) (Endo-1,4-beta-glucanase E-2)                                                        | <i>Thermobifida fusca</i> ( <i>Thermomonospora fusca</i> )                                                                                                       | CAA68609.1 | P08008 | 1L2W |
| AAF24127.1 | P23360 | 1GOK | Endo-1,4-beta-xylanase (Xylanase) (EC 3.2.1.8) (1,4-beta-D-xylan xylanohydrolase) (TAXI)                                                           | <i>Thermoascus aurantiacus</i>                                                                                                                                   | BAA07528.1 | Q57221 | 1PM4 |
| AAM77714.2 | Q8NJV3 | 1OLR | Endoglucanase                                                                                                                                      | <i>Humicola grisea</i>                                                                                                                                           | CNC69406.1 | Q57329 | 1RKX |
| CAA29568.1 | P09177 | 1MPP | Mucorpepsin (EC 3.4.23.23) (Mucor rennin)                                                                                                          | <i>Rhizomucor pusillus</i>                                                                                                                                       | CAF25344.1 | Q05608 | 2H7O |
| AAD45520.2 | Q9XBQ3 | 1PZ2 | Intracellular exo-alpha-(1->5)-L-arabinofuranosidase (ABF) (EC 3.2.1.55) (Intracellular arabinan exo-alpha-(1->5)-L-arabinosidase) (Arabinosidase) | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | CAF28566.1 | Q6EVP2 | 2IJR |
| AAC12257.1 | O66015 | 1KU0 | Lipase (EC 3.1.1.3)                                                                                                                                | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | CAH19996.1 | Q66ED7 | 2WWN |
| CAC11959.1 | P96084 | 1MT3 | Proline iminopeptidase (PIP) (EC 3.4.11.5) (Prolyl aminopeptidase) (PAP) (Tricorn protease-interacting factor F1)                                  | <i>Thermoplasma acidophilum</i> (strain ATCC 25905 / DSM 1728 / JCM 9062 / NBRC 15155 / AMRC-C165)                                                               | CAH21499.1 | Q66A71 | 2YJH |
| AAC45089.1 | O05209 | 1CZ4 | VCP-like ATPase                                                                                                                                    | <i>Thermoplasma acidophilum</i> (strain ATCC 25905 / DSM 1728 / JCM 9062 / NBRC 15155 / AMRC-C165)                                                               | AAN23066.1 | Q57323 | 3BB8 |
| BAB64339.1 | Q93HT9 | 1WL7 | Intracellular endo-alpha-(1->5)-L-arabinanase (ABN) (EC 3.2.1.99) (Endo-1,5-alpha-L-arabinanase)                                                   | <i>Geobacillus thermodenitrificans</i>                                                                                                                           | CAH22830.1 | Q665C6 | 3GG4 |
| BAA12070.1 | P71140 | 2C24 | Endoglucanase J                                                                                                                                    | <i>Clostridium thermocellum</i> ( <i>Ruminiclostridium thermocellum</i> )                                                                                        | CAH20441.1 | Q66D48 | 3ODG |

|            |        |      |                                                                                                                                                                            |                                                                                               |            |        |      |
|------------|--------|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|------------|--------|------|
| ABK52387.1 | P54583 | 1ECE | Endoglucanase E1 (EC 3.2.1.4) (Cellulase E1) (Endo-1,4-beta-glucanase E1) (Endocellulase E1)                                                                               | <i>Acidothermus cellulolyticus</i> (strain ATCC 43068 / 11B)                                  | CAF25413.1 | P40296 | 3UEP |
| CAA54291.1 | P00800 | 1FJ3 | Thermolysin (EC 3.4.24.27) (Thermostable neutral proteinase)                                                                                                               | <i>Bacillus thermoproteolyticus</i>                                                           | ACA68167.1 | B1JJ73 | 4AIH |
| AAC44232.1 | P77847 | 1PMH | Beta-1,4-mannanase (EC 3.2.1.78)                                                                                                                                           | <i>Caldicellulosiruptor saccharolyticus</i> ( <i>Caldocellum saccharolyticum</i> )            | CAF25423.1 | Q663I6 | 4D9V |
| BAD18071.1 | Q75UV1 | 1VC8 | Diadenosine hexaphosphate hydrolase (Ap6A hydrolase) (EC 3.6.1.61) (ATP-generating (di)nucleotide polyphosphate hydrolase) (ATP-generating Ap6A hydrolase) (Nudix protein) | <i>Thermus thermophilus</i>                                                                   | AAA27632.1 | P11922 | 1CWV |
| AAA24565.1 | P0A7Y4 | 1F21 | Ribonuclease HI (RNase HI) (EC 3.1.26.4) (Ribonuclease H) (RNase H)                                                                                                        | <i>Escherichia coli</i> (strain K12)                                                          | CAB94934.1 | Q7AZX0 | 4I2V |
| CAA43026.1 | P29253 | 1JL2 | Ribonuclease H (RNase H) (EC 3.1.26.4)                                                                                                                                     | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                               | CAG34980.1 | Q6ARP5 | 2PGE |
| BAD69943.1 | Q5SM23 | 1WF3 | GTPase Era                                                                                                                                                                 | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                               | CAG36506.1 | Q6AMB9 | 2PW9 |
| CAA55695.1 | P43522 | 1V3Y | Peptide deformylase (PDF) (EC 3.5.1.88) (Polypeptide deformylase)                                                                                                          | <i>Thermus thermophilus</i>                                                                   | CAG35507.1 | Q6AQ66 | 2UXQ |
| AGL50123.1 | O33835 | 1ZY9 | Alpha-galactosidase (EC 3.2.1.22)                                                                                                                                          | <i>Thermotoga maritima</i>                                                                    | AAZ25891.1 | Q484T9 | 2M89 |
| AAD36230.1 | Q9X0N8 | 1PBT | 6-phosphogluconolactonase (6PGL) (EC 3.1.1.31)                                                                                                                             | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | AAZ26870.1 | Q481E4 | 2JR2 |
| AAD36544.1 | Q9X1I7 | 1O0X | Methionine aminopeptidase (MAP) (MetAP) (EC 3.4.11.18) (Peptidase M)                                                                                                       | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | AAZ25333.1 | Q47XN7 | 2V27 |
| AGL50674.1 | Q9X264 | 1PW5 | NagD protein, putative (Putative NagD-like phosphatase)                                                                                                                    | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | AAP94017.1 | Q7WVY1 | 3CIA |
| AGL49592.1 | Q9WZD5 | 1J6O | Putative deoxyribonuclease YcfH (Uncharacterized protein)                                                                                                                  | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | AAZ24010.1 | Q47UF8 | 3CNB |
| AAD36531.1 | Q9X1H4 | 1NZ0 | Ribonuclease P protein component (RNase P protein) (RNaseP protein) (EC 3.1.26.5) (Protein C5)                                                                             | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | AAZ26474.1 | Q483U6 | 3EQZ |
| AAB89058.1 | O28085 | 1OI0 | Uncharacterized protein                                                                                                                                                    | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126) | AAZ24299.1 | Q488E0 | 3IC9 |
| AAB90868.1 | O29883 | 1Z0B | Archaeal Lon protease (EC 3.4.21.-) (ATP-dependent protease La homolog)                                                                                                    | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126) | AAZ25366.1 | Q47UI6 | 3IWG |

|            |        |       |                                                                                                                                                                                                                             |                                                                                                                                           |            |        |      |
|------------|--------|-------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
| AAC06997.1 | O67040 | 1T6C  | Exopolyphosphatase                                                                                                                                                                                                          | <i>Aquifex aeolicus</i> (strain VF5)                                                                                                      | AAZ28225.1 | Q484V4 | 3LO3 |
| AAB99612.1 | Q58989 | 1F5S  | Phosphoserine phosphatase (PSP) (PSPase) (EC 3.1.3.3) (O-phosphoserine phosphohydrolase)                                                                                                                                    | <i>Methanocaldococcus jannaschii</i> (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440) ( <i>Methanococcus jannaschii</i> ) | AAZ25958.1 | Q482G7 | 3LPX |
| BAA30713.1 | O59248 | 1X0T  | Ribonuclease P protein component 4 (RNase P component 4) (EC 3.1.26.5) (Rpp21)                                                                                                                                              | <i>Pyrococcus horikoshii</i> (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)                                             | AAZ24221.1 | Q47X65 | 3LYG |
| AAD54338.1 | Q7LYT7 | 1MW O | Alpha amylase (EC 3.2.1.1)                                                                                                                                                                                                  | <i>Pyrococcus woesei</i>                                                                                                                  | AAZ24666.1 | Q486I1 | 3LYX |
| AAL80661.1 | Q8U3D2 | 1U04  | Uncharacterized protein                                                                                                                                                                                                     | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                | AAZ25151.1 | Q47Z72 | 3O0Y |
| AAL80564.1 | E7FHX6 | 1DQ3  | Vitamin B12-dependent ribonucleoside-diphosphate reductase (B12-dependent RNR) (EC 1.17.4.1) (Ribonucleotide reductase) [Cleaved into: Endonuclease PI-PfuI (EC 3.1.-.-) (Pfu rnr-1 intein); Pfu rnr-2 intein (EC 3.1.-.-)] | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                | AAZ24954.1 | Q47ZP9 | 3QTH |
| AAD26328.1 | Q9X519 | 1EJJ  | 2,3-bisphosphoglycerate-independent phosphoglycerate mutase (23PGA-independent) (BPG-independent PGAM) (Phosphoglyceromutase) (iPGM) (EC 5.4.2.12)                                                                          | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                              | AAZ27345.1 | Q48A34 | 4O1W |
| BAA24446.1 | O52980 | 1IX5  | FKBP-type peptidyl-prolyl cis-trans isomerase (PPIase) (EC 5.2.1.8) (MtFK) (Rotamase)                                                                                                                                       | <i>Methanothermococcus thermolithotrophicus</i> ( <i>Methanococcus thermolithotrophicus</i> )                                             | AAZ28655.1 | Q48AL6 | 4PET |
| AAO40746.1 | Q84FH6 | 1ODE  | Chorismate mutase AroH (EC 5.4.99.5)                                                                                                                                                                                        | <i>Thermus thermophilus</i>                                                                                                               | AAZ25171.1 | Q489U8 | 4RHE |
| BAD71845.1 | Q53WB3 | 1V37  | Alpha-ribazole-5'-phosphate phosphatase                                                                                                                                                                                     | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                                                           | CAI86552.1 | Q3IGF7 | 2ZOO |
| AAD36114.1 | Q9X0C7 | 1QO2  | 1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino]imidazole-4-carboxamide isomerase (EC 5.3.1.16) (Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase)                                   | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                              | CAI86290.1 | P84612 | 3LIO |
| BAA31042.1 | O59580 | 1V7R  | Non-canonical purine NTP pyrophosphatase (EC 3.6.1.19) (Non-standard purine NTP pyrophosphatase)                                                                                                                            | <i>Pyrococcus horikoshii</i> (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)                                             | CAI86460.1 | Q3IL66 | 3LS2 |

|            |        |      |                                                                                                                                                            |                                                                                                    |            |        |      |
|------------|--------|------|------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|------------|--------|------|
|            |        |      | (Nucleoside-triphosphate diphosphatase)<br>(Nucleoside-triphosphate pyrophosphatase)<br>(NTPase)                                                           |                                                                                                    |            |        |      |
| CAC11654.1 | Q9HKT1 | 2ARS | Lipoate-protein ligase A subunit 1 (EC 2.7.7.63) (Lipoate--protein ligase subunit 1)                                                                       | <i>Thermoplasma acidophilum</i> (strain ATCC 25905 / DSM 1728 / JCM 9062 / NBRC 15155 / AMRC-C165) | ABE52840.1 | Q12UN6 | 3AQ1 |
| BAD70261.1 | P27000 | 1G59 | Glutamate--tRNA ligase (EC 6.1.1.17) (Glutamyl-tRNA synthetase) (GluRS)                                                                                    | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                    | AAA67362.1 | Q44048 | 1F1R |
| BAD69921.1 | Q5SM45 | 1IQ0 | Arginine--tRNA ligase (EC 6.1.1.19) (Arginyl-tRNA synthetase) (ArgRS)                                                                                      | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                    | BAB20416.1 | Q9LBQ9 | 1UG9 |
| CAC39622.1 | Q93RP5 | 1IQ0 | Arginine--tRNA ligase (EC 6.1.1.19) (Arginyl-tRNA synthetase)                                                                                              | <i>Thermus thermophilus</i>                                                                        | AAK16482.1 | Q9AGP8 | 1PJ5 |
| AAD36325.1 | Q9X0X7 | 1VKZ | Phosphoribosylamine--glycine ligase (EC 6.3.4.13) (GARS) (Glycinamide ribonucleotide synthetase) (Phosphoribosylglycinamide synthetase)                    | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                       | AAA18114.1 | P46881 | 1AV4 |
| AAD35259.1 | Q9WY13 | 1O5Z | Folylpolyglutamate synthase/dihydrofolate synthase                                                                                                         | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                       | AAS99880.1 | Q7X2H8 | 2JBV |
| BAB40336.1 | Q9AJM4 | 1VBL | Pectate lyase 47                                                                                                                                           | <i>Bacillus</i> sp. TS-47                                                                          | ABM01916.1 | A1SR01 | 4HS5 |
| AAS81077.1 | P16608 | 1UJP | Tryptophan synthase alpha chain (EC 4.2.1.20)                                                                                                              | <i>Thermus thermophilus</i> (strain HB27 / ATCC BAA-163 / DSM 7039)                                | ABM03255.1 | A1SUU0 | 4P3M |
| AAC07100.1 | O67135 | 1C3P | Acetoin utilization protein                                                                                                                                | <i>Aquifex aeolicus</i> (strain VF5)                                                               | ABL74378.1 | A1YIY3 | 3P4G |
| BAA30907.1 | O59452 | 1WR2 | Putative uncharacterized protein PH1788                                                                                                                    | <i>Pyrococcus horikoshii</i> (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)      | AAZ19911.1 | Q4FPZ7 | 2I9W |
| AAL64050.1 | Q8ZVK6 | 1XQO | N-glycosylase/DNA lyase (8-oxoguanine DNA glycosylase) (EC 3.2.2.-) (AGOG) (DNA-(apurinic or apyrimidinic site) lyase) (AP lyase) (EC 4.2.99.18) (Pa-AGOG) | <i>Pyrobaculum aerophilum</i> (strain ATCC 51768 / IM2 / DSM 7523 / JCM 9630 / NBRC 100827)        | AAZ18427.1 | Q4FU81 | 2PN2 |
| CAE00180.1 | Q70KY3 | 1GW0 | Laccase-1 (EC 1.10.3.2) (Benzenediol:oxygen oxidoreductase 1) (Diphenol oxidase 1) (Ligninolytic phenoloxidase) (Urishiol oxidase 1)                       | <i>Melanocarpus albomyces</i>                                                                      | AAZ18059.1 | Q4FV99 | 2RE7 |
| BAD76060.1 | Q5KZ26 | 1ZDR | Dihydrofolate reductase (EC 1.5.1.3)                                                                                                                       | <i>Geobacillus kaustophilus</i> (strain HTA426)                                                    | AAZ17906.1 | Q4FVQ2 | 3HN7 |
| BAD71304.1 | Q5SI93 | 1V98 | Thioredoxin                                                                                                                                                | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                    | AAZ19965.1 | Q4FPU3 | 3IF2 |
| BAD71988.1 | Q53WG9 | 1WJ9 | CRISPR-associated endoribonuclease Cse3                                                                                                                    | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                    | AAZ18050.1 | Q4FVA8 | 3KE3 |

|            |        |      |                                                                                                                                                                                               |                                                                                                                                                                  |            |        |      |
|------------|--------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
|            |        |      | (EC 3.1.-.-) (Cse3 endoRNase) (Cse3 endoribonuclease)                                                                                                                                         | 27634 / DSM 579)                                                                                                                                                 |            |        |      |
| AAD36074.1 | Q9X0A2 | 1VP5 | Oxidoreductase, aldo/keto reductase family                                                                                                                                                    | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | AAZ19650.1 | Q4FQQ8 | 3PF0 |
| AAD35190.1 | Q9WXV1 | 1VHN | tRNA-dihydrouridine synthase (EC 1.-.-.-)                                                                                                                                                     | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | ABE73807.1 | Q1QEU6 | 4NS4 |
| AGL49847.1 | Q9X022 | 1O2D | Alcohol dehydrogenase, iron-containing                                                                                                                                                        | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                                                                                     | ABQ94276.1 | A5WF35 | 4QBL |
| CAA73377.1 | O33839 | 1XJE | Ribonucleotide reductase                                                                                                                                                                      | <i>Thermotoga maritima</i>                                                                                                                                       | ACT35272.1 | C7E5X4 | 4RKC |
| AAL81406.1 | P24297 | 1BQ8 | Rubredoxin (Rd)                                                                                                                                                                               | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                                       | AAD13225.1 | Q9ZF99 | 1B8P |
| AAA22568.1 | P00344 | 1LDB | L-lactate dehydrogenase (L-LDH) (EC 1.1.1.27)                                                                                                                                                 | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | AAA88910.1 | P50921 | 1AW1 |
| AAB89026.1 | O28055 | 1SAU | Sulfite reductase, desulfovireidin-type subunit gamma (DsvC)                                                                                                                                  | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)                                                                    | CAM88673.1 | B1VBB0 | 4HMC |
| AAY81375.1 | Q55080 | 1F4T | Cytochrome P450 119 (EC 1.14.-.-) (Peroxidase) (EC 1.11.1.7)                                                                                                                                  | <i>Sulfolobus acidocaldarius</i> (strain ATCC 33909 / DSM 639 / JCM 8929 / NBRC 15157 / NCIMB 11770)                                                             | AAC62692.1 | O74049 | 2M1I |
| AAL81321.1 | Q8U1K9 | 1XHC | NADH oxidase /nitrite reductase                                                                                                                                                               | <i>Pyrococcus furiosus</i> (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)                                                                                       | CAB37062.1 | Q9Z4P0 | 1QO8 |
| CAA37630.1 | P11961 | 1B5S | Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex (EC 2.3.1.12) (Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex) (E2) | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | CAB38558.1 | P0C278 | 1E39 |
| AAD20315.1 | Q9X4D0 | 1D0Q | DNA primase (EC 2.7.7.-)                                                                                                                                                                      | <i>Geobacillus stearothermophilus</i> ( <i>Bacillus stearothermophilus</i> )                                                                                     | ABI71355.1 | O33731 | 2K3V |
| ABB58144.1 | Q06904 | 1T4Y | Adaptive-response sensory-kinase SasA (EC 2.7.13.3) ( <i>Synechococcus</i> adaptive sensor protein A)                                                                                         | <i>Synechococcus elongatus</i> (strain PCC 7942) ( <i>Anacystis nidulans</i> R2)                                                                                 | ABI70437.1 | Q087X8 | 2OOK |
| AAB85393.1 | O26981 | 1ILO | Thioredoxin                                                                                                                                                                                   | <i>Methanothermobacter thermautotrophicus</i> (strain ATCC 29096 / DSM 1053 / JCM 10044 / NBRC 100330 / Delta H) ( <i>Methanobacterium thermoautotrophicum</i> ) | ABI72942.1 | Q07YH2 | 2PFW |
| BAD70899.1 | Q5SJD8 | 1WG8 | Ribosomal RNA small subunit methyltransferase H (EC 2.1.1.199) (16S rRNA m(4)C1402 methyltransferase)                                                                                         | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                                                                                                  | ABI71165.1 | Q085A0 | 2PYX |



|            |        |      |                                                                                                                                                                                             |                                                                                               |            |        |      |
|------------|--------|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|------------|--------|------|
|            |        |      | (rRNA (cytosine-N(4)-)-methyltransferase RsmH)                                                                                                                                              |                                                                                               |            |        |      |
| BAC67244.1 | Q84BQ9 | 1UFK | Ribosomal protein L11 methyltransferase (L11 Mtase) (EC 2.1.1.-)                                                                                                                            | <i>Thermus thermophilus</i> (strain HB8 / ATCC 27634 / DSM 579)                               | ABI71819.1 | Q082J6 | 3BB9 |
| AAD36367.1 | Q9X119 | 1VLM | Diguanylate cyclase with PAS/PAC sensor (Uncharacterized protein)                                                                                                                           | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | ABI73370.1 | Q07X94 | 3D82 |
| AAD35573.1 | Q9WYV8 | 1NV8 | Release factor glutamine methyltransferase (RF MTase) (EC 2.1.1.297) (N5-glutamine methyltransferase PrmC) (Protein-(glutamine-N5) MTase PrmC) (Protein-glutamine N-methyltransferase PrmC) | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | ABI71883.1 | Q082D2 | 3FGE |
| AAD35939.1 | Q9WZW1 | 1MRZ | Riboflavin biosynthesis protein (EC 2.7.1.26) (EC 2.7.7.2)                                                                                                                                  | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | ABI73132.1 | Q07XY2 | 3LHO |
| AAD35910.1 | Q9WZT5 | 2AJR | Sugar kinase, pfkB family (Tagatose-6-phosphate kinase / 1-phosphofructokinase) (EC 2.7.1.144) (EC 2.7.1.56)                                                                                | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | ABI70397.1 | Q088B8 | 3LWU |
| AAD36360.1 | Q9X112 | 1T7L | 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase (EC 2.1.1.14) (Cobalamin-independent methionine synthase) (Methionine synthase, vitamin-B12 independent isozyme)      | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | BAL45195.1 | H1AFK5 | 3VND |
| AAD35358.1 | Q9WYA6 | 1J6R | Methionine synthase activation domain protein (EC 2.1.1.13) (Uncharacterized protein)                                                                                                       | <i>Thermotoga maritima</i> (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)                  | ABI72749.1 | Q07Z15 | 4CX9 |
| AAB89084.1 | O28126 | 1R89 | CCA-adding enzyme (EC 2.7.7.72) (CCA tRNA nucleotidyltransferase) (tRNA CCA-pyrophosphorylase) (tRNA adenylyl-/cytidyl- transferase) (tRNA nucleotidyltransferase) (tRNA-NT)                | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126) | AAL01059.1 | Q93CG9 | 3R87 |
| AAB91236.1 | O30245 | 1TQI | RIO-type serine/threonine-protein kinase Rio2 (AfRio2) (EC 2.7.11.1)                                                                                                                        | <i>Archaeoglobus fulgidus</i> (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126) | CAG19435.1 | Q6LTE1 | 4K46 |
| AAC07294.1 | O67322 | 1YE8 | Nucleoside-triphosphatase THEP1 (NTPase THEP1) (EC 3.6.1.15) (Nucleoside triphosphate phosphohydrolase)                                                                                     | <i>Aquifex aeolicus</i> (strain VF5)                                                          | CAA83122.1 | P41365 | 1LBS |



Table A1.4: Acidophile-Alkaliphile (A-B) dataset.

