Attaining Protein Thermostability – A Rationalised Approach

A Thesis
Submitted in Partial
Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

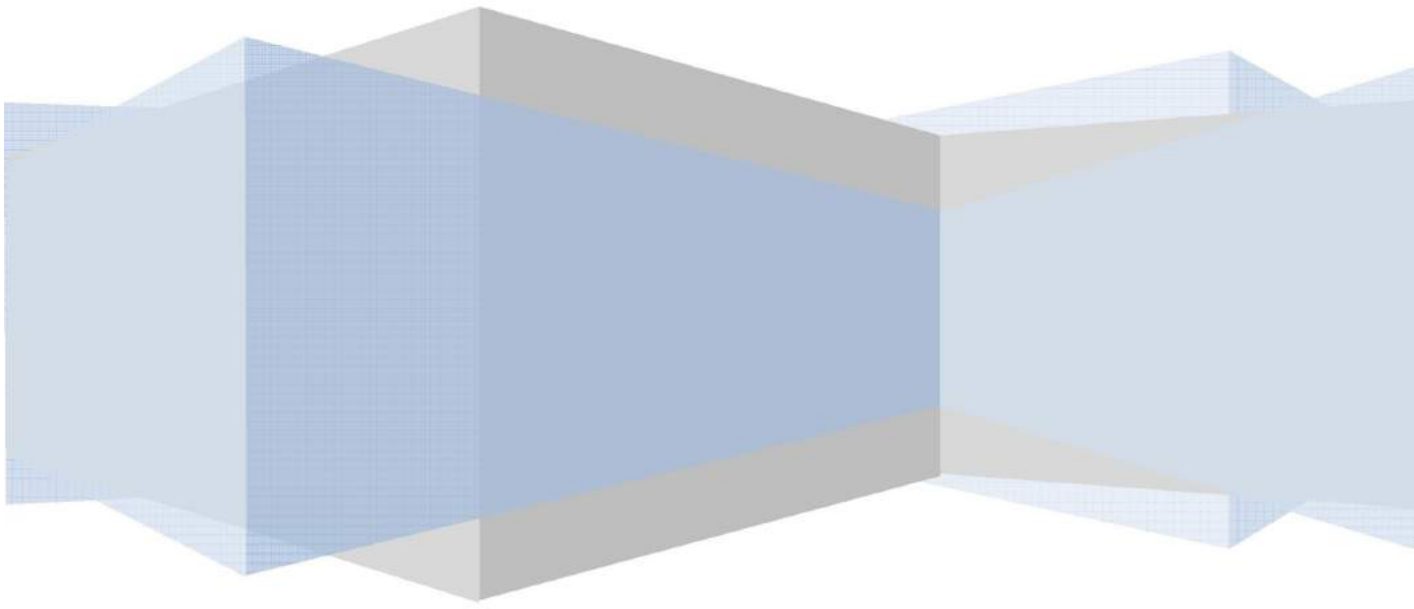
BY

DEBAMITRA CHAKRAVORTY

DEPARTMENT OF BIOSCIENCES & BIOENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI-781039, ASSAM, INDIA

JANUARY 2016

Dedicated to my family for their unconditional love and support…



INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

DEPARTMENT OF BIOSCIENCES & BIOENGINEERING

STATEMENT

I do hereby declare that the matter embodied in this thesis entitled ***"Attaining Protein Thermostability – A Rationalised Approach"***, is the result of investigations carried out by me in the Department of Biosciences & Bioengineering, Indian Institute of Technology Guwahati, India, under the guidance of Dr. Sanjukta Patra and co-supervision of Dr. Vishal Trivedi.

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 investigators.

January, 2016

Ms. Debamitra Chakravorty

Roll No. 10610619

ACKNOWLEDGEMENTS

This work is an outcome of persistent effort and a great deal of commitment. It has drawn intellectual support and generous help from experts from various fields. The list is endless and also the contributions. I take this opportunity to express my sincere thanks to everyone, who has been with me in this entire journey. First of all I would like to specially mention about my supervisor and mentor Dr. Sanjukta Patra who gave me the golden opportunity to work in such a nice environment where I could freely nurture my ideas. I am thankful for her constant inspiration and encouragement throughout these years.

I would also like to express my heartfelt gratitude to my co-supervisor Dr. Vishal Trivedi and my doctoral committee members; Dr. V.V Dasu, Dr. V.K Dubey and Dr. B. Pradhan for evaluating my work, giving critical comments, valuable guidance and inspirations. Without them nothing was possible.

I am heavily indebted to my lab members, Dr. P. Saravanan, Nivedita Singh, Mohd. Faheem Khan, Nitendra Yadav and Bhaskar Das who have provided me ongoing support, inspiration and an enjoyable working environment. I am especially thankful to Jai Vardhan and Deepanshu Goyal for assisting me in software development.

I take this opportunity to pay gratitude to all my seniors and friends. I am especially grateful to Dr. Nidhi Chaubey, she made my stay at IIT Guwahati a gratifying and memorable one. I owe my gratitude to the Department of Biotechnology and Central Instrument Facility, IIT Guwahati for providing me all supports and necessary facilities. My sincere gratitude remains for the Department Technical Assistants.

I also dedicate this work to my family members, especially my father, Prabir Kumar Chakravorty and mother, Susmita Chakravorty and my husband, Arnab Dawn. It is their unconditional love and patience which motivated me to achieve my goals.

My final words of acknowledgements are for Almighty God for giving me strength for establishing me in such a way so that I could complete my work properly.

Debamitra Chakravorty

January, 2016

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

DEPARTMENT OF BIOSCIENCES & BIOENGINEERING

CERTIFICATE

It is certified that the work described in this thesis, entitled "***Attaining Protein Thermostability – A Rationalised Approach"***, done by Ms Debamitra Chakravorty (Roll No: 10610619), for the award of degree of Doctor of Philosophy is an authentic record of the results obtained from the research work carried out under my supervision in the Department of Biosciences & Bioengineering, Indian Institute of Technology Guwahati, India, and this work has not been submitted elsewhere for a degree.

Dr. Sanjukta Patra
Associate Professor
Dept. of Biosciences & Bioengineering
(Thesis supervisor)

Dr. Vishal Trivedi
Associate Professor
Dept. of Biosciences & Bioengineering
(Co-Thesis supervisor)

CHAPTER I

Introduction and Literature Review

Prologue

Evolution is inevitable in nature and mutations are a way of attaining change for evolutionary purpose. In proteins natural mutation occurs such that it tries to maintain maximum stability while retaining functionality and better significance for existence indicating that nature rationalises the way of attaining its mutations. Thermophilic proteins are examples of such natural mutations that lead to stability of proteins at extreme of temperatures. Though these proteins find use in various industrial processes for example, detergent, paper, biofuel and chemical synthesis, but, it is rather difficult to extract and purify such proteins from their natural sources. Therefore it is important to know the rationale behind thermostability and whether it is interplay of factors at all hierarchies of protein organization. It is essential to know whether such enzymes can be produced recombinantly through protein engineering approaches. An in depth literature survey on this, uncovered that the trend followed had been performing random mutations via various methods and then screening the mutations by application of selection pressure to sort out the desired mutations. This method is known as directed evolution. A rationalised approach to thermostabilize proteins through protein engineering ceases to exist. Therefore this chapter aided in formulating the research question as whether engineering of *in vitro* mutations be rationalised by development of a predicting method so that proteins attain predicted enhanced thermostability along with retaining functionality and expression?

Outputs:

1. Debamitra Chakravorty, Sanjukta Patra. An insight into attaining thermostable proteins: process and challenge (Manuscript under preparation).

1.1. Introduction

Evolution is inevitable in nature and mutations are a way of attaining change for evolutionary purpose. In proteins natural mutation occurs such that it tries to maintain maximum stability while retaining functionality and better significance for existence indicating that nature rationalises the way of attaining its mutations. The present thesis is a research attempt to rationalise mutation attainment with focus on thermostable proteins and to develop a better model to classify thermostable proteins. Attaining stable mutation has always been moving towards to stable protein entity. Hence, it is essential to understand protein stability to attain mutations. Here by protein stability refers to the kinetic and thermodynamic stability of proteins in extreme environmental conditions like high temperature. Such proteins having a high (transition) temperature (T_m) are called thermostable (Turner et al. 2007). Thermostable proteins hold high priority in industries because they allow easy mixing, better substrate solubility, high mass transfer rate, and lowered risk of contamination (Turner 2007). Such enzymes are generally obtainable from thermophiles and hyperthermophiles that thrive at temperatures above 45°C in various geothermal milieu of the Earth. One exception is the thermostable lipase from *Candida antarctica* which is a psychrophile. Only a small portion of thermophilic microorganisms have been characterized (Raysenbach et al. 2002). Phylogenetically, they belong to bacteria and archae (Stetter, 1999). Fig 1.1 illustrates the distribution of hyperthermophiles among bacteria and archaea. In 2002, Reysenbach and Shock reported that hydrothermal milieu with temperature between 50°C and 90°C are dominated by bacteria. In hot springs with a temperature range of 60- 68°C, *Hydrogenobacter* are dominant. Thermophilic fungi have been found in anthropogenic, self-heated habitats like compost piles (Tansey et al. 1971). Highly thermophilic niches with temperatures above 90°C are dominated by archaea (Reysenbach and Shock, 2002). However, extreme thermophiles, those surviving above 65°C, comprise prokaryotes only (Madigan, 2003). The diversity of thermophiles is much greater in their natural environments than in environments that have been cultured, and these uncultured microbes belong to new families, classes,

orders, and even kingdoms. Hot springs, geysers, volcanoes and deep-sea hydrothermal vents have been popular haunting ground for the search of new types of thermophilic organism. Interestingly thermophiles are also found in the compost heaps and coal refuses piles where high temperature is due to biological activity (Table 1.1). The various sources of thermophilic proteins have been presented in Table 1.1.

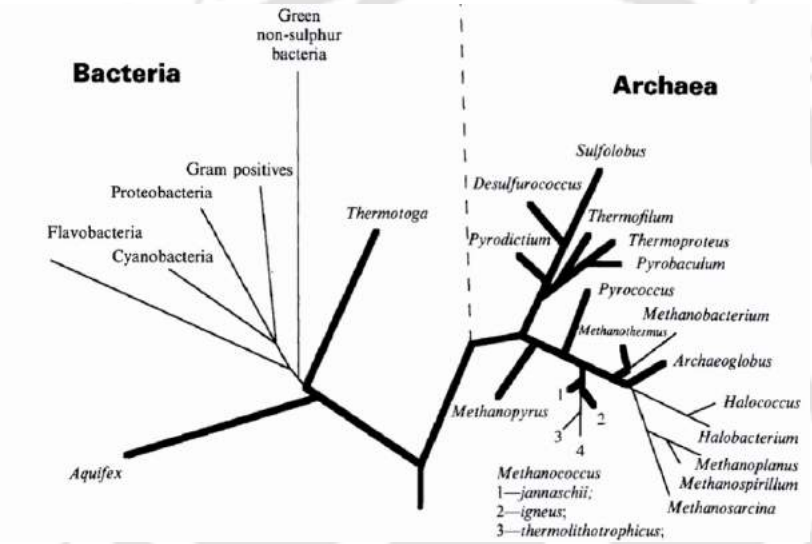


Fig. 1.1. Sources of some chosen thermostable enzymes. This figure has been adapted from Stetter KO. 1999. Extremophiles and their adaptation to hot environments. FEBS Letters. 452: 22-25.

Table 1.1. Sources of thermostable enzymes

Source	Microorganism	Enzyme
Hot spring	*Thermus sp*., *Bacillus thermoleovocans*	α-Amylase, lipase
Deep sea hydrothermal vent	*Staphilothermus marinus*, *Pyrococcus abyssi*	α -Amylase, Alkaline phosphatase
Compost of fermenting citrus peels, coffee and tea extract residues	*Bacillus strain*	Endochitinase
Compost	*Bacillus stearothermophilus CH-4*	β-N-acetylhexosaminidase
Sediments of hot springs	*Bacillus sp.* 3183	α-Amylase-like pullulanase
Garbage dump	*Bacillus circulans*	Xylanase

The vivid extreme milieu talked about here to be the natural sources of thermophiles are challenging to access. Thus, an easy approach for their isolation and large scale production will add to industrial economy. Therefore development of thermostable enzymes from their mesostable counterparts has become a trend through the strategies of directed evolution and site directed mutagenesis approaches. To attain this, physico-chemical parameters leading to thermostability has been an area of intense research. Irrespective of this fact there are very few commercially available thermostable enzymes as such environments are challenging to access. Thermostable proteins have been shown to possess increased number of intramolecular interactions such as hydrogen bonds and hydrophobic interactions, rigidity, and tighter core packing (Okada 2010). All the work narrows down conclusively to a single point that protein thermostability is the cumulative effect of multitudes of protein stabilizing factors. Each protein shows individual characteristics and patterns of intra molecular interactions to achieve the same. The end result is the absence of a guided approach for thermostabilizing proteins.

The features for thermostability are known but lack the information about how these features are globally balanced in order to render proteins thermostable. It can also be said here that the features have not yet been prioritized or ranked according to their contribution towards thermostabilizing proteins.

1.2. Protein attributes responsible for thermostability

Globular proteins stable at high temperature are examples of the remarkable work of nature. Thermophilies have adapted to temperature maxima by the evolution of a wealth of structural and functional factors right from their genome to proteome. It is vital to question the mechanism behind such stability. It is important for understanding the fundamentals of protein organization polypeptide chain and interesting for biotechnological issues, for preparing recombinant enzymes by protein engineering for their industrial use in high-temperature processes. Significant research had been carried out in determination of protein thermostabilizing factors.

Plethora of mechanisms has been reported by various researchers and thermostability has been attributed to be the synergistic effect of multitudes of factors stabilizing proteins. The factors that have been researched upon till date have been presented here. A vast literature exists, on the strategies for thermostability of proteins. Yano and Poulos 2003 and Trivedi et al. 2006, had carried out a detailed review for the same. Some of the common reported factors responsible for thermostability of proteins were increase in hydrophobicity of proteins (Haney et al. 1997; Sadeghi et al. 2006), better compactness of protein structure (Russell et al. 1997), increment of hydrogen bonding (Vogt and Argos 1997; Gromiha 2001; Sadeghi et al. 2006) and salt bridges (Kumar et al. 2000; Sadeghi et al. 2006). It has also been reported that protein thermo stabilize mainly by increasing ion pairs (Szilagyi and Zavodsky 2000). It is clear from the aforesaid that particular rule does not exist for thermostabilization of proteins and thermostability is said to be a compound effect of various structural factors governed by the milieu in which they

are actively present. The role of factors has been individually elaborated in the following subsections.

1.2.1. Genome

Till date large repertoire of thermophilic organisms have been identified. However only few genomes have been successfully sequenced. Search for genomes with the key words "thermophiles", "hyperthermophiles" and "thermostable" in NCBI genome search results in a total of 44 hits. Based on these genome studies it was demonstrated that variations in nucleotide composition can have very significant effects on the patterns of codon usage and thermostability (Stenico et al. 1994, Frank and Lobry, 1999, Sueoka and Kawanishi, 2000, Kanaya et al. 2001). Thermophiles and hyperthermophiles have been reported to have a high GC content (Bao et al. 2002; Saunders et al. 2003). They were also observed to have preference for purine-rich codons. Such codons generally code for charged amino acids (Paz et al. 2004). Additionally DNA was observed to possess positive supercoils resulting in greater stability (Madigan, 2000). In some hyperthermophiles, heat-resistant proteins associate with DNA and enhance its stability (Madigan, 2000). Recently, work carried out by Zeldovich et al. (2007) concludes 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. These amino acid residues increase the content of hydrophobic and charged amino acids, enhancing thermostability.

1.2.2. Proteome

Thermophiles are under constant threat of high temperature on their proteins. Hyperthermophilic proteins are more resistant to denaturation due to restriction on the flexibility of these proteins (Scandurra et al. 1998). Multitude of factors both in the primary, secondary and tertiary level of protein organization have been reported to contribute towards protein stability at elevated temperatures. Heat shock proteins and

chaperones and also aid in protein thermostability. Each of these factors has been described in the following sections.

Primary protein sequence: amino acid

Many researchers have attempted to study the effect of amino acids on protein thermostability (Spassov et al. 1995, Vogt and Argos, 1997, Vogt et al. 1997 and Szilagyi and Zavodszky, 2000). At temperatures greater than 100^{o}C the thermostability of amino acids have been reported by Jaenicke et al. (1998) as follows (V,L) > I> Y> K> H> M> T> S> W> (D, E, R, C). The question that arises here is that whether the same rule is applicable to proteins that are stable at <100^{o}C, or the psychrophilic proteins like *Candida Antarctica* lipase which show thermostability. An interesting rule was derived by Farias and Bonato (2003) who found that the E +K/Q +H ratio can distinguish hyperthermophiles (>4.5) from mesophiles (<2.5) and thermophiles (3.2-4.6). Trivedi et al. (2006) 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". Literature survey leads to fuzzy results. Table 1.2 summarizes the reported features. Conclusions drawn by authors employing study of individual proteins were excluded from this survey.

Table 1.2. Contribution of amino acid residues towards protein thermostability

Amino Acids	Positive Contribution	Negative Contribution
Ala	High helix propensity, hydrophobic interaction, higher aliphatic index	Nil
Cys	Disulphide bond	Thermolabile
Glu	Ionic interactions	Nil
Asp	Ionic interactions	Asp-Pro combination may be susceptible to hydrolysis of peptide bonds
Phe	Core hydrophobicity	Nil
Gly	Nil	Makes cavity in inner part of protein structure
His	Ionic interactions	Nil
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 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

Thus from Table 1.2 it is clear that the overall picture still lacks clarity and derivation of global rule is still elusive. In simple terms it can be said that derivation of "general rules" from comparative studies is difficult, because of the result of different stabilization features in each study (Burg and Eijsink, 2002).

Protein destabilization due to thermolabile residues is mainly due to deamidation, cleavage of peptide bonds and oxidation of Cys and Met residues. Deamidation of Asn residues is a spontaneous and non-enzymatic method which can occur in acidic, neutral, or alkaline conditions. Therefore proteins with high Asn residues are suspected to be less thermostable (Chakravarty and Varadarajan, 2000 and Kumar et al. 2000). Furthermore reports also suggest that asparagines containing peptides are more prone to deamidation at a faster rate than glutamine counterparts (Vijayarangakannan, 2005).

The cleavage of a peptide can destabilize proteins by disrupting a protein chain. Shirley 1995, 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. (3) Proteolysis by enzymes. Literature survey showed that protein having high frequency of Asp and Asn residues were less thermostable (Chakravarty and Varadarajan, 2000 and Kumar et al. 2000).

Oxidation of Cys and Met often result in protein destabilization. Cys has been reported to be involved in protein stability of extracellular proteins by formation of disulphide Bridge. 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 β-elimination (Whitaker and Feeney 1983). This has been shown to irreversibly inactivate of ribonuclease and lysozyme at pH 6-8 and 90-100°C (Ahern and Klibanov 1988). Similarly, Met oxidation has been associated with the inactivation of proteins (Swaim and Pizzo 1988). Many thermostable proteins have been reported to have low frequency of Met residues (Kumar et al. 2000; Xu et al. 1998; Mattos, 2002).

Intra-protein interactions

Enhancement of intra-protein or Van der Waals interactions have been time immemorial reported to be associated with increasing thermostability of proteins (Kumar et al. 2000; Trivedi et al. 2006). Such interactions and their possible contributions towards thermostabilizing proteins have been elaborately discussed in the next sub-sections. The available tools and softwares that can predict intra-molecular interactions have been presented in Table 1.3.

Table 1.3. Existing popular tools to predict intra-protein interactions

Tools	Properties	Reference
PIC web server	Hydrogen bond, Ionic, hydrophobic, aromatic, disulphide bonds	Tina et al. 2007
HBOND, HBAT, HBPLOT	Hydrogen bond	Petsko et al. 2004; Tiwari et al. 2007; Bikadi et al. 2007
ESBRI	Salt bridges	Costantini et al. 2008
VADAR	Hydrogen bond, accessible surface area	Willard et al. 2003
DiANNA, Disulfind, Dinosolve	Disulphide bonds	Ferre et al. 2006; Ceroni et al. 2006; Yaseen et al. 2013
CHpredict	Aromatic interactions	Kaur et al. 2006

Hydrogen bonds

Hydrogen bonding is the highest cited feature in literature for protein thermostabilization (Vogt et al. 1997). In thermostable proteins hydrogen bonds showed an increase of 11.7 hydrogen bonds per chain per 10°C rise in thermostability (Vogt et al. 1997). So how are hydrogen bonds (H-bonds) formed and how they can contribute towards protein stability? 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 H- bonds, it is said that whenever two non-hydrogen atoms with

opposite partial charges (donor (D)–accepter (A) pairs) were found to be within a distance of 3.5 Å, a hydrogen bond can form. The geometrical goodness of the hydrogen bond was assessed by computing the values of the following angles. Hydrogen bonds have good geometry if both these angles i.e., between vectors BD–D and D–A, BD is the atom covalently bonded to the donor (D) atom and between vectors D–A and A–BA, BA is the atom covalently bonded to the acceptor (A) atom, lie in the range 90–150° (Kumar et al. 2000). Additionally strongest, directional hydrogen bond, the hydrogen atom points directly to the acceptor atom. If it points more than 30º away the bond energy becomes much less (Watson et al. 2008). 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 (Khechinashvili et al. 2006). 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 higher the protein thermostability (Khechinashvili et al. 2006). 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 (Aparna et al. 2005). It was showed 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 (McDonald & Thornton 1994). Backbone-backbone H-bond are considered to have lower configurational entropy. The charge-transfer contribution to the hydrogen-bond energy increases and the angle decreases (Kolman et al. 1972). 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 (Guerois et al. 2002). 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. Whereas 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 thermophilic monomeric proteins (Kumar et al. 2000).

The other types that have been classified are charge charged and charge neutral hydrogen bonds. Charge neutral hydrogen bonds are more stabilizing as desolvation energy making for an H-bond residue is lower than that for an ion pair (Tanner et al. 1996). 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 showed that thermostable proteins showed a consistant increase in hydrogen bonds (Vogt et al. 1997). They can also be divided into: short strong hydrogen bonds: Distance 2-2.5Å. 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 (Watson, 2008; Panigrahi et al. 2008). Recently Srivastava et al. in 2014 showed that increase in hydrogen bond increases thermostability of *Bacillus subtilis* lipases through molecular dynamics simulations.

Electrostatic interactions

Electrostatic interactions have long been implicated in the thermostability of proteins (Perutz and Raidt, 1975; Perutz, 1978; Vogt and Argos, 1997; Jaenicke and Bohm, 1998; Szilagyi and Zavodszky, 2000; Petsko, 2001; Zhou, 2002). It was reported that thermophilic proteins tend to have more salt bridges and surface charge residues (Fukushima et al. 2003). 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 Angstroms (Costantini et al. 2008). A single salt

bridge can contribute 13–22 kJ/mol to the free energy of folding (Jaenicke et al. 1996) and unlike hydrophobic interactions (Privalov et al. 1988), they are relatively unaffected at extremely high temperatures. Elcock proposed that salt-bridge should be more stabilizing at high temperatures because the unfavorable desolvation penalty (Elcock et al. 1998) and the entropic cost of fixing two charged side-chains would decrease with temperatures. Furthermore, thermostability is achieved by upshifting or broadening the thermostability curve. A smaller ΔC_p (heat capacity) can increase the maximum ΔG_u: $\Delta G_u(T_s) = \Delta H_m - \Delta C_p (T_m - T_s)$, or in other words, the protein stability curve is up-shifted if ΔHm is increased or remains constant (Kumar et al. 2001). Salt bridges decrease ΔC_p thus results in the upshift of the thermostability curve (Chan et al. 2011). The effect of ionic interactions on thermostability has been studied by loss of function and gain of function mutations. Vetriani et al. (1998) reported that extensive ion-pair net-works may provide a general strategy for manipulating enzyme thermostability of multisubunit enzymes. 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 - π interaction is another form of electrostatic interaction responsible for protein Thermostability. Cation- π interactions are formed by the interactions between positively charged residues (Lys and Arg) and aromatic amino acids (Tyr,Trp and Phe). Gromiha et al. (2002) analyzed the influence of cation- π interactions to enhance the stability from mesophilic to thermophilic proteins. Tyr has a greater number of such interactions with Lys in thermophilic proteins. The influence of Phe in making cation- π interactions is higher in mesophiles than in thermophiles. Further, a network of cation- π interactions is maintained by Lys in thermophiles, whereas Arg plays a major role in mesophilic proteins. Moreover, atoms that have a substantial positive charge in both Lys and Arg make a more significant contribution for cation- π interactions than do cationic group atoms (Gromiha et al. 2002). The cation - π interaction between Arg19 and Tyr93 in the

protein indole-3-glycerol phosphate synthase from *Sulfolobus solfataricus* was reported to contribute towards stability (Knoechel et al. 1996).

Hydrophobic interactions

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 (±0.5) kcal mol^{-1} (Pace, 2011). An enhanced hydrophobic effect is one of the reported reasons for the slow unfolding of thermophilic proteins (Okada et al. 2010). Rathi et al. (2014) studied a set of 130 pairs of thermostable and mesophilic proteins and reported hydrophobic interactions as "key factors" for protein thermostability. Burg et al. (1994) increased thermostability of thermolysin-like neutral protease of *Bacillus stearothermophilus* by introducing Arg, Lys or bulky hydrophobic amino acids. Through their experiments it was showed 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 (Vieille et al. 2001).