| Acidophilic CDS |               |        |                                                                                                                              |            | Alkaliphilic CDS |               |        |                                                                                                                                                           |            |
|-----------------|---------------|--------|------------------------------------------------------------------------------------------------------------------------------|------------|------------------|---------------|--------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| ENA ID          | UniProtK B ID | PDB ID | Protein name                                                                                                                 | Optimum pH | ENA ID           | UniProtK B ID | PDB ID | Protein name                                                                                                                                              | Optimum pH |
| AAC49277.1      | Q01738        | 1D7B   | Cellobiose dehydrogenase (CDH) (EC 1.1.99.18) (Cellobiose-quinone oxidoreductase)                                            | 4.5        | AAD30993.1       | Q9X5C9        | 2NLO   | Quinate/shikimate dehydrogenase (NAD(+)) (QSDH) (EC 1.1.1.-) (EC 1.1.1.24)                                                                                | 9          |
| AAA33104.1      | Q02497        | -      | Laccase (EC 1.10.3.2) (Benzenediol:oxygen oxidoreductase) (Diphenol oxidase) (Ligninolytic phenoloxidase) (Urishiol oxidase) | 4.5        | EGR49477.1       | G0RH19        | -      | L-xylulose reductase (EC 1.1.1.10)                                                                                                                        | 9          |
| AAZ57111.1      | Q47KB1        | 5FW4   | Dye-decolorizing peroxidase Tfu_3078 (DyP) (EC 1.11.1.19) (Peroxidase Tfu_3078)                                              | 3.5        | AAL81115.1       | Q8U259        | -      | L-threonine 3-dehydrogenase (TDH) (EC 1.1.1.103) (L-threonine dehydrogenase)                                                                              | 10         |
| EHA48040.1      | A4QUT2        | 3UT2   | Catalase-peroxidase 2 (CP 2) (EC 1.11.1.21) (Peroxidase/catalase 2)                                                          | 5          | ABO67118.1       | A4IP64        | -      | Long-chain-alcohol dehydrogenase 1 (EC 1.1.1.192) (Alcohol dehydrogenase 1) (ADH1) (Fatty alcohol oxidoreductase 1) (Glycerol dehydrogenase) (EC 1.1.1.6) | 8          |
| BAA09601.1      | Q31MN3        | 1UB2   | Catalase-peroxidase (CP) (EC 1.11.1.21) (Peroxidase/catalase)                                                                | 5.5        | EJY57755.1       | D2WKD9        | -      | Farnesol dehydrogenase (EC 1.1.1.216) (NADP+-dependent farnesol dehydrogenase 1) (AaSDR-1)                                                                | 10         |
| AAB89022.1      | O28050        | -      | Catalase-peroxidase (CP) (EC 1.11.1.21) (Peroxidase/catalase)                                                                | 6          | AAW22052.1       | Q56S04        | -      | Shikimate dehydrogenase (NADP(+)) (SDH) (EC 1.1.1.25)                                                                                                     | 9          |
| CAB94692.1      | Q9LEH3        | -      | Peroxidase 15 (Prx15) (EC 1.11.1.7) (Anionic peroxidase)                                                                     | 5.5        | AAB89513.1       | Q28538        | -      | 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMG-CoA reductase) (EC 1.1.1.34)                                                                         | 10         |
| AAY81375.1      | Q55080        | 1F4T   | Cytochrome P450 119 (EC 1.14.-.-) (Peroxidase) (EC 1.11.1.7)                                                                 | 6          | CAB97430.1       | Q9K3J3        | -      | Malate dehydrogenase (EC 1.1.1.37)                                                                                                                        | 8.5        |

|            |        |      |                                                                                                                                                                                                                                                |     |            |        |      |                                                                                                                                                          |     |
|------------|--------|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|------------|--------|------|----------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| AAB32775.1 | P19021 | -    | Peptidyl-glycine alpha-amidating monooxygenase (PAM) [Includes: Peptidylglycine alpha-hydroxylating monooxygenase (PHM) (EC 1.14.17.3); Peptidyl-alpha-hydroxyglycine alpha-amidating lyase (EC 4.3.2.5) (Peptidylamidoglycolate lyase) (PAL)] | 6   | BAC77301.1 | Q7X3X5 | -    | Malate dehydrogenase (EC 1.1.1.37)                                                                                                                       | 9.5 |
| BAB40419.1 | P31213 | -    | 3-oxo-5-alpha-steroid 4-dehydrogenase 2 (EC 1.3.1.22) (5 alpha-SR2) (SR type 2) (Steroid 5-alpha-reductase 2) (S5AR 2) (Type II 5-alpha reductase)                                                                                             | 5.5 | BAA14099.1 | P39482 | -    | Glucose 1-dehydrogenase 1 (EC 1.1.1.47) (GLCDH-I)                                                                                                        | 8   |
| AEX54742.1 | O54435 | -    | Levansucrase (EC 2.4.1.10) (Beta-D-fructofuranosyl transferase) (Sucrose 6-fructosyl transferase)                                                                                                                                              | 4   | ADX85515.1 | F0NBH8 | 5JWJ | Protein-lysine N-methyltransferase (EC 2.1.1.-) (Archaeal protein lysine methyltransferase) (aKMT)                                                       | 8   |
| AAB36606.1 | Q43998 | 1W18 | Levansucrase (EC 2.4.1.10) (Beta-D-fructofuranosyl transferase) (Sucrose 6-fructosyl transferase)                                                                                                                                              | 5   | CAA57356.1 | Q48296 | -    | Ornithine carbamoyltransferase, catabolic (cOTCase) (EC 2.1.3.3)                                                                                         | 8.8 |
| CAD48195.1 | Q70XJ9 | -    | Levansucrase (EC 2.4.1.10)                                                                                                                                                                                                                     | 5.4 | AAO80551.1 | Q837U7 | 3TXX | Putrescine carbamoyltransferase (PTC) (PTCase) (EC 2.1.3.6) (Agmatine catabolism protein B) (Putrescine transcarbamoylase) (Putrescine transcarbamylase) | 8.5 |
| CAA41773.1 | P04830 | 3WMS | Cyclomaltodextrin glucanotransferase (EC 2.4.1.19) (Cyclodextrin-glycosyltransferase) (CGTase)                                                                                                                                                 | 5.7 | CAB59620.1 | Q9UTM7 | -    | Histone acetyltransferase type B catalytic subunit (EC 2.3.1.48)                                                                                         | 8   |
| CAA41772.1 | P31797 | 1CYG | Cyclomaltodextrin glucanotransferase (EC 2.4.1.19) (Cyclodextrin-glycosyltransferase) (CGTase)                                                                                                                                                 | 6   | AAB40877.1 | Q59601 | 1U9Y | 1-acyl-sn-glycerol-3-phosphate acyltransferase (1-AGP acyltransferase) (1-AGPAT) (EC 2.3.1.51) (Lysophosphatidic acid acyltransferase) (LPAAT)           | 9   |

|            |        |      |                                                                                                      |     |            |        |      |                                                                                                                                                                                                                              |     |
|------------|--------|------|------------------------------------------------------------------------------------------------------|-----|------------|--------|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| CAA48401.1 | P30920 | 1CGT | Cyclomaltodextrin glucanotransferase (EC 2.4.1.19) (Cyclodextrin-glycosyltransferase) (CGTase)       | 6   | ADE03756.1 | D4GS06 | -    | Citrate synthase (EC 2.3.3.16)                                                                                                                                                                                               | 8   |
| AAO00727.1 | Q6YDN9 | -    | Xyloglucan endotransglucosylase/hydrolase (EC 2.4.1.207) (BobXET16A)                                 | 5.5 | ACF75531.1 | B5APK2 | -    | Deoxyhypusine synthase (DHS) (EC 2.5.1.46) (Deoxyhypusine synthase from chromosome 34) (DHS34)                                                                                                                               | 9.6 |
| CAB79437.1 | Q9ZSU4 | -    | Xyloglucan endotransglucosylase/hydrolase protein 14 (At-XTH14) (XTH-14) (EC 2.4.1.207)              | 6   | AAB99374.1 | Q58761 | -    | Ribose-phosphate pyrophosphokinase (RPPK) (EC 2.7.6.1) (5-phospho-D-ribosyl alpha-1-diphosphate) (Phosphoribosyl diphosphate synthase) (Phosphoribosyl pyrophosphate synthase) (P-Rib-PP synthase) (PRPP synthase) (PRPPase) | 9.5 |
| BAA03923.1 | Q40144 | -    | Probable xyloglucan endotransglucosylase/hydrolase 1 (LeXTH1) (EC 2.4.1.207)                         | 6   | AAN00040.1 | Q9AFG9 | -    | N-acetylneuraminase cytidyltransferase (EC 2.7.7.43) (CMP-N-acetylneuraminic acid synthase) (CMP-NeuNAc synthase) (CMP-sialic acid synthase)                                                                                 | 9   |
| BAA13163.1 | P93349 | -    | Probable xyloglucan endotransglucosylase/hydrolase protein (EC 2.4.1.207)                            | 6   | AAT68144.1 | Q6DRD3 | -    | DNA polymerase beta (EC 2.7.7.7) (EC 4.2.99.-)                                                                                                                                                                               | 9   |
| AAY88919.1 | B2VPR8 | -    | Pectinesterase 2 (EC 3.1.1.11) (Pollen allergen Ole e 11.0102) (Ole e 11-2) (allergen Ole e 11.0102) | 5.5 | CAA96487.1 | P37967 | 1C7I | Para-nitrobenzyl esterase (EC 3.1.1.-) (Intracellular esterase B) (PNB carboxy-esterase) (PNBCE)                                                                                                                             | 8   |
| CAA46805.1 | P32947 | 1CLE | Lipase 3 (EC 3.1.1.3) (Cholesterol esterase)                                                         | 4   | AAC35293.1 | P27169 | 1V04 | Serum paraoxonase/arylesterase 1 (PON 1) (EC 3.1.1.2) (EC 3.1.1.81) (EC 3.1.8.1) (Aromatic esterase 1) (A-esterase 1) (K-45) (Serum arylalkylphosphatase 1)                                                                  | 8   |
| AAC08588.1 | O59952 | 1DT3 | Lipase (EC 3.1.1.3) (Triacylglycerol lipase)                                                         | 6   | AAA50466.1 | P22088 | 1HQD | Lipase (EC 3.1.1.3) (Triacylglycerol lipase)                                                                                                                                                                                 | 8   |
| BAA28619.1 | O59863 | 1UTE | Lysophospholipase (EC 3.1.1.5) (KlPLB) (Phospholipase B)                                             | 2   | AAA33878.1 | P61872 | 1TIC | Lipase (EC 3.1.1.3) (RDL) (Triacylglycerol lipase) (ROL)                                                                                                                                                                     | 8.2 |

|            |        |       |                                                                                                                                                          |     |            |        |      |                                                                                                                                                                                                                                                             |   |
|------------|--------|-------|----------------------------------------------------------------------------------------------------------------------------------------------------------|-----|------------|--------|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|
| AAA31139.1 | P09889 | 1XZW  | Tartrate-resistant acid phosphatase type 5 (TR-AP) (EC 3.1.3.2) (Tartrate-resistant acid ATPase) (TrATPase) (Type 5 acid phosphatase) (Uteroferrin) (UF) | 3   | AAN63044.1 | Q9M0D7 | -    | Phospholipase A2-gamma (EC 3.1.1.4) (Secretory phospholipase A2-gamma) (AtsPLA2-gamma)                                                                                                                                                                      | 8 |
| DAA07213.1 | P24031 | -     | Constitutive acid phosphatase (EC 3.1.3.2)                                                                                                               | 3   | AAM09694.1 | Q8QG87 | -    | Acidic phospholipase A2 BITP01A (svPLA2) (EC 3.1.1.4) (BinTX-I) (Phosphatidylcholine 2-acylhydrolase)                                                                                                                                                       | 8 |
| AAF19821.1 | Q9SE00 | 1N8N  | Purple acid phosphatase 1 (EC 3.1.3.2) (Manganese(II) purple acid phosphatase 1)                                                                         | 4.5 | CAB42689.2 | P47712 | 1BCI | Cytosolic phospholipase A2 (cPLA2) (Phospholipase A2 group IVA) [Includes: Phospholipase A2 (EC 3.1.1.4) (Phosphatidylcholine 2-acylhydrolase); Lysophospholipase (EC 3.1.1.5)]                                                                             | 9 |
| EAL88069.1 | Q8X176 | -     | Acid phosphatase (EC 3.1.3.2)                                                                                                                            | 5   | CAA27169.1 | P04058 | 1ACJ | Acetylcholinesterase (AChE) (EC 3.1.1.7)                                                                                                                                                                                                                    | 8 |
| AAC43149.1 | P0AE22 | 1DKL  | Class B acid phosphatase (CBAP) (EC 3.1.3.2)                                                                                                             | 6   | CAM21953.1 | Q9R0X4 | -    | Acyl-coenzyme A thioesterase 9, mitochondrial (Acyl-CoA thioesterase 9) (EC 3.1.2.-) (Acyl coenzyme A thioester hydrolase 2) (MTE-2) (Acyl-CoA thioester hydrolase 9) (Mitochondrial 48 kDa acyl-CoA thioester hydrolase 1) (Mt-ACT48.1) (Protein U8) (p48) | 8 |
| AAC74065.1 | P07102 | 1QW O | Periplasmic AppA protein [Includes: Phosphoanhydride phosphohydrolase (EC 3.1.3.2) (pH 2.5 acid phosphatase) (AP); 4-phytase (EC 3.1.3.26)]              | 2   | CAA60943.1 | P41903 | 1TBU | Peroxisomal acyl-coenzyme A thioester hydrolase 1 (EC 3.1.2.2) (Peroxisomal long-chain acyl-CoA thioesterase 1)                                                                                                                                             | 8 |
| AAB52507.1 | O00085 | -     | 3-phytase A (EC 3.1.3.8) (3 phytase A) (Myo-inositol hexakisphosphate phosphohydrolase A) (Myo-inositol-hexaphosphate 3-phosphohydrolase A)              | 4.5 | CAA60024.1 | O14734 | -    | Acyl-coenzyme A thioesterase 8 (Acyl-CoA thioesterase 8) (EC 3.1.2.27) (Choloyl-coenzyme A thioesterase) (HIV-Nef-associated acyl-CoA thioesterase) (PTE-2) (Peroxisomal acyl-coenzyme A thioester hydrolase 1) (PTE-1)                                     | 8 |

|            |        |      |                                                                                                                                             |     |            |        |      |                                                                                                                                                   |     |
|------------|--------|------|---------------------------------------------------------------------------------------------------------------------------------------------|-----|------------|--------|------|---------------------------------------------------------------------------------------------------------------------------------------------------|-----|
|            |        |      |                                                                                                                                             |     |            |        |      | (Peroxisomal long-chain acyl-CoA thioesterase 1) (Thioesterase II) (hACTE-III) (hACTEIII) (hTE)                                                   |     |
| CAD12029.1 | O00092 | -    | 3-phytase A (EC 3.1.3.8) (3-phytase A) (Myo-inositol hexakisphosphate phosphohydrolase A) (Myo-inositol-hexaphosphate 3-phosphohydrolase A) | 5.5 | CAA04665.1 | O42446 | -    | Deoxyribonuclease-1 (EC 3.1.21.1) (Deoxyribonuclease I) (DNase I)                                                                                 | 8   |
| AAI00885.1 | Q86SG7 | -    | Lysozyme g-like protein 2 (EC 3.2.1.-)                                                                                                      | 6   | AAC22072.1 | P44443 | -    | Ribonuclease E (RNase E) (EC 3.1.26.12)                                                                                                           | 8   |
| ADT80795.1 | E7CY70 | -    | Exo-alpha-(1->6)-L-arabinofuranosidase (ABF) (EC 3.2.1.-)                                                                                   | 4.7 | AAA24565.1 | P0A7Y4 | 1F21 | Ribonuclease HI (RNase HI) (EC 3.1.26.4) (Ribonuclease H) (RNase H)                                                                               | 8   |
| AAM53325.1 | Q9FGY1 | -    | Beta-D-xylosidase 1 (AtBXL1) (EC 3.2.1.-) (Alpha-L-arabinofuranosidase) (EC 3.2.1.55)                                                       | 5   | AAB90620.1 | O29634 | 1I39 | Ribonuclease HII (RNase HII) (EC 3.1.26.4)                                                                                                        | 8   |
| AAB47720.1 | P48845 | 10GM | Dextranase (EC 3.2.1.11) (Alpha-1,6-glucan-6-glucanohydrolase)                                                                              | 5   | BAE27230.1 | Q9CWY8 | 3KIO | Ribonuclease H2 subunit A (RNase H2 subunit A) (EC 3.1.26.4) (Ribonuclease HI large subunit) (RNase HI large subunit) (Ribonuclease HI subunit A) | 8   |
| BAB16369.1 | Q9GV16 | -    | Endoglycoceramidase (EGCase) (EC 3.2.1.123) (Glycosphingolipid-specific enzyme) (GSL-specific enzyme)                                       | 3   | AAD35996.1 | Q9X017 | 2ETJ | Ribonuclease HII (RNase HII) (EC 3.1.26.4)                                                                                                        | 8.5 |
| DAA09597.1 | P29029 | 2UY2 | Endochitinase (EC 3.2.1.14) (Soluble cell wall protein 2)                                                                                   | 2.6 | BAA79461.2 | Q9YET5 | -    | Ribonuclease HII (RNase HII) (EC 3.1.26.4)                                                                                                        | 9   |
| AAH36339.2 | Q9BZP6 | 2YBT | Acidic mammalian chitinase (AMCase) (EC 3.2.1.14) (Lung-specific protein TSA1902)                                                           | 4   | BAA77858.2 | P10442 | -    | Ribonuclease HII (RNase HII) (EC 3.1.26.4)                                                                                                        | 10  |
| AAK42272.1 | P95867 | -    | Malto-oligosyltrehalose trehalohydrolase (MTHase) (EC 3.2.1.141) (4-alpha-D-((1->4)-alpha-D-glucano)trehalose                               | 5   | AAA92562.1 | Q92058 | -    | Alkaline phosphatase, tissue-nonspecific isozyme (AP-TNAP) (TNSALP) (EC 3.1.3.1)                                                                  | 8.5 |

|            |        |      |                                                                                                                                                                                                  |     |            |        |      |                                                                                                                                                                                                                                                                                                                                                                                    |     |
|------------|--------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|------------|--------|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
|            |        |      | trehalohydrolase)<br>(Maltooligosyl trehalose<br>trehalohydrolase)                                                                                                                               |     |            |        |      |                                                                                                                                                                                                                                                                                                                                                                                    |     |
| CAB72931.1 | Q9P4W2 | -    | Endopolygalacturonase D<br>(PGD) (EC 3.2.1.15)<br>(Pectinase D)<br>(Polygalacturonase D)                                                                                                         | 4.2 | ABL96598.1 | A1YYW7 | 3Q3Q | Alkaline phosphatase PhoK (EC<br>3.1.3.1) (SPAP protein)                                                                                                                                                                                                                                                                                                                           | 9   |
| DAA08938.1 | P47180 | -    | Polygalacturonase (PG) (EC<br>3.2.1.15) (PGase SM)<br>(Pectinase)                                                                                                                                | 4.5 | AAC49121.1 | P0C5A3 | 5TCD | 3'(2'),5'-bisphosphate nucleotidase<br>(EC 3.1.3.7) (3'(2'),5-<br>bisphosphonucleoside 3'(2')-<br>phosphohydrolase) (DPNPase)                                                                                                                                                                                                                                                      | 8.5 |
| BAC22065.1 | Q8J0D2 | 1SQJ | Oligoxyloglucan reducing end-<br>specific cellobiohydrolase<br>(OXG-RCBH) (EC 3.2.1.150)                                                                                                         | 3.5 | AAV40908.1 | Q6UWR7 | -    | Ectonucleotide<br>pyrophosphatase/phosphodiesterase<br>family member 6 (E-NPP 6)<br>(NPP-6) (EC 3.1.4.-) (EC<br>3.1.4.38) (Choline-specific<br>glycerophosphodiester<br>phosphodiesterase)<br>(Glycerophosphocholine<br>cholinephosphodiesterase) (GPC-<br>Cpde)                                                                                                                   | 8.5 |
| BAD14925.1 | Q75W17 | -    | Furcatin hydrolase (FH) (EC<br>3.2.1.161)                                                                                                                                                        | 5   | AAB61536.1 | P97675 | -    | Ectonucleotide<br>pyrophosphatase/phosphodiesterase<br>family member 3 (E-NPP 3)<br>(B10) (Phosphodiesterase I beta)<br>(PD-Ibeta) (Phosphodiesterase<br>I/nucleotide pyrophosphatase 3)<br>(RB13-6 antigen) (CD antigen<br>CD203c) [Includes: Alkaline<br>phosphodiesterase I (EC 3.1.4.1);<br>Nucleotide pyrophosphatase<br>(NPPase) (EC 3.6.1.9)<br>(Nucleotide diphosphatase)] | 8.5 |
| BAC36900.1 | Q9EQQ9 | -    | Protein O-GlcNAcase (OGA)<br>(EC 3.2.1.169) (Beta-N-<br>acetylhexosaminidase) (Beta-<br>hexosaminidase) (Bifunctional<br>protein NCOAT)<br>(Meningioma-expressed<br>antigen 5) (N-acetyl-beta-D- | 5   | BAC86504.1 | Q6UWV6 | -    | Ectonucleotide<br>pyrophosphatase/phosphodiesterase<br>family member 7 (E-NPP 7)<br>(NPP-7) (EC 3.1.4.12) (Alkaline<br>sphingomyelin phosphodiesterase)<br>(Intestinal alkaline<br>sphingomyelinase) (Alk-SMase)                                                                                                                                                                   | 9   |



|            |        |      |                                                                                                                                     |     |            |        |      |                                                                                                                                                                                                       |     |
|------------|--------|------|-------------------------------------------------------------------------------------------------------------------------------------|-----|------------|--------|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
|            |        |      | glucosaminidase) (EC 3.2.1.52) (N-acetyl-beta-glucosaminidase)                                                                      |     |            |        |      |                                                                                                                                                                                                       |     |
| BAA75890.1 | P36924 | 1B90 | Beta-amylase (EC 3.2.1.2) (1,4-alpha-D-glucan maltohydrolase)                                                                       | 4.2 | AAA82110.1 | P47820 | -    | Angiotensin-converting enzyme (ACE) (EC 3.2.1.-) (EC 3.4.15.1) (Dipeptidyl carboxypeptidase I) (Kininase II) (CD antigen CD143) [Cleaved into: Angiotensin-converting enzyme, soluble form]           | 8.5 |
| BAA04815.1 | P16098 | 1B1Y | Beta-amylase (EC 3.2.1.2) (1,4-alpha-D-glucan maltohydrolase)                                                                       | 5   | BAA00055.1 | P07528 | 2UWF | Endo-1,4-beta-xylanase A (Xylanase A) (EC 3.2.1.8) (1,4-beta-D-xylan xylanohydrolase A)                                                                                                               | 9   |
| CAB46051.1 | O23553 | -    | Beta-amylase 3, chloroplastic (EC 3.2.1.2) (1,4-alpha-D-glucan maltohydrolase) (Beta-amylase 8) (Chloroplast beta-amylase) (CT-BMY) | 6   | AAF10269.1 | Q9RWH9 | 2BOO | Uracil-DNA glycosylase (UDG) (EC 3.2.2.27)                                                                                                                                                            | 8   |
| AAB02985.1 | Q43763 | -    | Alpha-glucosidase (EC 3.2.1.20) (Maltase)                                                                                           | 4   | AFP38807.1 | A0QV01 | -    | Uracil-DNA glycosylase (UDG) (EC 3.2.2.27)                                                                                                                                                            | 8   |
| BAB43946.1 | Q9C0Y4 | -    | Alpha-glucosidase (EC 3.2.1.20) (Maltase)                                                                                           | 4.5 | AAP98730.1 | Q9Z7D3 | -    | Uracil-DNA glycosylase (UDG) (EC 3.2.2.27)                                                                                                                                                            | 8   |
| BAA23616.1 | P56526 | -    | Alpha-glucosidase (EC 3.2.1.20) (Maltase)                                                                                           | 4.5 | CAA05572.1 | P26902 | 1HI9 | D-aminopeptidase (EC 3.4.11.-)                                                                                                                                                                        | 9   |
| AAK43151.1 | P0CD66 | 2G3M | Alpha-glucosidase (EC 3.2.1.20) (Maltase)                                                                                           | 4.5 | AAN31395.1 | Q8J2N2 | -    | Leucine aminopeptidase 1 (EC 3.4.11.-) (Leucyl aminopeptidase 1) (LAP1)                                                                                                                               | 8.5 |
| ABF94615.1 | Q8L7J2 | 3GNO | Beta-glucosidase 6 (Os3bglu6) (EC 3.2.1.21)                                                                                         | 4   | AAL21430.1 | Q9RF52 | -    | Peptidase B (EC 3.4.11.23) (Aminopeptidase B)                                                                                                                                                         | 8.5 |
| BAE87009.1 | Q25BW4 | -    | Beta-glucosidase 1B (EC 3.2.1.21) (Cellobiase 1B)                                                                                   | 5.5 | AAH07579.1 | Q9NQW7 | 3CTZ | Xaa-Pro aminopeptidase 1 (EC 3.4.11.9) (Aminoacylproline aminopeptidase) (Cytosolic aminopeptidase P) (Soluble aminopeptidase P) (sAmp) (X-Pro aminopeptidase 1) (X-prolyl aminopeptidase 1, soluble) | 8.5 |
| AAG51761.1 | Q9C525 | -    | Beta-glucosidase 21 (AtBGLU21) (EC 3.2.1.21) (Protein PHOSPHATE STARVATION-RESPONSE                                                 | 5.5 | AAB35614.1 | P27487 | 1J2E | Dipeptidyl peptidase 4 (EC 3.4.14.5) (ADABP) (Adenosine deaminase complexing protein 2) (ADCP-2) (Dipeptidyl peptidase                                                                                | 8   |

|            |        |      |                                                                                                    |     |            |        |      |                                                                                                                                                                                                                                                  |      |
|------------|--------|------|----------------------------------------------------------------------------------------------------|-----|------------|--------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
|            |        |      | 3.2)                                                                                               |     |            |        |      | IV) (DPP IV) (T-cell activation antigen CD26) (TP103) (CD antigen CD26) [Cleaved into: Dipeptidyl peptidase 4 membrane form (Dipeptidyl peptidase IV membrane form); Dipeptidyl peptidase 4 soluble form (Dipeptidyl peptidase IV soluble form)] |      |
| AAP54412.1 | Q9FXT4 | 1UAS | Alpha-galactosidase (EC 3.2.1.22) (Alpha-D-galactoside galactohydrolase) (Melibiase)               | 5   | BAD85453.1 | Q8NKS6 | -    | Archaeal Lon protease (EC 3.4.21.-) (LonTk)                                                                                                                                                                                                      | 9    |
| AAK43227.1 | Q97U94 | -    | Alpha-galactosidase (Alpha-Gal) (EC 3.2.1.22)                                                      | 5   | AAA29339.1 | P35045 | -    | Trypsin, alkaline A (EC 3.4.21.4)                                                                                                                                                                                                                | 10.5 |
| CAK44933.1 | A2QL72 | -    | Probable alpha-galactosidase A (EC 3.2.1.22) (Melibiase A)                                         | 1.8 | AAC66962.2 | O51558 | -    | Lon protease 2 (EC 3.4.21.53) (ATP-dependent protease La 2)                                                                                                                                                                                      | 9    |
| AAC60538.1 | P29853 | -    | Beta-galactosidase (EC 3.2.1.23) (Lactase-N) (Lactase) (Tilactase)                                 | 2   | BAA00951.1 | P12547 | -    | Alkaline protease 1 (ALP) (EC 3.4.21.63) (Aspergillopeptidase B) (Aspergillus proteinase B) (Elastase) (Elastinolytic serine proteinase) (Oryzin)                                                                                                | 9    |
| BAS82468.1 | Q10RB4 | -    | Beta-galactosidase 5 (Lactase 5) (EC 3.2.1.23)                                                     | 3   | BAB82381.1 | P81054 | 1G12 | Peptidyl-Lys metalloendopeptidase (MEP) (EC 3.4.24.20) (GfMEP)                                                                                                                                                                                   | 9.5  |
| AAN60229.1 | Q9LFA6 | -    | Beta-galactosidase 2 (Lactase 2) (EC 3.2.1.23)                                                     | 4   | CAE45336.1 | Q70I53 | 1ZZ0 | Histone deacetylase-like amidohydrolase (HDAC-like amidohydrolase) (HDAH) (EC 3.5.1.-)                                                                                                                                                           | 8    |
| AAB97862.1 | O52629 | -    | Beta-galactosidase (Lactase) (EC 3.2.1.23)                                                         | 4.5 | BAD29713.1 | Q6F4N1 | -    | Diacetylchitobiose deacetylase (EC 3.5.1.105) (N-acetylchitobiose deacetylase) (Tk-Dac)                                                                                                                                                          | 8    |
| AAA25244.1 | Q1G9Z4 | -    | Beta-galactosidase (Beta-gal) (EC 3.2.1.23) (Lactase)                                              | 5   | CAC29092.1 | Q99PX1 | -    | Chitin disaccharide deacetylase (EC 3.5.1.105) (Chitin oligosaccharide deacetylase)                                                                                                                                                              | 8.5  |
| AAH05430.1 | O09159 | -    | Lysosomal alpha-mannosidase (Laman) (EC 3.2.1.24) (Lysosomal acid alpha-mannosidase) (Mannosidase) | 4.2 | AAL55991.1 | Q9VIP7 | -    | Alkaline ceramidase (AlkCDase) (EC 3.5.1.23) (Alkaline N-acylsphingosine amidohydrolase) (Alkaline acylsphingosine                                                                                                                               | 8    |