Core packing is often linked to increased hydrophobicity and stability (Schumann et al. 1993). An increase in hydrophobicity, given that its buried will add to core stability due to increased *van der* Waals interactions. Programs such as Repacking of Cores (ROC) (Desjarlais and Handel, 1995) and protein simulated evolution (PROSE) (Hellinga et al. 1994) can be used to redesign protein cores using molecular force fields. Many proteins were successfully redesigned using these methods. However sometimes overpacking of the cores results in destabilization of the folded conformer (Ventura et al. 2002).

Aromatic interactions

Hydrophobic interactions between aromatic groups of phenylalanine and tyrosine less than 7Å distance results in π-π stacking and enhances protein stability.

Most of them are energetically favorable having potential energies between 0 and –2.0 kcal/mol. Such interactions also link secondary protein structures leading to overall stability (Burley et al. 1985). Surface exposed Tyr-Tyr and Phe-Phe pairs were observed to contribute –1.3 kcal/mol toward thermostabilization in Rnase from *Bacillus amyloliquefaciens* (Serrano et al. 1991). Irrespective of their potential in thermostabilizing proteins, unfortunately such interactions are very hard to engineer.

Disulphide Bonds

Covalent disulfide bonds between cysteine residues are an important tertiary structural feature that results in protein stability (Matsumura and Matthews 1991; Betz 1993; Darby and Creighton 1995). Thermostable proteins have been observed to possess such disulphide bridges. A disulphide bond leads to 2.5 - 3.5 kcal/mol of stabilization. This depends on the distance of the bonds (Braxton, 1996). They have also been observed to lead to stability by reducing the entropy of the denatured state (Betz, 1993).

Disulfide bonds have been shown to play important role in oligomerization (Reeds et al. 2013). Oligomerization has been regarded as important determinant of thermostability (Tanaka et al. 2004). 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 (Zavodszky et al. 2001). It was concluded that disulphide bridging stabilizes both native and denatured state (Zavodszky et al. 2001). The difference in stabilization between the two states determine the state of protein stability (Zavodszky et al. 2001). Reduction of five disulphide bonds in *Aspergillus niger* phytase have been linked to its destability (Wang et al. 2004). 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^{o}C increase in melting temperature and a three-fold increase in its half-life (Takagi et al. 1990).

Irrespective of the aforesaid, all thermostable proteins do not possess disulphide bridges. For example *Bacillus stearothermophilus* lipase (PDB Id: 1JI3) which is stable at temperatures greater than 80^{o}C lacks disulphide bonds. Disulphide mutants show decreased as well as increased stabilities (Matsumura and Matthews 1991). 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. This is vouched for by the fact that stability enhancement by insertion of novel disulfide bonds have not always been successful (Zavodszky et al. 2001).

Other Factors

Other factors that have been considered to contribute towards protein thermostability are many. To mention a few; deleted or shortened loops (Russell et al. 1997, 1998), greater rigidity by increasing Pro in proteins (Kumar et al. 2000), docking of N and C termini and anchoring of loose ends (Vieille et al. 2001) and metal coordination (Sujak et al. 2007) have been cited. It is understood that these features are reflections or extensions of the *van der waals* interactions in a protein structure. The question which arises here is that *'whether there are some other factors which can be important for protein thermostability yet to be explored'*. Furthermore, factors like metal coordination and docking of N- and C- terminal loops are qualitative in nature. They have been reported in specific cases of thermostable proteins. For example metal coordination of $Zn2^{+}$ has been predicted to plays a role in structural stability in bacterial lipases like *Bacillus stearothermophilus* (Tyndall et al. 2002). But such coordination can also be observed in mesostable proteins like lipases of *Staphylococcus epidermis*. Being qualitative in nature it is difficult to engineer such coordination to enhance stability.

1.3. Thermodynamic stability

"Function follows structure" is regarded as a truism in biology. Biological components like proteins attain their functional state through folding into their three dimensional tertiary structure by a hierarchical assembly process (Bolen et al. 2008). The native state of protein has been reported to be 5 to 10 kcal/mol greater in stability in comparison to its denatured state (Creighton, 1994). As defined by Anfinsen and Mirsky and Pauling before him, protein folding is inherently a thermodynamic problem. Later in Kauzmann's energy ledger approach it is said that proteins exist in equilibrium between the disordered and ordered conformations and unfolded to native equilibrium is exerted by a gradient in Gibbs free energy (Bolen et al. 2008). The driving force in folding was initially thought to be intramolecular hydrogen bonds and hydrophobicity. Nature's strategy operates predominantly on the backbone in the unfolded state, with much less involvement of the native state (Bolen et al. 2008). Stability is also a subject of various factors like organic osmolytes whose intracellular presence protects cells from environmental stress conditions and hence modulates protein folding. Proteins fold in water spontaneously because it is a poor solvent for unfolded proteins. From solution thermodynamics it is suggested that intramolecular hydrogen bonds are slightly more favored over water and backbone hydrogen bonds. Thus intra molecular hydrogen bonding is predicted to be the main force behind protein folding (Bolen et al. 2008).

1.4. Role of water

Water is involved in protein stability as it has high hydrogen bonding capacity rendering it to be a good solvent for many functional groups (Timasheff, 1995). Water molecules form hydrogen bonds, both in the folded and unfolded state of proteins. The calculation of the effect of hydrogen bonding water in protein stability is a complex issue. Several experimental studies show that the deletion of polar atoms that make hydrogen bonds with a partially or fully buried water molecule can have a large destabilising effect on the protein interaction. It results in hydrophobic effect

and thus destabilized proteins (Takano et al. 1997; Grantcharova et al. 2000; Covalt et al. 2001). Ser and Thr are known as the best residue for interacting with the waters surrounding protein structure (Mattos, 2002). The point to ponder is that as water is released at higher temperature, the local protein structure around water-binding site like Ser or Thr could be changed to be unstable enough to evoke protein instability (Denisov, 1999 and Nagendra et al. 1998). Thus, studies have shown that the thermophilic proteins have very low frequency of Ser compared with mesophilic proteins (Chakravarty and Varadarajan, 2000 and Kumar et al. 2000).

1.5. Approaches to develop thermostable 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 1.4). The approaches utilized to thermostabilize proteins have been presented in Table 1.4. 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 compromised stability and flexibility with the functional state of proteins (Jaenicke et al. 1998). 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.

Table 1.4. Present approaches for thermostabilizing proteins

Proposed features	Contributing factors
Helix stabilisation	Low frequency of Cβ-branched amino acids and specific amino acids at helical ends.
Stabilising interactions in folded protein	Disulfide bridges; Hydrogen bonds; Hydrophobic interactions; Aromatic interactions; Ion-pair networks charged residues; Docking of loose ends
Stabilising interactions between domains/subunits	Oligomer formation via *e.g.* ion pair networks
Dense packing	Increase core hydrophobicity.
Stable surface-exposed amino acids	Low level of surface amino acids prone to deamidation or oxidative degradation.
Approaches to introduce internal thermostability in mesophilic proteins	
Reducing length of or stabilising surface loops and turns	Structure-based site directed mutagenesis. Promising results reported for loop deletions; Proline-stabilisation of loops; Docking of loose ends.
Activity screen of diversified library at desired temperature	Directed evolution and other random methods utilized successfully in several cases
Approaches to develop thermostable proteins	
Diversifying specificity	Structure-based directed evolution
Improving activity at selected pH values	Directed evolution
Broadening temperature range for activity by introducing flexibility in active site	Structure-based directed evolution Patent by Diversa.

*This table has been adapted with permission from Turner et al. 2007 in Potential and utilization of thermophiles and thermostable enzymes in biorefining. Microb Cell Fact. 6:9.

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 (Lehmann et al. 2001). 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 (Lehmann et al. 2001). A single protocol ceases 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 (Spector et al. 2000). Thus to overcome the demerits of such approaches there is a need to develop a guided method

which can predict muatations as thermostabilizing so that such mutations can be carried out *in vitro* through site directed mutagenesis.

1.6. Available Thermostable Protein Databases

ProTherm is the only database available, possessing curated information about thermodynamic data of proteins and their mutants (Bava et al. 2004). The database excludes proteins whose thermodynamic data is unavailable. As all proteins do not possess thermodynamic data, the database is not fully curated for thermostable proteins. Moreover the database also lacks information regarding the physicochemical properties of the proteins. As physico-chemical properties are important determinants of protein thermostability, there is a need for a curated database which will provide comprehensive information about the same.

1.7. Theoretical prediction models of thermostability

To overcome the demerits of directed evolution approaches numerous *in silico* algorithms have been proposed which can predict whether conceptualized mutations will be thermostabilizing. These models have been developed by investigating protein features by comparing thermostable proteins with mesostable 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 (Rohl et al. 2004). Few are support vector machine (SVM) based (Capriotti et al. 2005) and further lesser are based on molecular dynamics (Benedix et al. 2009). Table presents the existing methods that have been used to predict protein thermostability.

Table 1.5. Existing popular softwares that predict stability of mutations

Tools	Salient Features	References
I-Mutant	Support Vector machine based, both sequence and structure can be used, single mutation	Capriotti, 2005
Cupsat	Sequence as input, single amino acid mutations	Parthiban et al. 2006
MUPRO	Support Vector machine based, sequence as input, Single mutation	Cheng et al. 2006
ERIS	Structure as input, multiple mutations	Yin et al. 2007
iPRESTAB	Machine learning based, single mutation	Huang et al. 2007
PoPMuSiC	Single mutation	Dehouck et al. 2009
WET-STAB	Machine learning based, multiple mutation	Huang et al. 2009
MUSTAB	Support Vector machine based, sequence as input, multiple mutations	Teng et al. 2010
AUTO–MUTE	Machine learning based, structure as input, single mutation	Masso et al. 2011
SDM	Sequence/structure as input, single mutation	Worth et al. 2011
iSTABLE	Support vector machine based, structure/sequence as input, single mutation	Chen et al. 2013
NeEMO	Machine learning based, structure as input,	Giollo et al. 2014
ENCoM	Neural Network based, single mutation	Frappier et al. 2014
iRDP	Ensemble of servers	Panigrahi et al. 2015

All the methods used for stability prediction presented in Table 1.5 employ machine learning methods on protein datasets to correctly classify thermostable proteins and discriminate between stabilizing and destabilizing mutations. They perform with higher accuracies than most of the statistical and molecular dynamics simulation methods. The latter also have the disadvantage of requiring high

computational power and proficiencies. There are various examples where machine learning approaches have been utilized. Such methods were based on support vector machines, neural networks and decision trees which can predict the effects of mutations on thermostability (Bava et al. 2004; Capriotti et al. 2005; Kumar et al. 2000). Large datasets of known primary, secondary, and tertiary structures of proteins were used to train the machine learning algorithms. Gromiha et al. analyzed the amino acid compositions of 3075 mesophilic and 1609 thermophilic proteins by logistic functions, neural networks, support vector machines, decision trees and found that charged residues as well as the hydrophobic residues have higher occurrence in thermophiles (Gromiha et al. 2008). In 2010, Prethermut software was developed, based on machine learning methods, to predict the effect of single- or multi-site mutations on protein thermostability (Tian et al. 2010). Ebrahimi et al. employed various supervised and unsupervised machine learning algorithms to find amino acid composition features that contribute to enzyme thermostability (Ebrahimi et al. 2011). They reported Gln content and frequency of hydrophilic residues as the most important protein features for thermostability. They also reported that the amino acid sequence is the main indicator of protein function but direct prediction of protein characteristics such as thermostability is not possible from the primary amino acid sequence (Ebrahimi et al. 2011). Consequently, methods to predict thermostability have focused on the three dimensional structures of proteins. From the aforementioned examples it is clear that bulk of the work done on prediction of protein thermostability is on the primary sequence and tertiary structures of proteins. Moreover though it has been reported that thermophiles can be distinguished by their pattern of synonymous codon usage for several amino acids (Lynn et al. 2002; Lobry et al. 2003), very less work related to model generation at the nucleotide and codon usage levels of thermophiles has been performed. It was also conclusively reported that at elevated temperature selective constraints at all three molecular levels: nucleotide content, codon usage and amino acid composition are important to stabilize thermophilic proteins (Lynn et al. 2002). Only recently Lu et al. developed a

hybrid fractal algorithm to predict thermophilic nucleotide sequences with an average accuracy of 0.945 (Lu et al. 2012).

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 (Seeliger et al. 2010). Khan and Vihinen recently evaluated and compared 11 online stability predictors and found that the predictions were only moderately accurate (Khan et al. 2010). 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 require several empirical parameters or heuristics such as patterning of residues for their calculations. Moreover statistical analysis based on Tm 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 (Kumar et al. 2000). Molecular dynamic simulations of mutation are several orders of magnitude complicated than that with a knowledge-based scoring function (Sleegier et al. 2010). 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 (Tian et al. 2010). For example, a predictive model weighted decision table method-WET-STAB was developed. It is a weighted decision table method for predicting protein thermostability change upon double mutation from amino acid sequences (Huang et al. 2009). However the accuracy drops to 0.57 when it is tested on the hypothetical reverse mutations (Li et al. 2012). The other model Protein Thermostability Random Forest model (PROTS-RF) is based on Random Forest algorithm and achieves an accuracy of 78.7% for multiple mutations (Li et al. 2012). The accuracy achieved until date creates limitation when greater than two mutations are to be performed. Additionally the cumulative effect of all the mutations on the physicochemical features or structural

changes associated with the same 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 thermostabilize proteins.

1.8. An insight from molecular dynamics simulation

Molecular dynamics (MD) simulations have been a recent and popular methodology to understand the rationale behind protein thermostability after machine learning approaches. To design thermostable proteins by 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 (Berendsen et al. 1995), NAMD (Phillips et al. 2005) and DESMOND (Bowers et al. 2006). Mutating highly flexible residues with ones that are more rigid can enhance their thermal stability. Manjunath et al. (2013) carried out 50 ns MD simulation of SAICAR synthetase from mesophilic and hyperthermophilic sources. They concluded that the thermophilic proteins were more rigid. Long distance interactions are lost in mesophilic proteins in contrast to that observed in the thermophilic counterparts. Paul et al. (2014) performed 10 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 mesophilic protein to have greater flexibility at higher temperatures. In another work, 16 thermostable mutants of *Bacillus subtilis*

lipase A were studied and a direct correlation was derived for structural rigidity and thermostability (Rathi et al. 2015). 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.

1.9. Origin of the work

Fig.1.2 depicts the origin of this research work. It illustrates that random protein engineering approaches are still utilized for thermostabilizing proteins from mesophilic sources with limited success. Thus, there is a need to develop a standard protocol which will employ site-directed mutagenesis to thermostabilize proteins from mesophilic sources with investment of much lesser time and money. It is evident from the aforementioned that though protein thermostability has been extensively studied for many decades, universal rules to render proteins thermostable do not exist. The questions that need to be answered are:

i. What are the statistically significant factors for thermostabilizing proteins?
ii. Can such factors be prioritized and a guided rule designed to render proteins thermostable?
iii. Which factor can be used for attaining mutations?
iv. Can multiple mutations be handled?
v. Can a model be generated to predict thermostabilizing mutations?

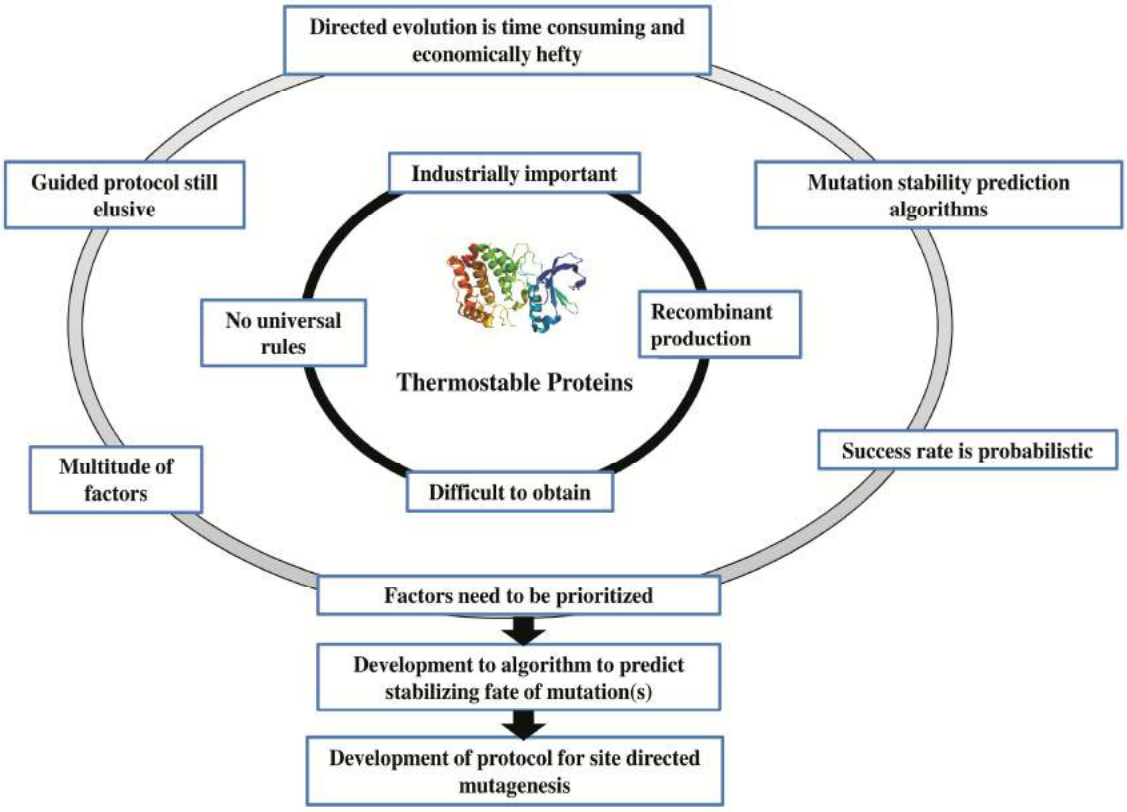


Fig.1.2. Schema representing the origin of the work.

1.10. Hypothesis

The hypothesis has been illustrated in Fig 1.3. Briefly it comprehends that thermostable proteins are important industrially and to attain thermostability, cumulative effect of all the available protein stabilizing features leading to thermostability needs to be enumerated. This can be done by analyzing statistically significant and non-correlated features that differentiate extremophilic and non extremophilic proteins. These factors then need to be prioritized or ranked according to their contribution towards thermostability. These numerical ranks can thus be used to develop a tool that can differentiate between extreme-stabilizing mutations. Such predicted mutations can then be carried out using site directed mutagenesis to obtain the desired thermostable proteins and further their applications studied.

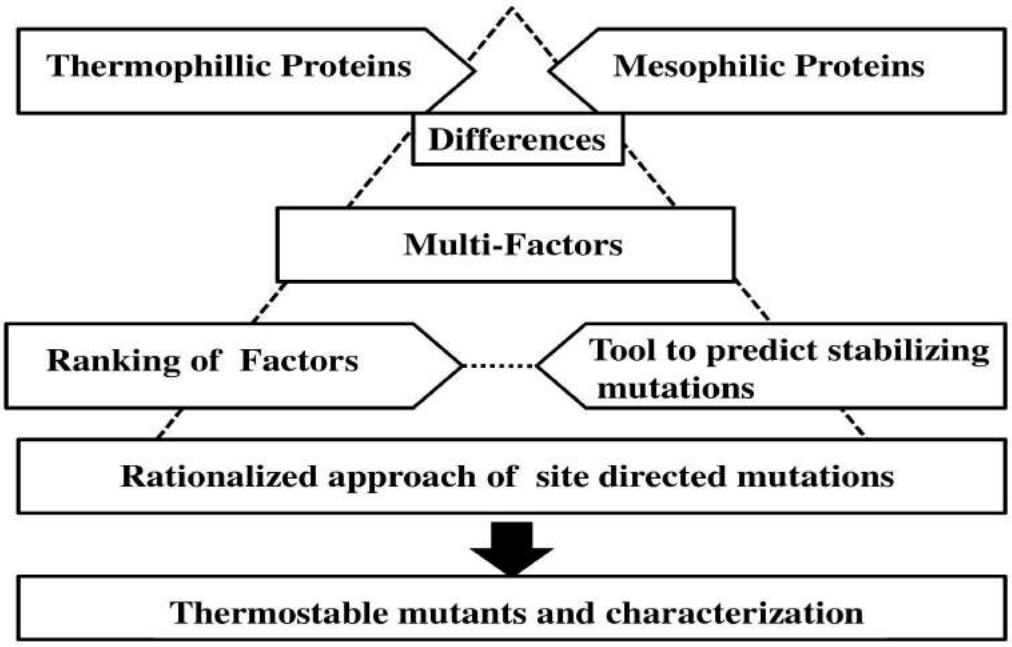


Fig. 1.3. The proposed hypothesis

Thus the developed method should be able to identify possible features contributing to thermostability, prioritise the features or rank them and identify thermosstabilizing mutants.

1.11. The relevance and expected outcome of the proposed study

A single design strategy to mutate amino acid residues which can render proteins thermostable at desired temperature is still a knowledge gap. Moreover cardinal points for each of these factors or a set of stabilization rules is still elusive. The edge of this work and its novelty is that it is based on the lacuna and the conclusion that protein stability has been attributed to the interplay of various

molecular factors which has not been prioritized and that a unified mechanism of protein stability ceases to exist. The work uses all possible knowledge available with respect to protein stability to develop a ranking model. The model will be able to predict whether mutations in protein will lead positively to thermostability. Its novelty lies in its basis of considering all the important intra-protein interactions altogether for developing the rank and this method is attractive due to its potential to consider multiple mutations at one time. This method is a low cost and time-saving technology in comparison to the current computational and experimental approaches.

The outcomes can be summarized as:

i. A Comprehensive database of thermostable proteins with their physicochemical properties.
ii. A ranking model for the thermostabilizing features in accordance to their importance in thermostabilizing proteins.
iii. Rules to mutate proteins to enhance their temperature stability through site directed mutagenesis.
iv. A tool to predict thermostabilizing mutation(s).
v. Recombinant proteins showing stability at high temperature.

1.12. Conclusions

It is expected that use of thermostable enzymes in industrial applications will increase with time, ultimately leading to wider availability and lower price. However, it is a tough job to access thermophilic environments and in laboratory the major problem in culturing thermophiles in solid media using agar as agar-based media exhibit syneresis at such high temperatures. Furthermore, recombinant production of thermostable enzymes in mesophilic hosts often results in improper codon usage and improper folding of the proteins which can result in reduced enzyme activity or low level of expression. Thus the goal should be increasing stability retaining optimum activity of such enzymes which can be achieved through recombinant production of

homologous mesostable enzymes which have undergone point mutations for thermostabilization. Lacuna derived from this survey shows that first, despite several statistical studies of primary sequences, general strategies in terms of preferred amino acid exchanges are yet to be expected. Second, single unique and global feature defining protein thermostability ceases to exist till date. Literature analysis shows that structure determines a proteins function and stability and thermal stability is the interplay of many factors. Third, though it was reported that very small 3D-structural alterations may suffice to cope with the various extreme conditions (Vieille et al. 2001), a single design strategy to mutate amino acid residues which can render proteins thermostable at desired temperature is still a knowledge gap. Fourth, work that has prioritized the factors in accordance with their role in contributing towards thermostability, ceases to exist. Fifth, Cardinal points for each of these factors or a set of stabilization rules is still elusive. Keeping the aforementioned in mind, this research work focuses in finding out the most critical factor(s) for thermostabilizing a protein. To develop a tool that can predict multiple stabilizing mutations and to derive global rules for protein thermostability, so that it can be utilized for enhancing thermal stability of proteins by targeted mutagenesis.

CHAPTER II

Creation of Thermostable Protein Database

Prologue

A thorough literature review highlights that a comprehensive and curated database is required for thermostable proteins. Thus, database creation has been one of the main focuses of this dissertation. A user friendly and web compatible relational database-Thermostable Protein Structural database was created. The same can be accessed through www.extreme-stabledb.in. The architecture of database was in Apache, PHP and MySql platform. The salient features include information about all the thermostable proteins available in Protein Data Bank. The database contains information about 378 thermostable proteins from 132 thermophilic and thermophilic organisms and 261 mutants. Details about their source organisms, phylogeny and enzyme classification were also collected. The database also has information about the engineered mutants which have played a pivotal role in understanding protein thermostability. Thermostability has been attributed to be the result of cumulative effect of multitudes of factors. Thus, data regarding protein features like their physicochemical properties and amino acid composition can also be searched for in the database. The database also provides information about the plethora of literature available about thermostable proteins and data which are dedicated to understand the rationale behind protein thermostability. The database has also been utilized to generate meaningful data by data refinement and collection of homologous mesostable protein structures and enumeration of their complementary features to that of the thermostable proteins. This data has been employed for an in-depth study about protein thermostability.

Outputs:

1. Thermostable protein structural database and Intra-Protein Interaction Enumerator.
2. Debamitra Chakravorty, Mohd. Faheem Khan, Sanjukta Patra. The creation of Extremostable Protein Database (Manuscript under preparation).

2.1. Introduction

Biological databases are computerized repository of organized and refined biological data with chief objective of easy retrieval of information. There are innumerable biological databases. High emphasis is on sequence and structural data of nucleotide and protein related databases. It is interesting to conclude here, that bulk of biological databases relates to enzymes which are proteins that support all biochemical reactions of life. The reason is well known as the development of biological web databases will provide very useful information and insights for biological systems (Zou et al. 2015). Global databases that have been developed are National Centre for Biotechnology Information (NCBI), Protein Data Bank (PDB) and UniProt. Huge amount of data are still being generated and deposited in such global databases. This necessitates data curation and channelizing it in the form of comprehensive repositories, so that it will be easily accessible and utilized for further knowledge development.

On similar lines large data have been generated for thermostable proteins over the past decades. This is evident from the fact that text query of "thermophile", "hyperthermophile" and "thermostable" results in 44 genomes in NCBI and 1280 protein structures in Protein Data Bank. In the direction of data curation related to protein thermodynamics, Gromiha et al. (1999) created the Protherm database for proteins and mutants. The database hosts only thermodynamic data of wild type and mutant proteins that are not exclusively thermostable. It is clear that a database dedicated solely to thermostable proteins ceases to exist until date. Further, in recent years theoretical predictions of protein stability and have led to the accumulation of thermostable mutants. They throw light on the mechanism for thermostability. However, such knowledge gets masked when they are deposited in global databases like PDB. Therefore, there is a need to curate such data so that a universal protocol for temperature stability can be obtained. This will aid in designing mutants from mesophilic proteins which will enhance protein stability. Such mutants are industrially important as thermostable proteins find use in paper, dairy, detergent and many other industries (Kumar et al. 2000). Additionally databases dedicated to

proteins lack information about their physicochemical properties like hydrogen bonds, salt bridges and ionic interactions. Thus, it is important to host a database which will have all such informations.

In this present chapter the development and integration of a curated database for thermostable proteins has been outlined. The database is a collection of data relevant to thermostable proteins. The database was built on an "entity relationship model" wherein data is arranged in tabulated form and each table is related to one another by primary and foreign keys. This is followed by building a user accessible web platform in HTML and CSS and a user friendly search engine which can be used to browse data in a meaningful format from the database. The immediate application is that these data can be further processed for acquiring knowledge about thermostable proteins. The overall schema has been illustrated in Fig. 2.1.

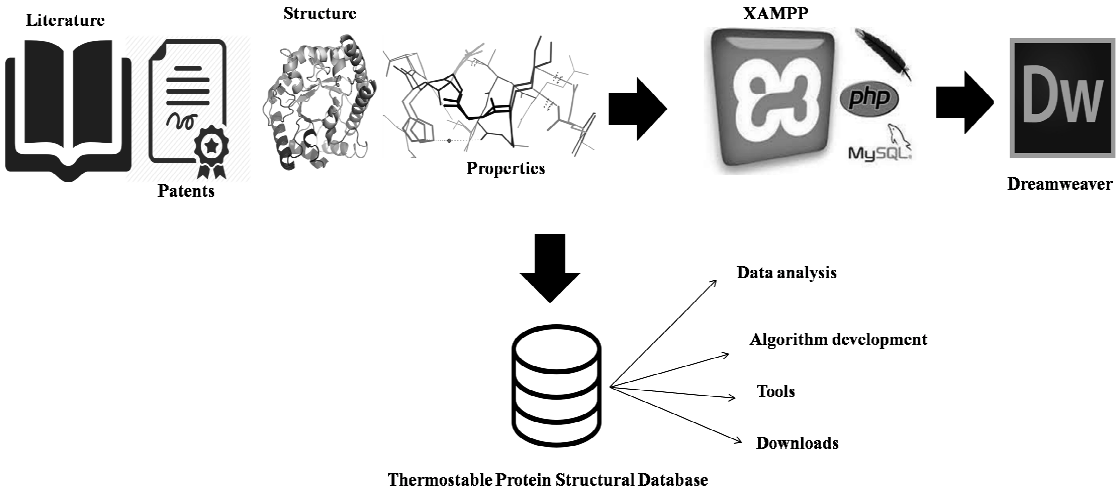


Fig. 2.1. The schema for thermostable protein database development.

2.2. Methodology

2.2.1. Data collection, database architecture and integration

Sequence and structures of all the available thermostable proteins from UniProt KB and the Protein Data Bank (PDB) with the key words search: "thermostable", "thermophilic" and "hyperthermophilic" were collected. As the main motive of this research was to correlate sequence and structural features to protein thermostability, sequences that do not have their crystallized structures in PDB were excluded from this study. Mutant structures that have been engineered with an increase or decrease in thermostability 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. Information regarding their temperature stability, source organism, the optimal growth temperature of the source organisms, phylogenetic classification, enzyme classification number (E.C No.), literature publications and patents were also collected from literature. Finally with all these information in hand, a thermostable protein structural database was created using MySQL, APACHE and PHP platforms. A web interface was created using Dreamweaver.

Servers and softwares used

Dreamweaver and XAMPP, PROTPARAM, COPID, MEME suite, VADAR, Intra Protein calculator (IPI) program (python) and PROMOTIF.

2.2.2. Data analysis

Classification

The source organisms highlighted the distribution of the thermostable proteins into different kingdoms of life. The distribution of the proteins in accordance to their temperature stability and enzyme class was also studied.

Amino acid composition analysis of all thermostable proteins

The FASTA sequences of the thermostable proteins were collected from Uniprot KB and ProtParam tool was used to generate percentage amino acid composition. The tool can be accessed through http://web.expasy.org/protparam/. The dipeptide composition was also enumerated with the aid of COPid server (Kumar et al. 2008). 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 thermostable proteins and a correlation was drawn with previous citations regarding the same.

Structural analysis of all thermostable proteins and feature generation

The PDB structures were used for feature generation. An algorithm was written to calculate intra-protein interaction from the 3-D coordinates obtained from the PDB files. The code was able to calculate the following attributes: The Hydrophobic interactions, Hydrogen bonds (Main chain to main chain, Main chain to side chain and side chain to side chain), Disulphide Bridges, Ionic Interactions, Aromatic-Aromatic Interactions, Aromatic-Sulphur Interactions and Cation-π Interactions. Promotif and Volume Area Dihedral Angle Reporter (VADAR) (Willard et al. 2003) were integrated with the aforementioned code to get the details about the 3-D structure and packaging details of the protein. Pre-decided cut-offs for the different interactions have been listed in Table 2.1. These cut-offs were obtained from literature.

Table 2.1. Intra-protein interactions important for protein stability and criteria for their calculation

Features	Qualifying Criteria	Importance in stability	Reference
Hydrophobic interactions	ALA, VAL, LEU, ILE, MET, PHE, TRP, PRO, TYR at a distance of 5 Å	Hydrophobic effect is the dominant driving force in protein folding	Kyte and Doolittle, 1982
Disulphide Bridges	Pairs of CYS within 2.2 Å	Covalent bond increases rigidity of protein	Darby et al. 1997
Hydrogen Bonds	Donor-acceptor distance cutoff (oxygen and nitrogen) is 3.50	Increased electrostatic strength	Overington et al 1990
Ionic Interactions	Ionic residue pairs falling with 6Å	Increased electrostatic strength	Vogt et al. 1997
Aromatic-Aromatic Interactions	Pairs of phenyl ring centroid that are eparated by a preferential distance of between 4.5 to 7 Å account for aromatic interactions	A pair of aromatic interaction contributes between –0.6 and –1.3 kcal/mol to the protein stability (Serrano et al. 1991)	Burley et al 1985
Aromatic-Sulphur Interactions	Interactions between the sulphur atoms of cysteine and methionine and the aromatic rings of phenylalanine, tyrosine and tryptophan within 5.3 Å	Play an important role in protein folding and stabilization	Reid et al. 1985
Cation-Pi Interactions	When a cationic side chain is near an aromatic side chain within 6 Å separation they account for cation-π interactions	Play an important role in protein folding and stabilization	Satyapriya et al. 2004

The tool has been integrated in the database and can be accessed to generate the aforementioned features of protein structures. The total dataset along with their features have been integrated in the database and thus can be accessed online.

All the features that were generated were normalized into their percentage scores according to eq 1.

$$\Phi_{\nu} = \frac{\Phi_{\alpha}}{N} \times 100 \qquad (1)$$

In eq. 1 Φ_{ν} stands for the normalized feature, Φ_{α} is the numerical attribute of the feature and N is the total number of atoms in a protein structure/total number of amino acid residues which form the protein.

2.2.3. Refinement of collected data and generated features

Out of the total proteins in our database, protein structures with a resolution greater 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.

Amino acid composition analysis of the refined dataset

To understand the role of amino acids in thermostability of proteins, comparison of amino acids composition between datasets of 127 pairs of thermostable and mesostable proteins were carried out. The percentage composition of amino acid residues were generated using the Expasy ProtParam tool.

Motif Discovery in refined data of thermostable and mesostable pairs

Multiple Em for motif elicitation (MEME) suite was employed to discover motifs in thermostable protein sequences. MEME discovers novel, ungapped motifs (recurring, fixed-length patterns) in the provided protein sequences (Timothy et al. 2009). To find motifs enriched in thermostable sequences, MEME was used in the discriminative mode i,e., the mesostable sequences were used as a negative control dataset. In this mode position-specific prior is calculated from the two sets of sequences followed by motif search. This approach has been reported to be based on the simple discriminative prior described by Narlikar et al. (2007). Position-specific prior assigns a probability, when a motif starts at each possible location in the sequence data. In the discriminative mode, words in the primary dataset occur frequently, but the words that are infrequent in the negative dataset are up-weighted.

Refinement of features and intra-protein interaction analysis of the refined dataset

After data refinement, it was necessary to analyze the difference in structural features between thermostable and mesostable protein dataset. Therefore features were weighted. Pearson correlation was calculated and features highly correlated with Pearson correlation greater than 0.9 were not considered for further analysis.

2.3. Results and Discussion

2.3.1. Data collection, database architecture and integration

Data collection

A total of 1280 proteins were collected from the PDB with the text search "thermostable", "thermophile" and "hyperthermophile". After removing redundancy, and separating mutated or engineered proteins, 378 proteins remained. Proteins having stability at >45^{o}C were only retained in the database as thermostability by definition relates to those proteins which have a temperature optimum >45^{o}C (Trivedi et al. 2006).

A separate dataset for 261 mutated proteins along with their 13 wild type counterparts were created. The datasets was manually curated to collect information regarding their temperature stability. Information regarding their enzyme classification, source organism, structural resolution, and literature citations were also accumulated. These datasets can be searched from the dynamic web interface that was developed using HTML, PHP and CSS to comprehensively search the required data. The database was named Thermostable Protein Structural Database (TPSD).

Database architecture

The thermostable protein structural database (TPSD) was built on an entity relationship model. The schema has been presented in Fig. 2.2. It can be accessed

through www.extreme-stabledb.in. The database works well with commonly available web browsers, such as Mozilla Firefox, Google Chrome and Microsoft Internet Explorer. Fig 2.3. is a schematic representation of the web interface and few snapshots of the HTML look of the database.

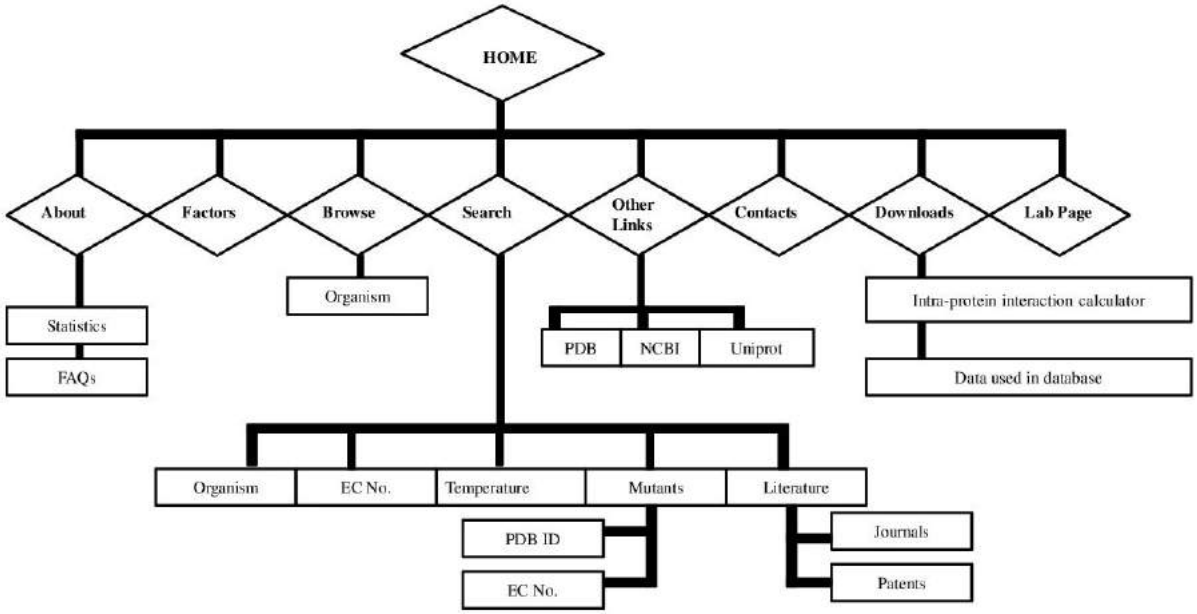


Fig. 2.2. Schematic of the web interface of thermostable protein structural database**.**

This is a curated database on thermostable proteins. It hosts literature and patent information available regarding protein thermostability. Further, it has a user friendly search and browse interface where search can be performed by user entered organism name, E.C No., temperature. Mutants can be searched by PDB ID and E.C No. Literature and patent information can also be browsed. It also has information about engineered proteins for thermostability. Third party databases like NCBI genome, PDB and UniProt KB can be directly searched from the database interface. All the PDB entries and their citations have been given direct link to the PDB and Pubmed sites. Furthermore, it harbors user friendly download links to all the datasets used to build this database. Such data can be useful to other researches about thermostable proteins. Additionally as thermostability have been researched upon by

comparison to mesostable proteins, a more refined dataset of thermostable proteins along with their homologous mesostable counterparts is also available for download. For the first time intra-protein interaction data have been integrated and can be searched for each individual entry in the database. A python tool to calculate percentage of 9 intra-protein interactions is freely available for download for academic purposes. The details about the tool are available in the forthcoming sections of this chapter.The database has also links to other webservers and tools required to analyze protein stability and perform protein engineering.

The database has been further utilized to better understand protein thermostability through classification, analysis of amino acid and dipeptide composition and intra-protein interactions.

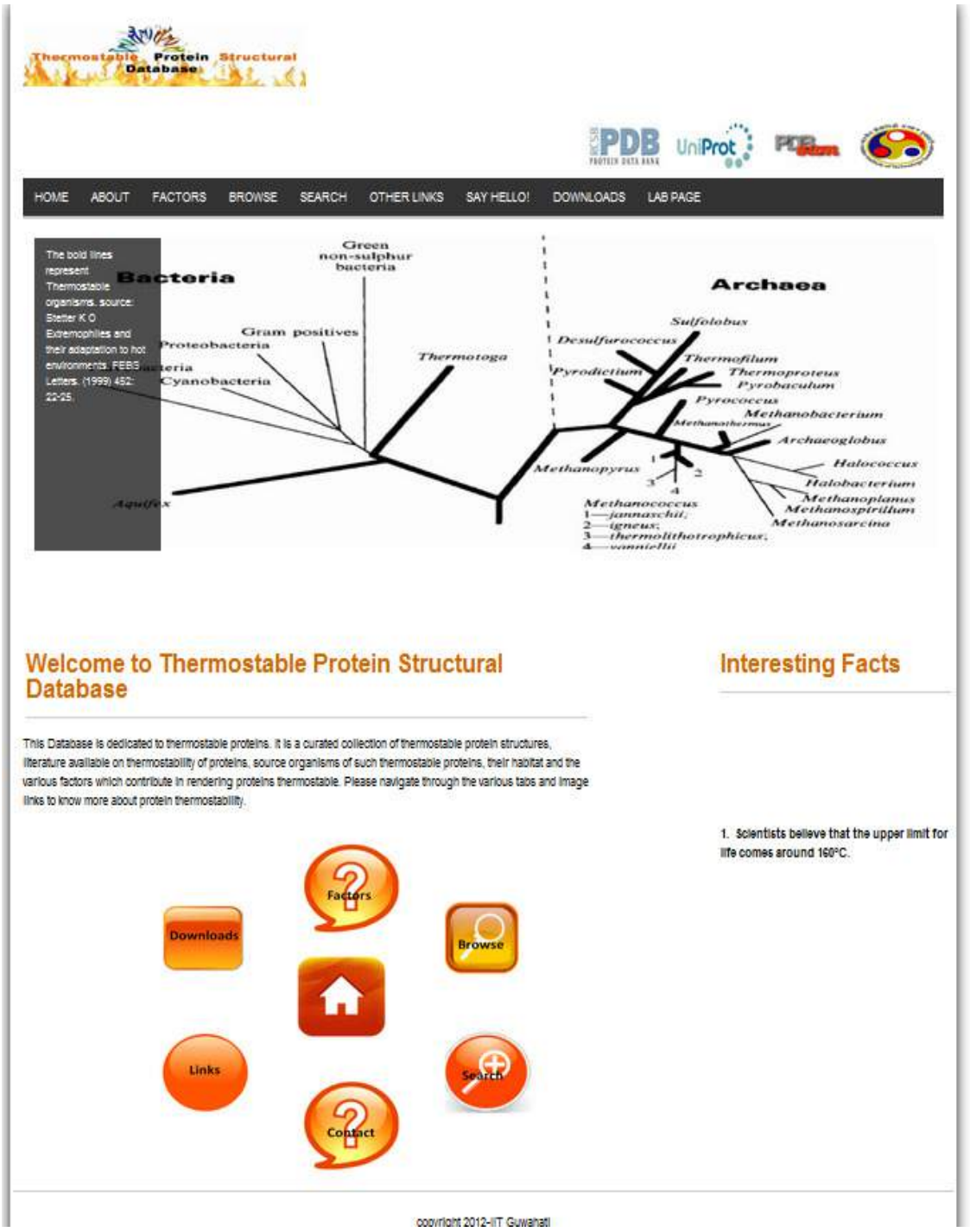


Fig. 2.3. Database snapshot of Thermostable Protein Structural Database

2.3.2. Data analysis

Classification

To bring out comprehensive information about the 378 thermostable proteins in the database, the proteins were grouped into various categories like phylogeny, enzyme class and temperature groups and pi-charts were constructed. Fig 2.4 is an illustration of the same. The enzyme class needs special mention as structural proteins were grouped as others. Many interesting observations were drawn. Most of the thermostable protein structures belong to bacteria (58%) followed by archaea and few are from eukaryotes. It was interesting to note that viruses too possess thermostable proteins. An example is an archaeal virus capsid protein (PDB Id: 2BBD) from *Sulfolobus* turreted icosahedral virus. The virus was isolated from an acidic hot spring (pH 2-4, 72-92^{o}C) in Yellowstone National Park (Khayat et al. 2005). Furthermore, 30% of the thermostable proteins having crystal structures belonging to Hydrolase and majority of the proteins have temperature stability range of 70-80^{o}C. The proteins are representation of 132 thermostable organisms in the database. The organisms having the highest optimum growth temperature of 100^{o}C is *Pyrobaculum aerophilum* and the most thermostable protein in the database is thermostable alcohol dehydrogenase from *Pyrobaculum aerophilum* having a melting temperature of 96.9^{o}C.

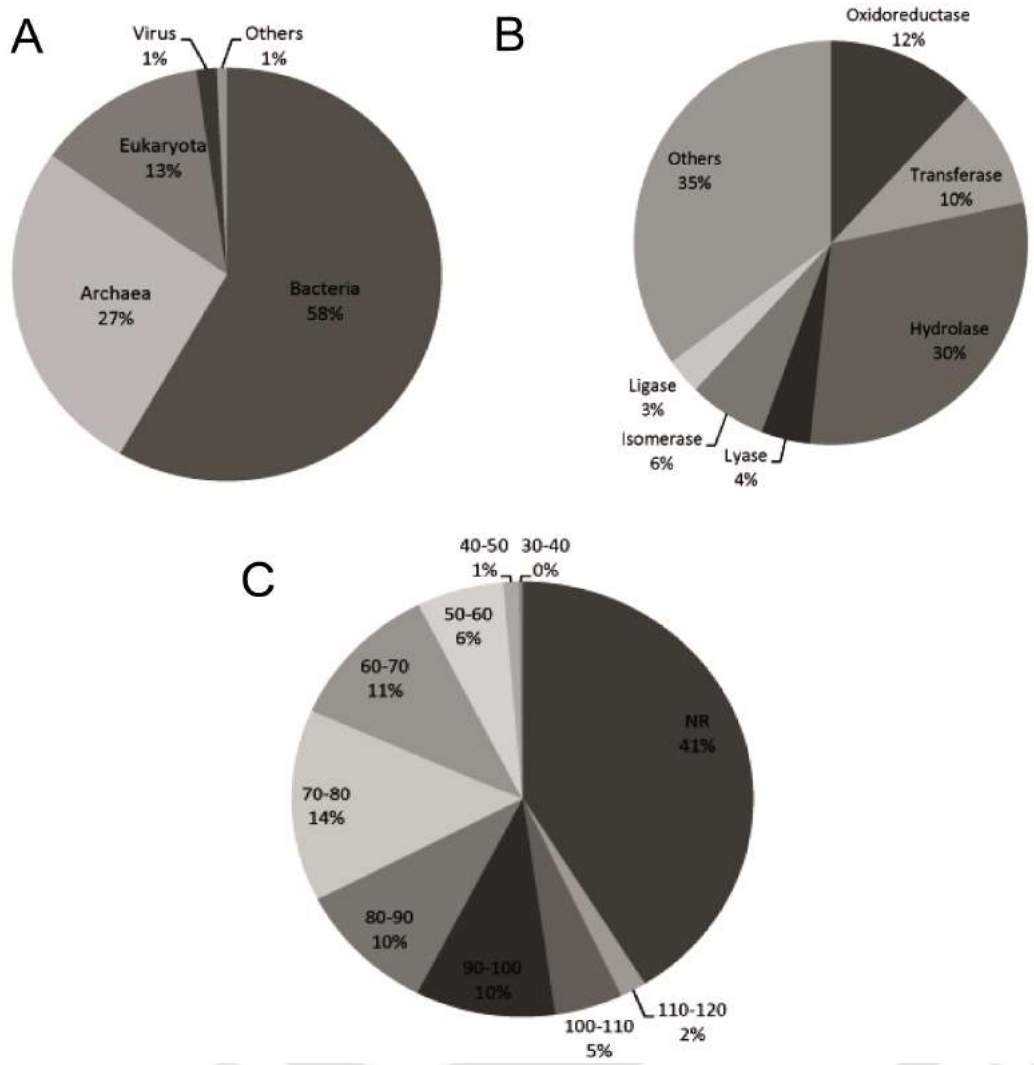


Fig. 2.4. A) Pi-charts illustrating classification of thermostable proteins in the kingdoms of classification. B) Classification according to the enzyme class they belong to. C) Classification according to the temperature stability group they belong to. NR denotes not reported. The numbers represent percentage of thermostable proteins. 'Others' in B represent structural proteins and unclassified proteins.

Amino acid composition analysis of all thermostable proteins

Amino acid composition analysis of the 378 thermostable proteins led to the conclusion that percentage of charged and non-polar amino acids were higher than

that of the polar residues (Fig. 2.5). Among charged residues and non polar residues were with the highest percentages. Most of the earlier work on thermostable proteins mentioned that charged residues were important for protein stability as they interact to form ionic interactions (Kumar et al. 2000; Trivedi et al. 2006). On similar lines non-polar residues were observed to result in hydrophobic interaction, filling voids and cavities and resulting in better core packing thus stabilizing proteins (Castagnoli et al 1994; Reed et al. 2013). Though Arg had been proposed to replace Lys in thermostable proteins based on its ability to maintain charge and provide an additional hydrogen bond (Folcarelli et al. 1996), Lys residues were found to have higher average than Arg residues.

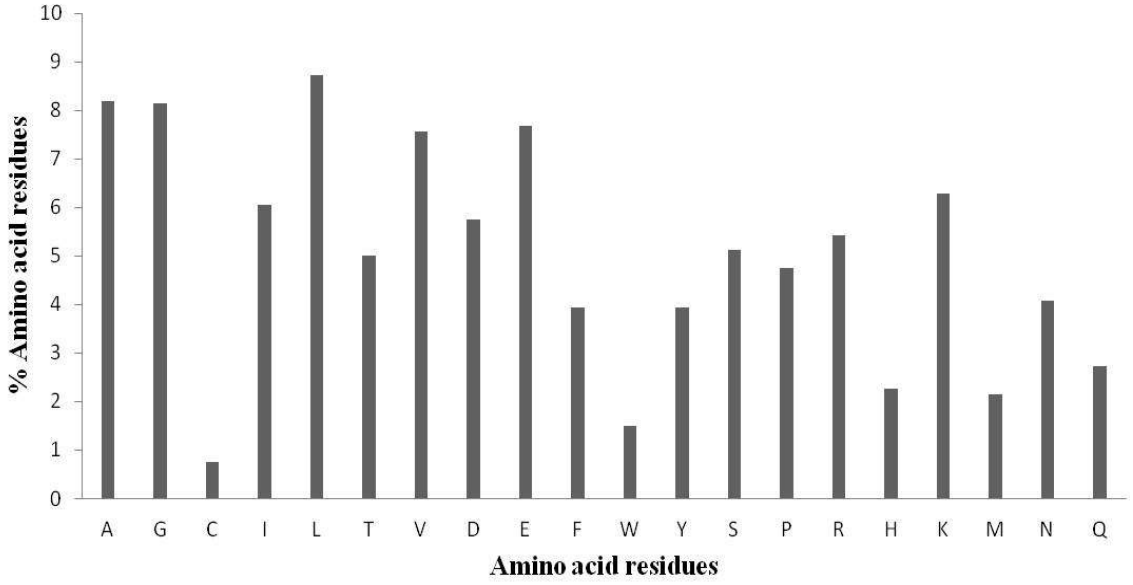


Fig. 2.5. Amino acid composition of the 378 thermostable protein data available in the database.