|            |        |      |                                                                                                                                                           |     |            |        |      |                                                                                                                                                                                                                                             |     |
|------------|--------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|-----|------------|--------|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
|            |        |      | alpha class 2B member 1)<br>(Mannosidase alpha-B)                                                                                                         |     |            |        |      | deacylase) (Protein brainwashing)                                                                                                                                                                                                           |     |
| AAD36485.1 | O33833 | 1UYP | Beta-fructosidase (EC 3.2.1.26) (Invertase) (Sucrase)                                                                                                     | 5.5 | AAH16828.1 | Q13510 | -    | Acid ceramidase (AC) (ACDase) (Acid CDase) (EC 3.5.1.23) (Acylsphingosine deacylase) (N-acylsphingosine amidohydrolase) (Putative 32 kDa heart protein) (PHP32) [Cleaved into: Acid ceramidase subunit alpha; Acid ceramidase subunit beta] | 8.5 |
| EAA35527.1 | Q7SFB0 | -    | Beta-glucuronidase (GlcAase) (EC 3.2.1.31) (Beta-D-glucuronoside glucuronosohydrolase)                                                                    | 6   | AAC74380.2 | P76038 | -    | Gamma-glutamyl-gamma-aminobutyrate hydrolase PuuD (Gamma-Glu-GABA hydrolase) (EC 3.5.1.94)                                                                                                                                                  | 8.5 |
| AAA27730.1 | Q08169 | 1FCQ | Hyaluronidase (Hya) (EC 3.2.1.35) (Allergen Api m II) (Hyaluronoglucosaminidase) (allergen Api m 2)                                                       | 3.8 | AAI10405.1 | O00519 | -    | Fatty-acid amide hydrolase 1 (EC 3.5.1.99) (Anandamide amidohydrolase 1) (Oleamide hydrolase 1)                                                                                                                                             | 8.5 |
| AAD13106.1 | O00089 | -    | Exo-1,4-beta-xylosidase xlnD (EC 3.2.1.37) (1,4-beta-D-xylan xylohydrolase xlnD) (Beta-xylosidase A) (Beta-xylosidase xlnD) (Xylobiase xlnD)              | 3.5 | AAB26961.1 | P97612 | 1MT5 | Fatty-acid amide hydrolase 1 (EC 3.5.1.99) (Anandamide amidohydrolase 1) (Oleamide hydrolase 1)                                                                                                                                             | 8.5 |
| ACF61038.1 | Q4AEG8 | -    | Exo-1,4-beta-xylosidase xlnD (EC 3.2.1.37) (1,4-beta-D-xylan xylohydrolase xlnD) (Beta-xylosidase A) (Beta-xylosidase xlnD) (Xylobiase xlnD)              | 5   | BAA86917.1 | Q9TUI8 | -    | Fatty-acid amide hydrolase 1 (EC 3.5.1.99) (Anandamide amidohydrolase 1) (Oleamide hydrolase 1)                                                                                                                                             | 8.5 |
| CAB75760.1 | A7WM73 | -    | Beta-hexosaminidase 1 (EC 3.2.1.52) (Beta-GlcNAcase 1) (Beta-N-acetylhexosaminidase 1) (Beta-hexosaminidase 2) (AtHEX2) (N-acetyl-beta-glucosaminidase 1) | 4   | AAB84217.1 | P46926 | 1NE7 | Glucosamine-6-phosphate isomerase 1 (EC 3.5.99.6) (Glucosamine-6-phosphate deaminase 1) (GNPDA 1) (GlcN6P deaminase 1) (Oscillin)                                                                                                           | 8   |
| AAC60521.1 | P49010 | -    | Chitooligosaccharidolytic beta-N-acetylglucosaminidase (EC 3.2.1.52) (Beta-GlcNAcase) (Beta-N-                                                            | 5.5 | AAN74434.1 | P56153 | 1YGZ | Inorganic pyrophosphatase (EC 3.6.1.1) (Pyrophosphate phospho-hydrolase) (PPase)                                                                                                                                                            | 8   |

|            |        |      |                                                                                                                                                                      |     |            |        |      |                                                                                                                                                                                                                                                                                                                            |     |
|------------|--------|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|------------|--------|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
|            |        |      | acetylhexosaminidase) (Beta-hexosaminidase)                                                                                                                          |     |            |        |      |                                                                                                                                                                                                                                                                                                                            |     |
| ABG66991.1 | Q0V8R6 | -    | Beta-hexosaminidase subunit alpha (EC 3.2.1.52) (Beta-N-acetylhexosaminidase subunit alpha) (Hexosaminidase subunit A) (N-acetyl-beta-glucosaminidase subunit alpha) | 5.5 | AAB39104.1 | P95765 | 1K20 | Probable manganese-dependent inorganic pyrophosphatase (EC 3.6.1.1) (Pyrophosphate phosphohydrolase) (PPase)                                                                                                                                                                                                               | 9   |
| AAC44672.1 | P96155 | -    | Beta-hexosaminidase (EC 3.2.1.52) (Beta-N-acetylhexosaminidase) (N-acetyl-beta-glucosaminidase)                                                                      | 5.8 | AAH49383.1 | Q8NFP7 | 3MCF | Diphosphoinositol polyphosphate phosphohydrolase 3-alpha (DIPP-3-alpha) (DIPP3-alpha) (hDIPP3alpha) (EC 3.6.1.52) (Diadenosine 5',5'''-P1,P6-hexaphosphate hydrolase 3-alpha) (Diadenosine hexaphosphate hydrolase (AMP-forming)) (EC 3.6.1.60) (Nucleoside diphosphate-linked moiety X motif 10) (Nudix motif 10) (hAps2) | 8.5 |
| BAB71803.1 | Q2UIM2 | -    | Alpha-L-arabinofuranosidase B (ABF B) (Arabinosidase B) (EC 3.2.1.55)                                                                                                | 4   | AAL80619.1 | P95479 | -    | Reverse gyrase [Includes: Helicase (EC 3.6.4.12); Topoisomerase (EC 5.99.1.3)]                                                                                                                                                                                                                                             | 8.8 |
| CAB13699.1 | Q45071 | 3C7E | Arabinoxylan arabinofuranohydrolase (AXH) (EC 3.2.1.55) (AXH-m2,3) (AXH-m23) (Alpha-L-arabinofuranosidase) (AF)                                                      | 5.6 | AAA25832.1 | Q52087 | -    | (S)-2-haloacid dehalogenase (EC 3.8.1.2) (2-haloalkanoic acid dehalogenase) (Halocarboxylic acid halidohydrolase) (L-2-haloacid dehalogenase)                                                                                                                                                                              | 10  |
| AAO84266.1 | Q841V6 | -    | Intracellular exo-alpha-(1->5)-L-arabinofuranosidase (ABF) (EC 3.2.1.55) (Intracellular arabinan exo-alpha-(1->5)-L-arabinosidase) (Arabinosidase)                   | 6   | AAS04374.1 | Q73Y99 | -    | Haloalkane dehalogenase (EC 3.8.1.5)                                                                                                                                                                                                                                                                                       | 8.6 |
| AAD45520.2 | Q9XBQ3 | 1PZ2 | Intracellular exo-alpha-(1->5)-L-arabinofuranosidase (ABF) (EC 3.2.1.55) (Intracellular arabinan exo-alpha-(1->5)-L-arabinosidase) (Arabinosidase)                   | 6   | CAB45532.2 | Q9XB14 | -    | Haloalkane dehalogenase (EC 3.8.1.5)                                                                                                                                                                                                                                                                                       | 9   |

|            |        |      |                                                                                                                                  |     |            |        |      |                                                                                                                                                                                      |     |
|------------|--------|------|----------------------------------------------------------------------------------------------------------------------------------|-----|------------|--------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| AAA67426.1 | Q00012 | -    | Mannan endo-1,4-beta-mannosidase A (EC 3.2.1.78) (Endo-beta-1,4-mannanase A)                                                     | 3   | AAH91479.1 | Q96GA7 | 2RKB | Serine dehydratase-like (L-serine deaminase) (L-serine dehydratase/L-threonine deaminase) (L-threonine dehydratase) (TDH) (EC 4.3.1.19) (Serine dehydratase 2) (SDH 2) (EC 4.3.1.17) | 8.3 |
| CAC81056.1 | Q8WPJ2 | 2C0H | Mannan endo-1,4-beta-mannosidase (EC 3.2.1.78) (Beta-mannanase) (Endo-beta-1,4-mannanase) (Man5A) (ManA)                         | 5.2 | AAM91425.1 | P45724 | -    | Phenylalanine ammonia-lyase 2 (EC 4.3.1.24)                                                                                                                                          | 8.4 |
| ACY46925.1 | P49425 | -    | Mannan endo-1,4-beta-mannosidase (EC 3.2.1.78) (Endo-(1,4)-beta-mannanase)                                                       | 5.4 | CAA68036.1 | Q43210 | -    | Phenylalanine ammonia-lyase (EC 4.3.1.24)                                                                                                                                            | 8.8 |
| CBF69494.1 | Q5AZ53 | -    | Mannan endo-1,4-beta-mannosidase C (EC 3.2.1.78) (Endo-beta-1,4-mannanase C)                                                     | 5.5 | AAA34179.2 | P35511 | -    | Phenylalanine ammonia-lyase (PAL) (EC 4.3.1.24)                                                                                                                                      | 8.8 |
| CAA49293.1 | P36217 | 1ENX | Endo-1,4-beta-xylanase 2 (EX 2) (Xylanase 2) (EC 3.2.1.8) (1,4-beta-D-xylan xylanohydrolase 2) (Alkaline endo-beta-1,4-xylanase) | 5   | CAA53581.1 | P45735 | -    | Phenylalanine ammonia-lyase (EC 4.3.1.24) (Fragment)                                                                                                                                 | 8.9 |
| BAE71133.1 | Q2PGY1 | -    | Endo-1,4-beta-xylanase A (Xylanase A) (EC 3.2.1.8) (1,4-beta-D-xylan xylanohydrolase A)                                          | 5   | CAA25817.1 | P00936 | -    | Adenylate cyclase (EC 4.6.1.1) (ATP pyrophosphate-lyase) (Adenylyl cyclase)                                                                                                          | 8.5 |
| CAG25554.1 | Q5ZNB1 | -    | Endo-1,4-beta-xylanase D (Xylanase D) (EC 3.2.1.8) (1,4-beta-D-xylan xylanohydrolase D)                                          | 5   | CAA65130.1 | Q59119 | -    | Adenylate cyclase (EC 4.6.1.1) (ATP pyrophosphate-lyase) (Adenylyl cyclase)                                                                                                          | 9.5 |
| CAC15487.1 | Q9HFH0 | 1TE1 | Endo-1,4-beta-xylanase C (Xylanase C) (EC 3.2.1.8) (1,4-beta-D-xylan xylanohydrolase C)                                          | 5.5 | CAB39935.1 | Q2PGG3 | -    | Serine racemase (AtSR) (EC 4.3.1.17) (EC 4.3.1.18) (EC 5.1.1.18) (D-serine ammonia-lyase) (D-serine dehydratase) (L-serine ammonia-lyase) (L-serine dehydratase)                     | 8.5 |
| ACY69861.1 | O43097 | 1YNA | Endo-1,4-beta-xylanase (Xylanase) (EC 3.2.1.8) (1,4-beta-D-xylan xylanohydrolase)                                                | 6   | BAG72134.1 | B5UAT8 | 4WD3 | L-arginine-specific L-amino acid ligase (EC 6.3.2.48) (L-amino acid ligase Riza)                                                                                                     | 9.5 |

|            |        |      |                                                                                                                                                             |     |            |        |      |                                                                                                                                            |     |
|------------|--------|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|------------|--------|------|--------------------------------------------------------------------------------------------------------------------------------------------|-----|
| ABW04217.1 | A8TGA1 | -    | Endo-1,4-beta-xylanase S20 (Xylanase S20) (EC 3.2.1.8) (1,4-beta-D-xylan xylanohydrolase S20)                                                               | 6   | BAD96917.1 | Q9UGM6 | 5EKD | Tryptophan--tRNA ligase, mitochondrial (EC 6.1.1.2) ((Mt)TrpRS) (Tryptophanyl-tRNA synthetase) (TrpRS)                                     | 9   |
| ABQ46657.1 | A5IKD4 | 4KC7 | Extracellular endo-alpha-(1->5)-L-arabinanase (ABN) (EC 3.2.1.99) (Endo-1,5-alpha-L-arabinanase)                                                            | 6   | AAO04282.1 | Q8CT69 | -    | Tryptophan--tRNA ligase (EC 6.1.1.2) (Tryptophanyl-tRNA synthetase) (TrpRS)                                                                | 9   |
| CAE46473.1 | Q70GH4 | -    | Tripeptidyl-peptidase sed3 (EC 3.4.14.-) (Sedolisin-C)                                                                                                      | 5   | BAB71770.1 | Q932V0 | -    | Alanine racemase, biosynthetic (EC 5.1.1.1)                                                                                                | 8   |
| BAG70211.1 | P42785 | 3N2Z | Lysosomal Pro-X carboxypeptidase (EC 3.4.16.2) (Angiotensinase C) (Lysosomal carboxypeptidase C) (Proline carboxypeptidase) (Prolylcarboxypeptidase) (PRCP) | 5.5 | AAM25327.1 | Q8R860 | -    | Alanine racemase 2 (EC 5.1.1.1)                                                                                                            | 11  |
| AAM27198.1 | Q8MZS4 | -    | Physarolisin (EC 3.4.21.103) (Physaropepsin) [Cleaved into: Physarolisin heavy chain; Physarolisin light chain]                                             | 1.7 | AAA25843.1 | Q00924 | -    | Hydantoin racemase (EC 5.1.99.5)                                                                                                           | 9.5 |
| CAB17078.1 | O24325 | -    | Vacuolar-processing enzyme (VPE) (EC 3.4.22.-) (Legumain-like proteinase) (LLP)                                                                             | 5.6 | AAP07457.1 | Q81IH1 | -    | DNA topoisomerase 3 (EC 5.99.1.2) (DNA topoisomerase III)                                                                                  | 9.8 |
| CAA05487.1 | P80884 | -    | Ananain (EC 3.4.22.31)                                                                                                                                      | 5   | BAD70107.1 | P59846 | 1J1Z | Argininosuccinate synthase (EC 6.3.4.5) (Citrulline--aspartate ligase)                                                                     | 8.5 |
| AAA46752.1 | P25783 | -    | Viral cathepsin (V-cath) (EC 3.4.22.50) (Cysteine proteinase) (CP)                                                                                          | 5   | AAL95357.1 | Q8REE6 | -    | Glutamate racemase (EC 5.1.1.3)                                                                                                            | 8.5 |
| AAT68959.1 | Q6DYE7 | -    | Renin (EC 3.4.23.15) (Angiotensinogenase)                                                                                                                   | 6   | CAA27559.1 | P04789 | 1AG1 | Triosephosphate isomerase, glycosomal (TIM) (Triose-phosphate isomerase) (EC 5.3.1.1)                                                      | 8   |
| EAK99117.1 | P0CY27 | 2QZW | Candidapepsin-1 (EC 3.4.23.24) (ACP 1) (Aspartate protease 1) (Secreted aspartic protease 1)                                                                | 2.5 | BAG09236.2 | B0M3E8 | -    | Bifunctional UDP-glucose 4-epimerase and UDP-xylose 4-epimerase 1 (EC 5.1.3.2) (EC 5.1.3.5) (UDP-D-xylose 4-epimerase) (UDP-L-arabinose 4- | 8.5 |



|            |        |      |                                                                                                                                        |     |            |        |      |                                                                                                                                                                                                                                                                                                                                                                                                                                                         |     |
|------------|--------|------|----------------------------------------------------------------------------------------------------------------------------------------|-----|------------|--------|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
|            |        |      |                                                                                                                                        |     |            |        |      | epimerase) (UDP-galactose 4-epimerase 1) (UDP-glucose 4-epimerase 1) (PsUGE1)                                                                                                                                                                                                                                                                                                                                                                           |     |
| BAG37496.1 | Q9NUN7 | -    | Alkaline ceramidase 3 (AlkCDase 3) (Alkaline CDase 3) (EC 3.5.1.-) (Alkaline dihydroceramidase SB89) (Alkaline phytoceramidase) (aPHC) | 4.5 | CAR79033.1 | B8ZV93 | -    | Tyrosine 2,3-aminomutase (MfTAM) (EC 5.4.3.6) (Tyrosine ammonia-lyase) (EC 4.3.1.23)                                                                                                                                                                                                                                                                                                                                                                    | 8.8 |
| AAH52321.1 | O08914 | -    | Fatty-acid amide hydrolase 1 (EC 3.5.1.99) (Anandamide amidohydrolase 1) (Oleamide hydrolase 1)                                        | 4.5 | CAR79034.1 | B8ZV94 | -    | Tyrosine 2,3-aminomutase (MxTAM) (EC 5.4.3.6) (Tyrosine ammonia-lyase) (EC 4.3.1.23)                                                                                                                                                                                                                                                                                                                                                                    | 8.8 |
| DAA09362.1 | P06169 | 1PVD | Pyruvate decarboxylase isozyme 1 (EC 4.1.1.-) (EC 4.1.1.1) (EC 4.1.1.43) (EC 4.1.1.74)                                                 | 5.7 | CAJ46694.1 | Q0VZ68 | -    | Tyrosine 2,3-aminomutase (EC 5.4.3.6) (Tyrosine ammonia-lyase) (EC 4.3.1.23)                                                                                                                                                                                                                                                                                                                                                                            | 8.8 |
| CAA59953.1 | Q12629 | 2VJY | Pyruvate decarboxylase (EC 4.1.1.1)                                                                                                    | 5.7 | DAA10303.1 | P53848 | 2BMB | Folic acid synthesis protein FOL1 [Includes: Dihydroneopterin aldolase (DHNA) (EC 4.1.2.25) (7,8-dihydroneopterin aldolase) (FASA) (FASB); 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase (HPPK) (EC 2.7.6.3) (2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase) (7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase) (PPPK) (FASC); Dihydropteroate synthase (DHPS) (EC 2.5.1.15) (Dihydropteroate pyrophosphorylase) (FASD)] | 8.5 |
| CAG62667.1 | Q6FJA3 | -    | Pyruvate decarboxylase (EC 4.1.1.1)                                                                                                    | 6   | AAC73957.1 | P75823 | 3WLX | Low specificity L-threonine aldolase (Low specificity L-TA) (EC 4.1.2.48)                                                                                                                                                                                                                                                                                                                                                                               | 8.5 |
| AAB16855.1 | Q9FFT4 | -    | Pyruvate decarboxylase 2 (AtPDC2) (EC 4.1.1.1)                                                                                         | 5.5 | BAA24794.1 | O50584 | -    | Low specificity L-threonine aldolase (Low specificity L-TA) (EC 4.1.2.48)                                                                                                                                                                                                                                                                                                                                                                               | 8   |

|            |        |      |                                                                                              |     |            |        |      |                                                                                                                                                                                                                       |     |
|------------|--------|------|----------------------------------------------------------------------------------------------|-----|------------|--------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| AAA23834.1 | P69910 | 1PMM | Glutamate decarboxylase beta (GAD-beta) (EC 4.1.1.15)                                        | 3.8 | AAA89107.1 | P51017 | -    | 4-hydroxy-2-oxovalerate aldolase (HOA) (EC 4.1.3.39) (4-hydroxy-2-keto-pentanoic acid aldolase) (4-hydroxy-2-oxopentanoate aldolase)                                                                                  | 8   |
| AED92414.1 | Q42521 | 3HBX | Glutamate decarboxylase 1 (GAD 1) (EC 4.1.1.15)                                              | 5.5 | AAA23098.1 | P31013 | 1TPL | Tyrosine phenol-lyase (EC 4.1.99.2) (Beta-tyrosinase)                                                                                                                                                                 | 8.5 |
| AAA64830.1 | Q42521 | -    | Inducible ornithine decarboxylase (ODC) (EC 4.1.1.17)                                        | 6   | ABF50844.1 | Q00645 | -    | Pectate lyase plyB (EC 4.2.2.2)                                                                                                                                                                                       | 8   |
| CAB15314.1 | O34714 | 1J58 | Oxalate decarboxylase OxdC (EC 4.1.1.2)                                                      | 4   | ADB78774.1 | D3JTC1 | -    | Pectate lyase A (EC 4.2.2.2) (Pectin lyase) (EC 4.2.2.10)                                                                                                                                                             | 9.5 |
| CAB13759.1 | O34767 | -    | Oxalate decarboxylase OxdD (EC 4.1.1.2)                                                      | 5   | CAB40884.1 | Q9X6Z2 | -    | Pectate lyase A (EC 4.2.2.2) (Pectin lyase) (EC 4.2.2.10)                                                                                                                                                             | 10  |
| AAL68399.1 | Q8L208 | -    | Alpha-acetolactate decarboxylase (EC 4.1.1.5)                                                | 6   | BAG12908.1 | B1B6T1 | -    | Pectate trisaccharide-lyase (EC 4.2.2.22) (Exopolygalacturonate lyase) (Pectate lyase) (Pel SWU)                                                                                                                      | 10  |
| AAT43070.1 | Q6L1T2 | -    | D-gluconate/D-galactonate dehydratase (GAD) (GNAD) (EC 4.2.1.140) (EC 4.2.1.39) (EC 4.2.1.6) | 6   | AAC04567.1 | O52195 | -    | Alginate lyase (EC 4.2.2.3) (Poly(beta-D-mannuronate) lyase)                                                                                                                                                          | 8.1 |
| CAA11481.1 | O50660 | -    | Alginate lyase (EC 4.2.2.3) (Poly(beta-D-mannuronate) lyase)                                 | 6   | CAB56499.1 | Q9LLR9 | -    | Epi-cedrol synthase (EC 4.2.3.39) (8-epicedrol synthase)                                                                                                                                                              | 8   |
| CAJ70135.1 | Q17ZY4 | -    | Proline racemase (EC 5.1.1.4)                                                                | 6   | ACO59490.1 | C1KKR1 | -    | D-tagatose 3-epimerase (DTE) (EC 5.1.3.31) (D-ribulose 3-epimerase) (EC 5.1.3.-) (Ketose 3-epimerase)                                                                                                                 | 9   |
| AAF61267.1 | Q9P9K9 | -    | DNA ligase (EC 6.5.1.1) (Polydeoxyribonucleotide synthase [ATP])                             | 6   | ABD30079.1 | Q2FZP6 | 4C12 | UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--L-lysine ligase (EC 6.3.2.7) (L-lysine-adding enzyme) (UDP-MurNAc-L-Ala-D-Glu:L-Lys ligase) (UDP-MurNAc-tripeptide synthetase) (UDP-N-acetylmuramyl-tripeptide synthetase) | 8.6 |

Table A1.5: Halophile-Nonhalophile (H-Nh) dataset.