Here it can be pointed out that to further understand the role of amino acids in protein thermostability a more comparable dataset of thermostable proteins along with their mesostable counterparts was required. Therefore, such a dataset was created and analyzed in the forthcoming sections of this chapter.

Structural analysis of all thermostable proteins and feature generation

A total of 25 protein structural secondary and tertiary features and intra-protein interactions reported until date to contribute towards thermostability have been considered. The features collected were percentage of residues forming hydrogen bonds, 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-π interactions, aromatic-aromatic interactions, salt bridge, packing volume, beta turns, beta hairpins, percentage of alpha helices, percentage of beta strands, percentage of beta sheets, percentage of loops, oligomerization state, disulphide bonds, total gamma turns, inverse gamma turns, fraction non polar Accessible surface area (ASA), fraction polar ASA, fraction charged ASA and metal coordination.

A python program was written for calculating percentage of protein features (eq 1) and the calculated features. The salient feature of the script is that it can perform batch processing of n number of protein structures. This will aid researchers in saving time for feature generation of protein structures. Furthermore, Volume Area Dihedral Angle Reporter (VADAR), PROMOTIF programs were employed for other feature generation. VADAR uses 5 different algorithms and programs for analyzing protein structures from their PDB coordinate data (Willard et al. 2003). PROMOTIF provides details about the 3D structural features such as secondary structures of a protein (Hutchinson et al. 1996).

These information have been integrated in the database and can be accessed through the search and browse options. This is a new feature exclusive to this database and none other database hosts this feature.

2.3.3. Refinement of collected data and generated features

Out of the total 378 proteins in our database, protein structures with a resolution greater than 2.5 Å were removed. The mesostable 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. After data cleansing we were left with 127 non-redundant thermophilic-mesophilic protein homologous structural pairs which formed our final dataset. The dataset has been presented in Table A1 in Appendix I. From the phylogenetic distribution of the final dataset of thermostable proteins, it was observed that the dataset consists of 3.1% eukaryotic proteins. Phylogeny was not kept as a predetermining factor for choosing the structural homologues. This is based on the fact that mechanism of protein structural stability is independent of phylogenetic diversity. Moreover some archaeal thermophilic proteins have higher homology to eukaryotic proteins than those from prokaryotes (Shin et al. 2014). Also in literature several examples are available where prokaryotic and eukaryotic homologous structural pairs have been compared (Vogt et al. 1996; Kumar et al. 2000; Sadeghi et al. 2006; Yokata et al. 2006).

Amino acid composition analysis of the refined data

The percentage of amino acid residues for thermostable and mesostable datasets were generated using ExPasy ProtParam tool. Fig. 2.6 is an illustration of the same. From Fig. 2.6 it is clear that charged and polar residues were much higher with a comparable difference, in thermostable proteins than other types of residues. Charged residues have been reported by many researchers to occur more frequently in thermostable proteins (Russel et al. 1995; Haney et al. 1996; Trivedi et al. 2006). It was interesting to note that among negatively charged residues the frequency of Glu was higher but Asp was lower than mesostable proteins. The rationale behind a lower percentage of Asp residues can be due to the liability of Asp-Pro bonds for peptide hydrolysis (Volkin et al. 1992). On the other side, positively charged Lys and Arg were higher in thermostable proteins. Thermolabile polar residues (Asn and Gln)

were lower in thermostable proteins. This is because Asn and Gln residues are prone to deamidation resulting in protein destabilization (Wright, 1991). Though Ala has been reported to have high helix propensity (Panja et al. 2015), its percentage was observed to be lower in thermostable proteins. Earlier studies on amino acid composition did not consider a dataset covering all the enzyme classes and structural protein pairs. Moreover though Ala can stabilize helices, they tend to destabilize hydrophobic cores in protein structures (Zhou et al. 1994). Non-polar residues like Val and Ile were found to me more in thermostable proteins. These residues are involved in hydrophobic interaction and core packing thus can enhance stability of proteins (Viellie et al. 2001; Panja et al. 2015).

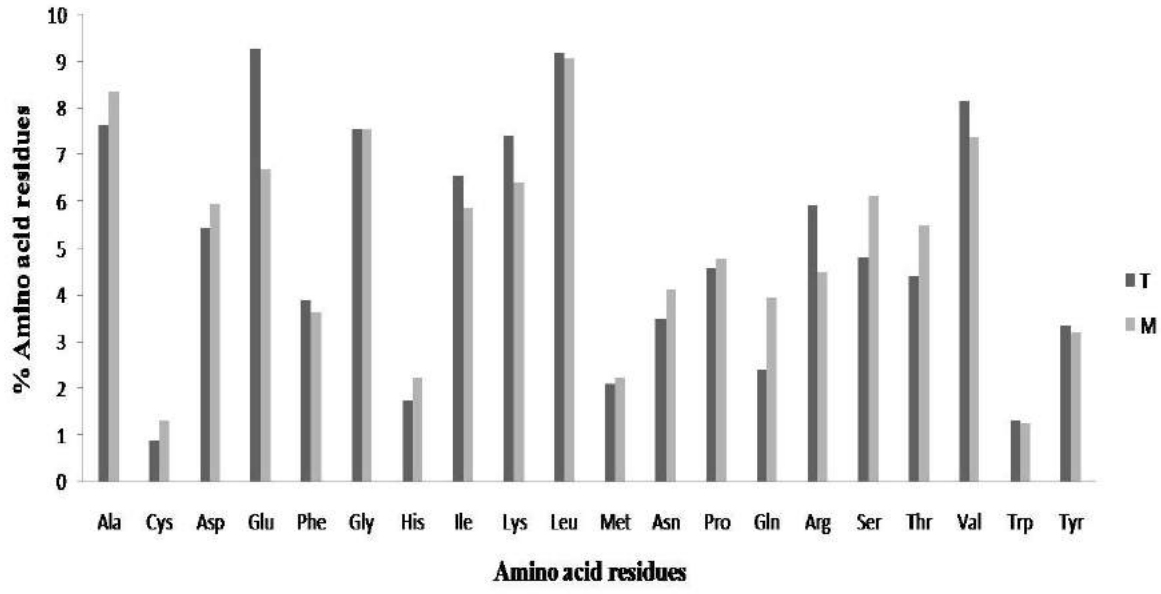


Fig.2.6. Amino acid composition analysis of thermostable and mesostable proteins. T denotes Thermostable proteins and M denotes mesostable proteins.

Dipeptide compositions were analyzed separately for the thermostable and mesostable protein datasets and a comparison was drawn. Composition >0.5 % were considered. The composition have been tabulated in Table A2-A3 in Appendix I. For the thermostable dataset the highest percentage of dipeptide was for EL (0.86%).

The others were AA, EA,KA, VA, LD, AE, GE, AG, KG, LG, EI, KI, VI, AK, EK, GK, IK, LK, VK, AL, EL, GL, IL, KL, RL, VL, AR, KR, AV, EV, GV, LV, NV, VV. This result corroborates previous results obtained by Ding et al. (2004). The highest percentage of dipeptide for the mesostable dataset was for AA (0.88%). The others were EA, GA, TA, VA, AE, LD, LE, LG, AL, DL, EL, GL, IL, KL, LL, LK, AT, AV, DV, LV, TV. Thus, it can be observed that combination of charged and non polar dipeptides were more in thermostable proteins.

Motif Discovery in refined data of thermostable and mesostable pairs

Sequence analysis to find out thermostability motif was performed using MEME suite using the homologous mesostable sequences as a negative control. MEME represents motifs and describes the probability of each possible letter at each position in the pattern (Timothy et al. 1994). A total of 5 motifs were discovered. These motifs were utilized for MAST search on a dataset of 2673 thermostable sequences collected from Uniprot. Motif alignment search tool (MAST) search with MEME motifs could identify only 72 proteins with the same motifs. The motifs were unsuccessful to identify all thermostable proteins in the refined dataset of 127 thermostable and mesostable proteins. Moreover the bigger disappointment was when they picked up 32 mesostable sequences. The generated motifs thus failed to detect all thermostable proteins as thermostable. Thus only structural features were considered for further studies implemented in the ensuing chapters of this dissertation work. It is also generally agreed that direct prediction of protein characteristics from the primary amino acid sequence is not possible as protein function is attributed to its tertiary structure (Ebrahimi et al. 2011).

Refinement of features and intra-protein interaction analysis of the refined dataset

Out of the 25 features few were discarded because they were highly correlated with Pearson correlation greater than 0.9. Finally we were left with a final set of 17 features. Table 2.2 presents the features that were chosen for analysis and the method

by which they were calculated for the raw and final dataset (127 pairs of homologous thermostable and mesostable proteins). All the features were the count of the number of intramolecular interactions in the proteins except the surface areas which were represented as the fractions out of the total accessible surface area of the proteins.

Table 2.3. The 17 statistically significant features responsible for protein thermostability

Sl. No	Number of Interactions	Abbreviation	Software/ Tools
1	Percentage of residues forming Hydrogen bonds	HB	VADAR
2	Main chain to main chain hydrogen bonds	MMH	Python script
3	Main chain to side chain hydrogen bonds	MSH	Python script
4	Side chain to side chain hydrogen bonds	SSH	Python script
5	Hydrophobic interactions	HI	Python script
6	Ionic interactions	II	Python script
7	Aromatic-sulphur interaction	ASI	Python script
8	Cation-pi interactions	CPI	Python script
9	Aromatic-aromatic interactions	AAI	Python script
10	Salt bridge	SB	Python script
11	Packing volume	PV	VADAR
12	Beta turns	BT	PROMOTIF
13	Total gamma turns	GT	PROMOTIF
14	Inverse gamma turns	IGT	PROMOTIF
15	Fraction non polar Accessible surface area (ASA)	NASA	VADAR
16	Fraction polar ASA	PASA	VADAR
17	Fraction charged ASA	CASA	VADAR

Calculation of intra-protein interaction in 127 pairs of thermostable and mesostable proteins brought forward some interesting observations. Fig. 2.7 is a graphical illustration of the contributions of intra-protein interactions in thermophilic

versus mesophilic proteins. The percentage mean of gamma turns, inverse gamma turns, cation-π interaction, ionic interactions, hydrophobic interactions, charged accessible surface area, main chain-main chain hydrogen bonds and aromatic interactions were >50% in thermostable proteins, out of the statistically significant 17 features related to protein stability analyzed. Such factors have been previously mentioned to enhance protein thermostability (Vieille et al. 2001; Kumar et al. 2000; Sadeghi et al. 2006). A protein tertiary structural feature; gamma turns for the first time was found to play important role in thermostabilizing proteins. A γ- 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 $(i)^{th}$ residue and NH group of $(i + 2)^{th}$ residue (Rose et al. 1985). Gamma turns have been classified into classic and inverse based on the dihedral angle values of the $(i + 1)^{th}$ residue (Rose *et al* 1985). The classic gamma turn gives rise to 180^{o}chain-reversal in proteins and is often observed at loop end of β-hairpins (Milner-White et al. 1986). The inverse γ- turns include a large proportion of weak hydrogen bonds according to the definition of hydrogen bonds (Kabsch and Sander 1983). It can be said here that tertiary structure which leads to an increase in hydrogen bonds can increase with the increment in protein thermal stability.

Another interesting observation was that though polar residues were found to be higher when comparing it with other amino acid residues in only thermostable proteins, such residues were less in thermostable proteins in comparison to their mesostable counterparts. This shows the importance of comparing thermostable and mesostable datasets in order to draw concrete conclusions about protein thermostability.

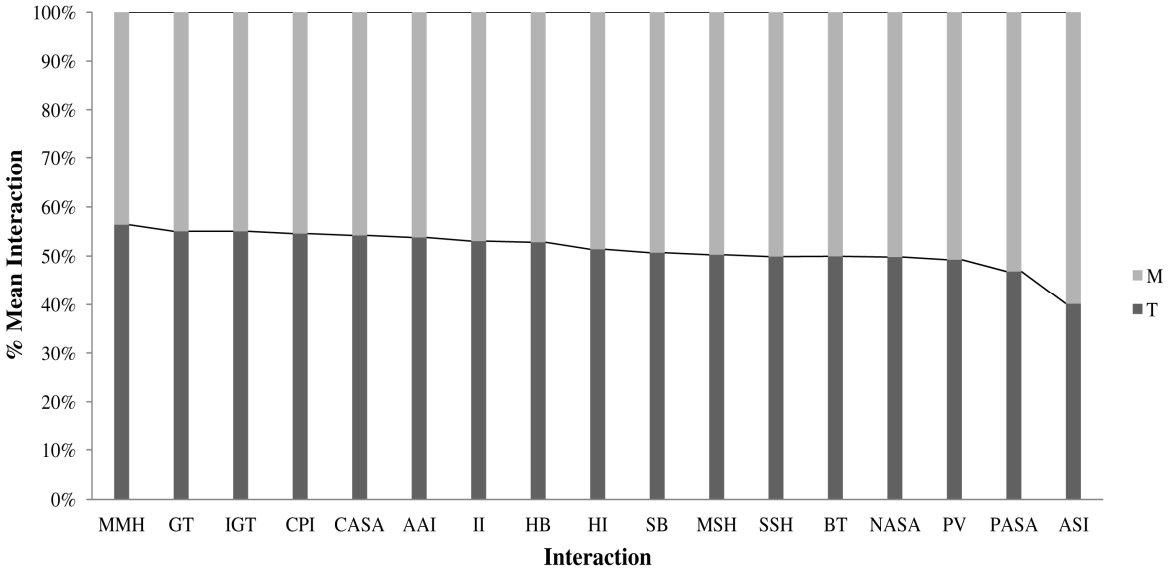


Fig. 2.7. Contribution of intra-protein interaction features towards protein thermostability in comparison to mesostable proteins.

Though features were found to be important contributors of thermostability, the elucidation of exact importance of each feature contributing towards thermostability was still fuzzy. It has been reported that thermostability of a protein is governed by multitudes of factors which are intrinsic and specific for individual protein (Chakrabarty and Varadarajan 2000; Kumar et al. 2001). For example the ornithine carbamoyl transferase from *Pyrococcus furiosus* is stabilized by hydrophobic interactions (Villeret et al. 1998), whereas glutamate dehydrogenase from the same organism is stabilized by electrostatic interactions (Karshinkoff and Landenstein 2001). Through this preliminary analysis of the datasets, thermostability stood out to be a multi criteria decision making problem. This necessitated the requirement of an algorithm which can rank these features in accordance to their importance towards contributing to protein thermal stability.

2.4. Conclusions

A curated database of thermostable proteins was successfully built on entity relationship model. The web interphase for the same can be accessed through www.extreme-stabledb.in. The salient features are that it is the only database on thermostable proteins created till date. The database hosts a total of 378 thermostable proteins from 132 thermophilic and thermophilic organisms and 261 mutants. Alcohol dehydrogenase from *Pyrobaculum aerophilum* is the most thermostable protein available in the database. The highest optimum growth temperature was observed for *Pyrobaculum aerophilum.* Around 14% of proteins have temperature optimum of 70-80^{o}C. Most of the thermostable protein structures available were found to belong to the hydrolase class.

The database also has information regarding the percentage of intra-protein interaction for each protein structure. A more refined dataset of 127 pairs of thermostable and mesostable proteins are available for download. This may assist further studies in protein thermostability. An intra-protein interaction calculator was developed in Python platform and is freely available for download.

Creation of the database and data analysis brought about many interesting observations. Amino acid composition analysis revealed that charged and hydrophobic residues were higher in thermostable proteins. Thermolabile residues (Asn, Gln) occurred less frequently and interestingly a lower percentage of Ala residues were enumerated. Furthermore, combination of charged and non-polar amino acid residue dipeptides (AA, EA,KA, VA, LD, AE, GE, AG, KG, LG, EI, KI, VI, AK, EK, GK, IK, LK, VK, AL, EL, GL, IL, KL, RL, VL, AR, KR, AV, EV, GV, LV, NV, VV), were observed to occur more frequently in thermostable proteins in comparison to their mesostable counterparts. This observation corroborates previous reports which said that charged and polar residues were important in enhancing protein thermostability.

Ionic, hydrophobic, aromatic and charged interactions were found to favour thermostable proteins. This is also the first report of γ-turns to be involved in stabilizing proteins at elevated temperatures. Conclusively it can be said that

thermostability is found to be a multi criteria decision making problem. The factors have been further analyzed and in the ensuing chapters of this thesis.

CHAPTER III

In silico Characterization of Thermostable Lipases: A Case Study from Database

Prologue

Thorough literature survey as presented in chapter II uncovered that majority of thermostable proteins belong to Hydrolase, populated by lipases. They are of high priority for industrial applications as they are endowed with the capability of carrying out diversified reactions at elevated temperatures. Extremophiles are their potential source. Sequence and structure annotation of thermostable lipases can elucidate evolution of lipases from their mesophilic counterparts with enhanced thermostability hence better industrial potential. In this chapter, characterization of thermostable lipases has been carried out both at sequence and structure levels. Present study shows that thermostabilizing features of lipases show overlaps to those reported about all other classes of proteins. Additionally, some new motifs were elucidated about a smaller fraction of thermostable lipases. Attempt was also made to study whether any additional factor(s) which have not yet been reported, can lead to thermostabilization of lipases.

Output:

1. Debamitra C, Saravanan P, Dubey VK and Sanjukta Patra. *In silico* Characterization of Thermostable Lipases. Extremophiles (2010) 15, 89-103.

3.1. Introduction

Literature survey reveals that majority of the thermostable proteins characterized till date belongs to the hydrolase class populated by lipases. The industrial potential of lipases intrigued the idea that further characterization of such enzymes would be instrumental. Therefore this chapter outlines the in depth characterization of thermostable lipases.

Lipases are known as glycerol ester or serine hydrolases and true lipases (E.C. No. 3.1.1.3) as triacylglycerol ester hydrolases and are widely distributed among the five kingdoms of life. The catalytic activity of lipases is hydrolysis of triacylglycerols to free fatty acids, diacylglycerols, monoacylglycerol and glycerol. Structurally lipases have a canonical alpha/beta hydrolase fold (Scharg et al. 1997). The active site is composed of the catalytic triad; Aspartate or Glutamate, Serine and Histidine residues. The active site Ser lies in the nucleophilic catalytic elbow and sequentially within the conserved pentapeptide GXSXG motif.

Lipases especially thermostable lipases from microbial origin are of high commercial interest mainly for biotechnological applications as they are stable in organic solvents, and do not involve cofactors, have broad substrate specificity and are highly enantioselective (Lee et al. 1999). Most importantly they are cost effective in production in industries such as food processing, detergent, organic synthesis, biopolymer synthesis, biodiesel production, pulp and oleochemical industries (Jaeger et al. 1998; Haki et al. 2003; Li et al. 2005). Two well-known commercially available thermostable lipases are from *Candida antarctica* (Novo Nordisk; Denmark, and Boehringer Mannheim) and *Burkholderia cepacia* (Amano, Fluka, and Boehringer Mannheim) application being organic synthesis and in detergent industries (Gunasekaran et al. 2004).

Thus, to satisfy the global requirement of the industrial process we should have more thermostable lipases on the shelf. In general lipases that are stable above 40^{o}C are required for industrial applications (Wang et al. 1995). Thermophilic microorganisms are found to be potential and alternative source of thermostable lipases (Brock, 1985). However, it is often impractical to use them directly due to low

yield of lipase. Alternatively, a direct approach could be screening of microorganisms for thermostable lipases and their sequence and structure analysis with comparison to that of mesostable counterparts with high expression level of lipases as this can pave the path for *in vitro* evolution from the mesophilic organisms with enhanced thermostability.

Significant research had been carried out in determination of thermostabilization factors for protein (Argos et al. 1979; Haney et al. 1997; Russell et al. 1997; Bogin et al. 1998; Grimsley et al. 1999; Chakravarty et al. 2002; Fukuchi et al. 2003; Yano et al. 2003; Szilagyi et al. 2000; Gromiha et al. 2001; Tyndall et al. 2002; Saraboji et al. 2005; Sadegi et al. 2006; Trivedi et al. 2006). In Chapter I the details about the factors responsible for protein thermostability have been presented. However, only little work has been carried out on determination of factors responsible for thermostability in sequence and structure level of lipases. Comparing thermostable *Bacillus stearothermophilus* P1 and mesostable *Chromobacterium viscosum* lipases it was reported that ligand stabilization, salt bridges, aromatic clusters and Pro residues are responsible for its thermostability (Tyndall et al. 2002). It was concluded that combinations of factors contribute to the thermostability of lipases as in other thermostable proteins.

In this chapter *in silico* study for unraveling additional factors leading to better thermostabilization of lipases has been outlined. The chapter correlates all the thermostabilization factors at sequence and structure level, laying special emphasis on the role of tertiary structure and distribution of amino acids in thermostabilization of lipases and to bring out possible ways that will assist in protein engineering of mesostable lipases to render them thermostable. The theme of this chapter is illustrated by Fig 3.1.

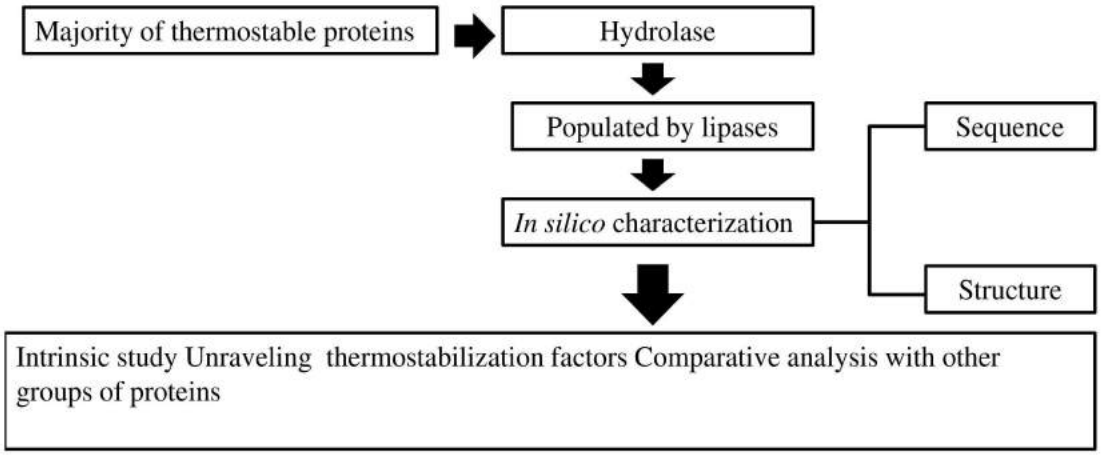


Fig.3.1. Schematic representation of the aim to study thermostability of lipases.

3.2. Methodology

3.2.1. Sequence collection and characterization

To draw a concrete conclusion for the factors contributing to thermostability of lipases twenty three available thermostable lipase sequences were collected from UNIPROT Knowledgebase release 15.12. Respective families were assigned to the lipases through BLAST search in Lipase Engineering Database (LED) (Jurgen et al. 2000) (Table 3.1). Seven well characterized mesostable true lipase sequences that shared homology with the thermostable lipases were chosen for comparison with their thermostable counterparts. The lipases that were chosen were from the following organisms: *Rhizopus oryzae* (ROL; P61872; 1TIC; abH23.01), *Pseudomonas sp. B-11* (PLB; O52270, abH04.01), *Bacillus sphaericus 205y* (BSL205y; Q8VQP2, abH04.04), *Staphylococcus epidermis* (SEL; Q9Z4M7; abH15.01), *Rhizomucor meihei* (RML; P19515; 4TGL; abH23.01), *Bacillus pumilus* (BPL; B8Y3H3; abH18.01), *Chromobacterium viscosum* (CVL; Q05489; 1CVL; abH15.02) and *Bacillus subtilis* (BSL; P37957; 1I6W; abH18.01).

Table 3.1. List of organisms producing thermostable lipases with their properties

Sl No.	Source	Lipases/ (abbreviations)	Optimum range ^{o}C	Family	Reference
1	Bacteria	*Bacillus stearothermophilus* P1 (BSLP)	30-65	abH15.1	Joel et al. 2002
2	Bacteria	*Bacillus thermocatenulatus* (BTL)	60–80	abH15.1	Quyen et al. 2002
3	Bacteria	*Geobacillus sp.* TW1 (GLTW1)	60- 70	abH15.1	Lee et al. 2005
4	Bacteria	*Geobacillus thermoleovorans* YN (GThLYN)	60-70	abH15.1	Soliman et al. 2007
5	Bacteria	*Bacillus stearothermophilus* L1 (BSLL)	60–65	abH15.1	Ahn et al. 2004
6	Bacteria	*Bacillus licheniformis* (BLL)	55-70	abH18.1	Horani 2004
7	Bacteria	*Bacillus strain* 42 (BL42)	70-80	abH15.1	Etaweel et al. 2005
8	Bacteria	*Bacillus thermoleovorans lipase* ID-1 (BThLID)	75	abH15.1	Rathi et al. 2000
9	Bacteria	*Geobacillus thermoleovorans* IHI-91 (GThLIH1)	65	abH15.1	Soliman et al. 2007
10	Bacteria	*Geobacillus zalihae strain* T1(GZLT)	65	abH15.1	Leow et al. 2004
11	Bacteria	*Staphylococcus xylosus* (SXL)	45-60	abH15.1	Horchani et al. 2008
12	Bacteria	*Staphylococcus aureus* (SAL3)	55	abH15.1	Horchani et al. 2008
13	Fungi	*Aspergillus terreus* (ATL)	15-90	abH23.1	Yadav et al. 1998
14	Fungi	*Aspergillus niger* F044 (ANL)	45-60	abH23.1	Yadav et al. 1998
15	Bacteria	*Pseudomonas cepacia* (PCL)	60	abH15.2	Sugihara et al. 1992
16	Bacteria	*Pseudomonas sp.*KW1-56 (PLKW1)	60	abH15.2	Izumi et al. 1990
17	Fungi	*Candida antarctica* Lipase A (CALA)	>90	abH18.1	Maria et al. 2005
18	Fungi	*Thermomyces lanuginose* (TLL)	45	abH23.1	Hayashi et al. 1987
19	Fungi	*Rhizopus chinensis* (RCL)	30-50	abH23.1	Sun et al. 2009
20	Bacteria	*Pseudomonas fluorescens* SIK W1 (PFLSIK)	60	abH24.1	Chung et al. 1991
21	Fungi	*Candida Antarctica* Lipase B (CALB)	30- 40	abH37.01	Suen et al. 2004
22	Bacteria	*Thermoanaerobacter thermohydrosulfuricus* (TTL)	75-90	abH27.01	Royter et al. 2005,
23	Bacteria	Thermostable *Bacillus subtilis* lipase (TBSL)	65	abH18.01	Ahmad et al.2008

3.2.2. Multiple Sequence Alignment (MSA) of thermostable and mesostable lipases

Through MSA it was intend to bring out the possible amino acid residues conserved in thermostable lipases and absent in mesostable ones which may be responsible for their enhanced thermostability. Thus, thorough analysis of thermostable and mesostable lipases was carried out by MSA using Parallel PRRN; progressive (amino acid content) with iterative refinement, with default parameters (http://align.genome.jp/prrn/) using PAM 250 matrix due to the extensive differences in the length of the lipase sequences ranging from 181 to 453 amino acid residues.

3.2.3. Study of percentage amino acid composition of thermostable and mesostable lipases

A detailed comparison of thermostable and mesostable lipases were carried out with respect to the percentage amino acid composition of the sequences as this would shed light on the quantitative estimation of the twenty different amino acid residue composition of thermostable lipases in comparison to the mesostable lipases. This was performed using the webserver COPid which is Composition Based Protein Identification (Kumar et al. 2008) and can be accessed through www.imtech.res.in/raghava/copid/index.html.

3.2.4. Structural characterization by tree based annotation of thermostable and mesostable lipases

To get a complete picture on the thermostabilizing factors of lipases, sequence analysis was inadequate, thus structural comparison was performed using the available PDB structures (8 for thermostable and 4 for mesostable lipases) by tree based classification. The other 18 lipase sequence collected, unfortunately lack crystal structures in PDB till date.

Tree based annotation was performed of the thermostable-mesostable lipase pairs for structural comparison as large differences were encountered in the length of the lipase sequences, altered alignment was observed in MSA for important lipase motifs with respect

to bacterial and fungal lipases and poor secondary structure alignment was noticed for many thermostable-mesostable lipase pairs belonging to different classes as classified by LED. Moreover, PDB structures gave huge RMSD deviations as assigned by CE Calculate (Shindyalov et al. 1998). Due to all the aforesaid it was concluded that structural comparison of very divergent structures like that of fungal and bacterial thermostable-mesostable pairs would not solve the purpose of figuring out the possible minute structural differences which can lead to thermostability. Comparison of very similar structures by classification of lipases into subfamilies would yield better results. So classifying the thermostable and mesostable lipases into individual subfamilies was tried for simplification of the task of structural annotation using SCI-PHY (Brown et al. 2007), PIRSF (Wu et al. 2004), SECATOR (Wicker et al. 2001) and CLUSS (Kelil et al. 2008). Finally, it was agreed upon to use the latest CLUSS 2 version 1.0 (IJCBDD 2008) which can be accessed through http://prospectus.usherbrooke.ca/CLUSS/Server/Index.html. The criteria to be fulfilled for classification of thermostable and mesostable lipases was possible by using only CLUSS because it utilizes nonaligned protein sequences as its input. The other classification programs utilized for the same purpose required MSA as their input for classification. This took care of the differences in length of the sequences leading to poor multiple alignments. Moreover CLUSS having high sensitivity towards very similar and divergent sequences solved our problem of classifying the thermostable and mesostable lipases which show both properties of high similarity and divergence among themselves. From the resultant subfamilies, representatives for mesostable and thermostable lipases were chosen for further comparison on the basis of their least RMSD deviation from the rest of the family members and the availability of their PDB structures.

3.2.5. Structural analysis

To reach to a consummate conclusion of the structural factors leading to thermostability of lipases and to further support obtained results, the different structural features of the representatives of each subfamily were compared by structural superimposition performed using PyMol v0.99 (Delano 2006). Adding to this qualitative

and quantitative assessment of the representative lipase structures were performed using web server VADAR (Volume, Area, Dihedral Angle Reporter) (Willard et al. 2003).

3.2.6. Study of structurally important residues of thermostable-mesostable lipases

This study was intended to find out the most structurally important residues of thermostable-mesostable lipase pairs of subfamilies, lipases having greater than 80% sequence similarity were chosen because comparison of lipase structure having less than 80% sequence similarity yielded false positive results. SRide server (Gromiha et al. 2004, Magyar et al. 2005) was employed to obtain information on structurally important residues which can be held responsible for thermostability. Additionally, HotSpot Wizard 1.4 (Pavelka et al. 2009), which represent an easy way to perform several structural and evolutionary analysis at once with minimum difficulty and can be accessed at (http://loschmidt.chemi.muni.cz/hotspotwizard/index.jsp) was employed for the same purpose. Moreover the mutational effect of the structurally important residues were checked with the CUPSAT server (Prathiban et al. 2006) which predicts protein stability changes upon point mutations.

3.3. Results and Discussion

3.3.1. Sequence characterization

Multiple sequence alignment (MSA) revealed some interesting facts about the active site, oxyanion hole and lid domain of thermostable lipases which could be contributory to temperature stability. The MSA has been presented in Appendix II (Fig. A 2.1).

Active site residues

Though it has been reported that Ala replacing Gly of the pentapeptide sequence leads to thermal stability of *Bacillus* lipases because the side chain of the Ala residue stabilizes the loop conformation by tight packing, contributing to thermostability (Jeong et al. 2002) it is reported here, the presence of the same signature sequence in mesostable *Bacillus* lipases ensuring that Ala residue alone cannot account for the thermostability of this class of lipases. It has been said that lack of Cys residues in hyperthermostable lipase from *Thermoanaerobacter thermohydrosulfuricus* may account for its thermostability (Royter et al. 2009) due to high sensitivity of free Cys residues to oxidation at elevated temperatures. However, these findings indicate that mesostable lipases also lack Cys residues, it can be concluded that lack of free Cys residues cannot be the only factor leading to thermostabilization. It is known that the electrical potentials generated by the charged titratable residues have an important role in catalysis since they can enhance substrate binding, stabilization of the transition states of substrate and efficient product release (Peterson et al. 2001) leading to protein stability. This fact is also justified by the observation that the presence of more titratable amino acid residues near the active site Ser of thermostable lipases can lead to functional stability of the active conformation of lipases at elevated temperature.

Oxyanion hole

MSA revealed that the oxyanion hole consensus for lipases that was reported as [PNTSVL]-[VIF]-[VIFL]-[VIFLM]-[VLCAISQ]-H-G (Philip 2002) whether thermostable or mesostable was highly conserved. No differences could be attributed to the sequence pattern to account for thermostability in relation to the oxyanion hole of lipases. However, all the thermostable lipases studied had the GX class of the oxyanion hole signature and unlike some mesostable *Bacillus* lipases never fell into the GGGX class. GX and GGGX class had been attributed the status of oxyanion hole class by the Lipase Engineering Database (Jurgen 2000).

The lid of lipases

The lid of lipases was given special emphasis as it is involved in interfacial activation generating a possibility that it can be target site for enhancing thermostability of lipases. It was observed through MSA that in the lid of thermostable lipases *Staphylococcus* lipases lid domain possess less positively charged residues than mesostable Staphylococcus epidermis lipase. Moreover the presence of Trp residue in the lid domain of *Aspergillus terrus, Aspergillus niger, Rhizomucor meihei* and *Staphylococcus* lipases ensures the fact that not only thermostable but mesostable lipases also possess this Trp residue unlike the Trp89 in the lid of thermostable *Thermomyces lanuginosa* lipase (Zhu et al. 2001). By this analysis Tyr224 cannot be strongly promoted to play a crucial role in thermostabilization of *Bacillus* lipases as per as a previous report (Wu et al. 2009) because it is present in *Bacillus* thermoalkalophilic, thermostable *Staphylococcus* lipases along with mesostable *Staphylococcus epidermis* lipase. Val or Ala and Pro residues were found at comparable positions in thermoalkalophilic *Bacillus* lipases like Val137 and Asn138 in thermostable *Pseudomonas cepacia* lipase (Santarossa et al. 2005). Another notable fact from MSA was that *Bacillus* thermostable lipases have poly Ala in the form of di- or tripeptide in their lid domain. The observations through sequence characterization justifies that the presence of less positively charged residues in the lid may be responsible for greater negative potential of the active site of thermostable Staphylococcus lipases leading to its stability. Even though it has been said by Zhu et al. that Trp89 in the lid *Thermomyces lanuginosa* due to its hydrophobicity and π-cation interactions contributes to free movement of the lid at elevated temperatures and thus to thermostability (Zhu et al. 2001), by this study we can say that as Trp residue is present in lid of mesostable lipases also. Hence, ruling out its sole contribution to thermostability of fungal and bacterial thermostable lipases. Another previous report stated that Tyr224 strengthens hydrophobic interaction between the helix7 and the lid helix6 and strengthens the tight packing and stability of the active site residues in *Geobacillus* sp. lipase (Wu et al. 2009). But MSA highlighted presence of Tyr224 residue near the lid domain of not only thermoalkalophilic *Bacillus* and thermostable *Staphylococcus* lipases, also in mesostable lipase also not supporting that this is crucial for rendering the lipases thermoactive. As it is well known

that hydrobhobicity and rigidity are among the two important factors which leads to stability of proteins, presence of Val or Ala and Pro at comparable positions in thermoalkalophilic *Bacillus* lipases like that reported for *Pseudomonas cepacia* lipase in the lid helix (Santarossa et al. 2005) can enhance thermoadaptivity by making the active site more hydrophobic and the lid rigid.

This result clearly supports this point as we noticed the presence of poly Ala residues in the lid helix in *Baciullus* thermoalkalophilic lipases which can lead to their thermostability to a large extent. It has been said that Ala residue is a good helix stabilizer as it leads to the formation of peptide hydrogen bonds due to its small side chain which is well accommodated in the helices (Rohl et al. 1999). Moreover stability of lid helix at elevated temperature can be critical for thermo activeness of lipases.

Ion binding

It has been observed that Ca^{2+} and Zn^{2+} are tightly bound ions in lipases. The role of Ca^{2+} in increasing thermal stability of lipases was reported as it restricts conformational flexibility of certain helices and loops (Invernizzi et al. 2009) and stabilizes the catalytic His residue through hydrogen bonding as studied in *Pseudomonas* and *Burkholderia* lipases (Kim et al. 1996). It has also been reported that Calcium ion removal causes protein unfolding and aggregation (Invernizzi et al. 2009) though the role of Ca^{2+} is not as critical in *Bacillus* lipases as in *Pseudomonas* lipases (Jeong et al. 2002). However, in this study Ca^{2+} coordinating amino acid residues were found in all thermostable and mesostable lipases thus Ca^{2+} coordination alone cannot render lipases thermostable. Similarly, Zn^{2+} coordination in lipases was considered to solely play role in structural stability, not taking part in catalysis and thus can lead to thermostability (Tyndall et al. 2002). Irrespective of this fact it was found that Zn^{2+} coordinating residues are not only present in *Bacillus* and *Staphylococcus* thermostable lipases as well as in comparable position in mesostable *Staphylococcus epidermis* lipase. Thus, it can be assumed that Zn^{2+} coordination, although playing a major role, cannot alone lead to thermoadaptivity. Interestingly, it can be pointed out here that if both Zn^{2+} and Ca^{2+} coordination is present in lipases, as in thermostable bacterial lipases, it can increase their stability at elevated temperatures due to their cooperative involvement in stabilization of lipase structures.

P-loop motif

P-loop motif is present in many thermostable enzymes like T4 Bacteriophage polynucleotide kinase (Wang et al. 2001) and *Pyrococcus horikoshii Clp1* (PhoClp1) ; a thermostable 5′-OH polynucleotide kinase active at 55°C-85°C (Jain et al.2009). Moreover a P-loop like motif with Arg to Lys replacement is observed in protein tyrosine phosphatases (Zang et al. 1998). We also witnessed the presence of such a P-loop like motif in thermoalkalophilic *Bacillus* lipases in line with the Zn^{2+} binding residues. As it was reported that Zn^{2+} binding causes thermostability of lipases (Fujii et al. 1996), this P-loop like motif can thus by our data analysis, be considered as a conserved pattern in thermoalkalophilic *Bacillus* lipases. So it can be theorized to lead to stronger coordination of the metal ion enhancing thermostability in *Bacillus* lipases.

The AXXXA and GXXXG motifs

By manual counting of the 30 thermostable and mesostable lipase sequences for the AXXXA and GXXXG motifs which are involved in thermostable protein structure stabilization (Kleiger et al. 2002)**,** it was found that frequency of occurrence of AXXXA motif is more for thermostable *Bacillus* and *Pseudomonas* lipases than their mesostable counterparts (Appendix II Table A2.1). The pronounced role of AXXXA motif in thermostabilization of proteins had been attributed for the presence of this motif in helices which leads to improvement in inter-helix interaction and stabilization of the folded state of many proteins. As stated by Kleiger et al. 2005 this motif is increased in twenty-four fully sequenced genomes; *Aquifex aeolicus* having the greatest occurrence. This motif can lead to dimerization of proteins causing better stability by strong van der Waals interaction in thermostable proteins (Kleiger et al. 2002, Leonov et al. 2005). Thus, it can be concluded that the higher frequency of this motif in thermostable lipases stabilizes the structures at elevated temperature.

3.3.2. Comparison of amino acid composition of thermostable and mesostable lipases

Differences in amino acid composition had been associated with thermostable and mesostable proteins. It was found some notable differences among thermostable and mesostable lipases on the basis of their percentage amino acid compositions (Figure 3.2). The exact role of amino acids in protein thermostability has been a long study. Studies carried out have shown the involvement of amino in thermostability of proteins and we stretch these findings being contributory to thermoadaptive properties of lipases. It had been shown that presence of more charged residues leading to thermostability of proteins as they are involved in electrostatic interactions which stabilizes the secondary structure of protein (Fukuchi et al. 2001, Silver et al. 2003). An example is Arg as it shows charge resonance of the guanidium group which gives it the possibility to form more salt bridges and its side chain can form maximum of five hydrogen bonds preferably to carbonyl oxygen of peptide bond (Feller et al. 1997). We are in complete agreement with the same as our data showed the presence of more charged residues in thermostable lipases than mesostable ones. Cys residues are known to play a dual role by both increasing thermostability when present in disulphide bridges and decreasing thermostability when available in free form as it is highly sensitive to oxidation at elevated temperature (Vieille et al. 2005). Though the data on thermostable lipases show lesser percentage of Cys residues the result becomes controversial due to the presence of conserved Cys residues in some thermostabe lipase sequence not involved in disulphide bridge formation and fungal thermostable lipases having a much higher percentage of Cys residues. Among other amino acid residues the relatively high percentage of Ala was noticeable in our data clearly shown in Fig. 3.2. Ala being a small non polar residue has been credited to be a good helix former (Shalongo 1994) as the small side chain of Ala does not shield the backbone from solvent, allowing water to interact with the peptide carbonyl groups in a polyalanine helix (Luo et al. unpublished work). Thus, the higher percentage of Ala in thermostable lipases attribute towards thermostabilization. Moreover, thermolabile amino acid residues like Asn, Gln and Met should be less in thermostable lipases as they tend to undergo oxidation and deamidation at elevated temperatures (Fukuchi et al. 2001, Russel et al. 1997). This is also

endorsed by the result which shows lesser percentage of thermolabile amino acid residues in thermostable lipases. Gly is the smallest amino acid residue and is flexible which aids in relaxation of steric hindrance of thermophilic ezymes and increases stability (Vieille et al. 1996). The high percentage of Gly observed in bacterial thermostable lipases observed accounts for this fact but lesser percentage of Gly in thermostable lipases than mesostable ones prominating its role as helix breaker. Thus, it can be said that Gly shows preference for bacterial thermostable lipases and further structural analysis showed that their frequency is much greater in loops than helices. Moreover, higher Pro percentage for thermostable lipases shows its pronounced role in enhancing thermostability as Pro is more rigid than other amino acids and reduces the entropy of the main chain polypeptide decreasing the chance of unfolding at elevated temperatures (Sælensminde et al. 2008). Greater percentage of hydrophobic; Val and Ile (Gromiha et al. 2008) and aromatic residues have been long assumed to lead to better thermostability of protein as weakly polar interaction made by the aromatic ring of residues like Phe are of an enthalpic importance compared to that of hydrogen bonding (Feller et al. 1997). However, contradicting the results of Gromiha et al. 2008, Val and Ile shows increasing trend in mesostable lipases. Data from the present study, shows greater percentage aromatic residues in thermostable lipases which is in agreement with the aforesaid.

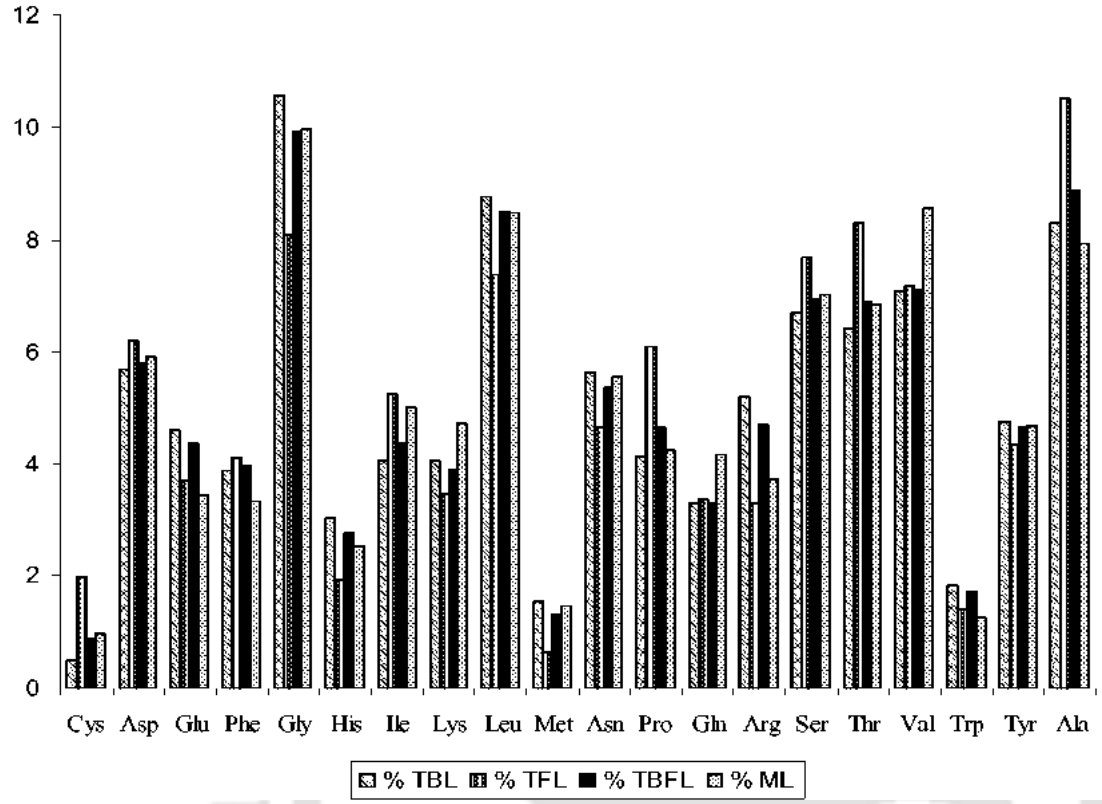


Fig 3.2. Average % amino acid composition of bacterial, fungal thermostable and mesostable lipases. % TBL (Percentage of bacterial thermostable lipases); % TFL (Percentage of fungal thermostable lipases); % TBFL (Percentage of bacterial and fungal thermostable lipases); % ML (Percentage of mesostable lipases).

3.3.3. Structural analysis of lipases by subfamily tree annotation

Subfamily clustering by CLUSS 2 ver. 1.0 (IJCBDD 2008) separated thermostable and mesostable lipases into eleven subfamilies. The subfamilies have been illustrated by phenogram in Figure 3.3. Thermoalkalophilic *Bacillus* lipases completely separated into a separate subfamily (subfamily 6 in our classification). Rest of the ten subfamilies showed mixed occurrence of thermostable and mesostable lipases (Fig. 3.3). Each individual subfamily showed marked differences in their thermostability, pH stability, substrate specificity, sequence and structure level.

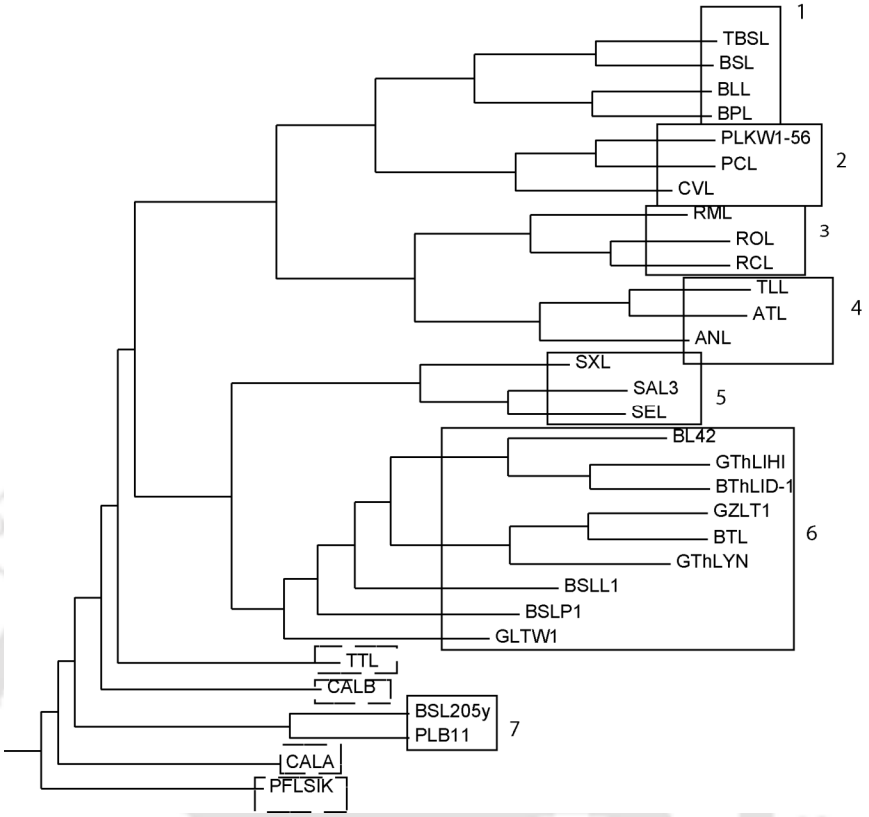


Fig. 3.3. Phenogram showing subfamily clustering of thermostable and mesostable lipases into 11 subfamilies. The boxes represent subfamily1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11. Subfamilies 8, 10, 11 are the orphan subfamilies represented by dashed box and subfamily9 contains two bacterial mesostable lipase sequences.

Subfamily1 was represented by minimum α/β hydrolase lipases having two thermostable-mesostable pairs. Similar features were that all the family members lack Cys residues in their sequences, they lack lid domain, are alkalophilic with pH range of 9-11and show preference for medium to short chain length fatty acid substrate. Moreover, the thermostable lipases show temperature range of 55-70^{o}C.

Subfamily2 was dominated by *Pseudomonas* lipases; two thermostable and one mesostable. The thermostable lipases showed temperature stability range of 55-60^{o}C. All the family members showed near neutral pH optimum, funnel shaped substrate binding cavity (Pleiss et al. 1998), varied substrate specificity, absence of Zn^{2+} binding.

Subfamily3 and Subfamily4 were fungal filamentous lipases with PDB structure available for only one mesostable and one thermostable member of each subfamily respectively. The former family showed slightly alkalophilic pH stability range, substrate specificity for medium chain fatty acids; hydrophobic, crevice-like binding site (Pleiss et al. 1998) and one member was moderately thermostable though the latter was dominated by all thermostable lipases having wide range of pH stability and showing 1,3 regiospecificity (Macrae et al. 1985) for their substrate.

Subfamily5 comprised of all *Staphylococcus* lipases; two thermostable and one mesostable. Sequences showed Zinc binding motif and P-loop motif in two members. This subfamily was dominated with members showing short chain length substrate specificity.

Subfamily6 was the largest of all comprising of all *Bacillus* thermoalkalophilic lipases showing medium chain length substrate specificity, and presence of Zinc binding.

Subfamily 7,8,10 and 11 contained orphan sequences and hence were not considered for the present study.

Subfamily 9 consisted of two mesostable *Bacillus* and *Pseudomonas* lipases lacking any structural information. Distinguishing characteristic from other subfamily being their GGGX type oxyanion hole signature.