| Halophilic CDS |              |        |                                                                                                                                                                                                                                                                        | Non-halophilic CDS     |            |              |        |
|----------------|--------------|--------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|------------|--------------|--------|
| ENA ID         | UniProtKB ID | PDB ID | Protein name                                                                                                                                                                                                                                                           | Organism name          | ENA ID     | UniProtKB ID | PDB ID |
| AAA83029.1     | Q44532       | 1A8P   | Ferredoxin--NADP reductase (FNR) (Protein X) (EC 1.18.1.2)                                                                                                                                                                                                             | Azotobacter vinelandii | AAD35744.1 | Q9WZC8       | 2ZBU   |
| CAA36681.1     | P20708       | 1GHJ   | Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex (EC 2.3.1.61) (2-oxoglutarate dehydrogenase complex component E2) (OGDC-E2) (Dihydrolipoamide succinyltransferase component of 2-oxoglutarate dehydrogenase complex) | Azotobacter vinelandii | CAA25284.1 | P0AFG6       | 1BAL   |
| AAB03239.1     | P52197       | 1E0C   | Thiosulfate sulfurtransferase (EC 2.8.1.1) (Rhodanese-like protein)                                                                                                                                                                                                    | Azotobacter vinelandii | ABO08125.1 | A3MU08       | 3W6U   |
| SFX65938.1     | Q44529       | 1H9J   | Molybdate transport system regulatory protein (Potential molybdenum-pterin-binding-protein)                                                                                                                                                                            | Azotobacter vinelandii | AAB85058.1 | O26652       | 4MM1   |
| CAA30987.1     | P10802       | 1DPB   | Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex (EC 2.3.1.12) (Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex) (E2)                                                                          | Azotobacter vinelandii | BAB96685.1 | P06959       | 1QJO   |
| BAA11169.1     | P16100       | 1ITW   | Isocitrate dehydrogenase [NADP] (IDH) (EC 1.1.1.42) (Oxalosuccinate decarboxylase)                                                                                                                                                                                     | Azotobacter vinelandii | AAM21720.1 | Q877G8       | 3IYF   |
| AAA64715.1     | Q44534       | 1LRV   | Leucine rich repeat variant (ORF 2)                                                                                                                                                                                                                                    | Azotobacter vinelandii | AAN70724.1 | Q88CM1       | 4IW3   |
| AAA64735.1     | P00324       | 1YOB   | Flavodoxin 2                                                                                                                                                                                                                                                           | Azotobacter vinelandii | AAD35573.1 | Q9WYV8       | 1NV8   |
| AAA22121.1     | P22759       | 1SOF   | Bacterioferritin (BFR) (EC 1.16.3.1) (Cytochrome b-557.5)                                                                                                                                                                                                              | Azotobacter vinelandii | BAE77955.1 | P0ABD3       | 1BCF   |
| CAA46044.1     | P30663       | 2GJ3   | Nitrogen fixation regulatory protein (EC 2.7.13.3)                                                                                                                                                                                                                     | Azotobacter vinelandii | AAD36730.1 | Q9X1Z1       | 4HLU   |
| SFX65599.1     | Q9F5X9       | 1P90   | Nitrogen fixation protein NifX (Nitrogenase gamma subunit)                                                                                                                                                                                                             | Azotobacter vinelandii | AAN66417.1 | Q88PQ6       | 3O74   |
| AAD04920.1     | Q9ZFH0       | 2ML1   | Poly(beta-D-mannuronate) C5 epimerase 6 (EC 5.1.3.-) (Mannuronan epimerase 6)                                                                                                                                                                                          | Azotobacter vinelandii | AAC74543.1 | P31992       | 1GYJ   |

|            |        |      |                                                                                                                        |                                                    |            |        |      |
|------------|--------|------|------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|------------|--------|------|
| AAA87310.1 | Q44493 | 2AGM | Poly(beta-D-mannuronate) C5 epimerase 4 (EC 5.1.3.-) (Mannuronan epimerase 4)                                          | Azotobacter vinelandii                             | AAA81026.1 | P39838 | 1SR2 |
| AAN86140.1 | Q8GHE2 | 2QLE | Green fluorescence protein                                                                                             | Azotobacter vinelandii                             | AAC75021.1 | P76329 | 1XVI |
| CAA36679.1 | P18925 | 3LAD | Dihydrolipoyl dehydrogenase (EC 1.8.1.4) (Dihydrolipoamide dehydrogenase) (E3 component of pyruvate complex)           | Azotobacter vinelandii                             | AAK25432.1 | Q9A2T6 | 4FMT |
| ACO79120.1 | C1DM33 | 3VS8 | Type III polyketide synthase                                                                                           | Azotobacter vinelandii (strain DJ / ATCC BAA-1303) | BAA15508.1 | P77247 | 1TE2 |
| ACO80442.1 | P84253 | 2OGX | Molybdenum storage protein subunit beta (Mo storage protein subunit beta) (MoSto subunit beta)                         | Azotobacter vinelandii (strain DJ / ATCC BAA-1303) | BAA16530.1 | P77295 | 3CUO |
| AAA64709.1 | P00459 | 1DE0 | Nitrogenase iron protein 1 (EC 1.18.6.1) (Nitrogenase Fe protein 1) (Nitrogenase component II) (Nitrogenase reductase) | Azotobacter vinelandii                             | AAD36848.1 | Q9X2A5 | 2E54 |
| AAA64710.1 | P07328 | 1FP4 | Nitrogenase molybdenum-iron protein alpha chain (EC 1.18.6.1) (Dinitrogenase) (Nitrogenase component I)                | Azotobacter vinelandii                             | AAD35449.1 | Q9WYJ7 | 2X7V |
| AAA16869.1 | P00214 | 1A6L | Ferredoxin-1 (Ferredoxin I) (FdI)                                                                                      | Azotobacter vinelandii                             | CAA68816.1 | P17846 | 1AOP |
| AAS42356.1 | Q734F7 | 2O3L | Uncharacterized protein                                                                                                | Bacillus cereus (strain ATCC 10987 / NRS 248)      | AAA97028.1 | P0AF61 | 2LLZ |
| AAS41083.1 | Q739H9 | 2P1A | Uncharacterized protein                                                                                                | Bacillus cereus (strain ATCC 10987 / NRS 248)      | AAB03001.1 | P32131 | 1OLT |
| AAS41379.1 | Q738D3 | 3BJA | Transcriptional regulator, MarR family, putative                                                                       | Bacillus cereus (strain ATCC 10987 / NRS 248)      | AAD36224.1 | Q9X0N2 | 1ZOR |
| AAS43782.1 | Q72YY9 | 3BLN | Uncharacterized protein                                                                                                | Bacillus cereus (strain ATCC 10987 / NRS 248)      | CAC12316.1 | Q9HIY5 | 2OHC |
| AAS40261.1 | Q73BT6 | 3CC8 | Uncharacterized protein                                                                                                | Bacillus cereus (strain ATCC 10987 / NRS 248)      | AGL49465.1 | Q9WZ13 | 3CNI |
| AAS40871.1 | Q73A38 | 3D6W | Methyl-accepting/DNA response regulator, putative                                                                      | Bacillus cereus (strain ATCC 10987 / NRS 248)      | AAA97187.1 | P13036 | 1KMO |
| AAS43556.1 | Q72ZL3 | 3DI5 | Uncharacterized protein                                                                                                | Bacillus cereus (strain ATCC 10987 / NRS 248)      | AGL49846.1 | Q9X021 | 1VLA |
| AAS40968.1 | Q739U3 | 3DUL | O-methyltransferase, putative                                                                                          | Bacillus cereus (strain ATCC 10987 / NRS 248)      | ACL96011.1 | B8GZM2 | 1W25 |
| AAS41117.1 | Q739E5 | 3FGG | Uncharacterized protein                                                                                                | Bacillus cereus (strain ATCC 10987 / NRS 248)      | AAD35678.1 | Q9WZ62 | 4PRS |
| AAS43117.1 | Q731F0 | 3GB0 | Peptidase T (EC 3.4.11.-)                                                                                              | Bacillus cereus (strain ATCC 10987 / NRS 248)      | CAC11977.1 | Q9HJW5 | 4MMN |

|            |        |      |                                                                            |                                                                                                 |            |        |      |
|------------|--------|------|----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|------------|--------|------|
| AAS39848.1 | Q73CZ7 | 3GRD | Uncharacterized protein                                                    | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | AAD35809.1 | Q9WZI6 | 1VL4 |
| AAS41789.1 | Q736M3 | 3H41 | NLP/P60 family protein                                                     | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | AAD36075.1 | Q9X0A3 | 2F4P |
| AAS40951.1 | Q739W0 | 3JU7 | DegT/dnrJ/eryC1/strS family protein                                        | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | AGL50311.1 | Q9X1A0 | 1Z85 |
| AAP11787.1 | Q816E8 | 3BVS | Uncharacterized protein                                                    | Bacillus cereus (strain ATCC 14579 / DSM 31 / JCM 2152 / NBRC 15305 / NCIMB 9373 / NRRL B-3711) | AGL50173.1 | Q9X0X1 | 1VQ3 |
| AAS39861.1 | Q73CY4 | 3K0Z | Lipoprotein, putative                                                      | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | AAD36207.1 | Q9X0L5 | 1VP4 |
| AAS45086.1 | Q74NK7 | 3KW0 | Uncharacterized protein                                                    | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | AAD35665.1 | Q9WZ49 | 2CE7 |
| AAS42332.1 | Q734I1 | 3L5Z | Transcriptional regulator, GntR family                                     | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | CAA47308.1 | P0A9T0 | 1PSD |
| AAS43902.1 | Q72YM0 | 3N73 | Uncharacterized protein                                                    | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | ABO07588.1 | A3MSH1 | 4DHD |
| AAS44225.1 | Q72XQ0 | 3QNQ | Permease IIC component                                                     | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | AGL50195.1 | Q9X0Z3 | 2NZC |
| AAS39998.1 | Q73CJ8 | 3RAO | Bacterial luciferase family protein                                        | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | AAD36713.1 | Q9X1X9 | 2P61 |
| AAS44920.1 | Q74P24 | 4EI7 | Plasmid replication protein RepX                                           | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | BAD86474.1 | Q5JDA3 | 3W4S |
| AAS42357.1 | Q734F6 | 4ESF | Uncharacterized protein                                                    | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | AAD35400.1 | Q9WYE8 | 1ZH8 |
| AAS39434.1 | Q73E61 | 4EUY | Uncharacterized protein                                                    | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | BAE76108.1 | P75691 | 1UUF |
| AAS39312.1 | Q73EI2 | 4JG2 | Phage-related protein                                                      | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | BAD84190.1 | P77933 | 1WN7 |
| AAS41032.1 | Q739M9 | 4JH1 | Metallothiol transferase FosB (EC 2.5.1.-) (Fosfomycin resistance protein) | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | CAC11653.1 | Q9HKT2 | 3R07 |
| AAS41902.1 | Q736B0 | 4ME9 | Transcriptional regulator, TetR family                                     | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | BAA08115.1 | Q52428 | 1B8A |
| AAS39177.1 | Q73EW7 | 4QVU | Uncharacterized protein                                                    | Bacillus cereus (strain ATCC 10987 / NRS 248)                                                   | AAD35946.1 | Q9WZW8 | 1WW1 |
| AJI07030.1 | Q4MK66 | -    | Purine nucleoside phosphorylase DeoD-type (PNP) (EC 2.4.2.1)               | Bacillus cereus G9241                                                                           | BAD84374.1 | Q5JFM9 | 3A11 |
| AJI07684.1 | Q4MW04 | 3D19 | Uncharacterized protein                                                    | Bacillus cereus G9241                                                                           | BAA02590.1 | P24215 | 4EAC |

|            |        |      |                                                                                                                                          |                                                                                           |            |        |      |
|------------|--------|------|------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|------------|--------|------|
| AJI02883.1 | Q4MWP8 | 3DBY | Uncharacterized protein                                                                                                                  | Bacillus cereus G9241                                                                     | AAD36489.1 | Q9X1D6 | 3CIN |
| EAL15945.1 | Q4MV79 | 4FK7 | Putative ADP-ribosyltransferase Certhrax (EC 2.4.2.-) (Toxin Certhrax)                                                                   | Bacillus cereus                                                                           | AAN66586.1 | Q88P91 | 4FCZ |
| AEK80404.1 | G1ED17 | 3ATG | Beta-1,3-glucanase (EC 3.2.1.39)                                                                                                         | Cellulosimicrobium cellulans (Arthrobacter luteus)                                        | AGL48947.1 | Q9WXN1 | 1GUI |
| ABE59821.1 | Q1QUN7 | 3AY3 | NAD-dependent epimerase/dehydratase                                                                                                      | Chromohalobacter salexigens (strain DSM 3043 / ATCC BAA-138 / NCIMB 13768)                | AAC75909.1 | P66899 | 4D9G |
| ABE60087.1 | Q1QTX1 | 3LWD | 6-phosphogluconolactonase (EC 3.1.1.31)                                                                                                  | Chromohalobacter salexigens (strain DSM 3043 / ATCC BAA-138 / NCIMB 13768)                | AAN68372.1 | Q88J85 | 4C76 |
| ABE59828.1 | Q1QUN0 | 3MZN | D-glucarate dehydratase (EC 4.2.1.40)                                                                                                    | Chromohalobacter salexigens (strain DSM 3043 / ATCC BAA-138 / NCIMB 13768)                | ABE42492.1 | Q12G50 | 3UM9 |
| ABE60052.1 | Q1QU06 | 4JCI | 4-hydroxyproline 2-epimerase (4Hyp 2-epimerase) (4HypE) (EC 5.1.1.8)                                                                     | Chromohalobacter salexigens (strain DSM 3043 / ATCC BAA-138 / NCIMB 13768)                | AAN65932.1 | Q88R33 | 2HLJ |
| ABE60319.1 | Q1QT89 | 3BSM | D-galactonate dehydratase family member ManD (EC 4.2.1.-) (D-gluconate dehydratase) (EC 4.2.1.39) (D-mannonate dehydratase) (EC 4.2.1.8) | Chromohalobacter salexigens (strain DSM 3043 / ATCC BAA-138 / NCIMB 13768)                | AAK22519.1 | Q9AAR4 | 3VCN |
| ABE58021.1 | Q1QZT7 | 4N8G | TRAP dicarboxylate transporter, DctP subunit                                                                                             | Chromohalobacter salexigens (strain DSM 3043 / ATCC BAA-138 / NCIMB 13768)                | AAB85363.1 | Q04926 | 1IIO |
| ABE59826.1 | Q1QUN2 | 4P1L | Solute-binding protein Csa_2479                                                                                                          | Chromohalobacter salexigens (strain DSM 3043 / ATCC BAA-138 / NCIMB 13768)                | ABE45021.1 | Q128M1 | 4MHF |
| ABE58039.1 | Q1QZR9 | 4UAB | Twin-arginine translocation pathway signal                                                                                               | Chromohalobacter salexigens (strain DSM 3043 / ATCC BAA-138 / NCIMB 13768)                | BAA03950.1 | P36659 | 2KQX |
| AAG20406.1 | P00216 | 1EOZ | Ferredoxin                                                                                                                               | Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium) | CAC12556.1 | Q9HIA5 | 2L5T |
| AAG20524.1 | Q9HMP7 | 1MOJ | DNA protection during starvation protein (EC 1.16.-.-) (Bacterioferritin DpsA)                                                           | Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium) | AAT48134.1 | P76187 | 1OG6 |
| BAB17308.1 | P61136 | 2AZ1 | Nucleoside diphosphate kinase (NDK) (NDP kinase) (EC 2.7.4.6) (Nucleoside-2-P kinase)                                                    | Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium) | AAA58136.1 | P0CE47 | 1D8T |
| AAG19753.1 | Q9HPW4 | 1MOG | Uncharacterized protein                                                                                                                  | Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium) | ABE46141.1 | Q124A1 | 4XK2 |
| CAB37866.1 | B0R2U4 | 1E12 | Halorhodopsin (HR)                                                                                                                       | Halobacterium salinarum (strain ATCC 29341 / DSM 671 / R1)                                | AAN67002.1 | Q88N37 | 1RE5 |
| CAP14567.1 | B0R748 | 4AF1 | Peptide chain release factor subunit 1                                                                                                   | Halobacterium salinarum (strain ATCC                                                      | CAA04517.1 | O33832 | 2P3N |



|            |        |      |                                                                                          |                                                                                                                                    |            |        |      |
|------------|--------|------|------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
|            |        |      | (Translation termination factor aRF1)                                                    | 29341 / DSM 671 / R1)                                                                                                              |            |        |      |
| CAP14037.1 | B0R5M0 | 1MOG | Dodecin                                                                                  | Halobacterium salinarum (strain ATCC 29341 / DSM 671 / R1)                                                                         | ABE42204.1 | Q12GY8 | 4IQG |
| BAM20977.1 | I4DST7 | 4FBZ | Deltarhodopsin (Fragment)                                                                | Haloterrigena thermotolerans                                                                                                       | AAD36468.1 | Q9X1B7 | 4Q2E |
| BAA06680.1 | P94854 | 4JR8 | Cruxrhodopsin-3 (COP-3) (CR-3)                                                           | Haloarcula vallismortis (Halobacterium vallismortis)                                                                               | CAA55388.1 | P0A7E9 | 2BND |
| AAG18851.1 | Q9HSF6 | 4E19 | Ribonuclease HI (Halo-RNase HI) (RNase HI) (EC 3.1.26.4)                                 | Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium)                                          | AAN67924.1 | Q88KH9 | 3RCM |
| CAA49774.1 | P02945 | 1AP9 | Bacteriorhodopsin (BR) (Bacterioopsin) (BO)                                              | Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium)                                          | BAC92138.1 | Q7NDN8 | 2XQ3 |
| AHZ22740.1 | B8ZYW1 | 4OZJ | Ammonium transporter (Nitrogen regulatory protein P-II) (PII family protein glnK2)       | Haloferax mediterranei (strain ATCC 33500 / DSM 1411 / JCM 8866 / NBRC 14739 / NCIMB 2177 / R-4) (Halobacterium mediterranei)      | AAK24710.1 | Q9A4T4 | 2RAG |
| ADE04659.1 | P15093 | 1VDR | Dihydrofolate reductase HdrA (DHFR A) (hDHFR-1) (EC 1.5.1.3)                             | Haloferax volcanii (strain ATCC 29605 / DSM 3757 / JCM 8879 / NBRC 14742 / NCIMB 2012 / VKM B-1768 / DS2) (Halobacterium volcanii) | BAC91042.1 | Q7NGR7 | 3GD5 |
| ELY26392.1 | D4GTC1 | 4B45 | Tubulin-like protein CetZ2 (Cell-structure-related euryarchaeota tubulin/FtsZ homolog 2) | Haloferax volcanii (strain ATCC 29605 / DSM 3757 / JCM 8879 / NBRC 14742 / NCIMB 2012 / VKM B-1768 / DS2) (Halobacterium volcanii) | AAC76989.1 | P13009 | 1BMT |
| ELY33433.1 | D4GVD7 | 4B46 | Tubulin-like protein CetZ1 (Cell-structure-related euryarchaeota tubulin/FtsZ homolog 1) | Haloferax volcanii (strain ATCC 29605 / DSM 3757 / JCM 8879 / NBRC 14742 / NCIMB 2012 / VKM B-1768 / DS2) (Halobacterium volcanii) | BAE77097.1 | P0A7J0 | 1G57 |
| ABH10498.1 | Q9P9L2 | 4BGU | Malate dehydrogenase (EC 1.1.1.37)                                                       | Haloferax volcanii (strain ATCC 29605 / DSM 3757 / JCM 8879 / NBRC 14742 / NCIMB 2012 / VKM B-1768 / DS2) (Halobacterium volcanii) | AAA53657.1 | P30178 | 2G8Y |
| CBA13558.1 | E1VBK4 | 3HGM | TRAP-T-associated universal stress protein TeaD (UspA domain transporter regulator TeaD) | Halomonas elongata (strain ATCC 33173 / DSM 2581 / NBRC 15536 / NCIMB 2198 / 1H9)                                                  | AAD36761.1 | Q9X220 | 4KR6 |
| AE008691.1 | Q3ITX1 | 5B0W | Halorhodopsin                                                                            | Natronomonas pharaonis (strain ATCC 35678 / DSM 2160 / CIP 103997 / NBRC 14720 / NCIMB 2260 /                                      | AAC74259.1 | P0AEZ3 | 3Q9L |

|            |        |      |                                                                     |                                                                                                           |            |        |      |
|------------|--------|------|---------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|------------|--------|------|
|            |        |      |                                                                     | Gabara (Halobacterium pharaonis)                                                                          |            |        |      |
| ABC44045.1 | Q2S0Y5 | 2KCQ | Mov34/MPN/PAD-1 family                                              | Salinibacter ruber (strain DSM 13855 / M31)                                                               | BAE76119.1 | P25524 | 1K6W |
| ABC44923.1 | Q2S6C5 | 2KCL | Tetratricopeptide repeat domain protein                             | Salinibacter ruber (strain DSM 13855 / M31)                                                               | AAC76779.1 | P05793 | 1YRL |
| ABC45135.1 | Q2S289 | 3NEP | Malate dehydrogenase (EC 1.1.1.37)                                  | Salinibacter ruber (strain DSM 13855 / M31)                                                               | AAD35644.1 | Q9WZ29 | 2ANU |
| ABC45176.1 | Q2S0R5 | 3USH | Uncharacterized protein                                             | Salinibacter ruber (strain DSM 13855 / M31)                                                               | AAN66010.1 | Q88QV5 | 1XTO |
| AAG19484.1 | Q9HQM9 | 2GF4 | Uncharacterized protein                                             | Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium)                 | AGL50683.1 | Q9X273 | 3AMC |
| AAG19211.1 | Q9HRE7 | 2LOK | Uncharacterized protein                                             | Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium)                 | AAA23738.1 | P0AAI9 | 1MLA |
| AAV45400.1 | Q5V502 | 2LNU | Uncharacterized protein                                             | Haloarcula marismortui (strain ATCC 43049 / DSM 3752 / JCM 8966 / VKM B-1809) (Halobacterium marismortui) | AAB95620.1 | O32583 | 1F0Z |
| AAV47110.1 | Q5V043 | 3B73 | PhiH1 repressor-like                                                | Haloarcula marismortui (strain ATCC 43049 / DSM 3752 / JCM 8966 / VKM B-1809) (Halobacterium marismortui) | AAB03058.1 | P0A6F3 | 1BO5 |
| AAV45155.1 | Q5V5P7 | 3BWL | HTR-like protein                                                    | Haloarcula marismortui (strain ATCC 43049 / DSM 3752 / JCM 8966 / VKM B-1809) (Halobacterium marismortui) | AAF08344.2 | Q9RQQ9 | 4Q20 |
| AAV45003.1 | Q5V649 | 3CRJ | Transcription regulator                                             | Haloarcula marismortui (strain ATCC 43049 / DSM 3752 / JCM 8966 / VKM B-1809) (Halobacterium marismortui) | AAD36356.1 | Q9X108 | 1UP4 |
| AAV45548.1 | Q5V4K4 | 3QTA | Chemotaxis protein CheC                                             | Haloarcula marismortui (strain ATCC 43049 / DSM 3752 / JCM 8966 / VKM B-1809) (Halobacterium marismortui) | AAD36860.1 | Q9X2B7 | 4XTK |
| CAA76423.1 | O59651 | 1ITK | Catalase-peroxidase 2 (CP 2) (EC 1.11.1.21) (Peroxidase/catalase 2) | Haloarcula marismortui (strain ATCC 43049 / DSM 3752 / JCM 8966 / VKM B-1809) (Halobacterium marismortui) | AAB84713.1 | O26309 | 2KCP |
| AAV47867.1 | Q5UXY6 | 4PKK | Bacteriorhodopsin-I (HmBRI)                                         | Haloarcula marismortui (strain ATCC 43049 / DSM 3752 / JCM 8966 / VKM B-1809) (Halobacterium marismortui) | CAC12009.1 | Q9HJT3 | 3BF5 |
| AAG19753.1 | Q9HPW4 | 1MOG | Uncharacterized protein                                             | Haloarcula marismortui (strain ATCC 43049 / DSM 3752 / JCM 8966 / VKM B-1809) (Halobacterium marismortui) | AAN66545.1 | Q88PD0 | 4AB4 |

|            |        |      |                                                                                              |                                                                                                                                    |            |        |      |
|------------|--------|------|----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|------------|--------|------|
| CAP14037.1 | B0R5M0 | 1MOG | Dodecin                                                                                      | Halobacterium salinarum (strain ATCC 29341 / DSM 671 / R1)                                                                         | AAA96989.1 | P37351 | 1NN4 |
| CAP15544.1 | B0R9W3 | 2X98 | Alkaline phosphatase (EC 3.1.3.1)                                                            | Halobacterium salinarum (strain ATCC 29341 / DSM 671 / R1)                                                                         | AAK25702.1 | Q9A226 | 3CZB |
| ADE02783.1 | D4GVB0 | 2M19 | Small archaeal modifier protein 3 (SAMP3) (Ubiquitin-like small archaeal modifier protein 3) | Haloferax volcanii (strain ATCC 29605 / DSM 3757 / JCM 8879 / NBRC 14742 / NCIMB 2012 / VKM B-1768 / DS2) (Halobacterium volcanii) | AAT48238.1 | P32664 | 1VK6 |
| CAC48389.1 | D4GTL2 | 3OYX | Malate synthase (MSH) (EC 2.3.3.9)                                                           | Haloferax volcanii (strain ATCC 29605 / DSM 3757 / JCM 8879 / NBRC 14742 / NCIMB 2012 / VKM B-1768 / DS2) (Halobacterium volcanii) | CAC12192.1 | Q9HJA6 | 3CTA |
| ADE04519.1 | D4GUF6 | 2L83 | Small archaeal modifier protein 1 (SAMP1) (Ubiquitin-like small archaeal modifier protein 1) | Haloferax volcanii (strain ATCC 29605 / DSM 3757 / JCM 8879 / NBRC 14742 / NCIMB 2012 / VKM B-1768 / DS2) (Halobacterium volcanii) | BAA16282.2 | P15042 | 2OWO |
| ADE03392.1 | D4GZE7 | 2L32 | Small archaeal modifier protein 2 (SAMP2) (Ubiquitin-like small archaeal modifier protein 2) | Haloferax volcanii (strain ATCC 29605 / DSM 3757 / JCM 8879 / NBRC 14742 / NCIMB 2012 / VKM B-1768 / DS2) (Halobacterium volcanii) | AAA23881.1 | P0AFB5 | 1R62 |
| ACV49264.1 | C7P2E9 | 3U1D | Uncharacterized protein                                                                      | Halomicrobium mukohataei (strain ATCC 700874 / DSM 12286 / JCM 9738 / NCIMB 13541) (Haloarcula mukohataei)                         | AAD36653.1 | Q9X1S2 | 1VKW |
| ACV46878.1 | C7NZX0 | 4KVH | Uncharacterized protein                                                                      | Halomicrobium mukohataei (strain ATCC 700874 / DSM 12286 / JCM 9738 / NCIMB 13541) (Haloarcula mukohataei)                         | AAC73352.1 | P75677 | 2HJJ |

Table A1.6: Barophile-Nonbarophile (B-Nb) dataset.