Subfamily annotation clearly showed which thermostable-mesostable lipase structures were to be compared for further detailed analysis for structural role in thermostability. For this purpose structural superimposition was performed of representative thermostable-mesostable lipase pairs from each intra and inter subfamily to grip concrete data for their differences in temperature stability. Two intra subfamily structures from subfamily1 (BSL: 1i6w with TBSL: 3d2c) and subfamily2 (PCL: 3lip with CVL: 1cvl) having 95% and 84% sequence identity showed RMSD of 0.217 and 0.32 respectively. Additionally two inter subfamily structures from bacteria (BSLL1: 1KU0 with CVL: 1cvl) and fungus (TLL: 1tib with RML: 4tgl) having 24% and 34% sequence similarity showed RMSD of 14.49 and 0.97 respectively when superimposed. The overall structural differences observed in thermostable lipases w.r.t mesostable ones were more γ-turns with difference in their structural positions being nearer to N or C terminus of helices or strands as the γ-turns near helix7 of TBSL. In PCL Pro233 replaces Leu233 in CVL

forming a γ-turn which may play a role to increase conformational stability. The other thermoadaptivity defining trends observed were lesser β-branched residues (Val, Ile, Thr) in helices, longer β-strands, shorter loops, frequent distribution of Ala residues in α-helices, e.g. Ala (dipeptide or tripeptide) residues in lid domain of Bacillus thermoalkalophilic lipases, more statistically favoured residues at N-cap and C-cap position (Asp, Asn, Ser, Glu, Gln, Ala, Arg, Lys) of α-helices, e.g. His in N-cap position of helix 1 in CVL is replaced by Arg in thermostable PCL and more Ser residue in N-cap end of helix1 in TBSL than BSL, greater preference for negatively charged amino acid residues in amino terminus and positively charged residues in carboxy terminus as seen in TBSL and BSL having Glu and Ala residues in amino terminus respectively. Moreover, amino terminus Asp91 of helix 6 in TBSL is not observed in BSL. Increase in protein conformational stability by optimizing β-turn sequences was also observed in thermostable lipases by increased preference for Pro, Gly and aromatic residues and avoidance of Ala residues in beta turns. This trend however was not noticed to be strictly followed in their mesostable counterparts. Structural analysis by VADAR program showed that thermostable lipases possess more charged-neutral hydrogen bonding pair and greater exposed polar surface are. Structural comparison of thermostable and mesostable lipase pairs by tree based annotation revealed many factors that can altogether be held responsible for thermostability of lipases. To the best of knowledge it is reported for the first time that increment in inverse γ-turn and their presence near the amino or carboxyl end of helices and strands plays an important role in enhancement of thermostability of lipases. To add to this, this finding being based on the study of many different thermostable lipases also supports the assumption drawn long back by Paupitt et al. in 1988 that a γ-turn may contribute to stability of Thermolysin when compared to neutral protease from Bacillus (Pauptit et al. 1988). As it was considered that one intramolecular hydrogen bond contributes about 2.1-6.3 KJ/mol to the free energy of stabilization (Stark et al. 1992) it can be said that increase in inverse γ-turns can lead to enhancement of thermostability of lipases as they are sharp and tight turns comprising of three amino acid residues which leads to increase in intramolecular hydrogen bonding because they induce hydrogen bonding between i (C=O) and i+2 (N—H) residues in the polypeptide chain. Adding to the

aforesaid the report is also supported by the information provided by Milner that γ-turns near the amino or carboxyl terminal end of helices or strands are involved in stronger hydrogen bonding (Milner-White 1990). Thus, it can be said that the presence of inverse γ-turns near the end of helices and strands in thermostable lipase structures leads to thermal stabilization of lipases. In another report by Jong et al. it was said that γ-turns contribute very less to conformational entropy (Gong et al. 2007). Based on this property of γ-turns it can be said that increase in inverse γ-turns in thermostable lipases will lead to their lesser conformational entropy at elevated temperatures and thus to thermostability.

Lesser beta branched residues in helices was another trend observed for thermostable lipases. Present obervation is in agreement with the report that beta branched residues affects helix stability as their side chains are not well accommodated in the helix thus disturbing the helix propensity (Wang et al. 2003). Thus it can be said that this can lead to decrease in thermostability of lipases. Hence, thermostable lipases through evolution have chosen to avoid beta branched residues in their helices.

Increase in strand length upto seven residues near β-hairpins was considered to increase protein conformational stability (Stanger et al. 2001). Similar trend was observed in thermostable lipases thus longer β-strands near β-hairpins in thermostable lipases can be assumed to contribute to their temperature stability because they make the structure more compact leading to their conformational stability at elevated temperature. It is also believed that presence of shorter loops in thermostable protein structures plays a role in their thermoadaptive properties by increasing compactness and reduction of the entropy of unfolding in proteins leading to their stability (Li et al. 2005). Present study also came up with similar data analysis with regard to thermostable lipases, strongly supporting the aforesaid. Another observation which could be given credit for increasing thermostability of lipases was the presence of greater frequency of Ala residues in α- helices of thermostable lipases and the absence of the same trend in mesostable lipases. This can be said because Ala is credited with the title of "best helix forming residue" (Vieille et al. 2001) as it increases helix propensity because its side chain gets well accommodated in helices so can lead to stabilization of the structure of lipases at elevated temperatures. In addition to the above mentioned reasons it was reported that the occurrence of residues

which can form hydrogen bonds in N-cap and C-cap positions like Asp and Glu in helices can stabilize protein structures by stabilization of helix dipole and hydrogen bonding (Li et al. 2005). This trend was observed in thermostable lipases and thus it can be said that N-cap and C-cap residues are important for making lipases thermostable. It is well known that negatively charged amino acid residues in amino terminus and positively charged residues in carboxy terminus in proteins stabilize their helix dipole (Eijsink et al. 1992). Therefore stabilization of helix dipole leads to protein stability. This trend was followed by maximum of the thermostable lipases studied. This indicates that proper N- and C-terminal residues also play a crucial role in thermostabilization of lipases. It is also thought that some β-turn residues like Gly relieve steric strain of the protein losing thermostability (Trevino et al. 2007). Aromatic residues and Pro remain partially buried and increases hydrophobic interactions. Pro also has a restricted φ-angle leading to further increase in rigidity. This is entropically favorable at certain turn positions and leads to increase in protein rigidity (Trevino et al. 2007). Between thermophilic and mesophilic lipase structures it was also observed that significant difference in the presence and distribution frequency of these residues which is more for the thermostable lipases. Thus, these residues and their distribution is another reason for increase in thermal stability of lipases. Besides, all the above mentioned factors which can be held responsible for thermostabilizing lipases, two other noticeable trends in some thermostable lipases that differed from mesostable ones were increase in exposed polar surface area, increase in hydrophobic contact by amino acid substituition and change in torsion angle and surface exposure of amino acid residues. This observation can be rightly justified to also contribute to thermostability of lipases as it was showed that increase in the number of hydrophobic interaction (Branden et al. 1999) and increase in polar surface area leads to better protein stability by enhancement of ionic interactions and hydrogen bonding with surrounding water (Suvd et al. 2001).

3.3.4. HotSpot Wizard and CUPSAT analysis of structurally important residues

HotSpot Wizard analysis of intra subfamily thermostable-mesostable structures (TBSL-BSL and PCL-CVL) having greater than 80% sequence similarity highlighted the following interesting facts. The structurally important residues observed in TBSL and not in BSL with very low mutability rate were Leu143, Thr117, Ile169, Lys88, Asn138, Tyr86, and Asp118. These residues show maximum destabilizing mutation as predicted by CUPSAT and compared with BSL. The interatomic interaction of these residues as studied for both BSL and TBSL lipases in Analysis of Interatomic Contacts of Structural Units (CSU) (Sobolev et al. 1999) accessed through PDB gave the following information about the contribution of each of these residues towards thermostabilization. Leu143, Ile169, Lys88 showed enhanced hydrophic-hydrophobic contact. Thr117, Ile169, Lys88, Asn138, Tyr86 show increase in hydrogen bonding. Asp118 revealed differences in interaction with residues involved in hydrogen bond formation. Also, increase in charge-neutral interaction was observed for Leu143 and Ile169.

The structurally important residues in PCL with low mutability rate which may be responsible for enhancement of thermostability were observed to be Val267, Leu315, Pro304, Gly295, Thr245, Asn202, Gly60, Ala272, Thr280, Tyr282, Arg309, Lys316, and Ile232. Mutation of these residues can lead to very low to moderate stabilization of the lipase as predicted by CUPSAT. Only Thr245, Asn202, Arg309 which are buried in the core, will stabilize the structure when mutated with hydrophobic residues. Residues Val267, Leu315, Pro304, Ala272, and Ile232 increases hydrophobic-hydrophobic contacts. Gly295, Ala272 shows increment in hydrogen bonds. Greater charged neutral hydrogen bonding was observed for Gly60 and Lys316. The amino acid substitution in PCL responsible for increase in hydrogen bonding, hydrophobic contact and charged neutral hydrogen bonding are Ser280→Thr280, Thr309→Arg309, Ser267→Val267, Lys315→Leu315, Leu316→Lys316, Ser202→Asn202, Ser280→Thr280, His282→Tyr282. These mutations show stabilization of the structure when cross validated with CUPSAT. The residues predicted by Scride for structural stabilization of PCL absent from CVL are Ile11 and Ala105 which cannot lead to stabilizing mutation as predicted by

CUPSAT and Hotspot wizard. These are hydrophobic and buried in the core, can be responsible for thermostability of PCL. Although no such substitution for the structurally important residues was observed in TBSL in comparison to BSL.

3.4. Conclusions

The motive of this chapter was an in depth analysis of factors that lead to thermostabilization of lipases. Lipases were chosen because most of the thermostable enzymes reported till date is lipases. Additionally they possess high industrial priority. By sequence and structure analysis of thermostable lipases with an added approach of tree based annotation it can be concluded that each thermostable lipase adopts its own strategy in relation with the three dimensional arrangement of amino acid residues to increase its stability at elevated temperatures. Conclusively, the present study has come up with a set of strategies that can enhance thermostability of lipases. Increasing the titrable amino acid residues near the active site Ser can enhance thermostability. Increment of charged residues in surface and decreasing β-branched residues in helices can make mesostable lipases thermostable. However, the strategy that can be employed to increase thermal stability of bacterial lipases may not be applicable for fungal lipases as thermostabilization factors show a changing trend among these two distinct phylogenetic groups of lipase. An example is increase in frequency of Gly residues in loops of only *Bacillus* lipases leading to their thermostabilization. Decreasing the percentage of free Cys residues without disturbing the conserved ones can lead to enhanced thermostability. Increase of poly Ala residues in helices especially in the lid of Bacillus lipases, mutating thermolabile residues in helices with residues having high helix propensity, more Pro residue in loops and turns and more aromatic residues in surface and core which increases hydrophobicity in lipases can also enhance their thermo stability. Increasing AXXXA motif in helices and increase in strand length upto seven residues near β-hairpins were found contributory to increase lipase stability. Moreover mutating residues in N-cap and C-cap positions with Asp and Glu which can stabilize helix dipole, increase in exposed polar surface area and increase in hydrophobic contact by amino acid substitution can cause lipase thermostabilization. The most important fact to be endorsed leading to enhancement of temperature stability is the

increment in inverse γ-turn near the amino or carboxyl end of helices and strands which has been reported by us for the first time.

This result is in concordance with Chapter II where γ-turn was found to increase in thermostable proteins. Therefore in addition of it being a new characterized factor, an attempt can be made to increase γ-turns in mesostable proteins for rendering them thermostable.

CHAPTER IV

Rationalizing Protein Thermostability by Multiple Feature Ranking for Model Generation

Prologue

From previous chapters it was observed that classifying thermostable proteins and predicting thermostabilizing mutations is still a challenging job. This is due to the fact that numerous factors related to protein thermostability have been proposed. The mechanism of stability has been attributed to the cumulative effect of all such factors. However, a guided approach to attain thermostabilizing mutations is still elusive. Through this work for the first time, a two-step hybrid approach to derive thermostabilizing mutations has been developed. Thermostabilizing protein attributes have been ranked according to their role in contributing towards thermostability and a model has been generated which can predict multiple mutations to lead to thermostability. The method employed is analytical hierarchical process to derive at the mutations and site-directed mutagenesis to engineer such mutations *in vitro*. The novelty of the present work is the production of thermostable proteins in a rationalised and predictable way. The ranking was developed based on an elaborate analysis of a set of quantitative structural features on a final dataset of 127 pairs of thermostable and mesostable protein structures. Ionic interaction and main-chain to main-chain hydrogen bonds were the features showing the highest priority vectors for thermostability. The ranks were used to develop a tool- RankProt. It can be used for predicting multiple point mutation(s) that can contribute positively in thermostabilizing proteins. The higher ranked proteins were found to be better in thermostability when compared to their mesostable counterparts. Further, a rank value of 0.54 for a protein undergoing mutations was predicted to be thermostabilizing and the accuracy of the method was calculated to be 91%.

Outputs:

1. RankProt- tool to predict thermostabilizing mutations.
2. Development of a two-step hybrid approach to attain thermostabilizing mutations (Manuscript under preparation).

4.1. Introduction

In 1969 T. D. Brock and colleagues discovered *Thermus aquaticus* from which thermostable enzyme Taq polymerase having an optimum temperature stability of 95°C was extracted. Since then attaining thermostability of proteins has been a constant challenge for researchers, while the desire to attain them has been getting stronger always. This is because thermostable proteins find use in many industrial applications such as paper, detergent, and biofuel industries (Lehman et al. 2001). Irrespective of this fact, there are very few commercially available thermostable enzymes, because engineering mesophilic proteins for thermostability has not been a trivial job. Therefore available methods to successfully engineer proteins for thermostability are very few. One such method that is conventionally used to mutate proteins towards thermostability, is called directed evolution. The other well-known methods are called "sequence consensus" and "rational design" approaches. Former involves changing the amino acid at a specific position to that most frequently observed in the sequence homologs. The latter, rational design method relies on the structure of a protein (Tsai et al 2009). These approaches are practically limited as they require ample time and are also labor intensive. Further, the mutations done are randomly selected and therefore the success of the method becomes a chance event. Another drawback is that the methods involve application of selection pressure of high temperatures, to achieve at the mutations, followed by colony screening for the desired trait and only repeated round of random mutagenesis leads to success (Lehmann et al. 2001; Tsai et al 2009). For example, in 2008, Stéphane Emond isolated 60,000 amylosucrase variants from two libraries generated by the MutaGen random mutagenesis method. They were then screened through an *in vivo* selection procedure leading to the isolation of more than 7000 active variants. These clones were then further screened for increased thermostability using an automated screening process. This experiment yielded three improved variants (two double mutants and one single mutant) showing 3.5- to 10-fold increased half-lives at 50°C compared to the wild-type enzyme (Emond et al. 2008). Therefore it is clear that today to successfully engineer thermostable proteins a better method to discriminate between

thermophilic and mesophilic proteins is necessary and a rationalised model to predict such mutations is required so that predicted mutations can be carried out *in vitro* through site directed mutagenesis.

To achieve at a prediction and classification model it is necessary to gain in-depth knowledge about protein thermostability. In literature, till date, numerous factors have been discovered to be responsible for enzyme thermostability (Vogt et al. 1997; Ventriani et al. 1998; Haney et al. 1999; Thompson and Eisenberg, 1999; Cambillau and Claverie, 2000; Szilagyi and Zavodszky, 2000). These factors have been elaborated in Chapter I of this thesis. Conclusively, all these work state that thermostability has been attributed to be the cumulative effect of all protein stabilizing factors. It has also been concluded that a single protocol ceases to exist which can render proteins thermostable through protein engineering approaches. Further, the factors responsible for thermostability vary from one protein to another (Kumar et al. 2000; Gromiha et al. 2008; Ebrahimi et al. 2011). Therefore, at present, it is necessary to rank such factors in accordance to their importance in contributing towards protein Thermostability and generalize them through all six protein classes. This ranking is necessary as it will be instrumental to develop a prediction model and a guided protein engineering protocol to render mesostable proteins thermostable.

In this direction, up to now, numerous *in silico* algorithms have been proposed which can predict important thermostabilizing features and also predict whether conceptualized mutations will be stabilizing. The lacuna is that, a method to rank features according to their priority and predict whether predicted mutations thermostabilizing, has not been yet developed. The available models have been developed by investigating protein features by comparing thermostable proteins with mesostable proteins and prediction of thermostabilizing mutations relied on amino acid sequence composition and the three dimensional structures of proteins. Bulk of these methods achieves favored mutations either through ancestral or consensus methods (Malccolm et al. 1990; Li et al. 2012; Wijma et al. 2013). Others (Vogt and Argos 1997; Szilagyi and Zavodszky 2000) compare structural features of homologous proteins setting priority to intramolecular interactions and the results have been used to design mutants with enhanced stability (Petsko 200; Eijsink et al.

2004). The major problems that researchers faced in development of such prediction models with high accuracy are high level of sequence and structure similarity between thermosphilic and mesophilic proteins. This raises complexity in evaluating and ranking the most effective factor(s) which can render them to be stable (Vielle et al. 2001; Luke et al. 2007). Other complexities are that most of the *in silico* algorithms used to predict such stabilizing mutations are dependent on matrices of melting temperature or alterations of folding free energies between wild type proteins and their mutants (Li et al. 2012). It has been reported that the accurate prediction of the thermodynamic consequences caused by mutations still remains challenging (Seeliger et al. 2010). Other models are mostly knowledge based (Rohl et al. 2004). Few are support vector machine (SVM) based (Capriotti et al. 2005) and further less are based on molecular dynamics (Benedix et al. 2009). Furthermore, the favored mutations are related to the global stability of a protein (Wijma et al. 2013). They do not comprehend whether they are parametrically thermostabilizing or not. Additionally, majority of them require complex computational power and proficiencies. Machine learning approaches often require training and testing set and the level of accuracy reached is 90% until date (Ebrahimi et al. 2011). Molecular dynamic simulations of mutation are several orders of magnitude complicated than that with a knowledge-based scoring function (Sleegier et al. 2010). The methods also require several empirical parameters or heuristics such as patterning of residues or rotamers for their calculations. Most of the times to attain Thermostability, multiple mutations are required. However, only few algorithms can predict the effect of multiple mutations because multi-site mutations are expected to have more complex effect on protein thermostability as compared to single point mutations (Tian et al. 2010). The accuracy achieved until date creates limitation when greater than two mutations are to be performed. Another lacuna is that all these methods give multiple choices of possible stabilizing mutations and fail to select which point mutation (single, multiple) or combination of mutations will thermostabilize proteins. In short they are unable to rank or prioritize the plausible mutations based on their effect on protein stability. To summarize the aforementioned observations it can be said that firstly, a rational design through site directed mutagenesis still ceases to exist and a

robust scoring function which can predict the behavior of the mutations is still elusive but essential (Potapov et al. 2009; Li et al. 2010). Secondly to achieve the same a deeper understanding of the mechanisms underlying protein thermostability is still a prerequisite (Eijsink et al., 2004). Thirdly, most of the work carried out to mutate proteins to render them thermostable are based on random mutagenesis and directed evolution strategies.

These aforementioned observations called for development of a better prediction model for thermostabilizing mutations with the capability of priotizing thermostability features. The methods and results of development of such a prediction model have attempted through this dissertation work and presented in this chapter. The schematic of the attemt has been illustrated in Fig. 4.1.

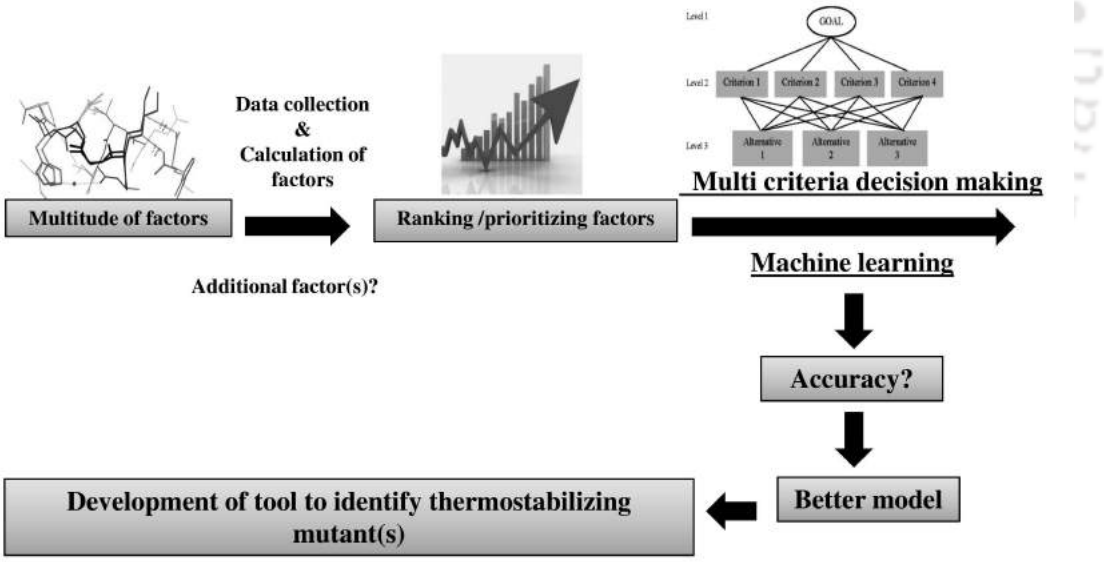


Fig. 4.1. Schema to identify and generate thermostable mutants.

4.2. Materials and Methods

4.2.1. Datasets for feature generation

Initially key word search of "thermostable", "thermophilic" and "hyperthermophilic" in Protein Data Bank (PDB) resulted in 1280 structures. The details about data collection have been presented in Chapter II of this dissertation work. The data was cleaned for redundancy. Partially sequenced proteins and putative sequences were removed from the analysis. Thus only 378 proteins remained. Mesostable counterparts were chosen by BLAST search. Since the interest was in detecting small differences in composition, only structural protein pairs with sequence similarity >70% were retained. The dataset consisted of both eukaryotic and prokaryotic protein homologous pairs as phylogeny was not kept as a predetermining factor for choosing the structural homologues. This is based on the fact that mechanism of protein structural stability is independent of phylogenetic diversity (Shin et al. 2014). Moreover, some archaeal thermophilic proteins have higher homology to eukaryotic proteins than those from prokaryotes (Shin et al. 2014). Also in literature several examples are available where prokaryotic and eukaryotic homologous structural pairs have been compared (Vogt et al. 1996; Kumar et al. 2000; Sadeghi et al. 2006). Therefore, the final training set contained 127 non redundant homologous thermophilic- mesophilic protein pairs. Numerical features were generated using various software and tools. The details have been presented in Chapter II. 25 structural features were generated using the developed python tool (Intra-protein Interaction calculator), Volume Area Dihedral Angle Reporter (VADAR) and Promotif. The features were count of the intramolecular interactions and the number of secondary structures in proteins. They were normalized with respect to the number of atoms or the length of the protein sequence as the case may be. All the features were converted to their percentage scores. The datasets were then subjected to non parametric two tail Kolmogorov Smirnov test of significance at 95% confidence level. Features with $p < 0.05$ were retained and considered significant and

highly correlated features with Pearson correlation greater than 0.9 were discarded. The two sets in each case were assigned name of TP for thermostable proteins and MP for mesostable proteins.

4.2.2. Classification of thermostable proteins through machine learning algorithms

The final dataset with their features were imported into Rapid Miner (RapidMiner 5.3.000, Rapid-I GmbH, Stochumer Str. 475, 44227 Dortmund, Germany) and the thermostable and mesostable proteins (categorised as T and M) were set as the label attribute. The dataset was then analyzed for feature importance through different weighting algorithms. Further, the next goal was model generation which has the ability to distinguish between thermostable and mesostable proteins. This is because only knowledge about important features for thermostabilizing proteins is not enough to design mutations which can render proteins to be thermostable. For the same, the dataset was further subjected to unsupervised, lazy modeling and supervised algorithms. Numerous algorithms were applied as there are no gold standard methods which will give predicted and high accuracy results. Furthermore, it was necessary to analyze which dataset generates the best model for thermostability prediction. The details have been presented in the following sub-sections.

Application of attribute weighting to enumerate important thermostabilizing features

Attribute weighting is a multi criteria analysis method and was used to find out important features that contribute towards thermostability of proteins. For performing the same, 11 different algorithms of attribute weightings were applied separately on the three datasets. Weight by Information gain, Information Gain ratio, Rule, Deviation, Chi-squared statistic, Gini index, Uncertainty, Relief, SVM (Support Vector Machine) and PCA (Principle Component Analysis) were calculated. The

results were obtained with a value between 0 and 1 for each protein feature, which revealed the importance of that attribute in regard to a label attribute. All attributes with weight greater than 0.5 were selected. Attribute weighting presented with important features and manageable attributes but alone was insufficient in generating models for protein thermostability. Thus the dataset was further subjected to unsupervised and supervised clustering algorithms.

Application of unsupervised clustering for model generation for protein thermostability

Unsupervised clustering algorithms panel data into clusters according to various criteria so that objects within a cluster are similar to each other but are dissimilar to objects in other clusters (Han et al. 2001, Ebrahimi et al. 2011). Most of the methods are associated with their own merits and demerits. The success of the method in discovering structures or patterns on its own depends on the type of data and algorithm used. Thus, it is important to apply more than one clustering algorithms to arrive at the most accurate model. k-Means, k-Means (kernel), k-Medoids, Support Vector Clustering, DBSCAN and Expectation Maximization Clustering (EMC) were applied on the final dataset with split validation of 70% testing and 30% non overlapping training datasets. k-Means is one of the simplest parametric unsupervised learning algorithms (MacQueen, 1967). It classifies a given data set through a certain number of k-clusters. The aim is to divide M points with N dimensions into k clusters by minimizing the within cluster sum of squares (Hartigan et al. 1979). It finds k centroids and assigns every object to the nearest centroid, where the coordinate of each centroid is the mean of the coordinates of the objects in the cluster. However, it is sensitive to the presence of outliers (Park et al. 2009). A modification of k-Means is the k-Means kernel which uses kernel method instead of the Euclidean distance. In k-Medoids representative objects called medoids are considered instead of centroids and it is less sensitive to the presence of outliers (Park et al. 2009). Support Vector Clustering is non parametric and data points are mapped by means of a Gaussian kernel to a high dimensional feature space (Ben-Hur et al. 2001). DBSCAN is a

density-based notion of clusters which is designed to discover clusters of arbitrary shape (Ester et al. 1996). As biological datasets can have missing values, Expectation Maximization algorithm estimates probabilistic parameter in models with incomplete data (Do et al. 2008). Sometimes clustering methods fail to correctly cluster data-points into their correct classes, thus it is necessary to employ other machine learning methodologies to get to the most accurate model. Therefore, the dataset was further subjected to supervised clustering methods.