| Barophilic CDS |              |        |                                                        |                       | Non-barophilic CDS |              |        |
|----------------|--------------|--------|--------------------------------------------------------|-----------------------|--------------------|--------------|--------|
| EBI-ENA ID     | UniProtKB ID | PDB ID | Protein name                                           | Organism name         | EBI-ENA ID         | UniProtKB ID | PDB ID |
| CCE70035.1     | Q9V0Y8       | 1H64   | SnRNP SM-Like Protein                                  | Pyrococcus abyssi     | AAL81996.1         | Q8TZV6       | 1SQR   |
| CAB49216.1     | Q9V1Y2       | 4JIU   | Putative uncharacterized protein                       | Pyrococcus abyssi     | AAN55697.1         | Q8EDS4       | 2NR5   |
| CAB49735.1     | Q9V0G8       | 2V7F   | RPS19E SSU Ribosomal Protein S19E                      | Pyrococcus abyssi     | AAL80184.1         | Q8U4M1       | 2NNW   |
| CAB49760.1     | Q9V0E4       | 1YZ6   | Probable translation initiation factor 2 alpha subunit | Pyrococcus abyssi     | AAL81254.1         | Q8U1S5       | 4W8X   |
| CCE69434.1     | Q9V2L3       | 2QEN   | Walker-Type Atpase                                     | Pyrococcus abyssi     | AAL81023.1         | Q8U2E0       | 2PK8   |
| CAB49082.1     | Q9V2B6       | 2DVY   | Restriction endonuclease PabI                          | Pyrococcus abyssi     | AAN53089.1         | Q8EKT7       | 2XUT   |
| CAB49050.1     | Q9V2E8       | 2VLD   | UPF0286 Protein PYRAB01260                             | Pyrococcus abyssi     | AAL80320.1         | P83194       | 1QXJ   |
| CCE70558.1     | Q9UZK4       | 4CVN   | Putative Adenylate Kinase                              | Pyrococcus abyssi GE5 | AAN56751.1         | Q8EAX1       | 2IIU   |
| CAB50232.1     | P77918       | 1ML4   | Aspartate Transcarbamoylase                            | Pyrococcus abyssi     | AAL81019.1         | Q8U2E3       | 1SGW   |
| CAB49000.1     | Q9V2J8       | 2BFW   | Glga Glycogen Synthase                                 | Pyrococcus abyssi     | AAL81248.1         | Q8U1T1       | 4W8V   |
| CCE71000.1     | Q9UYG2       | 2VUG   | PAB1020                                                | Pyrococcus abyssi GE5 | AAL81235.1         | Q8U1U4       | 2QM3   |
| CCE69622.1     | Q9V228       | 3P8T   | AsnS-like asparaginyl-tRNA synthetase related protein  | Pyrococcus abyssi     | AAN54243.1         | Q8EHP5       | 2R76   |
| CCE70489.1     | Q9UZR7       | 2JJQ   | Uncharacterized Rna Methyltransferase Pyrab10780       | Pyrococcus abyssi     | CAA86244.1         | P40555       | 3WHJ   |
| CAB49614.1     | Q9V0T9       | 1ZQ1   | Glutamyl-tRNA(Gln) amidotransferase subunit D          | Pyrococcus abyssi     | AAL82171.1         | Q8TZE8       | 4NJE   |
| CAB49304.1     | Q9V1P5       | 4A8E   | Probable Tyrosine Recombinase XERC-Like                | Pyrococcus abyssi     | AAN55220.1         | Q8EF26       | 2JUW   |
| CCE71245.1     | Q9UXT7       | 2IVN   | O-Sialoglycoprotein Endopeptidase                      | Pyrococcus abyssi     | CAA56040.1         | P29603       | 1SIZ   |
| CAB49389.1     | Q9V1G0       | 1KJZ   | eIF2gamma                                              | Pyrococcus abyssi     | AAL81841.1         | Q8U082       | 2D74   |
| CCE69911.1     | Q9V1A5       | 2AUS   | pseudouridine synthase                                 | Pyrococcus abyssi     | AAL81909.1         | Q7LWY0       | 2EY4   |
| BAI70419.1     | D2YZL2       | 3VMK   | 3-isopropylmalate dehydrogenase                        | Shewanella benthica   | AAN57206.1         | Q8E9N3       | 3VKZ   |
| CAB49532.1     | Q9V119       | 2PNZ   | Probable exosome complex exonuclease 1                 | Pyrococcus abyssi     | AAL80594.1         | Q8U3J6       | 1XNE   |
| BAE48285.1     | Q33E90       | 2ZE0   | Alpha-glucosidase                                      | Geobacillus sp.       | AAF65616.1         | Q9P9M8       | 1XI9   |

|            |        |      |                                                      |                               |            |        |      |
|------------|--------|------|------------------------------------------------------|-------------------------------|------------|--------|------|
| CAB50304.1 | Q9UYV8 | 3IVZ | Nitrilase                                            | Pyrococcus abyssi             | AAN56824.1 | Q8EAP9 | 2OX6 |
| CAB49588.1 | P62008 | 1PXW | LSU ribosomal protein L7AE                           | Pyrococcus abyssi             | AAL81491.1 | Q8U160 | 2HVV |
| CCE70299.1 | Q9V099 | 1YK4 | Rubredoxin                                           | Pyrococcus abyssi             | AAL81406.1 | P24297 | 1BQ8 |
| CAB49044.1 | Q9V2F4 | 2LCJ | Pab polC intein                                      | Pyrococcus abyssi             | CAB57529.1 | Q9UXG1 | 2M45 |
| CCE69600.1 | Q9V250 | 3WOV | Oligosaccharyl transferase                           | Pyrococcus abyssi             | AAL80661.1 | Q8U3D2 | 1U04 |
| CAB49895.1 | Q9V011 | 1MKH | C-terminal domain of Methionyl-tRNA synthetase       | Pyrococcus abyssi             | AAL81814.1 | Q8U0A9 | 1ZCO |
| CAB50248.1 | Q9UZ14 | 1Y2Q | Threonyl-tRNA synthetase                             | Pyrococcus abyssi             | AAN56890.1 | Q8EAI7 | 3H2D |
| CAB50343.1 | Q9UYR9 | 1YR6 | ATP(GTP)binding protein                              | Pyrococcus abyssi             | AAL80151.1 | Q8U4Q3 | 2FYH |
| CAB49062.1 | Q9V2D6 | 2QMI | Pbp related beta-lactamase                           | Pyrococcus abyssi             | AAN56434.1 | Q8EBR3 | 1T0A |
| CAB50155.1 | Q9UZA4 | 3BK7 | ABC transporter ATP-binding protein                  | Pyrococcus abyssi             | AAL81895.1 | Q8U046 | 1YD7 |
| CAB49352.1 | Q9V1J7 | 3LGA | SAM-dependent methyltransferase                      | Pyrococcus abyssi             | AAL81253.1 | Q8U1S6 | 3UNG |
| CAB50625.1 | P0CL77 | 4FLT | DNA polymerase 1                                     | Pyrococcus abyssi             | AAN54832.1 | Q8EG33 | 4LMH |
| CAA90888.1 | P0CL76 | 4FLU | Pyrococcus abyssi B family DNA polymerase            | Pyrococcus abyssi             | AAN53784.1 | Q8EIX3 | 4PU3 |
| ABN58504.1 | A5Y2E1 | 4HR1 | Putative uncharacterized protein                     | Pyrococcus abyssi virus 1     | AAL80819.1 | Q8U2Y3 | 3CAX |
| CAB49513.1 | Q9V138 | 4MNO | Translation initiation factor 1A                     | Pyrococcus abyssi             | AAL81250.1 | Q8U1S9 | 4RDP |
| AAB99223.1 | Q58610 | 4QHF | Uncharacterized protein MJ1213                       | Methanocaldococcus jannaschii | AAN54831.1 | Q8EG34 | 4LM8 |
| BAD85153.1 | Q5JIB9 | 3J15 | Protein pelota                                       | Pyrococcus abyssi             | AAL80952.1 | Q8U2K6 | 3RJZ |
| CAB50592.1 | Q9UY20 | 2V94 | 30S Ribosomal Protein S24E                           | Pyrococcus abyssi             | AAL81252.1 | Q8U1S7 | 3W2V |
| CAB50204.1 | Q9UZ55 | 4CZ8 | NA <sup>+</sup> /H <sup>+</sup> Antiporter, Putative | Pyrococcus abyssi GE5         | AAL80661.1 | Q8U3D2 | 1U04 |

Table A1.7: Python script for 1-9 scale ranking

```

class Person(object):
    def __init__(self, name, height):
        self.name = name
        self.height = height

names = ['HI', 'DS', 'MMH', 'MSH', 'SSH', 'II', 'AAI', 'ASI', 'CPA', 'SB', 'HB', 'PASA', 'CASA', 'PV', 'AH', 'BS']
heights = [1.10696005277567, 2.69909269709271, 0.976045729173392, 1.3340092584508, 1.34841757916474,
            1.03393022018831, 2.17159680009998, 1.30441635482715, 1.18909490901365, 1.13008726634559,
            0.964636387503282, 1.19531329597555, 0.867774420946626, 1.29096008466275, 0.825164070420332,
            1.2930168930356]

people = [Person(n, h) for (n, h) in zip(names, heights)]

_min = Person('Kleiner', min(heights))
_max = Person('Magnusson', max(heights))

scale_min, scale_max = 1, 9

attrs = ['height']
for attr in attrs:
    x, X = getattr(_min, attr), getattr(_max, attr)
    if X - x <= 0:
        raise StandardError('b0rked input')
    for person in people:
        rel_in_0_1 = (getattr(person, attr) - x) / (X - x)
        setattr(person, 'rel_{}'.format(attr), int(round(rel_in_0_1 * (scale_max - scale_min) + scale_min)))

for person in people:
    print """name: {p.name}
          rel height: {p.rel_height},
          """.format(p=person)

```



## Appendix II

## THERMOPHILE-MESOPHILE CODON

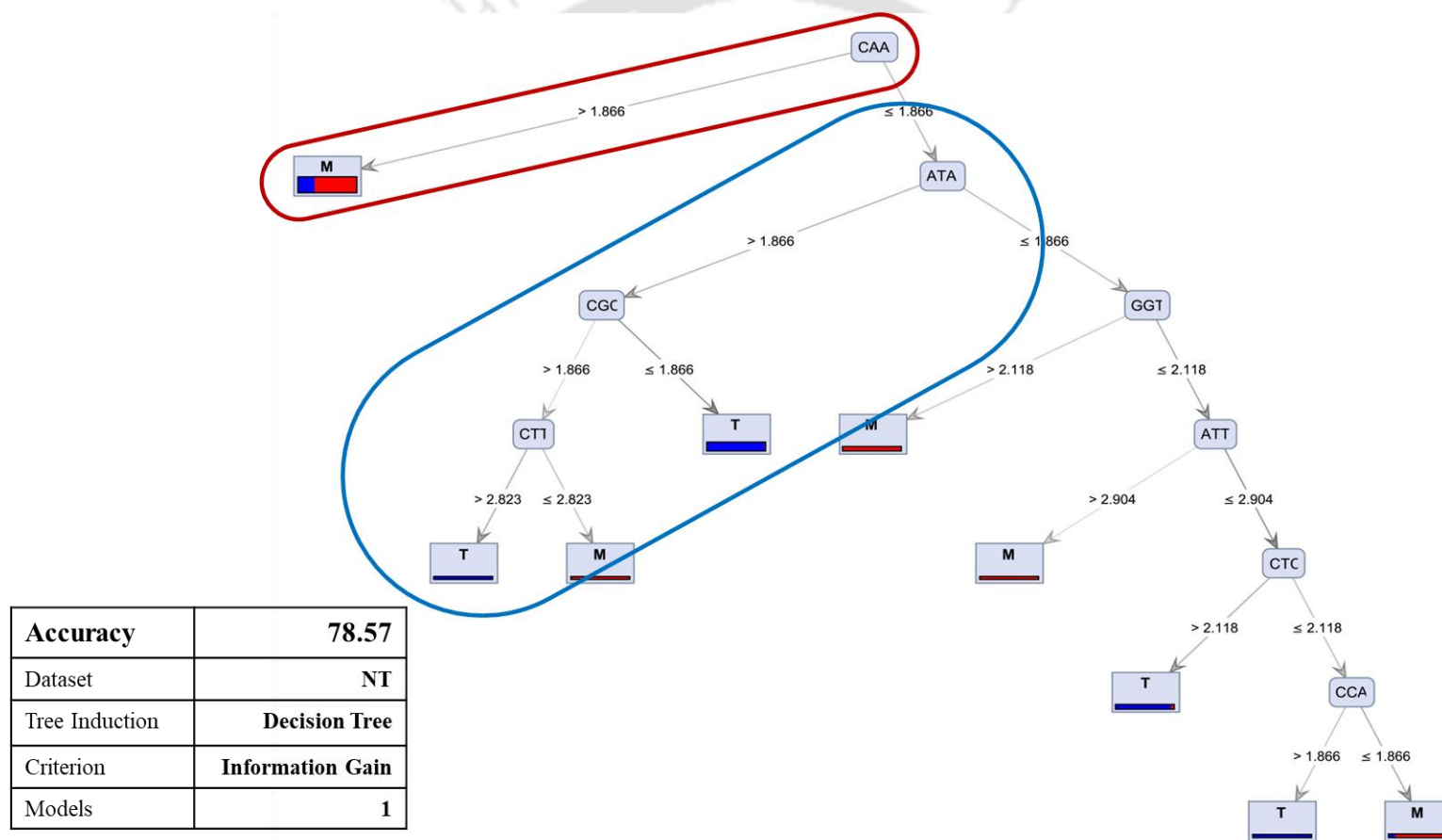


Figure A2.1: Decision tree prediction for codon in T-M dataset.

## PSYCHROPHILE-MESOPHILE CODON

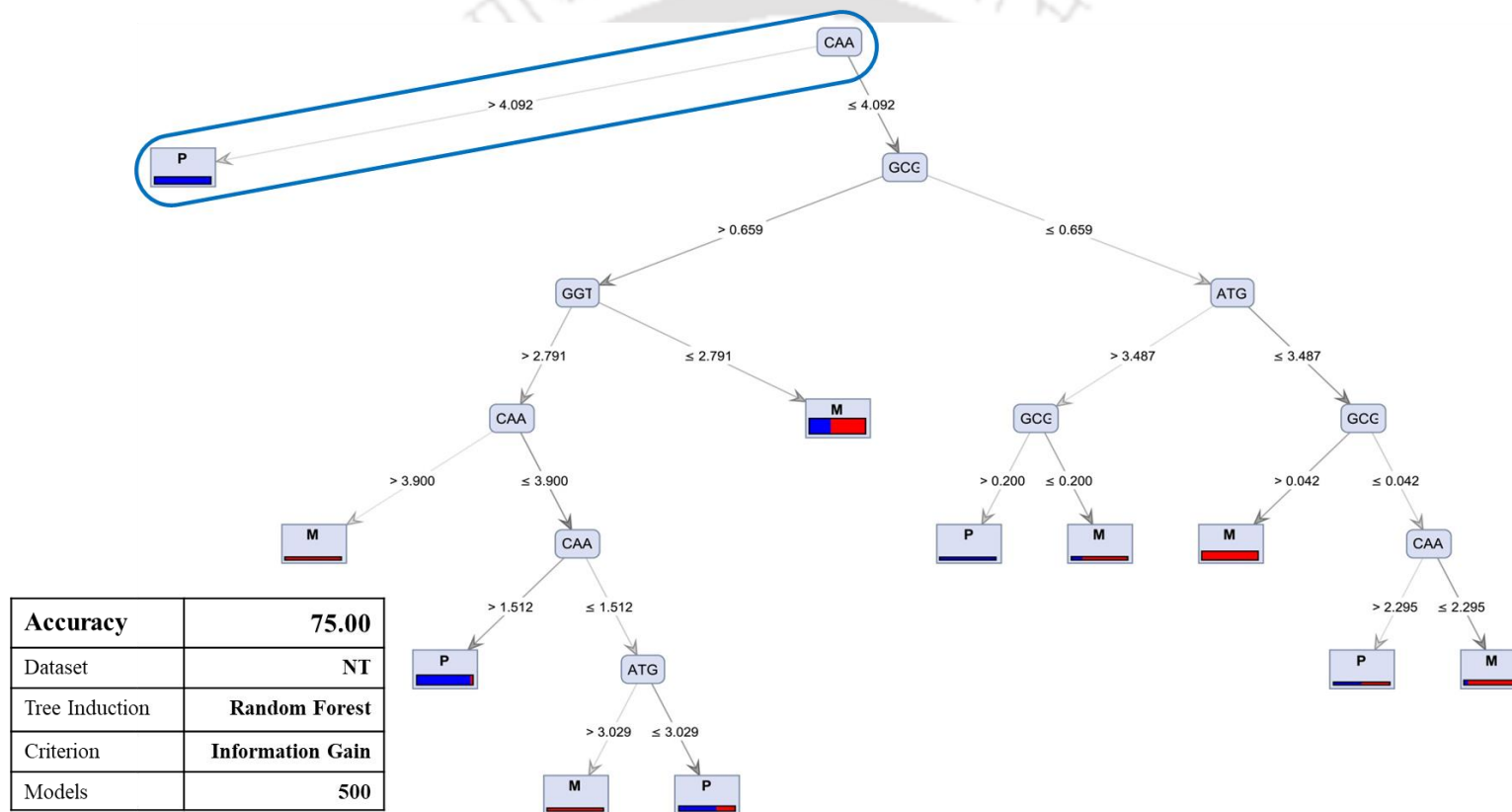


Figure A2.2: Decision tree prediction for codon in P-M dataset.

## PSYCHROPHILE-THERMOPHILE CODON

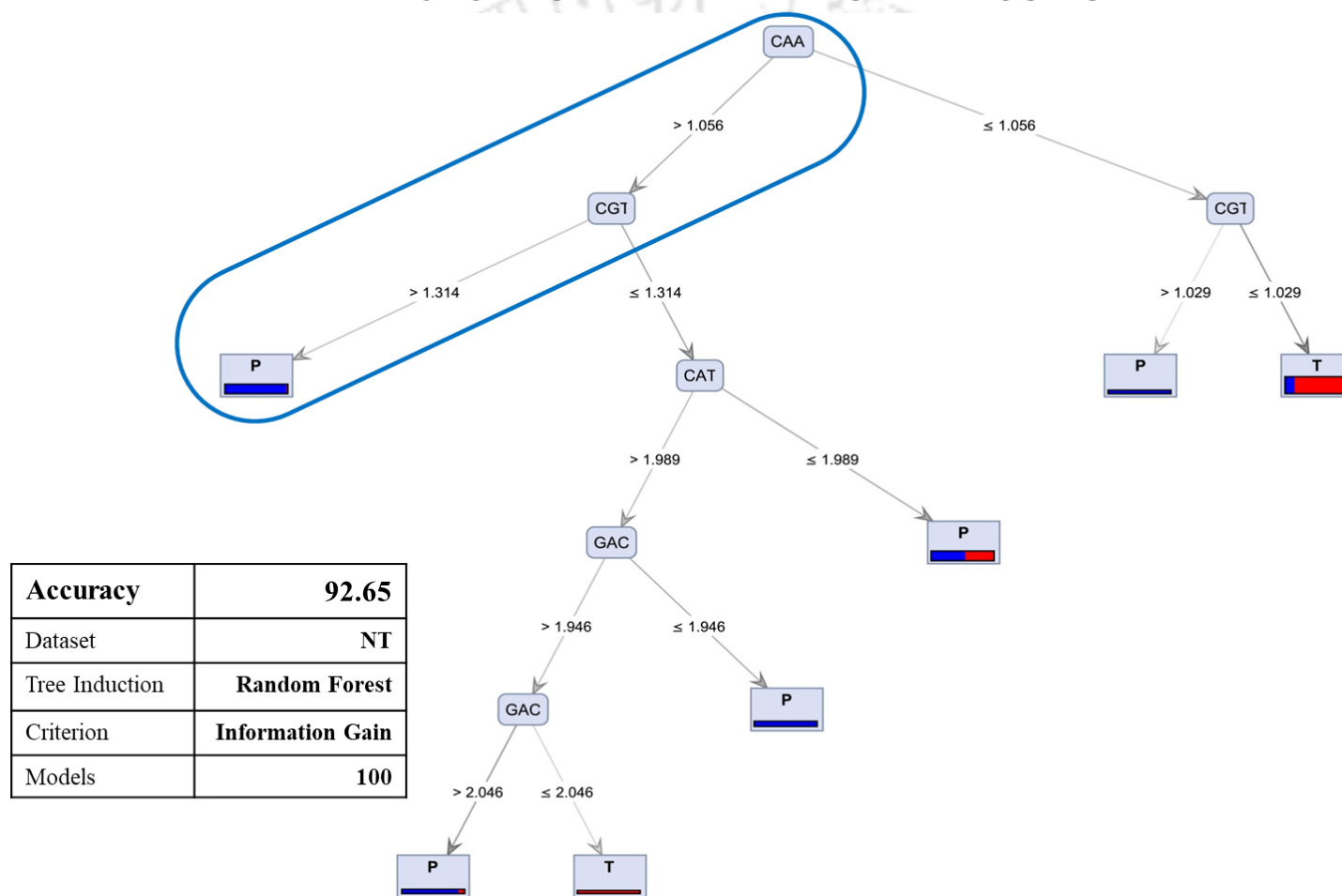


Figure A2.3: Decision tree prediction for codon in T-P dataset

## ALKALIPHILE-ACIDOPHILE CODON

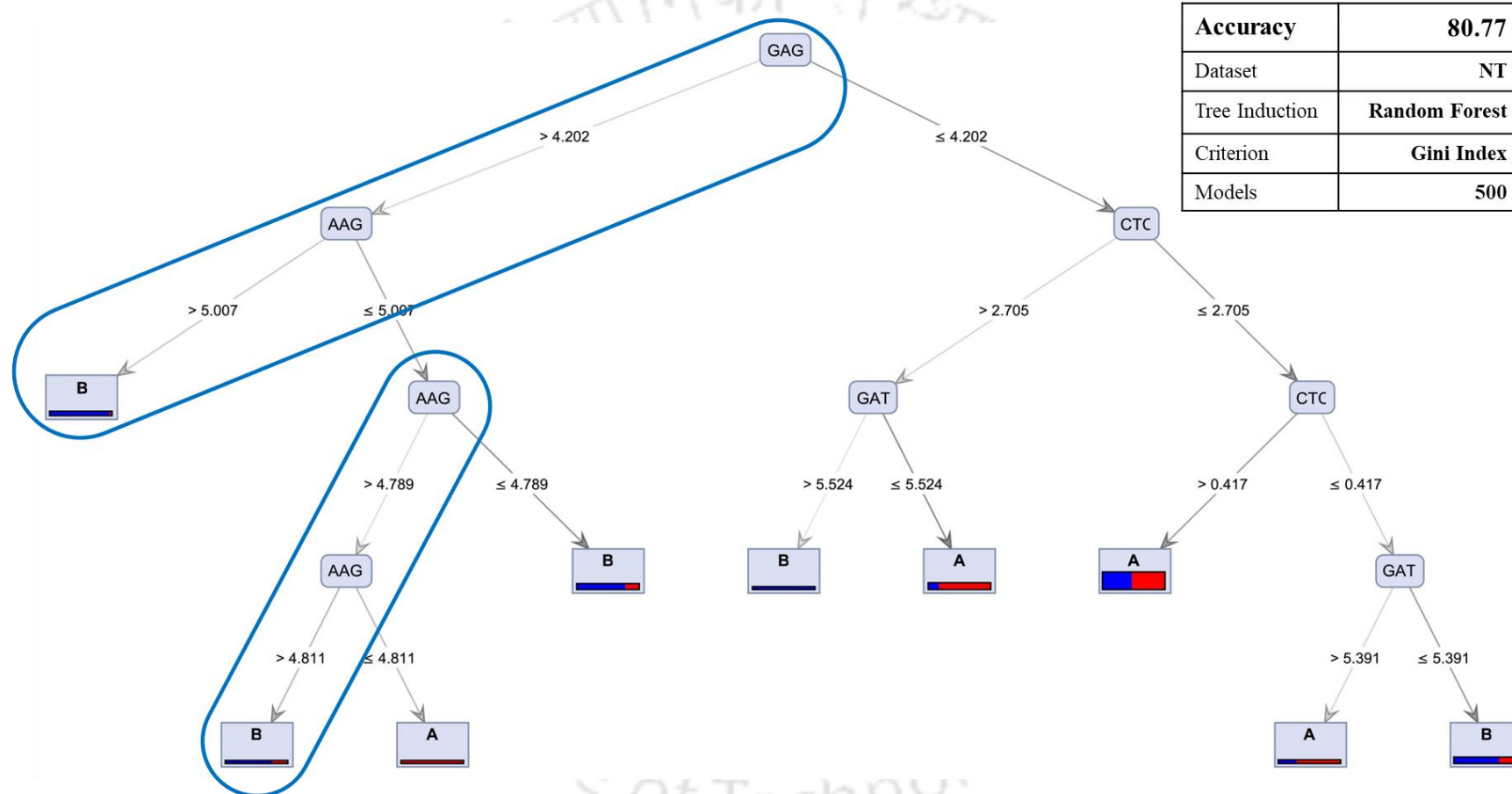


Figure A2.4: Decision tree prediction for codon in A-B dataset.

## HALOPHILE-NON-HALOPHILE CODON

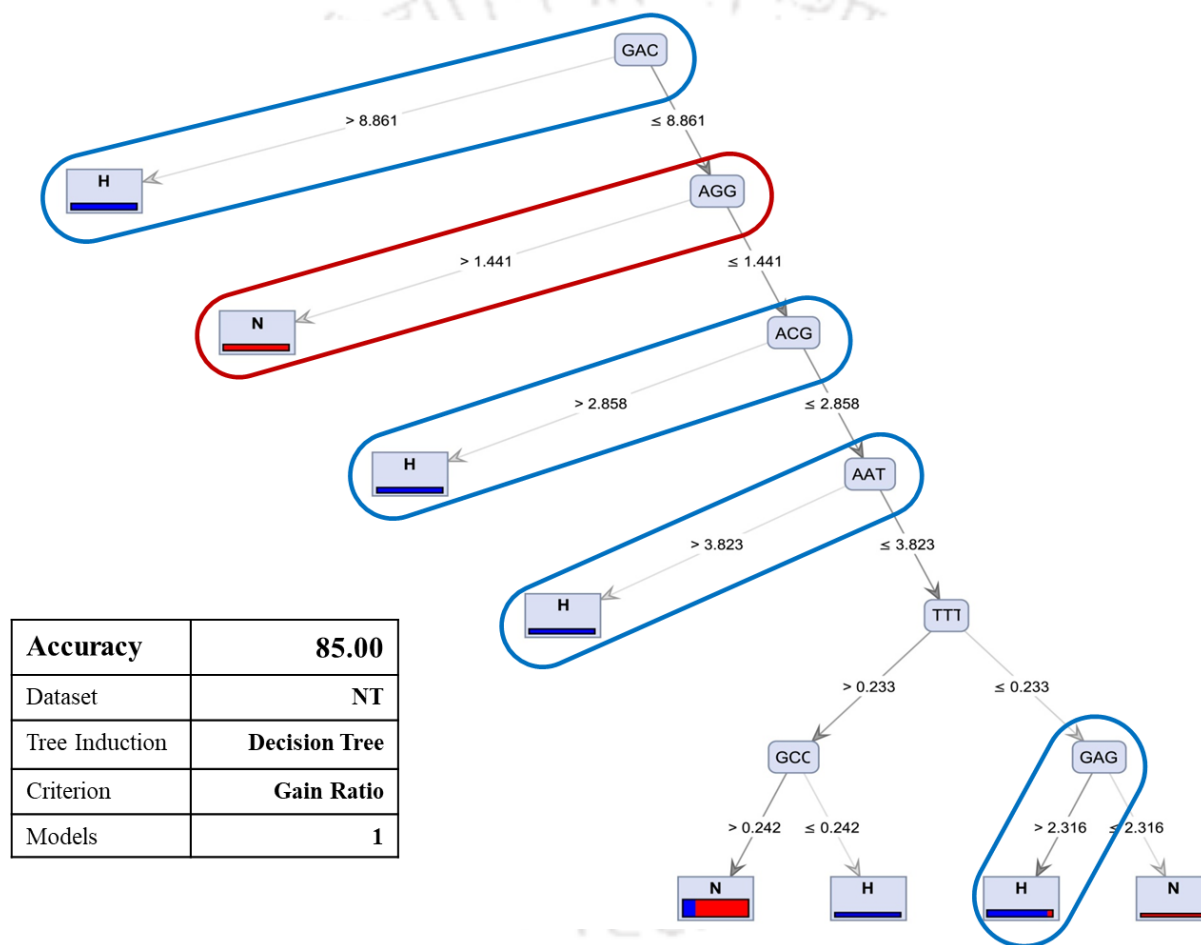


Figure A2.5: Decision tree prediction for codon in H-Nh dataset.