Application of supervised clustering for model generation for protein thermostability

Supervised clustering is also known as model based clustering method. Unlike unsupervised clustering, other than finding groups of objects, they find characteristic descriptions for each group, where each group represents a concept or class (Rokach et al. 2010). The most frequently used induction methods are decision trees, neural networks and lazy modeling. In decision trees, the data is represented by a hierarchical tree, where each leaf refers to a concept and contains a probabilistic description of that concept. Three tree inductions models (Decision tree, Random forest, Decision stump) were run on the dataset. Each tree induction model ran with the following four different criteria: Gain Ratio, Information Gain, Gini Index and Accuracy. Additionally, ID3, CHAID and a weight-based parallel decision tree model, was run with 11 different weighting criteria (information gain; information gain ratio; rule; deviation; correlation; chi-squared statistics; gini index; uncertainty; relief; SVM; PCA).

As prediction accuracies may differ based on datasets, lazy modeling algorithms were also employed. K-Nearest Neighbor (K-NN) and Naive Bayes algorithms were applied as lazy learners on the final dataset and validated with split validation of 70% testing and 30% non overlapping training datasets. K-NN is an intuitive and non-parametric classification algorithm (Huang et al. 2002). It searches for k- nearest neighbors of the data point and transfers their functions by weighted averaging, such that nearer neighbors have larger influence to prediction than the

farther ones (Rosen et al. 2011). The datasets were further modeled through Naive Bayes algorithm as it is known as probabilistic classifiers which applies Bayes' theorem. It assumes that the values of a particular feature are independent of all the other features. Such classifiers have been demonstrated to be able to achieve a high degree of prediction accuracy (Kohonen et al. 2009). Furthermore, in this study feed forward neural network (Neural net and Perceptron) were applied on the dataset. 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 (Rokach et al. 2010).

4.2.3. Application of Multicriteria Decision Making algorithm

The next goal was to rank or prioritize protein structural features according to their contribution towards thermostability. Machine learning methods can classify thermostable proteins but cannot prioritize thermostability factors by ranking them in accordance to their importance in rendering proteins thermostable. Therefore, Multi Criteria Decision Making approach-Analytical Hierarchical Process was employed. The steps involved have been discussed as follows.

Hierarchical clustering

The said problem was arranged as a hierarchy. It involved four phases, first is the hierarchical structuring of complexity into homogeneous clusters of factors with objectives, criteria and alternatives. Secondly, deriving at weights for criteria and alternatives and representing those with numbers. Thirdly, using the numbers to calculate the priorities of the criteria and alternatives and lastly completing the

synthesis of these results to determine the most important alternative (Saaty, 2008). The hierarchy of thermostability is composed of three tiers. First tier is the goal and second is the criteria. The third tier was the alternatives. The alternatives are a set of wild type or engineered proteins. Fig 4.2 illustrates the hierarchical clustering of the said problem.

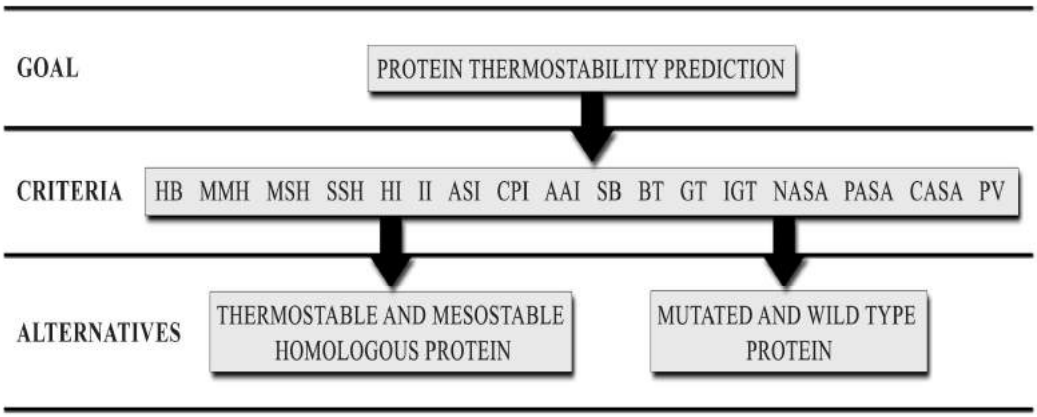


Fig. 4.2. The hierarchical clustering for predictioin of thermostability of proteins.

Deriving at weights of features and the pairwise comparison matrix

As the features are numeric in nature, there arose a need to derive pro-rata weight for the thermostability datasets. According to Saaty, 2008 a scale of numbers that indicates how many times more important or dominant one feature is over another feature is a pre-requisite for making comparisons. Thus to prioritize the criteria their weights were derived. This led to the formation of a positive reciprocal pairwise comparison matrix. As the thermostable protein dataset consisted of structural homologues the pro-rata weights were derived through the following formula:

$$\Delta_\iota = \Phi_\nu TP - \Phi_\nu MP \quad (1)$$

Where, the symbols have the following meanings: ι= 1,..,n where n = number of features; Δ_ι is the difference in the normalized feature, $\Phi_v TP$ stands for the normalized feature of the thermostable protein dataset and $\Phi_v MP$ stands for the normalized feature of the mesostable protein dataset. Then the differences in the features were represented by vectors, where each difference in attribute, Δ_ι takes the value of 1 if the feature of type ι is positive, and 0 if negative or there is no difference. This formed our difference matrix (Appendix III Table A3.1). Further the number of proteins in the difference matrix having the value of 1 for each of the 17 features was summed and converted to their percentage scores w.r.t., the total number of proteins. This gave the percentage weight of the number of protein showing increase in a feature w.r.t. the mesostable protein dataset.

In the next step all the percentage weights were scaled down to a 1-9 interval scale weight by a python script which uses the equation:

$$\text{Weight } (W_i) = \left[\frac{(\xi_\iota - \alpha)}{(\beta - \alpha)} \times 8\right] + 1 \qquad (2)$$

Where, W_i is the derived weight in the 1-9 scale, i= 1,..,n where n = number of features; ξ is the value of the weight for feature i, α is the minimum value in the weight for feature and β is the maximum value in the weight of feature. This gave the relative importance of each feature w.r.t. each other. This ratio was supplied to the 17x17 pairwise comparison matrices (Appendix III Table A3.2).

Further step was column wise normalization of the matrices so that the sum of each column was 1. This was followed by calculating the sum of the rows. This gave the priority vectors or the eigen vectors of the matrix. The priority vectors were indicators of the relative importance of each feature over the others for their positive contribution towards protein stability in elevated temperatures. Conclusively it can be pointed out here that higher the value more is its impact towards rendering proteins to be stable in such extreme conditions.

The next step was to calculate the consistency of the matrix from formula 3. The purpose for doing this was to make sure that the original weights given to the features were consistent.

$$\text{Consistency index (CI)} = \frac{\lambda max\text{-}N}{N\text{-}1} \quad (3)$$

Where, λmax is the consistency measure of each row in the second matrix which is calculated as the dot product of initial matrix with the eigen vectors which is then divided by N, where N = total number of features. Further consistency ratio is derived according to formula 4.

$$\text{Consistency ratio (CR)} = \frac{CI}{RI} \quad (4)$$

According to Alonso et al. RI is a random index and is equal to 1.6086 for N=17 and and1.6181 for N=18 (Alonso et al. 2006). A matrix is accepted as consistent if and only if CR < 0.1 (Alonso et al. 2006).

4.2.4. Development of RankProt

The aforementioned steps were simplified and automated through the development of robust software written in python. To achieve the same the developed python program for calculating intra protein interactions along with VADAR and Promotif and the eigen vectors for thermostability, calculated through AHP were embedded into an algorithm to rank the mutations. The principle for deriving at ranks was by matrix multiplication of features in the test set by the priorities/eigen vectors of the features. Normalized feature matrix were generated for the test set and multiplied with the previously calculated eigen vectors for thermostability. This gives the dot product of the matrix and is called the ranks of the proteins. This aided in predicting the rank to be considered for any protein undergoing point or multiple mutations as thermostable. Figure 4.3 is an illustration of the typified algorithm used for developing RankProt.

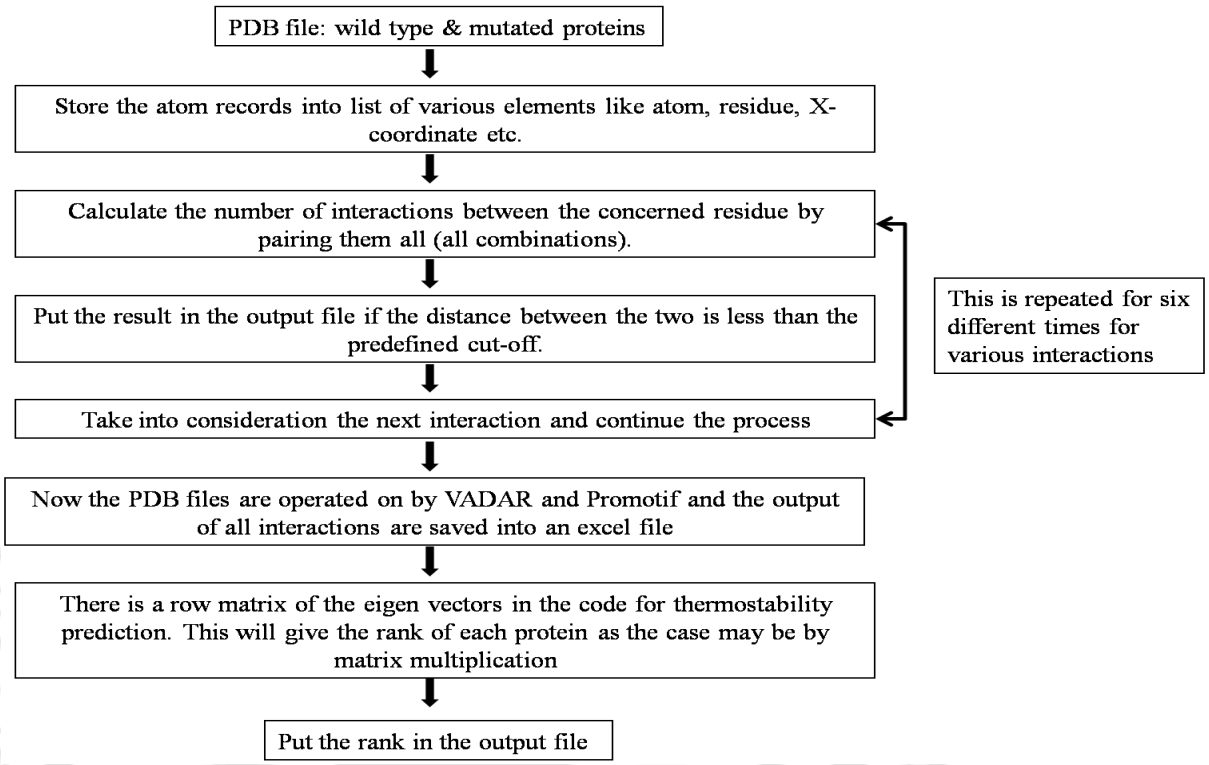


Fig. 4.3: The algorithm of RankProt.

The protocol to run RankProt is that the user submitted test set should consist of two protein structures in .pdb format. One is the wild type and the other mutated. For deriving at the mutated structure, *in silico* mutations can be performed through Chimera or the required mutated FASTA file of protein sequence can be subjected to homology modeling. The mutations should be carefully chosen for deriving at stable mutations which do not disintegrate the overall protein structure and activity. The following norms if followed will lead to successful mutations.

(1) Only residues which show high mutability propensity should be chosen. The mutations should not be done on stabilizing centre residues of a protein. These conditions can be checked through the webservers, HotSpot Wizard and Stride respectively.

(2) The mutated residues should not belong to the active site pocket of the proteins.

(2) The mutations should be done on the surface exposed areas of a protein or in their loop regions. Such areas do not hamper the overall stability of a protein.

Thus if wild type and mutated structure are available, it can be predicted whether the mutation will lead to thermostability by the ranks given by RankProt to the wild type and mutated structures. If and only if, such stabilizing mutations increase the number of the highest prioritized features they will lead to thermostabilization of proteins. Thus if the rank of the mutated structure is higher than the reference structure, then such mutations will qualify as thermostabilizing. The instructions to run RankProt have been provided with the software package available in the DVD attached along with this dissertation.

4.2.5. Performance and validation

Ranking proteins and mutations

To be able to rank proteins as thermostable/mesostable, 100 thermostable-mesostable protein pairs were chosen randomly as alternatives from the dataset. These were processed through RankProt to derive at the ranks. A second blind test was carried out with a set of 40 mesostable proteins with another mesostable counterpart chosen randomly from the dataset. The proteins were assigned ranks and the mean rank value and the difference in rank value were calculated for the thermostable-mesostable and mesostable-mesostable datasets. Furthermore, to validate RankProt other blind tests were carried out.

i. All the mutated thermostable structures available for *Bacillus subtilis* lipases were retrieved from RCSB PDB and they were ranked with respect to their wild type mesostable structure (PDB ID: 1i6w).
ii. Mutated structures that led to gain of stability of bacteriophage T4 lysozyme (PDB ID: 2lzm) were ranked.
iii. Mutated structures of human lysozyme (PDB ID: 1lz1) that led to loss of stability were ranked.

Accuracy

The final performance of RankProt was ranked by accuracy (formula 7) which are estimated according to Mirseska et al. 2009 using false positive (FP), False Negative (FN), True Positive (TP) and True Negative (TN) values (Mirseska et al. 2009).

$$\textbf{Accuracy:} \quad \frac{TP + TN}{TP + TN + FP + FN} \qquad (7)$$

Accuracy is the global representation of performance of the method or it shows the number of sample for which class of T or M is classified correctly. TP is the number of thermophilic proteins that our proposed model identifies them correctly as thermostable proteins. TN is the number of mesostable proteins that are identified correctly as mesostable proteins. FP is the number of mesostable proteins that the model identifies them wrongly as thermophilic proteins. FN is the number of thermostable proteins that system identifies them wrongly as mesophilic proteins.

4.3. Results and Discussion

4.3.1. Classification of thermostable proteins through machine learning algorithms

Datasets for feature generation

Machine learning models were built by analyzing features that can render proteins thermostable. For feature generation, 127 homologous and non redundant dataset of thermostable proteins were derived at after data cleansing for redundancy and high sequence homology. The features were further subjected to pre-pruning by application of Kolmogorov Smirnov two-tail test for significance. Only features which had a p value <0.05 at 95% confidence level were selected as the final set of

features to be analyzed by machine learning methodologies. This led to the generation of 17 features (Chapter II, Table 2.2).

Application of attribute weighting to enumerate important thermostabilizing features

Interesting observations could be made when attributes were weighted by the 11 different weighting algorithms. The results have been presented in Table 4.1.

Table 4.1. Results of attribute weighting

Features	Details	No. of weighting algorithms
MMH	Main chain-main chain Hydrogen bond	9
PASA	Polar accessible surface area	7
CASA	Charged accessible surface area	7
II	Ionic interactions	5

From Table 4.1 it can be observed that Main chain-main chain hydrogen bonds, polar accessible surface area, charged accessible surface area and ionic interactions were the attributes given weights by 10 of the weighting algorithms. Increment in main chain-main chain hydrogen bonds and ionic interactions has been reported to increase themostability of proteins (Sadeghi et al. 2006). A study by Vogt et al. showed that in 13 thermostable protein families out of a total 16, 11.7 internal hydrogen bonds increased per chain with every 10^{o}C rise in temperature. Again surface exposed charged and polar amino acids can contribute to the increase in ion pair networks and hydrophilic properties of thermostable proteins respectively. Both properties have been implicated to increase thermostability of proteins (Sadeghi et al. 2006; Ebrahimi et al. 2011). It has been reported that increase in ion pairs gave 70% consistency with the addition rate of 1.8 per 10^{o}C rise per chain (Vogt et al. 1997).

Unsupervised clustering to generate model for protein thermostability

The performance of the used algorithms varied. None of them could correctly classify thermostable proteins and mesostable proteins in two distinguishable groups. SVC, DBSCAN and EMC could not assign single thermostable protein into the correct class. Previously EMC models when applied on 2090 thermostable protein sequences with 800 amino acid attributes resulted in 100% accuracy of clustering proteins into correct groups (Ebrahimi et al. 2011). However, this failed when applied on smaller datasets. Thus it can be concluded unsupervised clustering algorithms are biased towards larger dataset of thermostable and mesostable proteins and fails when a more non redundant dataset is chosen. It also shows different prediction accuracies with different types of dataset chosen. It has also been reported that their success depends on the attribute weighting algorithms which chooses the dataset on which such clustering algorithm is applied (Ebrahimi et al. 2011). This result shows the importance of applying different machine learning algorithms on different types of feature space and datasets before arriving at a conclusive statement about the best prediction model.

Supervised clustering to generate model for protein thermostability

As unsupervised clustering could not yield results with high performance accuracy, the dataset was analyzed through supervised methods. All the models generated through supervised methods were validated through 10 fold cross validation. The results have been presented in Table 4.2.

Table 4.2. Predicted accuracy of model generation by supervised clustering

Model for protein thermostability prediction	% Accuracy
k-NN	61.84
Naïve Bayes	76.06
SVM (anova kernel)	80.26
ANN (2 hidden layer, 40 neuron in each layer)	**84.21**
Decision tree (Random forest, 10 trees, information gain)	80.26

Lazy modeling to generate model for protein thermostability

The performance accuracy of the lazy modeling algorithms (k-Nearest Neighbour, Naïve Bayes) have been presented in Table 4.2. As it is assumed that great speed and high accuracy is achieved when analyzing very large datasets, literature shows low accuracy of prediction when Naïve Bayes was used as the prediction model (Wu et al. 2009). Thus to further increase the accuracy of the models the dataset was subjected to supervised learning methods.

Decision Trees to generate model for protein thermostability

The dataset was then analyzed through decision trees which are simple to read and understand. Most of the decision trees were without roots and leaves and thus were discardedThe 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 (Wu et al. 2009). Random forest decision tree with 10 trees and information gain criteria gave the highest accuracy of 80.26%. Fig. 4.3 is an illustration of the decision tree obtained by Random Forest algorithm. The tree showed that the percentage of main chain to main chain hydrogen bonds >3.773%, followed by γ- turns > 1.232% and aromatic aromatic interactions > 0.227% were the decisive feature for thermostabilizing proteins. It has been said that lowering of configurational entropy stabilizes a protein. Main chain hydrogen bonds have been reported to stabilize proteins by lowering configurational entropy. Comparatively higher configurational entropy is assumed for two nearby residues that are not involved in main chain to main chain H-bonds (Guerois et al. 2002). This results

corroborate previous study which reported such bonds to be higher in thermostable proteins (Sadeghi et al. 2006). γ- turns 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. This study further extends the aforementioned observation to be true for thermostabilizing proteins from all classes. Aromatic aromatic interactions stabilize proteins as a typical aromatic-aromatic interaction has energy of between -1 and -2 kilocalories per mole and contributes between -0.6 and -1.3 kcal/mol to protein stability (Burley et al. 1985; Serrano et al. 1991).

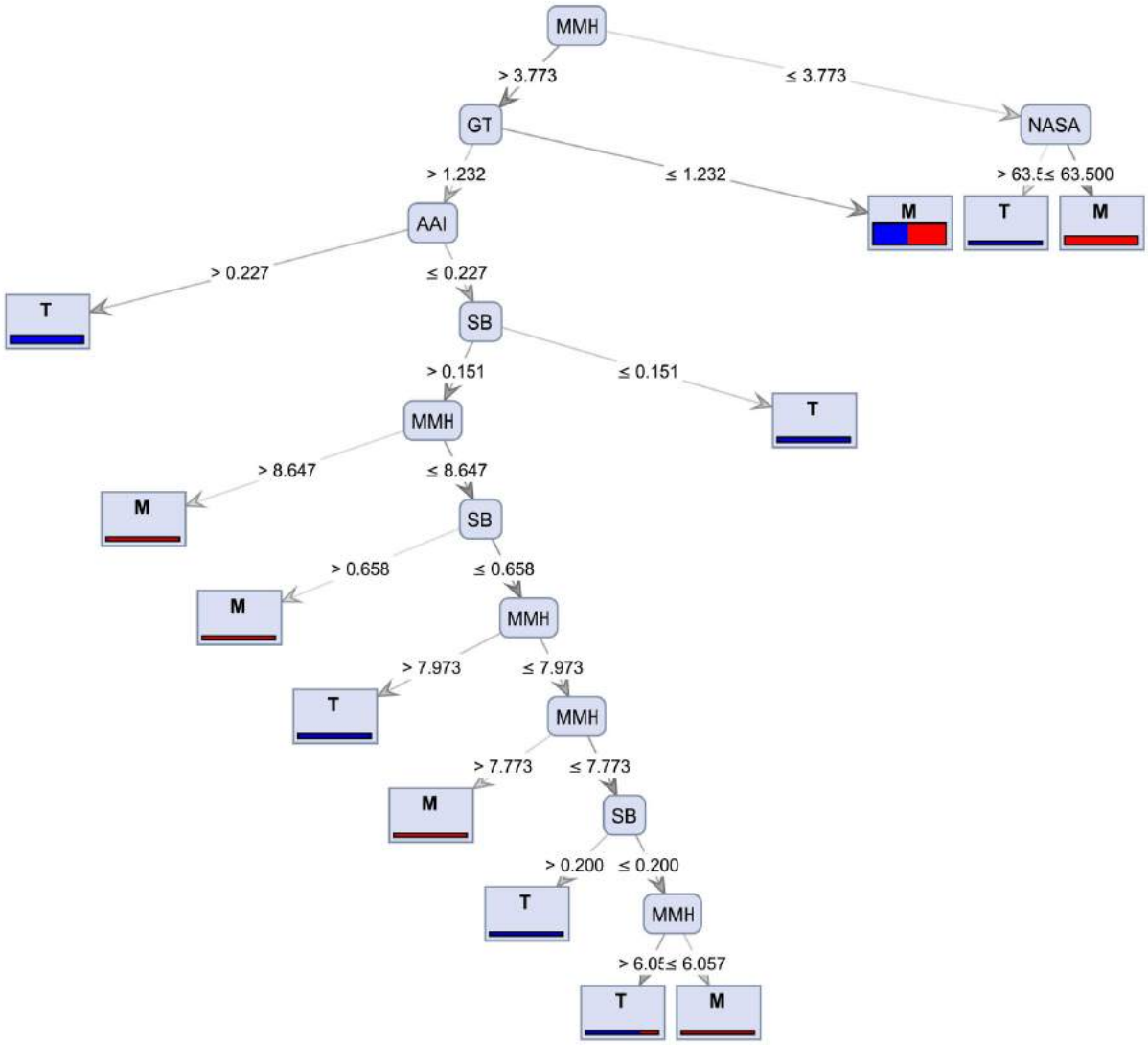


Fig.4.3. Decision tree of contributing structural features. Increase in MMH, GT and AAI classifies a protein as thermostable, whereas decrease in MMH and NASA classifies a protein as mesostable.

Support vector machines and Artificial Neural Networks to generate model for protein thermostability

Support vector clustering with anova kernel gave an accuracy of 80.26%. The highest accuracy was obtained with ANN (2 hidden layer, 40 neuron in each layer).

The result shows that supervised clustering performs better than unsupervised clustering and lazy modeling for classifying thermostable proteins.