## BAROPHILE and NON-BAROPHILE CODON

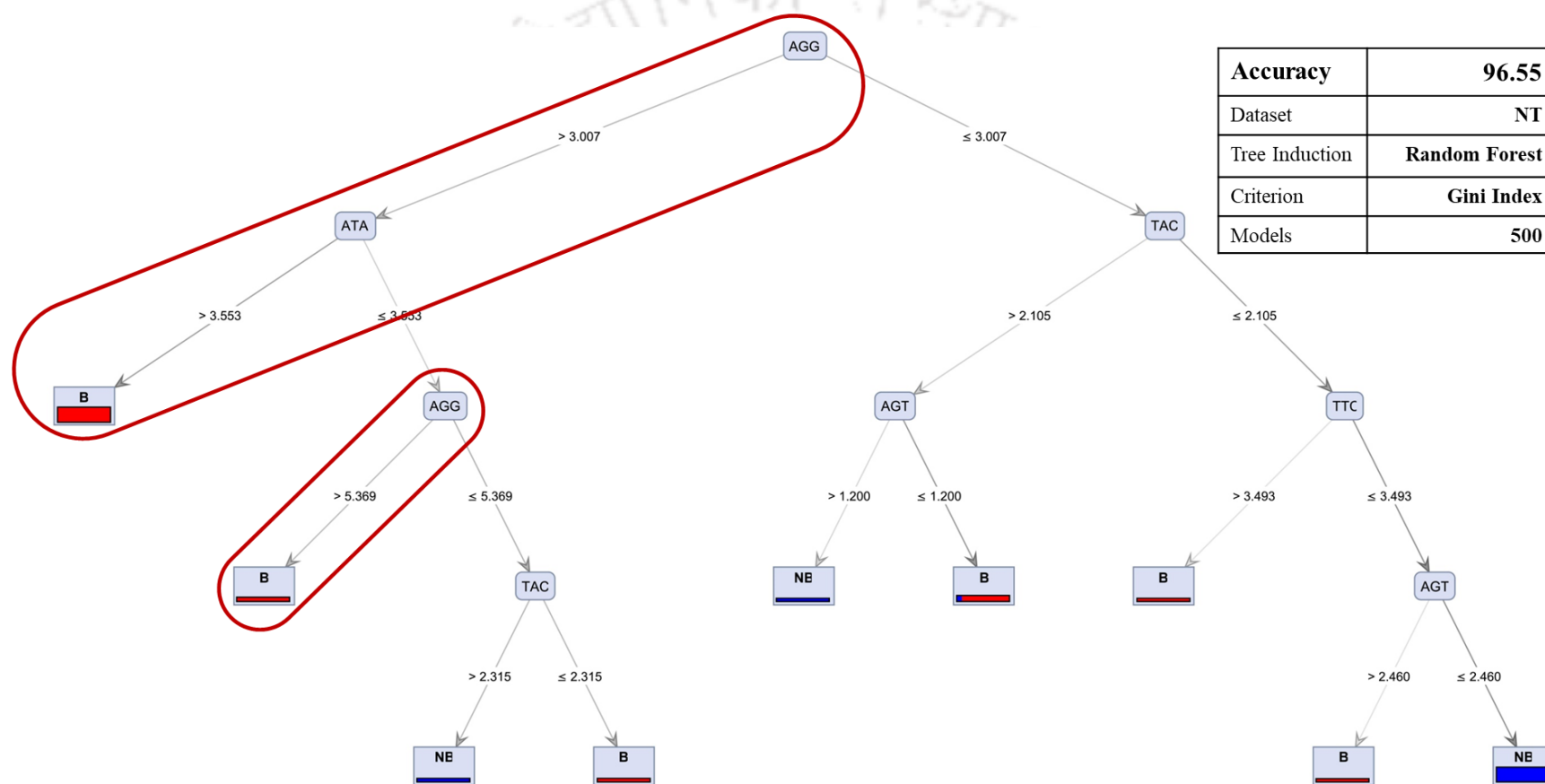
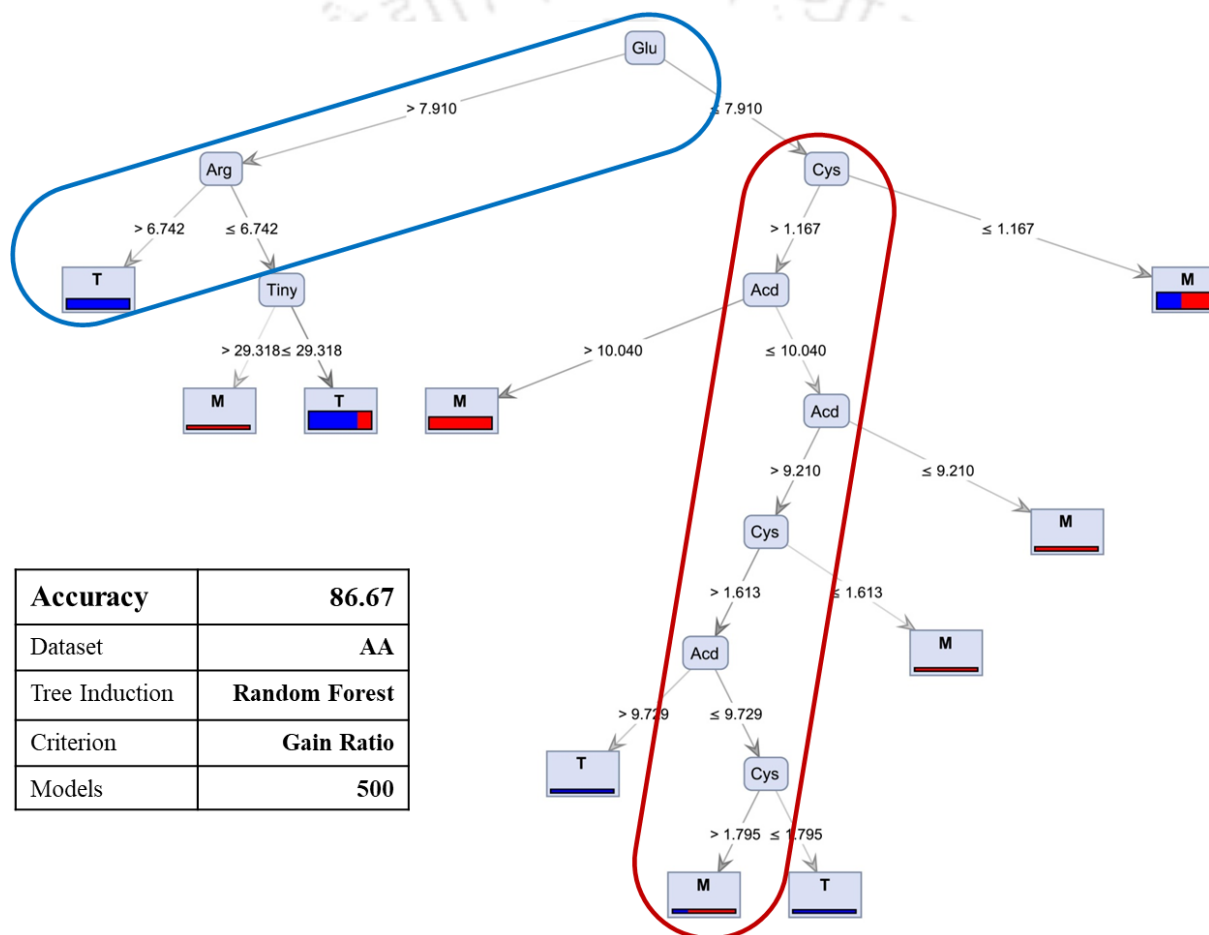


Figure A2.6: Decision tree prediction for codon in B-Nb dataset.



## THERMOPHILIC-MESOPHILIC PROTEIN



|                |               |
|----------------|---------------|
| Accuracy       | 86.67         |
| Dataset        | AA            |
| Tree Induction | Random Forest |
| Criterion      | Gain Ratio    |
| Models         | 500           |

Figure A2.7: Decision tree prediction for amino acid (AA) in T-M dataset.

## PSYCHROPHILIC-MESOPHILIC PROTEIN

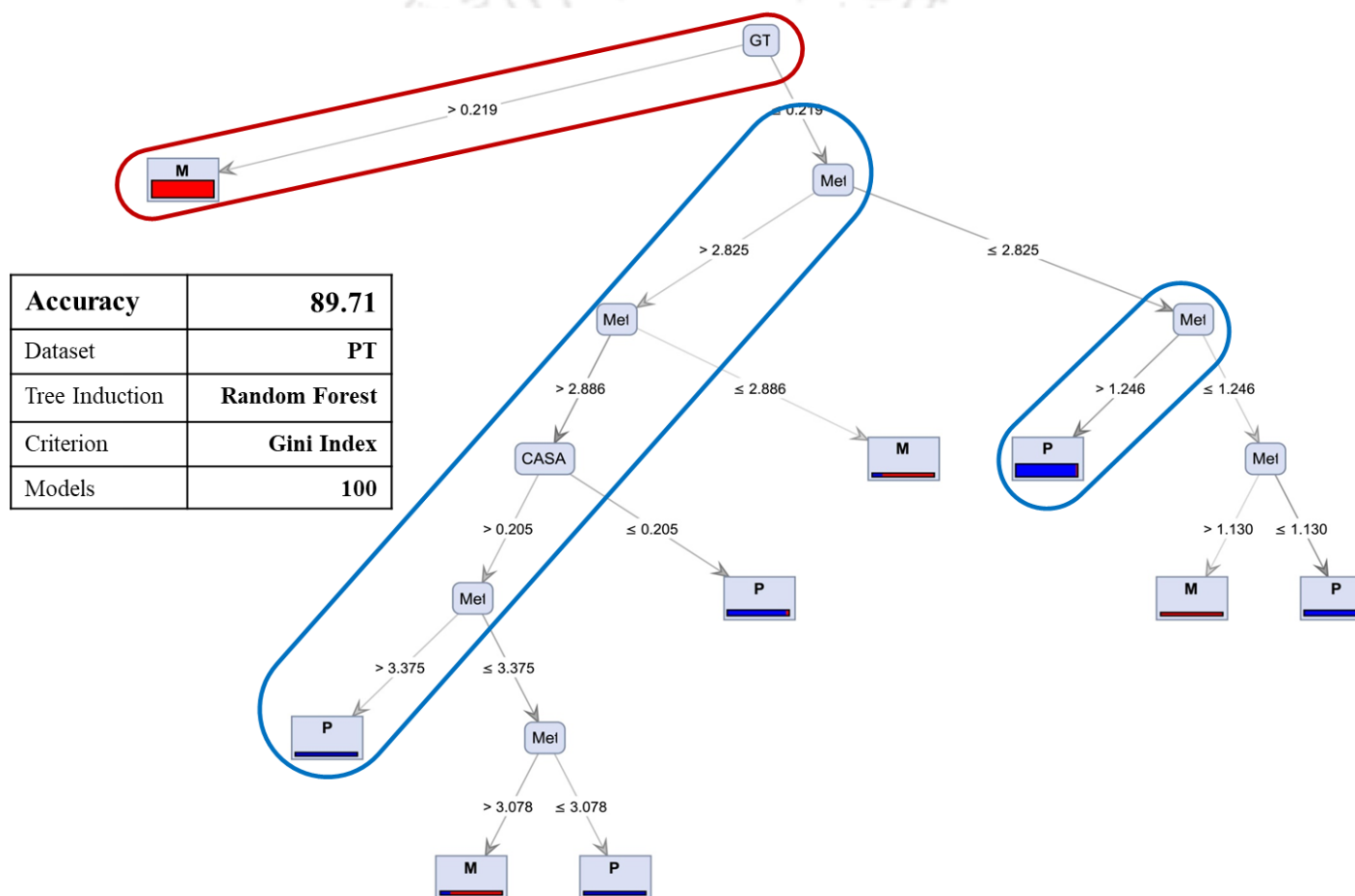
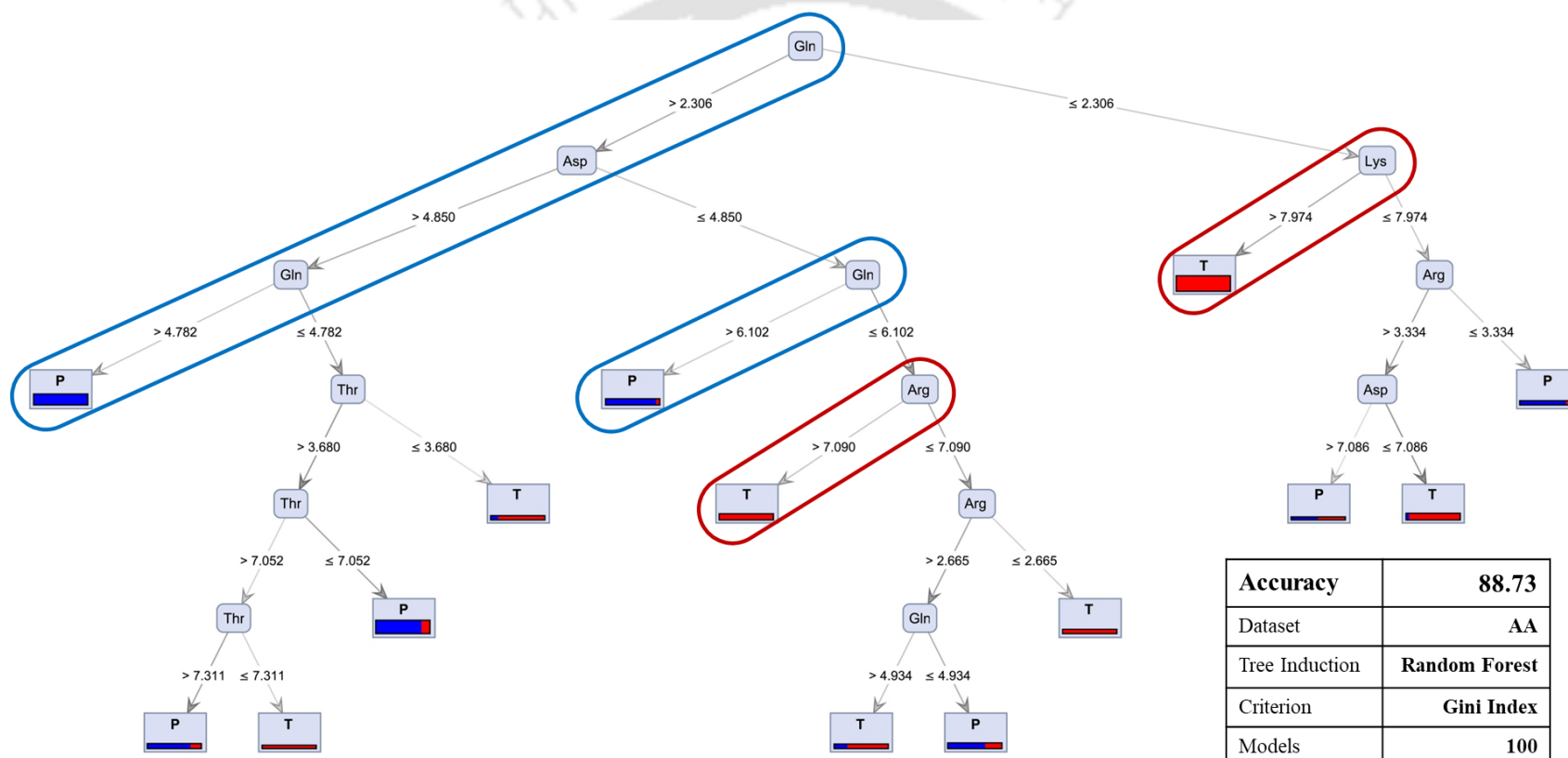


Figure A2.8: Decision tree prediction for amino acid and structural attribute (AA+ST) in P-M dataset.

## PSYCHROPHILIC-THERMOPHILIC PROTEIN



|                |               |
|----------------|---------------|
| Accuracy       | 88.73         |
| Dataset        | AA            |
| Tree Induction | Random Forest |
| Criterion      | Gini Index    |
| Models         | 100           |

Figure A2.9: Decision tree prediction for amino acid (AA) in T-P dataset.

## PSYCHROPHILIC-THERMOPHILIC PROTEIN

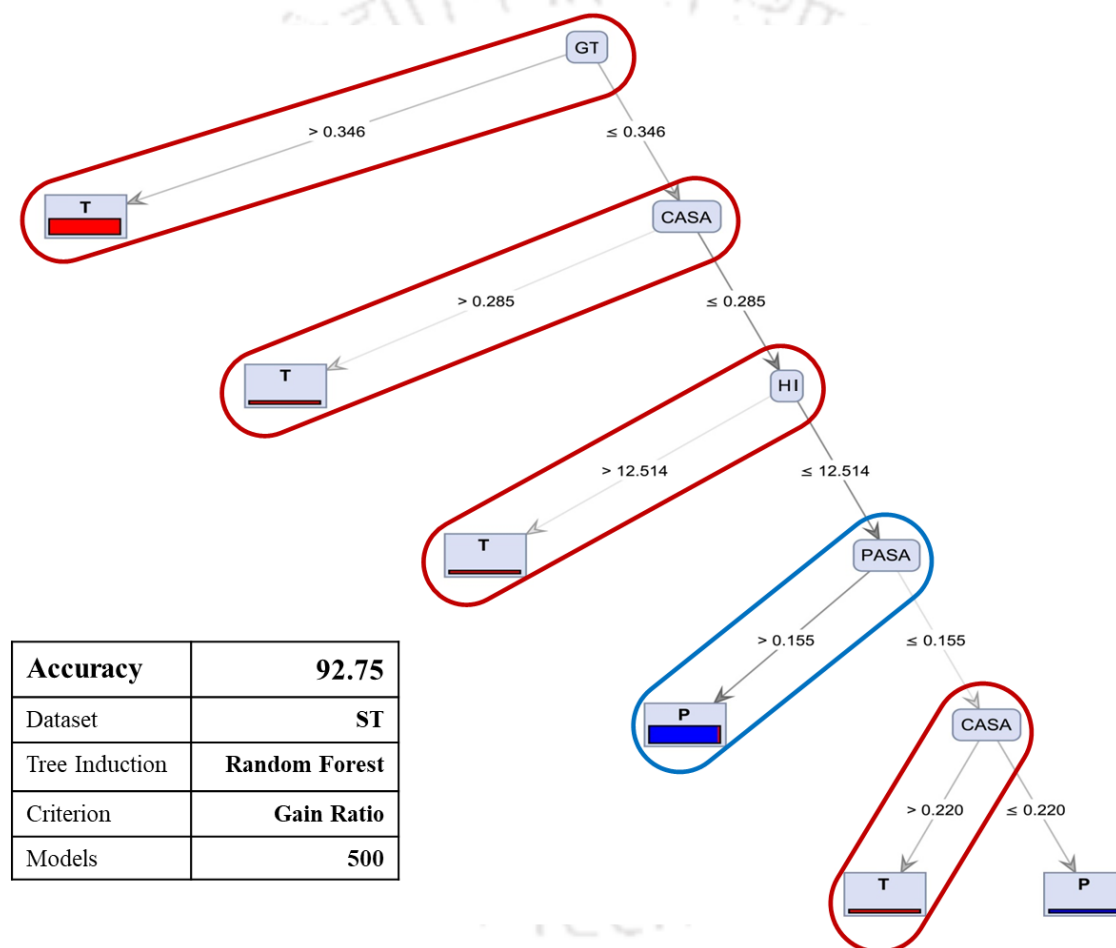


Figure A2.10: Decision tree prediction for structural attribute (ST) prediction in T-P dataset.

## ALKALIPHILIC-ACIDOPHILIC PROTEIN

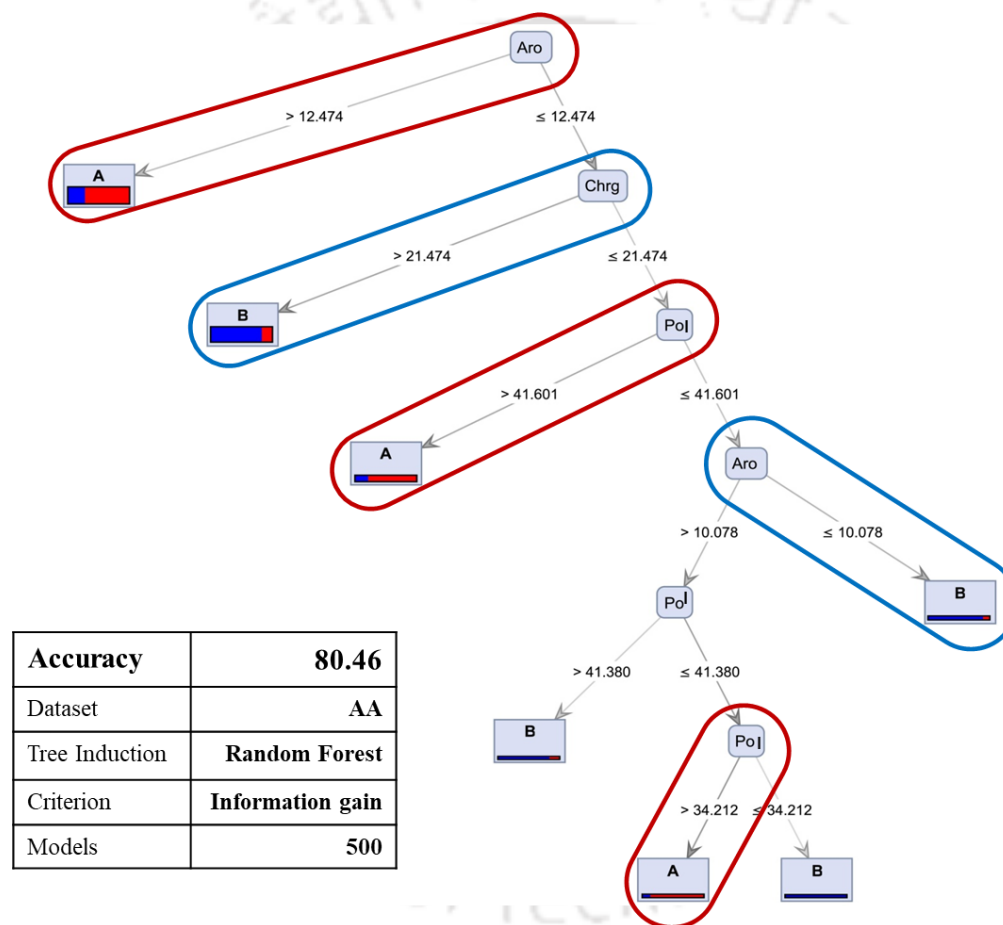


Figure A2.11: Decision tree prediction for amino acid (AA) in A-B dataset.

## ALKALIPHILIC-ACIDOPHILIC PROTEIN

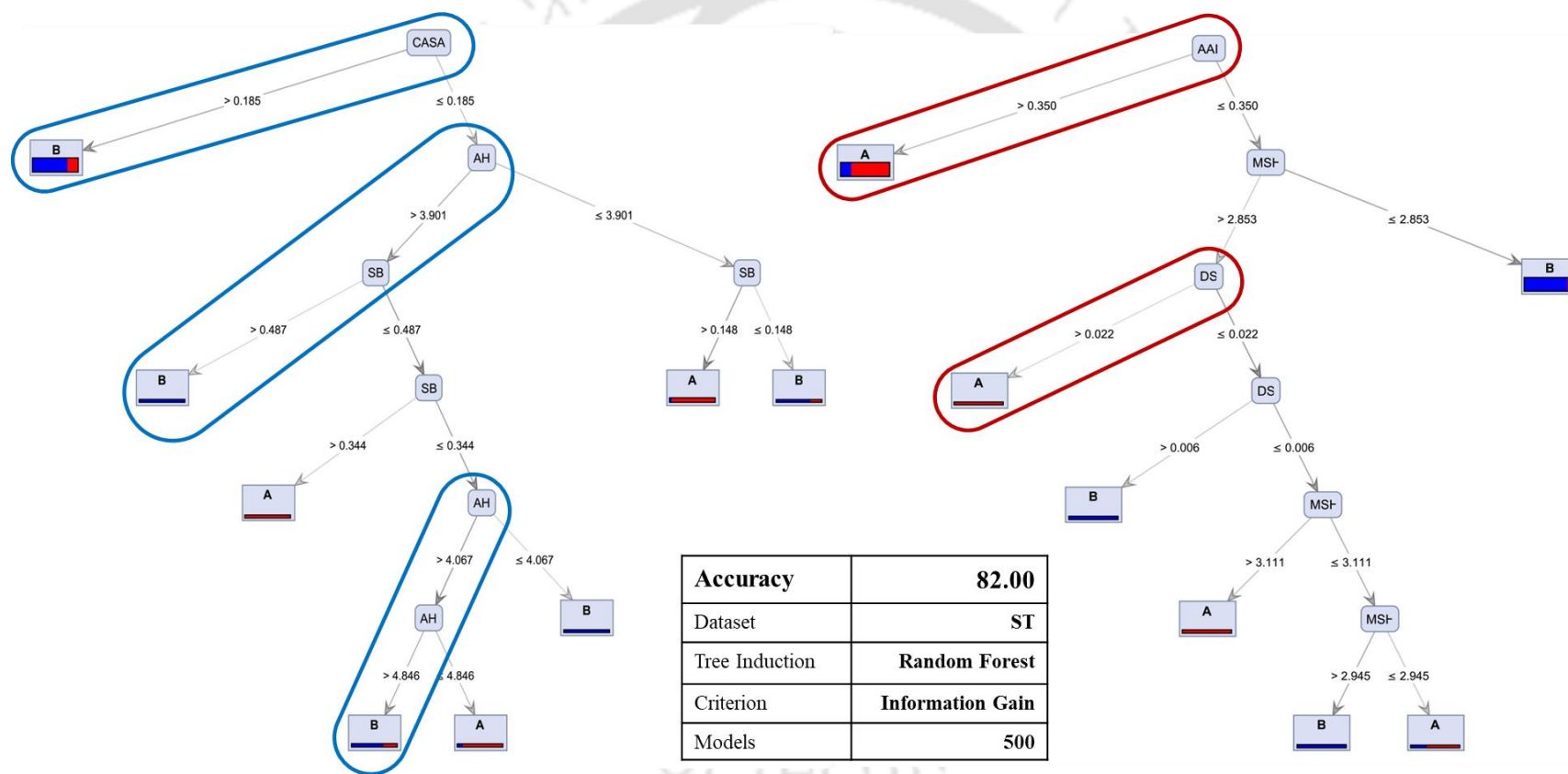


Figure A2.12: Decision tree prediction for amino acid (AA) in A-B dataset.



## HALOPHILIC-NON-HALOPHILIC PROTEIN

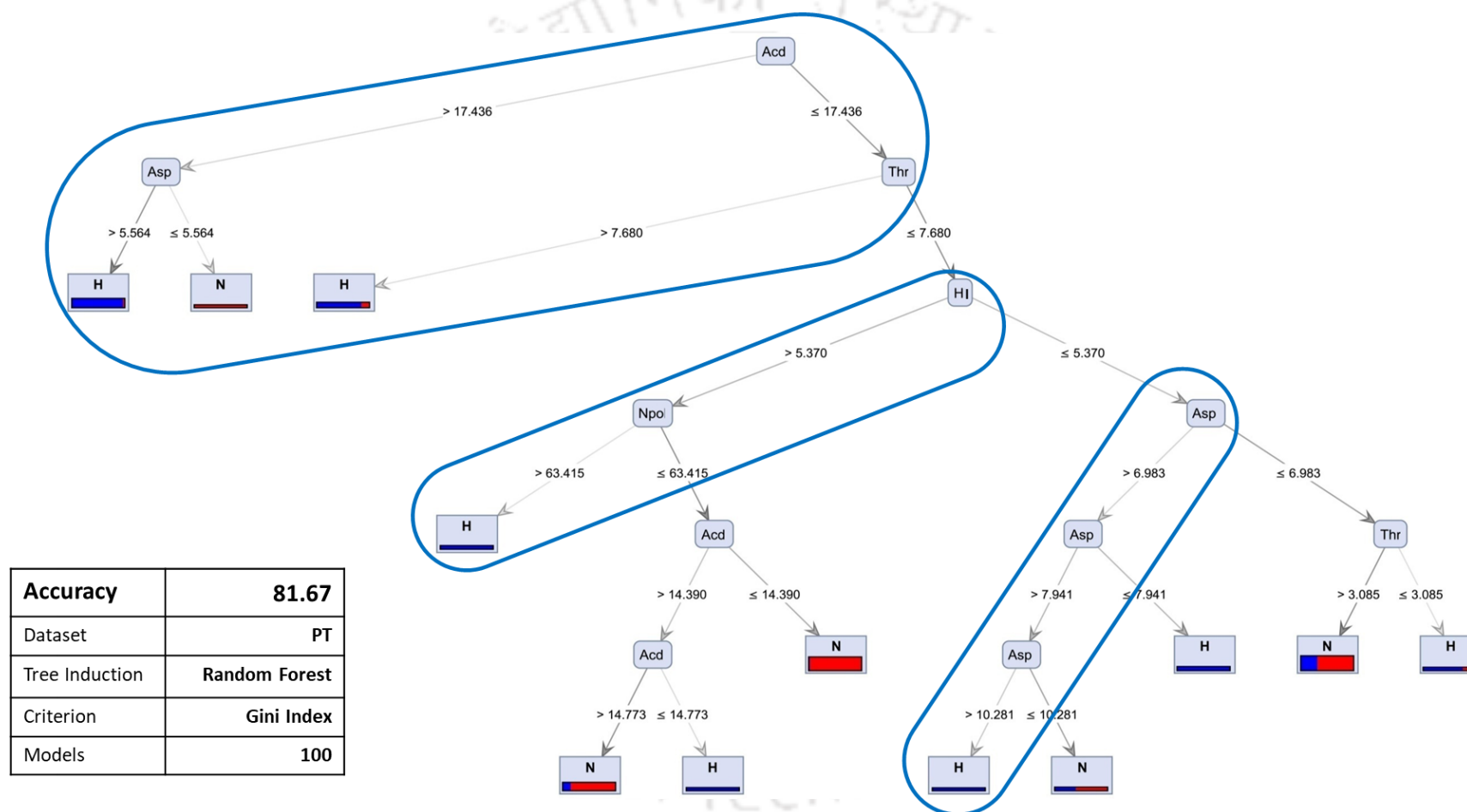


Figure A2.13: Decision tree prediction for amino acid and structural attribute (AA+ST) in H-Nh dataset.

## BAROPHILIC and NON-BAROPHILIC PROTEIN

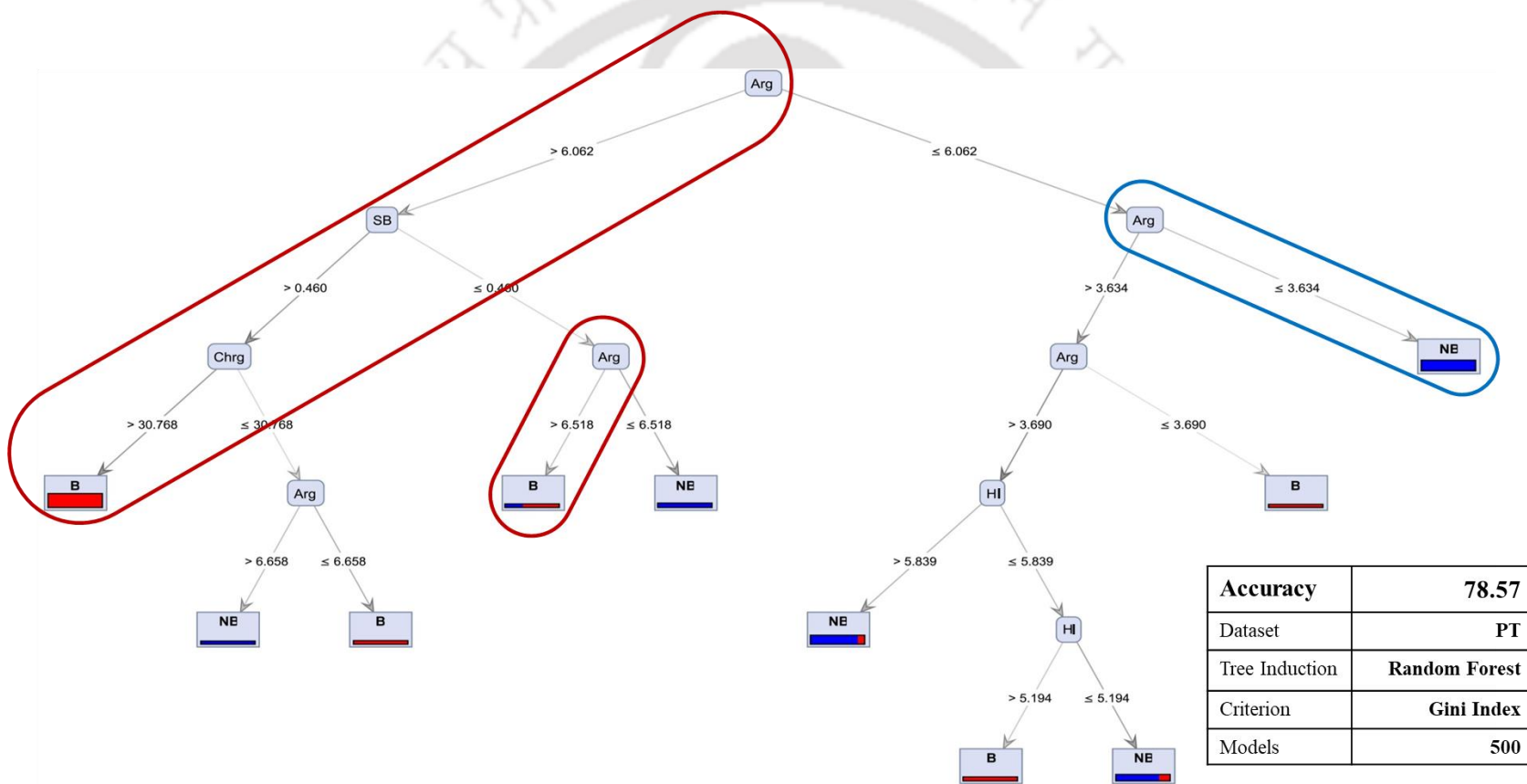


Figure A2.14: Decision tree prediction for amino acid and structural attribute (AA+ST) in B-Nb dataset.

## THERMOPHILIC-MESOPHILIC PROTEIN

Rank by 1-9 scale weight using a python script

| Attribute  | Ala | Cys | Asp | Glu | His | Ile | Lys | Asn | Gln | Arg | Ser | Thr | Val | Tiny | Sml | Pol | Chrg | Bsc | Acd |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|------|-----|-----|
| Rank (1-9) | 4   | 2   | 4   | 9   | 3   | 6   | 7   | 3   | 1   | 8   | 3   | 3   | 6   | 4    | 4   | 5   | 7    | 7   | 7   |

19x19 Pairwise Reciprocal Matrix

| Attribute | Ala  | Cys  | Asp  | Glu  | His  | Ile  | Lys  | Asn  | Gln  | Arg  | Ser  | Thr  | Val  | Tiny | Sml  | Pol  | Chrg | Bsc  | Acd  | $\lambda$ | Priority |
|-----------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-----------|----------|
| Ala       | 1.00 | 2.00 | 1.00 | 0.45 | 1.34 | 0.67 | 0.57 | 1.34 | 4.00 | 0.50 | 1.34 | 1.34 | 0.67 | 1.00 | 1.00 | 0.80 | 0.57 | 0.57 | 0.57 | 19.03     | 0.04     |
| Cys       | 0.50 | 1.00 | 0.50 | 0.23 | 0.67 | 0.34 | 0.29 | 0.67 | 2.00 | 0.25 | 0.67 | 0.67 | 0.34 | 0.50 | 0.50 | 0.40 | 0.29 | 0.29 | 0.29 | 19.16     | 0.02     |
| Asp       | 1.00 | 2.00 | 1.00 | 0.45 | 1.34 | 0.67 | 0.57 | 1.34 | 4.00 | 0.50 | 1.34 | 1.34 | 0.67 | 1.00 | 1.00 | 0.80 | 0.57 | 0.57 | 0.57 | 19.03     | 0.04     |
| Glu       | 2.25 | 4.50 | 2.25 | 1.00 | 3.00 | 1.50 | 1.29 | 3.00 | 9.00 | 1.13 | 3.00 | 3.00 | 1.50 | 2.25 | 2.25 | 1.80 | 1.29 | 1.29 | 1.29 | 19.01     | 0.10     |
| His       | 0.75 | 1.50 | 0.75 | 0.34 | 1.00 | 0.50 | 0.43 | 1.00 | 3.00 | 0.38 | 1.00 | 1.00 | 0.50 | 0.75 | 0.75 | 0.60 | 0.43 | 0.43 | 0.43 | 19.03     | 0.03     |
| Ile       | 1.50 | 3.00 | 1.50 | 0.67 | 2.00 | 1.00 | 0.86 | 2.00 | 6.00 | 0.75 | 2.00 | 2.00 | 1.00 | 1.50 | 1.50 | 1.20 | 0.86 | 0.86 | 0.86 | 19.02     | 0.06     |
| Lys       | 1.75 | 3.50 | 1.75 | 0.78 | 2.34 | 1.17 | 1.00 | 2.34 | 7.00 | 0.88 | 2.34 | 2.34 | 1.17 | 1.75 | 1.75 | 1.40 | 1.00 | 1.00 | 1.00 | 19.02     | 0.08     |
| Asn       | 0.75 | 1.50 | 0.75 | 0.34 | 1.00 | 0.50 | 0.43 | 1.00 | 3.00 | 0.38 | 1.00 | 1.00 | 0.50 | 0.75 | 0.75 | 0.60 | 0.43 | 0.43 | 0.43 | 19.03     | 0.03     |
| Gln       | 0.25 | 0.50 | 0.25 | 0.12 | 0.34 | 0.17 | 0.14 | 0.34 | 1.00 | 0.13 | 0.34 | 0.34 | 0.17 | 0.25 | 0.25 | 0.20 | 0.14 | 0.14 | 0.14 | 19.12     | 0.01     |
| Arg       | 2.00 | 4.00 | 2.00 | 0.89 | 2.67 | 1.34 | 1.14 | 2.67 | 8.00 | 1.00 | 2.67 | 2.67 | 1.34 | 2.00 | 2.00 | 1.60 | 1.14 | 1.14 | 1.14 | 19.01     | 0.09     |
| Ser       | 0.75 | 1.50 | 0.75 | 0.34 | 1.00 | 0.50 | 0.43 | 1.00 | 3.00 | 0.38 | 1.00 | 1.00 | 0.50 | 0.75 | 0.75 | 0.60 | 0.43 | 0.43 | 0.43 | 19.03     | 0.03     |
| Thr       | 0.75 | 1.50 | 0.75 | 0.34 | 1.00 | 0.50 | 0.43 | 1.00 | 3.00 | 0.38 | 1.00 | 1.00 | 0.50 | 0.75 | 0.75 | 0.60 | 0.43 | 0.43 | 0.43 | 19.03     | 0.03     |
| Val       | 1.50 | 3.00 | 1.50 | 0.67 | 2.00 | 1.00 | 0.86 | 2.00 | 6.00 | 0.75 | 2.00 | 2.00 | 1.00 | 1.50 | 1.50 | 1.20 | 0.86 | 0.86 | 0.86 | 19.02     | 0.06     |
| Tiny      | 1.00 | 2.00 | 1.00 | 0.45 | 1.34 | 0.67 | 0.57 | 1.34 | 4.00 | 0.50 | 1.34 | 1.34 | 0.67 | 1.00 | 1.00 | 0.80 | 0.57 | 0.57 | 0.57 | 19.03     | 0.04     |
| Sml       | 1.00 | 2.00 | 1.00 | 0.45 | 1.34 | 0.67 | 0.57 | 1.34 | 4.00 | 0.50 | 1.34 | 1.34 | 0.67 | 1.00 | 1.00 | 0.80 | 0.57 | 0.57 | 0.57 | 19.03     | 0.04     |
| Pol       | 1.25 | 2.50 | 1.25 | 0.56 | 1.67 | 0.84 | 0.72 | 1.67 | 5.00 | 0.63 | 1.67 | 1.67 | 0.84 | 1.25 | 1.25 | 1.00 | 0.72 | 0.72 | 0.72 | 19.06     | 0.05     |
| Chrg      | 1.75 | 3.50 | 1.75 | 0.78 | 2.34 | 1.17 | 1.00 | 2.34 | 7.00 | 0.88 | 2.34 | 2.34 | 1.17 | 1.75 | 1.75 | 1.40 | 1.00 | 1.00 | 1.00 | 19.02     | 0.08     |
| Bsc       | 1.75 | 3.50 | 1.75 | 0.78 | 2.34 | 1.17 | 1.00 | 2.34 | 7.00 | 0.88 | 2.34 | 2.34 | 1.17 | 1.75 | 1.75 | 1.40 | 1.00 | 1.00 | 1.00 | 19.02     | 0.08     |
| Acd       | 1.75 | 3.50 | 1.75 | 0.78 | 2.34 | 1.17 | 1.00 | 2.34 | 7.00 | 0.88 | 2.34 | 2.34 | 1.17 | 1.75 | 1.75 | 1.40 | 1.00 | 1.00 | 1.00 | 19.02     | 0.08     |

Relative Priority Vectors

$$\lambda_{\max} = 19.04$$

$$CI = \frac{\lambda_{\max} - n}{n - 1} ; n = 19$$

$$\text{Consistency Index (CI)} = 0.0021$$

$$\text{Random Index (RI) for 19} = 1.63$$

$$CR = \frac{CI}{RI}$$

$$\text{Consistency Ratio (CR)} = 0.0013$$

RESULT: Matrix is consistent (CR < 0.1)

Figure A2.15: Reciprocal pairwise comparison matrix and consistency calculation for T-M dataset of protein.

## PSYCHROPHILIC-MESOPHILIC PROTEIN

Rank by 1-9 scale weight using a python script

| Attribute | Ala | His | Met | Asn | Ser | Thr | Trp | Tiny | Ali | Aro | Bsc | Acid | NPASA | CASA | GT |
|-----------|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|-----|------|-------|------|----|
| Rank 1-9  | 8   | 9   | 9   | 8   | 8   | 8   | 8   | 8    | 8   | 8   | 7   | 7    | 8     | 7    | 1  |

15x15 Pairwise Reciprocal Matrix

| Attribute | Ala  | His  | Met  | Asn  | Ser  | Thr  | Trp  | Tiny | Ali  | Aro  | Bsc  | Acid | NPASA | CASA | GT   | $\lambda$ | Priority |
|-----------|------|------|------|------|------|------|------|------|------|------|------|------|-------|------|------|-----------|----------|
| Ala       | 1.00 | 0.89 | 0.89 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.14 | 1.14 | 1.00  | 1.14 | 8.00 | 15.003    | 0.07     |
| His       | 1.13 | 1.00 | 1.00 | 1.13 | 1.13 | 1.13 | 1.13 | 1.13 | 1.13 | 1.13 | 1.29 | 1.29 | 1.13  | 1.29 | 9.00 | 15.010    | 0.08     |
| Met       | 1.13 | 1.00 | 1.00 | 1.13 | 1.13 | 1.13 | 1.13 | 1.13 | 1.13 | 1.13 | 1.29 | 1.29 | 1.13  | 1.29 | 9.00 | 15.010    | 0.08     |
| Asn       | 1.00 | 0.89 | 0.89 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.14 | 1.14 | 1.00  | 1.14 | 8.00 | 15.003    | 0.07     |
| Ser       | 1.00 | 0.89 | 0.89 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.14 | 1.14 | 1.00  | 1.14 | 8.00 | 15.003    | 0.07     |
| Thr       | 1.00 | 0.89 | 0.89 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.14 | 1.14 | 1.00  | 1.14 | 8.00 | 15.003    | 0.07     |
| Trp       | 1.00 | 0.89 | 0.89 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.14 | 1.14 | 1.00  | 1.14 | 8.00 | 15.003    | 0.07     |
| Tiny      | 1.00 | 0.89 | 0.89 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.14 | 1.14 | 1.00  | 1.14 | 8.00 | 15.003    | 0.07     |
| Ali       | 1.00 | 0.89 | 0.89 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.14 | 1.14 | 1.00  | 1.14 | 8.00 | 15.003    | 0.07     |
| Aro       | 1.00 | 0.89 | 0.89 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.14 | 1.14 | 1.00  | 1.14 | 8.00 | 15.003    | 0.07     |
| Bsc       | 0.88 | 0.78 | 0.78 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 1.00 | 1.00 | 0.88  | 1.00 | 7.00 | 15.006    | 0.06     |
| Acid      | 0.88 | 0.78 | 0.78 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 1.00 | 1.00 | 0.88  | 1.00 | 7.00 | 15.006    | 0.06     |
| NPASA     | 1.00 | 0.89 | 0.89 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.14 | 1.14 | 1.00  | 1.14 | 8.00 | 15.003    | 0.07     |
| CASA      | 0.88 | 0.78 | 0.78 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 1.00 | 1.00 | 0.88  | 1.00 | 7.00 | 15.006    | 0.06     |
| GT        | 0.13 | 0.11 | 0.11 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.14 | 0.14 | 0.13  | 0.14 | 1.00 | 15.003    | 0.01     |

Relative Priority Vectors

$$\lambda_{\max} = 15.005$$

$$CR = \frac{CI}{RI}$$

$$CI = \frac{\lambda_{\max} - n}{n - 1} ; n = 15$$

$$\text{Consistency Ratio (CR)} = 0.0002$$

$$\begin{aligned} \text{Consistency Index (CI)} &= 0.0003 \\ \text{Random Index (RI) for 15} &= 1.5838 \end{aligned}$$

**RESULT: Matrix is consistent (CR < 0.1)**

Figure A2.16: Reciprocal pairwise comparison matrix and consistency calculation for P-M dataset of protein.

## PSYCHROPHILIC-THERMOPHILIC PROTEIN

Rank by 1-9 scale weight using a python script

| Attributes | NPASA | PASA | CASA | PV | GT | HI | MMH | MSH | II | CPI | SB |
|------------|-------|------|------|----|----|----|-----|-----|----|-----|----|
| Rank (1-9) | 8     | 9    | 6    | 8  | 1  | 7  | 8   | 9   | 6  | 6   | 6  |

11x11 Pairwise Reciprocal Matrix

| Attributes | NPASA | PASA | CASA | PV   | GT   | HI   | MMH  | MSH  | II   | CPI  | SB   | $\lambda$ | Priority |
|------------|-------|------|------|------|------|------|------|------|------|------|------|-----------|----------|
| NPASA      | 1.00  | 0.89 | 1.34 | 1.00 | 8.00 | 1.14 | 1.00 | 0.89 | 1.34 | 1.34 | 1.34 | 11.03     | 0.11     |
| PASA       | 1.13  | 1.00 | 1.50 | 1.13 | 9.00 | 1.29 | 1.13 | 1.00 | 1.50 | 1.50 | 1.50 | 11.03     | 0.12     |
| CASA       | 0.75  | 0.67 | 1.00 | 0.75 | 6.00 | 0.86 | 0.75 | 0.67 | 1.00 | 1.00 | 1.00 | 11.03     | 0.08     |
| PV         | 1.00  | 0.89 | 1.34 | 1.00 | 8.00 | 1.14 | 1.00 | 0.89 | 1.34 | 1.34 | 1.34 | 11.03     | 0.11     |
| GT         | 0.13  | 0.12 | 0.17 | 0.13 | 1.00 | 0.14 | 0.13 | 0.12 | 0.17 | 0.17 | 0.17 | 11.03     | 0.01     |
| HI         | 0.88  | 0.78 | 1.17 | 0.88 | 7.00 | 1.00 | 0.88 | 0.78 | 1.17 | 1.17 | 1.17 | 11.03     | 0.09     |
| MMH        | 1.00  | 0.89 | 1.34 | 1.00 | 8.00 | 1.14 | 1.00 | 0.89 | 1.34 | 1.34 | 1.34 | 11.03     | 0.11     |
| MSH        | 1.13  | 1.00 | 1.50 | 1.13 | 9.00 | 1.29 | 1.13 | 1.00 | 1.50 | 1.50 | 1.50 | 11.03     | 0.12     |
| II         | 0.75  | 0.67 | 1.00 | 0.75 | 6.00 | 0.86 | 0.75 | 0.67 | 1.00 | 1.00 | 1.00 | 11.03     | 0.08     |
| CPI        | 0.75  | 0.67 | 1.00 | 0.75 | 6.00 | 0.86 | 0.75 | 0.67 | 1.00 | 1.00 | 1.00 | 11.03     | 0.08     |
| SB         | 0.75  | 0.67 | 1.00 | 0.75 | 6.00 | 0.86 | 0.75 | 0.67 | 1.00 | 1.00 | 1.00 | 11.03     | 0.08     |

Relative Priority Vectors

$$\lambda_{\max} = 11.033$$

$$CI = \frac{\lambda_{\max} - n}{n - 1} ; n = 11$$

$$\text{Consistency Index}(CI) = 0.0033$$

$$\text{Random Index (RI) for 11} = 1.5141$$

$$CR = \frac{CI}{RI}$$

$$\text{Consistency Ratio}(CR) = 0.0022$$

**RESULT: Matrix is consistent (CR < 0.1)**

Figure A2.17: Reciprocal pairwise comparison matrix and consistency calculation for T-P dataset of protein.

## ALKALIPHILIC-ACIDOPHILIC PRTOEIN

Rank by 1-9 scale weight using a python script

| Attributes | HI | DS | MMH | MSH | SSH | II | AAI | ASI | CPI | SB | HB | PASA | CASA | PV | AH | BS |
|------------|----|----|-----|-----|-----|----|-----|-----|-----|----|----|------|------|----|----|----|
| Rank (1-9) | 6  | 1  | 7   | 5   | 5   | 7  | 2   | 5   | 5   | 6  | 7  | 5    | 8    | 5  | 9  | 5  |

| 16x16 Pairwise Reciprocal Matrix | Attributes | HI   | DS   | MMH  | MSH  | SSH  | II   | AAI  | ASI  | CPI  | SB   | HB   | PASA | CASA | PV   | AH   | BS   | $\lambda$ | Priority | Relative Priority Vectors |
|----------------------------------|------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-----------|----------|---------------------------|
|                                  | HI         | 1.00 | 6.00 | 0.86 | 1.20 | 1.20 | 0.86 | 3.00 | 1.20 | 1.20 | 1.00 | 0.90 | 1.20 | 0.75 | 1.20 | 0.67 | 1.20 | 16.05     | 0.07     |                           |
|                                  | DS         | 0.17 | 1.00 | 0.15 | 0.20 | 0.20 | 0.14 | 0.50 | 0.20 | 0.20 | 0.17 | 0.14 | 0.20 | 0.13 | 0.20 | 0.12 | 0.20 | 15.86     | 0.01     |                           |
|                                  | MMH        | 1.17 | 7.00 | 1.00 | 1.40 | 1.40 | 1.00 | 3.50 | 1.40 | 1.40 | 1.17 | 1.00 | 1.40 | 0.88 | 1.40 | 0.78 | 1.40 | 16.01     | 0.08     |                           |
|                                  | MSH        | 0.83 | 5.00 | 0.71 | 1.00 | 1.00 | 0.71 | 2.50 | 1.00 | 1.00 | 0.83 | 0.71 | 1.00 | 0.63 | 1.00 | 0.56 | 1.00 | 16.03     | 0.06     |                           |
|                                  | SSH        | 0.83 | 5.00 | 0.71 | 1.00 | 1.00 | 0.71 | 2.50 | 1.00 | 1.00 | 0.83 | 0.71 | 1.00 | 0.63 | 1.00 | 0.56 | 1.00 | 16.02     | 0.06     |                           |
|                                  | II         | 1.17 | 7.00 | 1.00 | 1.40 | 1.40 | 1.00 | 3.50 | 1.40 | 1.40 | 1.17 | 1.00 | 1.40 | 0.88 | 1.40 | 0.78 | 1.40 | 16.02     | 0.08     |                           |
|                                  | AAI        | 0.33 | 2.00 | 0.29 | 0.40 | 0.40 | 0.29 | 1.00 | 0.40 | 0.40 | 0.33 | 0.30 | 0.40 | 0.25 | 0.40 | 0.22 | 0.40 | 16.07     | 0.02     |                           |
|                                  | ASI        | 0.83 | 5.00 | 0.71 | 1.00 | 1.00 | 0.71 | 2.50 | 1.00 | 1.00 | 0.83 | 0.71 | 1.00 | 0.63 | 1.00 | 0.56 | 1.00 | 16.02     | 0.06     |                           |
|                                  | CPI        | 0.83 | 5.00 | 0.71 | 1.00 | 1.00 | 0.71 | 2.50 | 1.00 | 1.00 | 0.83 | 0.71 | 1.00 | 0.63 | 1.00 | 0.56 | 1.00 | 16.02     | 0.06     |                           |
|                                  | SB         | 1.00 | 6.00 | 0.86 | 1.20 | 1.20 | 0.86 | 3.00 | 1.20 | 1.20 | 1.00 | 0.86 | 1.20 | 0.75 | 1.20 | 0.67 | 1.20 | 16.02     | 0.07     |                           |
|                                  | HB         | 1.17 | 7.00 | 1.00 | 1.40 | 1.40 | 1.00 | 3.50 | 1.40 | 1.40 | 1.17 | 1.00 | 1.40 | 0.88 | 1.40 | 0.78 | 1.40 | 16.02     | 0.08     |                           |
|                                  | PASA       | 0.83 | 5.00 | 0.71 | 1.00 | 1.00 | 0.71 | 2.50 | 1.00 | 1.00 | 0.83 | 0.71 | 1.00 | 0.63 | 1.00 | 0.56 | 1.00 | 16.03     | 0.06     |                           |
|                                  | CASA       | 1.33 | 8.00 | 1.14 | 1.60 | 1.60 | 1.14 | 4.00 | 1.60 | 1.60 | 1.33 | 1.14 | 1.60 | 1.00 | 1.60 | 0.89 | 1.60 | 16.02     | 0.09     |                           |
|                                  | PV         | 0.83 | 5.00 | 0.71 | 1.00 | 1.00 | 0.71 | 2.50 | 1.00 | 1.00 | 0.83 | 0.71 | 1.00 | 0.63 | 1.00 | 0.56 | 1.00 | 16.03     | 0.06     |                           |
|                                  | AH         | 1.50 | 9.00 | 1.29 | 1.80 | 1.80 | 1.29 | 4.50 | 1.80 | 1.80 | 1.50 | 1.29 | 1.80 | 1.13 | 1.80 | 1.00 | 1.80 | 16.02     | 0.10     |                           |
|                                  | BS         | 0.83 | 5.00 | 0.71 | 1.00 | 1.00 | 0.71 | 2.50 | 1.00 | 1.00 | 0.83 | 0.71 | 1.00 | 0.63 | 1.00 | 0.56 | 1.00 | 16.03     | 0.06     |                           |

$$\lambda_{\max} = 16.018$$

$$CI = \frac{\lambda_{\max} - n}{n - 1} ; n = 16$$

$$\text{Consistency Index (CI)} = 0.0012$$

$$\text{Random Index (RI) for 16} = 1.5978$$

$$CR = \frac{CI}{RI}$$

$$\text{Consistency Ratio (CR)} = 0.0008$$

**RESULT: Matrix is consistent (CR < 0.1)**

Figure A2.18: Reciprocal pairwise comparison matrix and consistency calculation for A-B dataset of protein.



## HALOPHILIC-NON-HALOPHILIC PROTEIN

Rank by 1-9 scale weight using a python script

| Attributes | Ala | Asp | Ile | Asn | Arg | Thr | Tiny | Sml | Ali | Npol | Pol | Bsc | Acd | HI | CPI | SB | NPASA |
|------------|-----|-----|-----|-----|-----|-----|------|-----|-----|------|-----|-----|-----|----|-----|----|-------|
| Rank (1-9) | 7   | 9   | 3   | 5   | 2   | 9   | 7    | 6   | 5   | 5    | 6   | 3   | 8   | 4  | 3   | 1  | 3     |

17x17 Pairwise Reciprocal Matrix

| Attributes | Ala  | Asp  | Ile  | Asn  | Arg  | Thr  | Tiny | Sml  | Ali  | Npol | Pol  | Bsc  | Acd  | HI   | CPI  | SB   | NPASA | λ     | Priority |
|------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|-------|-------|----------|
| Ala        | 1.00 | 0.78 | 2.33 | 1.40 | 3.50 | 0.78 | 1.00 | 1.17 | 1.40 | 1.40 | 1.17 | 2.33 | 0.88 | 1.75 | 2.33 | 7.00 | 2.33  | 17.00 | 0.08     |
| Asp        | 1.29 | 1.00 | 3.00 | 1.80 | 4.50 | 1.00 | 1.29 | 1.50 | 1.80 | 1.80 | 1.50 | 3.00 | 1.13 | 2.25 | 3.00 | 9.00 | 3.00  | 17.01 | 0.10     |
| Ile        | 0.43 | 0.33 | 1.00 | 0.60 | 1.50 | 0.33 | 0.43 | 0.50 | 0.60 | 0.60 | 0.50 | 1.00 | 0.38 | 0.75 | 1.00 | 3.00 | 1.00  | 17.01 | 0.03     |
| Asn        | 0.71 | 0.56 | 1.67 | 1.00 | 2.50 | 0.56 | 0.71 | 0.83 | 1.00 | 1.00 | 0.83 | 1.67 | 0.63 | 1.25 | 1.67 | 5.00 | 1.67  | 17.00 | 0.06     |
| Arg        | 0.29 | 0.22 | 0.67 | 0.40 | 1.00 | 0.22 | 0.29 | 0.33 | 0.40 | 0.40 | 0.33 | 0.67 | 0.25 | 0.50 | 0.67 | 2.00 | 0.67  | 17.04 | 0.02     |
| Thr        | 1.29 | 1.00 | 3.00 | 1.80 | 4.50 | 1.00 | 1.29 | 1.50 | 1.80 | 1.80 | 1.50 | 3.00 | 1.13 | 2.25 | 3.00 | 9.00 | 3.00  | 17.01 | 0.10     |
| Tiny       | 1.00 | 0.78 | 2.33 | 1.40 | 3.50 | 0.78 | 1.00 | 1.17 | 1.40 | 1.40 | 1.17 | 2.33 | 0.88 | 1.75 | 2.33 | 7.00 | 2.33  | 17.00 | 0.08     |
| Sml        | 0.86 | 0.67 | 2.00 | 1.20 | 3.00 | 0.67 | 0.86 | 1.00 | 1.20 | 1.20 | 1.00 | 2.00 | 0.75 | 1.50 | 2.00 | 6.00 | 2.00  | 17.01 | 0.07     |
| Ali        | 0.71 | 0.56 | 1.67 | 1.00 | 2.50 | 0.56 | 0.71 | 0.83 | 1.00 | 1.00 | 0.83 | 1.67 | 0.63 | 1.25 | 1.67 | 5.00 | 1.67  | 17.00 | 0.06     |
| Npol       | 0.71 | 0.56 | 1.67 | 1.00 | 2.50 | 0.56 | 0.71 | 0.84 | 1.00 | 1.00 | 0.83 | 1.67 | 0.63 | 1.25 | 1.67 | 5.00 | 1.67  | 17.01 | 0.06     |
| Pol        | 0.87 | 0.67 | 2.00 | 1.20 | 3.00 | 0.67 | 0.87 | 1.00 | 1.20 | 1.20 | 1.00 | 2.00 | 0.75 | 1.50 | 2.00 | 6.00 | 2.00  | 17.03 | 0.07     |
| Bsc        | 0.43 | 0.33 | 1.00 | 0.60 | 1.50 | 0.33 | 0.43 | 0.50 | 0.60 | 0.60 | 0.50 | 1.00 | 0.38 | 0.75 | 1.00 | 3.00 | 1.00  | 17.01 | 0.03     |
| Acd        | 1.14 | 0.89 | 2.67 | 1.60 | 4.00 | 0.89 | 1.14 | 1.33 | 1.60 | 1.60 | 1.33 | 2.67 | 1.00 | 2.00 | 2.67 | 8.00 | 2.67  | 17.00 | 0.09     |
| HI         | 0.57 | 0.45 | 1.33 | 0.80 | 2.00 | 0.45 | 0.57 | 0.67 | 0.80 | 0.80 | 0.67 | 1.33 | 0.50 | 1.00 | 1.33 | 4.00 | 1.33  | 17.00 | 0.05     |
| CPI        | 0.43 | 0.33 | 1.00 | 0.60 | 1.50 | 0.33 | 0.43 | 0.50 | 0.60 | 0.60 | 0.50 | 1.00 | 0.38 | 0.75 | 1.00 | 3.00 | 1.00  | 17.01 | 0.03     |
| SB         | 0.15 | 0.11 | 0.33 | 0.20 | 0.50 | 0.11 | 0.15 | 0.17 | 0.20 | 0.20 | 0.17 | 0.33 | 0.13 | 0.25 | 0.33 | 1.00 | 0.33  | 17.13 | 0.01     |
| NPASA      | 0.43 | 0.33 | 1.00 | 0.60 | 1.50 | 0.33 | 0.43 | 0.50 | 0.60 | 0.60 | 0.50 | 1.00 | 0.38 | 0.75 | 1.00 | 3.00 | 1.00  | 17.01 | 0.03     |

Relative Priority Vectors

$$\lambda_{\max} = 17.017$$

$$CR = \frac{CI}{RI}$$

$$CI = \frac{\lambda_{\max} - n}{n - 1}; n = 17$$

$$\text{Consistency Ratio (CR)} = 0.0007$$

$$\text{Consistency Index (CI)} = 0.0011$$

$$\text{Random Index (RI) for 17} = 1.6086$$

RESULT: Matrix is consistent (CR < 0.1)

Figure A2.19: Reciprocal pairwise comparison matrix and consistency calculation for H-Nh dataset of protein.

## BAROPHILIC and NON-BAROPHILIC PROTEIN

Rank by 1-9 scale weight using a python script

| Attributes | Gln | Arg | Ser | Tiny | Chrg | Bsc | Acd | HI | II | CPI | SB | BSH | BST |
|------------|-----|-----|-----|------|------|-----|-----|----|----|-----|----|-----|-----|
| Rank (1-9) | 1   | 9   | 1   | 3    | 6    | 6   | 5   | 4  | 6  | 5   | 6  | 5   | 5   |

13x13 Pairwise Reciprocal Matrix

| Attributes | Gln  | Arg  | Ser  | Tiny | Chrg | Bsc  | Acd  | HI   | II   | CPI  | SB   | BSH  | BST  | $\lambda$ | Priority |
|------------|------|------|------|------|------|------|------|------|------|------|------|------|------|-----------|----------|
| Gln        | 1.00 | 0.12 | 1.00 | 0.34 | 0.17 | 0.17 | 0.20 | 0.25 | 0.17 | 0.20 | 0.17 | 0.20 | 0.20 | 13.18     | 0.02     |
| Arg        | 9.00 | 1.00 | 9.00 | 3.00 | 1.50 | 1.50 | 1.80 | 2.25 | 1.50 | 1.80 | 1.50 | 1.80 | 1.80 | 13.00     | 0.15     |
| Ser        | 1.00 | 0.12 | 1.00 | 0.34 | 0.17 | 0.17 | 0.20 | 0.25 | 0.17 | 0.20 | 0.17 | 0.20 | 0.20 | 13.18     | 0.02     |
| Tiny       | 3.00 | 0.34 | 3.00 | 1.00 | 0.50 | 0.50 | 0.60 | 0.75 | 0.50 | 0.60 | 0.50 | 0.60 | 0.60 | 13.02     | 0.05     |
| Chrg       | 6.00 | 0.67 | 6.00 | 2.00 | 1.00 | 1.00 | 1.20 | 1.50 | 1.00 | 1.20 | 1.00 | 1.20 | 1.20 | 13.01     | 0.10     |
| Bsc        | 6.00 | 0.67 | 6.00 | 2.00 | 1.00 | 1.00 | 1.20 | 1.50 | 1.00 | 1.20 | 1.00 | 1.20 | 1.20 | 13.01     | 0.10     |
| Acd        | 5.00 | 0.56 | 5.00 | 1.67 | 0.84 | 0.84 | 1.00 | 1.25 | 0.84 | 1.00 | 0.84 | 1.00 | 1.00 | 13.04     | 0.08     |
| HI         | 4.00 | 0.45 | 4.00 | 1.34 | 0.67 | 0.67 | 0.80 | 1.00 | 0.67 | 0.80 | 0.67 | 0.80 | 0.80 | 13.04     | 0.06     |
| II         | 6.00 | 0.67 | 6.00 | 2.00 | 1.00 | 1.00 | 1.20 | 1.50 | 1.00 | 1.20 | 1.00 | 1.20 | 1.20 | 13.01     | 0.10     |
| CPI        | 5.00 | 0.56 | 5.00 | 1.67 | 0.84 | 0.84 | 1.00 | 1.25 | 0.84 | 1.00 | 0.84 | 1.00 | 1.00 | 13.04     | 0.08     |
| SB         | 6.00 | 0.67 | 6.00 | 2.00 | 1.00 | 1.00 | 1.20 | 1.50 | 1.00 | 1.20 | 1.00 | 1.20 | 1.20 | 13.01     | 0.10     |
| BSH        | 5.00 | 0.56 | 5.00 | 1.67 | 0.84 | 0.84 | 1.00 | 1.25 | 0.84 | 1.00 | 0.84 | 1.00 | 1.00 | 13.04     | 0.08     |
| BST        | 5.00 | 0.56 | 5.00 | 1.67 | 0.84 | 0.84 | 1.00 | 1.25 | 0.84 | 1.00 | 0.84 | 1.00 | 1.00 | 13.04     | 0.08     |

Relative Priority Vectors

$$\lambda_{\max} = 13.039$$

$$CI = \frac{\lambda_{\max} - n}{n - 1} ; n = 13$$

$$\text{Consistency Index (CI)} = 0.003$$

$$\text{Random Index (RI) for 13} = 1.555$$

$$CR = \frac{CI}{RI}$$

$$\text{Consistency Ratio (CR)} = 0.0021$$

**RESULT: Matrix is consistent (CR < 0.1)**

Figure A2.20: Reciprocal pairwise comparison matrix and consistency calculation for B-Nb dataset of protein.

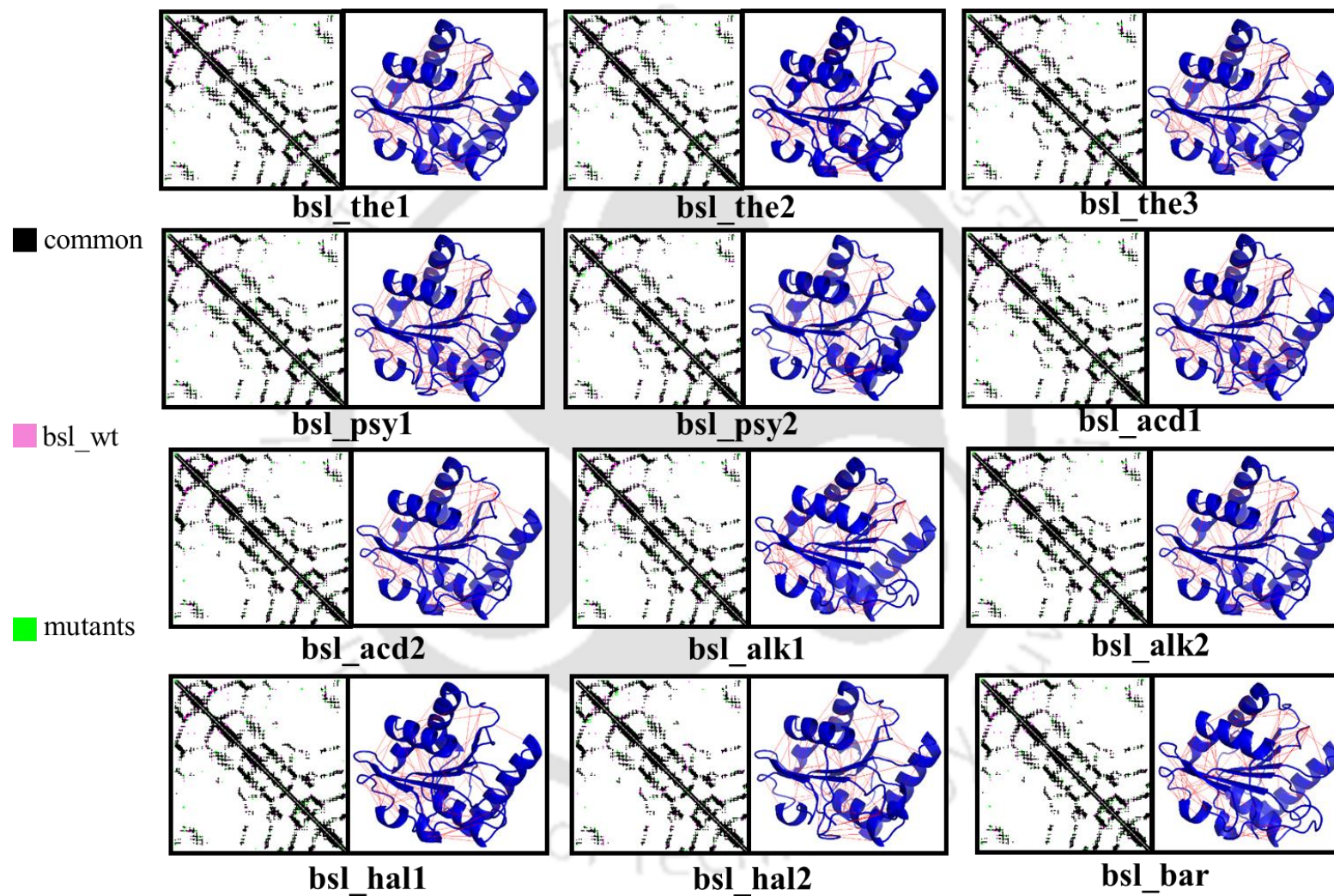


Figure A2.21: Validation of generated wild-type and mutant lipases through Contact map analysis using CMView 1.1.1.



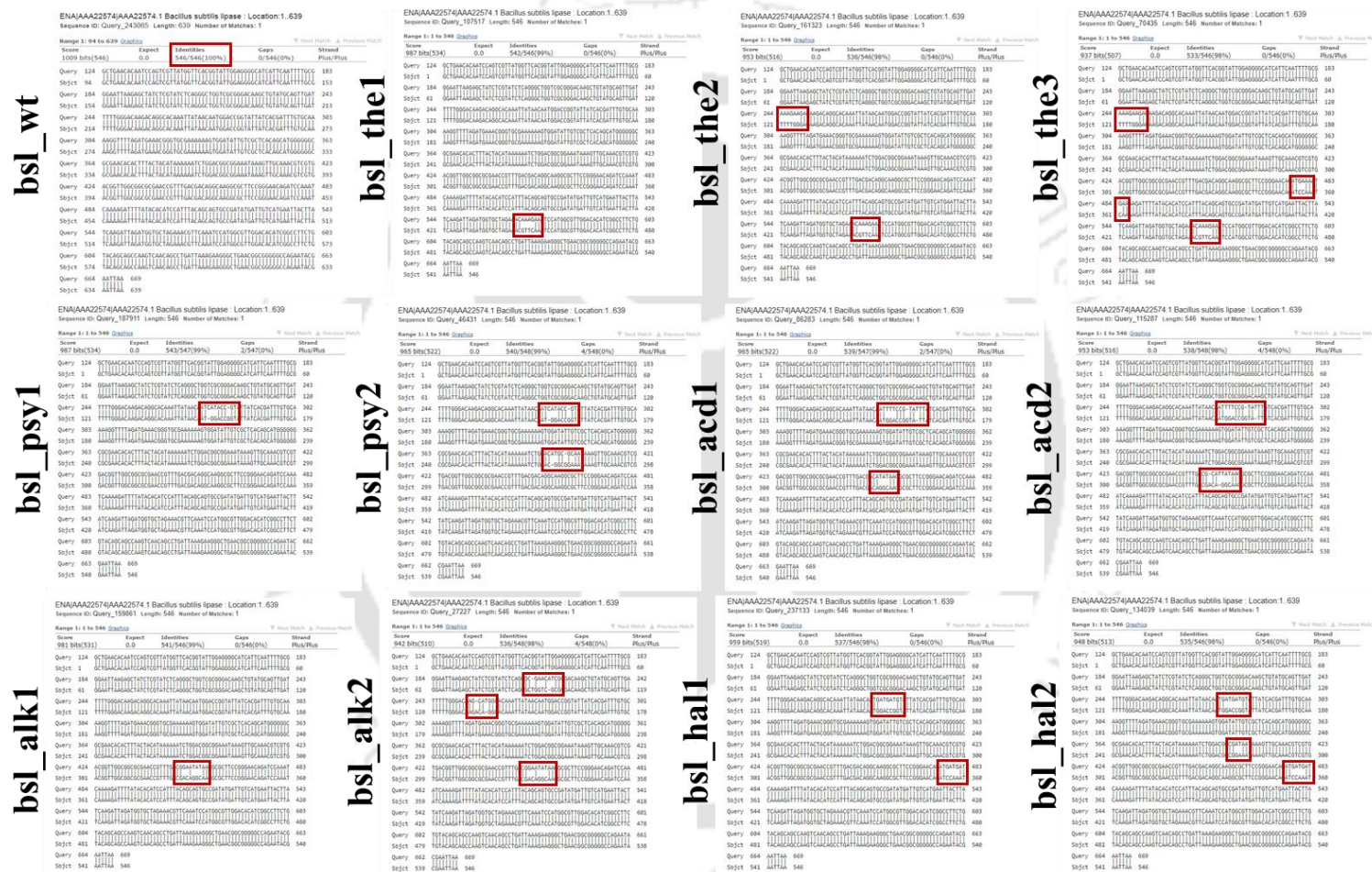


Figure A2.22: Confirmation of cloning and mutations in lipase by sequencing

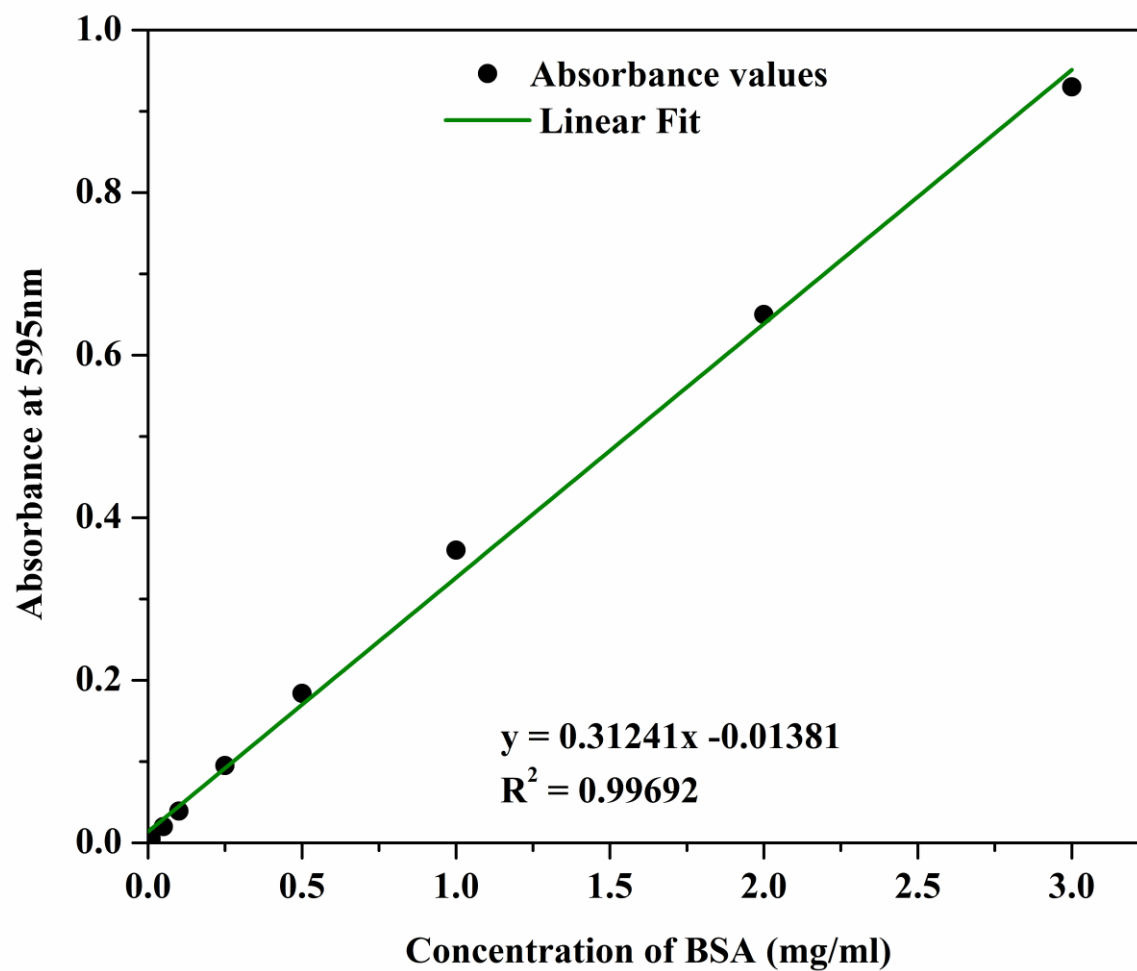


Figure A2.23: Standard curve of BSA for protein estimation