4.3.2. Multicriteria decision making to rank thermostabilizing features

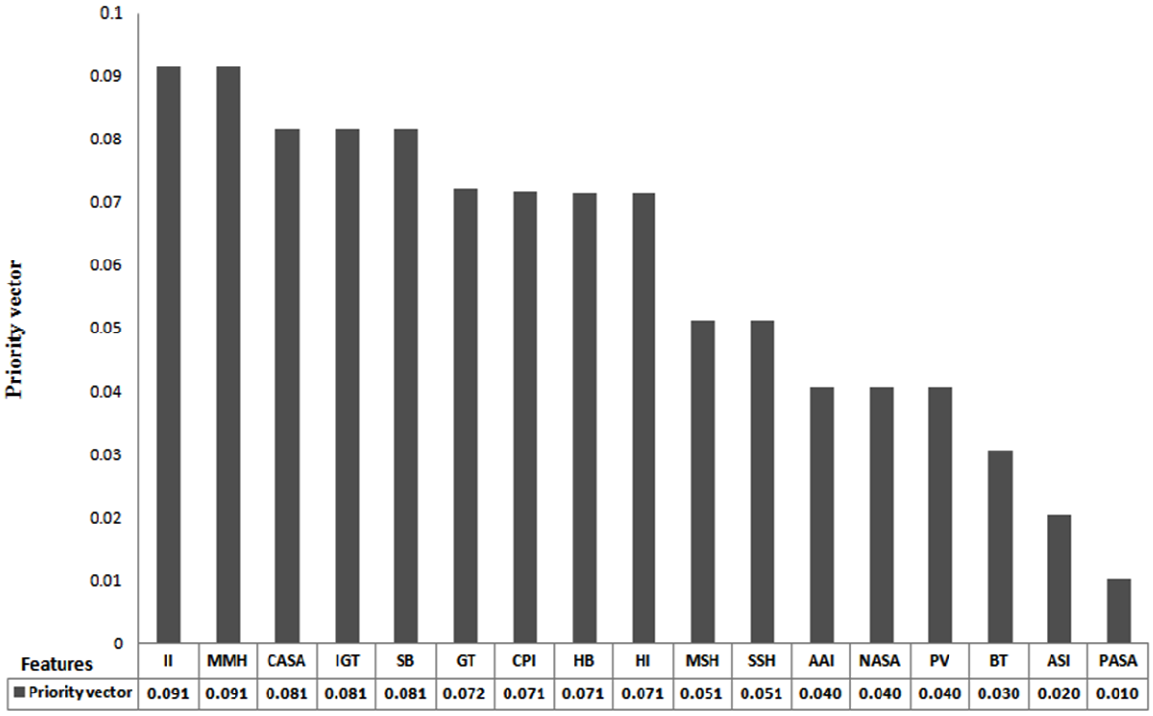
The application of AHP for ranking thermostabilizing features

Earlier methods were not able to rank thermostability features therefore, AHP was the method of choice because it is a structured technique for dealing with complex decisions, based on mathematics and psychology and it is an example of heuristic algorithm (Moore et al. 2001). Prediction of Thermostability was the goal and the criteria level represented all the 17 features. The alternative level comprised of the proteins to be categorized as thermostable/mesostable proteins based on feature ranks. The second step was the construction of a pairwise comparison matrix for prioritizing the features in accordance to their contribution towards thermostabilizing proteins.

Generation of feature weights for thermostability factors

In order to form a pairwise comparison matrix the features were given weights. The weights for thermostability were derived at by the formation of a difference matrix by subtracting the values of the features of the Mesostable Protein (MP) set from the Thermostable Protein (TP) set. Then the matrix was represented by 0 and 1 vectors (Appendix III Table A3.1) where 0 is attributed to a decrease or equality in the feature and 1 to an increase corresponding to the TP set. From the difference matrix the percentage weight i.e., the number of proteins showing an increase for each feature or having the value of 1, in the total set of 127 proteins were calculated. Finally the weights for both the cases were scaled down to a 1-9. Through this method the priority of all the features according to its importance in contributing

towards thermostability was thus assigned. The method prioritized thermal stability as presented in Figure 4.5.



Features	II	MMH	CASA	IGT	SB	GT	CPI	HB	HI	MSH	SSH	AAI	NASA	PV	BT	ASI	PASA
Priority vector	0.091	0.091	0.081	0.081	0.081	0.072	0.071	0.071	0.071	0.051	0.051	0.040	0.040	0.040	0.030	0.020	0.010

Fig. 4.5. The priority vectors generated through AHP model.

Through this method the priority of all the features according to its importance in contributing towards thermostability of proteins were assigned. The CR of the pairwise comparison matrix of thermostability was calculated to be 0.002. Thus according to Saaty, 2008 the aforementioned judgement to derive at the priority vectors can be accepted as consistent as the consistency ratio (CR) value is less than 0.10 (Saaty et al. 2008).

Ranking obtained for features contributing to thermostability

Ionic interactions and main-chain to main-chain hydrogen bonds were the highest ranked features related to thermostability of proteins. Ionic interactions and

hydrogen bonds have been reported as major factors contributing towards thermostability (Ladenstein et al. 1998; Vetriani et al. 1998). Ionic interactions are long-range interactions between charged residues and upon unfolding of a protein energy is required for breaking up these coulombic interactions. In addition, the amount of energy required for breaking a charge-charge interaction raises with temperature, due to a decrease of the dielectric constant of water (Pace et al. 2000). Thus the desolvation penalty paid on the formation of ionic interaction decreases, favoring stability (Elcoc, 1998). Mutations in charge-charge interactions are also beneficial as they are located in the surface and are less likely to affect protein activity. Vetriani et al. in 1998 studied glutamate dehydrogenases from *Pyrococcus furiosus* and *Thermococcus litoralis* whose optimal growth temperatures are 100°C and 88°C by homology-based modeling and direct structure comparison. They suggested that extensive ion-pair networks may provide a general strategy for manipulating enzyme thermostability (Vetriani et al. 1998).

The results also rank main-chain to main-chain hydrogen bonds or oxygen amide hydrogen bonds as most important contributor towards thermostability of proteins. In such hydrogen bonds the donor (NH) and the acceptor (CO) atoms come from the backbone. They have also been reported by Sadegi et al. through the analysis of high-quality dataset containing 60 structures of thermophilic proteins and their mesophilic homologues, to be the most important factor contributing towards protein thermostability (Sadeghi et al. 2006). This class of the hydrogen bond also has been reported to be the main factor in the secondary and tertiary structure formation of proteins (Sadeghi et al. 2006). It has been said that lowering of configurational entropy stabilizes a protein. Backbone-backbone H-bonds are considered to have lower configurational entropy. Comparatively higher configurational entropy is assumed for two nearby residues that are not involved in backbone-backbone H-bonds (Guerois et al. 2002). Another interesting observation was that overall percentage of hydrogen bonds in a protein occupied the seventh rank. This indicates that all type of hydrogen bonds do not equally contribute towards thermostabilizing proteins. Thus it is clear that when mutating a protein, increasing the oxygen amide hydrogen bonds will yield better result. It was also interesting to note that inverse γ-

turns, salt bridges and charged exposed surface area were ranked as the third, fourth and fifth major factor contributing towards thermostability. In γ- turns hydrogen bond forms between one main chain carbonyl oxygen to the main chain N-H group 2 residues along the chain. Such hydrogen bonds are also known as short strong hydrogen bonds with a distance <2.7Å.

When classical and inverse γ- turns were considered together as a feature for protein thermostability, it was ranked sixth. This indicates that classical gamma turns is not that much important in thermostabilizing proteins. They are also rare in protein structures. Another notable feature was that non polar accessible surface area, aromatic aromatic interactions, beta turns, aromatic sulphur interactions and polar accessible surface area occupied the lowest ranks. This indicates that very few thermostable proteins show increase in these factors when compared to the mesostable proteins. The rationale behind this is unknown and yet to be explored. It has been always reported that single dominating stabilization mechanism has not evolved in proteins, rather, their stability results from a multitude of local improvements of interactions (Ladenstein 2008). Efforts to prioritize all such factors thus have been for the first time attempted.

RankProt: Validation and Accuracy

To validate the thermostability predictor, the first blind test involving a set of 100 thermostable/mesostable pairs were randomly chosen from the dataset as alternatives (Appendix III, Table A3.3). Table 4.3 represents the accuracy and ranking capacity of the predictor.

Table 4.3. Validation and accuracy of RankProt

Blind test set	No. of proteins	Homology	Accuracy%	Mean Rank	Rank Difference
1	100 TP	Homologous	91	0.54	0.09
	100 MP	Homologous		0.45	
2	40 MP	Homologous		0.49	0.003
	40 MP	Homologous		0.50	

The accuracy of RankProt for thermostability was calculated to be 91%. In a second blind test, when the mesostable-mesostable protein pairs were compared rank value difference was calculated to be 0.003. This is much less than the rank value difference of the thermostable-mesostable protein pairs. Here it can be conclusively said that this method could correctly differentiate between thermostable proteins from their mesostable proteins. The higher rank value difference for the thermostable-mesostable set also strongly indicates that the priority set to the 17 intra-protein interactions are correct and this priority vectors can thus be conveniently utilized to distinguish among thermostabilizing and non-thermostabilizing mutations. Any mutation in a protein which leads to a rank of 0.54 or higher when compared to its mesostable homologue through this method will predict thermostabilization of the protein.

The third blind test, involved 5 thermostable mutants of *Bacillus subtilis* lipases with optimum temperature >50°C. These were ranked against their mesostable wild type counterpart (1i6w) through RankProt. The results have been presented in Table 4.4.

Table 4.4. Ranking of thermostable mutants of *Bacillus subtilis* lipases

Sl. No	Mutant	Ranks	T °C	No of mutations	Mutations	Method of mutation
1	1t2n	0.546	55	3	L114P, A132D, N166Y	Error-prone PCR
	1i6w	0.453	35	0	Wild Type	
2	1t4m	0.541	55	2	A132D, N166Y	Error-prone PCR
	1i6w	0.458	35	0	Wild Type	
3	3d2c	0.573	60	9	A15S, F17S, A20E, N89Y, G111D, L114P, A132D, I157M, N166Y	Directed evolution
	1i6w	0.426	35	0	Wild Type	
4	3qmm	0.566	78	12	A15S,F17S,A20E,N89Y,G111D,L114P,A132D,M134E,M137P,I157M,S163P,N166Y	Directed evolution
	1i6w	0.433	35	0	Wild Type	
5	3qzu	0.537	60	7	R33Q, D34N, K35D, K112D, M134D, Y139C, I157M	Iterative saturation mutagenesis with randomization sites chosen on the basis of the highest B-factors
	1i6w	0.462	35	0	Wild Type	

*T-Temperature

It is clear that RankProt was successful to assign higher ranks to the thermostable mutants. As the ranks are relative and obtained by matrix normalization with the column sums equal to 1, the rank value of the mesostable protein differs from case to case and is not a constant. Thus the rank value of the subject protein is compared upon this rank and is case dependent. Thus it can be conclusively said that the tool RankProt can correctly identify thermostabilizing mutants. Moreover it can evaluate the role of multiple mutations at one go. It can also provide information

regarding the changes in the physicochemical properties of the protein which led to the enhancement of thermal stability of the mutated enzyme structure.

The fourth blind test involved a total of 104 mutated and wild type structures of bateriophage T4 lysozyme were retrieved from PDB. The wild type structure, 2lzm has Tm of 41.9 (Matsumura et al. 1989). The mutants were all gain of function mutations. The stability was reported to increase in the mutants. Therefore to test RankProt the mutants and the wild type were ranked (Appendix III Table A3.4). RankProt performed extremely well in identifying the stable mutants by providing them with higher rank corresponding to the wild type. A snippet of the Table has been presented here in Table 4.5. Out of the 104 proteins enumerated, 99 were provided with the correct rank.

Table 4.5. Snippet of the ranking obtained for bacteriophage T4 lysozyme and its mutants

Mutant	*T ^{o}C	Rank_Mutant	Rank_wild type
1dya	53.08	0.52	0.43
1dyb	53.08	0.5	0.45
1dyc	68.3	0.52	0.43
1dyd	64.7	0.5	0.45
1dye	53.08	0.51	0.44
1dyf	53.08	0.5	0.44
1dyg	53.08	0.51	0.44
1l00	66.3	0.5	0.45
1l02	53.6	0.51	0.44
1l03	42	0.5	0.42
1l04	42	0.52	0.43
1l06	53.6	0.5	0.42

*T-Temperature

The fifth blind test involved a total of 47 structures of loss of stability mutants of human lysozyme (PDB ID: 1lz1). They were ranked against the wild type which has a melting temperature of 64.9^{o}C (Takano et al. 1997). Therefore according to

concept the mutants should be ranked lower than the wild type. Out of 47 structures, 42 were given lower ranks in comparison to the wild type protein. Table 4.6 is a snippet of the full table provided in Appendix III Table A3.5. Therefore RankProt is able to efficiently recognize thermo stabilizing mutants.

Conclusively it can be said that in ranking predicted mutations with RankProt, the resultant rank of thermostable mutations is expected to be higher and also the rank of thermostable proteins is expected to be higher than its mesostable protein. Therefore when a protein with its mutated structure is considered as alternatives for ranking, the mutated protein will receive a higher rank if and only if the mutation contributes positively towards thermostability by increasing the highest ranked features through AHP. Therefore RankProt serves both the purpose of identifying thermostable mutations and classifying thermostable proteins.

Table 4.6. Snippet of the ranking obtained for human lysozyme and its mutants

Mutant	T°C	Rank_Mutant	Rank_wild type
2bqb	46	0.39	0.41
2bqc	47	0.38	0.41
2bqd	42.9	0.39	0.41
2bqe	44.3	0.39	0.41
2bqf	46	0.39	0.41
2bqg	46	0.39	0.41
2bqh	49.5	0.39	0.41
2bqi	39.9	0.39	0.41
2bqj	42.2	0.39	0.41
2bqk	44.4	0.39	0.41
2bqm	47	0.39	0.41
2bqn	44	0.39	0.41

*T-Temperature

4.5. Conclusions

Directed evolution approaches often do not provide a minimalist design for obtaining a desired property in proteins. Furthermore ranking or prioritizing thermostabilizing protein features has not been performed as it is complicated

because of the interplay of each of the factors in protein stability. Moreover literature still lacks information on which factors is most important to render proteins stable at such extreme milieu. In this work we have tried to prioritize a statistically significant set of 17 physicochemical features for thermostability prediction through a novel multi criteria decision making approach; Analytical Hierarchical Process. This work was also successful to come up with a scoring model which can differentiate thermophilic/mesophilic proteins. A set of 127 structural homologous thermostable/mesostable protein structures formed the final dataset. The problem was decomposed into hierarchies and the factors ranked with the aid of eigen vectors. Ionic interaction and main chain to main chain hydrogen bonding were given the highest priority for conferring thermal stability. This resulted in the development of a tool RankProt using the priority vectors for thermostability. Further the tool was validated through blind tests. A random set of 100 proteins were ranked and the thermostable proteins were assigned an average rank value of 0.54. The accuracy of the method was calculated to be 91%.

Three case studies to check the efficiency of RankProt were carried out with 5 thermostable mutants of *Bacillus subtilis* lipase, 100 thermostable mutants bacteriophage T4 lysozyme and 47 loss of stability mutants of human lysozyme. In all the three cases RankProt successfully recognized thermostabilizing mutants. Thus it can be conclusively said that this method can successfully identify thermostabilizing mutations. Moreover the edge of this method is that multiple combinations of mutations can be prioritized at a single go with higher rank assigned to the more stabilizing ones. In summary an efficient ranking model has been developed which can be used stand-alone on Centos Linux platform as VADAR is not web compatible (Fortran program) and also requires F77 compiler. The software package can be downloaded on request and the download link is available on the web interface of Thermostable Protein Structural Database.

Chapter V

Attaining Plausible Mutations to Enhance Protein Thermostability

Prologue

From the previous chapter, it was observed that higher ranked proteins in comparison to their mesostable counterparts were predicted to be better in thermostability when compared to their mesostable counterparts. Further a rank value of 0.54 for a protein undergoing mutations was predicted to be thermostabilizing and the accuracy of the method was calculated to be 91%. The ranks were used to develop a tool called RankProt. To further validate the method, in this chapter, RankProt has been utilized for predicting multiple point mutation(s) that can contribute positively in thermostabilizing proteins. RankProt was used to predict double mutations for mesostable *Bacillus subtilis* lipase A with rank value <0.54 (mut 1) and another double mutant with rank value >0.54 (mut 2). The resulting double mutants were further evaluated to be stabilizing by performing 30ns molecular dynamics simulations, molecular docking and intra-protein contact map analysis of the mutant and wild type structures. In all these study mut 2 was found to be better in stability than mut 1, thus validating RankProt. The features that increased in mut 2 causing enhanced thermostability have been elaborated. Furthermore molecular docking with (C8) substrate gave lower binding energy for mutants ensuring intactness of catalytic pocket. Contact map analysis highlighted that number of unique contacts in predicted thermostable mutants is much higher than the wild type structure. It can be conclusively said here that RankProt successfully predicted thermostabilizing mutations in *Baclillus subtilis* lipase and can be further utilized to predict thermostabilizing mutations in any test protein.

Outputs:

1. Clone of thermostable mutant of *Bacillus subtilis* lipase.

5.1. Introduction

Untill now, the factors that enhance thermostability have been ranked and RankProt has been developed to predict thermostabilizing mutations. To further validate RankProt, in this chapter, wild type *Bacillus subtilis* lipase A protein stable at 35^{o}C (Acharya et al. 2004), was chosen as the model enzyme for carrying out mutations. In the preceding chapter, the factors that enhance thermostability have been ranked and RankProt has been developed to predict thermostabilizing mutations. In Chapter III, thermostable lipases have been studied in details and their industrial importance has been brought forward. Thus the study was instrumental in making the choice for a model enzyme to perform mutations. It was also observed that most of the thermostable proteins are lipases. Therefore, to further validate RankProt, in this chapter, wild type *Bacillus subtilis* lipase A protein stable at 35^{o}C (Acharya et al. 2004), was chosen as the model enzyme for carrying out mutations. *Bacillus subtilis* lipase is an industrially important enzyme especially because *Bacillus subtilis* lipase is that it lacks lid and thus will not operate in oil water interphase (Dartois et al. 1992; Acharya et al. 2004; Westers et al. 2005; Srivastava et al. 2014). Earlier thermostable structures of *Bacillus subtilis* lipase were obtained by directed evolution approaches. A total of 12 progressive mutations were performed with the conclusion that accumulation of mutations increases temperature stability of the lipase (Acharya et al. 2004; Ahmad et al. 2008; Kamal et al. 2011). Further crystallization of the structures led to the elucidation of the factors that contributed in increasing their temperature stability. Therefore the aim was to validate RankProt in designing thermostabilizing mutations that can be performed through site- directed mutagenesis to increase thermostability of the lipase.

To further study the features that lead to thermostability after mutation is performed, was studied through contact map analysis and molecular dynamics simulations. Contact map analysis provides the study of residue interaction network. Studying such interactions can provide insights to protein folding (Vendruscolo et al. 1997; Bagler and Sinha, 2007) and stability (Brinda and Vishveshwara, 2005). To date, globular proteins have been most effectively studied using contact maps (Bagler

et al. 2005). Furthermore molecular dynamics has been employed to support the results drawn from RankProt, contact map and physicochemical feature analysis. Further it was employed, to gain insights into the effect of mutations on the dynamics of the lipase. The schema of the present study has been illustrated in Fig. 5.1.

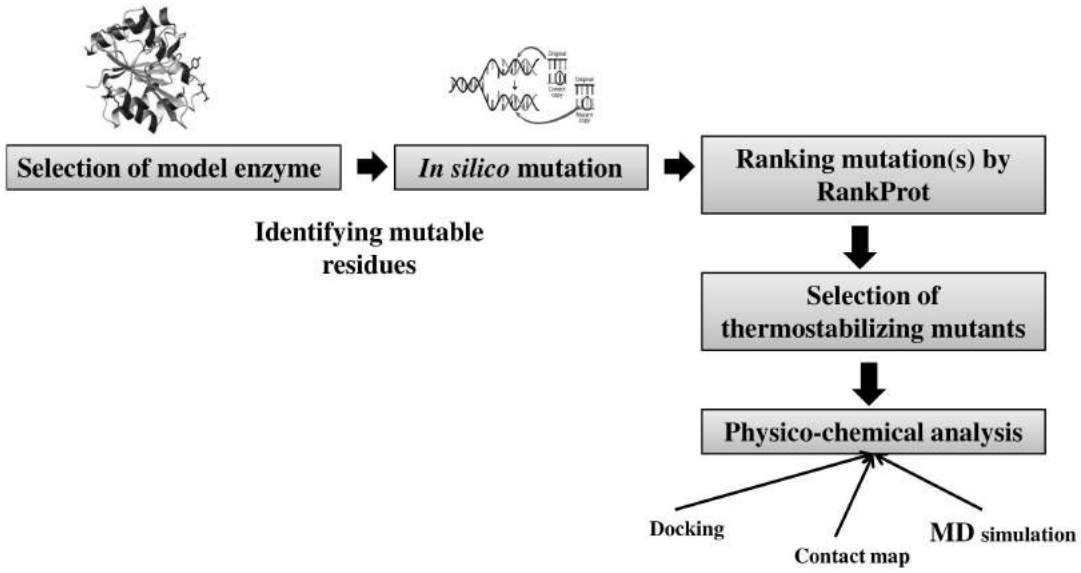


Fig. 5.1. Rationalised approach to attain protein thermostability.

5.2. Methodology

5.2.1. Selecting model enzyme for experimentation

As a model for validating RankProt, *Bacillus subtilis* lipase A was chosen. Additional comparative analysis was further performed for physicochemical features of 4 already available engineered thermostable structures of *Bacillus subtilis* lipase (1t2n, 1t4m, 3d2c, 3qmm).

5.2.2. *In silico* mutagenesis

The sequence for the wild type lipase of *Bacillus subtilis* 168 (1i6w) sequence was collected from Uniprot Knowledgebase release 2010_06 (Uniprot ID: P37957). Analysis of stabilizing center residues was done through Scide (Dosztanyi et al. 2003) and Sride (Magyar et al. 2005). Stabilizing mutations were carried out in mesostable *Bacillus subtilis* 168 lipase (PDB: 1i6w). Mutations were carried out through CHIMERA and homology modeling through I-Tasser software, using the wild type structure as template. Mutational stability of the residues was also analyzed through the developed RankProt tool, I-Mutant2 (Capriotti et al. 2005), Cupsat (Parthiban et al. 2005) and ERIS (Yin et al. 2007) web servers. Physicochemical characterization was performed using Vadar ver 1.8 web interfaces, Promotif and the developed python tool for calculating intra-protein interaction (IPI enumerator). Docking with Triacylglycerol (C8) substrate was performed through Autodock. Wild type and mutated structures were superimposed by PyMol V0.99. The ranks of the mutated structures were calculated through RankProt tool. Only those mutations that led to an increase in the rank of the mutated structures w.r.t. the wild type were chosen for further *in vitro* validation.

5.2.3. Contact map analysis

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 (Barah and Sinha, 2008; Bhavani et al. 2011). Calculation and visualization of protein contact maps were performed using CMView (Vehlow et al. 2011). The network is a graph where each residue corresponds to a node and two nodes are connected by an edge if and only if the two residues are in contact. Two residues are considered to be in contact if they are spatially close in the three dimensional structure. The contact type and the distance cutoff are provided by the tool (Vehlow et al. 2011). The contact type defines a subset of atoms of the residue. Two residues i and j are then in contact if two atoms out of this subset, one from i and one from j, are not further apart

than the distance cutoff. Therefore, combined contact map and 3D structure visualization was performed through PyMol software. Contact maps of the 2 structures that have been mutated with the aid of RankProt were also compared to the wild type structure. As controls of this experiment, comparative analysis of 4 engineered structures of *Bacillus subtilis* lipase (1t2n, 1t4m, 3d2c, 3qmm) and their wild type structure (1i6w) was performed. Number of unique contacts in each structure was calculated.

From previous chapters it was deduced that hydrogen bonds play the most important role in protein thermostability. Therefore, HB-plot tool (Bikadi et al. 2007) was employed to analyze the network of hydrogen bonds in wild type and mutated structures. The program was instrumental in this study as it can distinguish between main chain – main chain, main chain – side chain and side chain – side chain hydrogen bonding interactions (Bikadi et al. 2007).

5.2.4. Molecular dynamics simulation

All MD simulations were carried out using GROMACS-4.5.3 in conjunction with the OPLSA force field. The starting structures of wild type (WT) and the two mutated structures were then subjected to mutation. After adding the hydrogens, the protein structures were solvated with TIP3P water molecules in a cubic solvent box with a minimal distance between the protein surface and box of 0.9 nm. Ions were added to the system by replacing 5 solvent atoms with Cl^{-1} ions. Energy minimized using the steepest descent algorithm, until it converged with a force tolerance of 100 kJ mol^{-1} nm^{-1}. After minimization, each system was equilibrated to the desired temperature through a stepwise heating protocol in the NVT ensemble followed by 500 ps equilibration in the NPT ensemble with position restraints on the protein molecule followed by an equilibration step of 500ps without position restraint on the protein. Finally, a production simulation was performed for each system at four different temperatures, i.e., at 320, 330, 350 K for 30 ns under periodic boundary conditions without any restraints on the protein. The temperature and pressure (1 bar) were controlled by using a velocity-rescale thermostat (Bussi et al. 2007) and a

Parrinello–Rahman barostat (Parrinello et al. 1981) with a temperature and pressure coupling time constant of 1.0 ps. Nonbonded interactions were calculated using the particle-mesh Ewald (Essman et al. 1995) method with a cutoff of 0.9 nm. The LINCS algorithm (Hess et al. 1997) was used to constrain all bonds involving hydrogen atoms during simulation.

Analysis of MD Simulation Trajectories

Secondary structure analysis was performed using the DSSP51 program. Other analyses such as root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF), radius of gyration, solvent accessible surface area (SASA), hydrogen bonds, and salt bridges were performed using tools within the GROMACS simulation package. RMSD calculation was done using the starting structure of each simulation as a reference. For hydrogen bond calculations, a donor–acceptor cutoff distance of 0.35 nm and acceptor–donor–hydrogen bond angle cutoff of 30° were considered. The visual analysis of structures and preparation of figures was carried out using Pymol, VMD and Xmgrace.

5.3. Results and Discussion

5.3.1. Selecting model enzyme for experimentation

Bacillus subtilis lipase was the model enzyme on which thermostabilizing mutations were predicted. The other enzymes used in this study as control with their temperature stability and features rendering them thermostable have been presented in Table 5.1. The reported features in the Table 5.1 have been presented after analysis of these crystallized structures from the work of Acharya et al. 2004 and Srivastava et al. 2014.

Table 5.1. Comparative analysis of reported and predicted features in thermostability of *Bacillus subtilis* lipases

Mutant	T °C	No of mutations	Mutations	Reported Features
1t2n	55	3	L114P, A132D, N166Y	Anchoring of C-ter to rest of the protein
1t4m	55	2	A132D, N166Y	Improved solvent interaction, stacking interaction
3d2c	60	9	A15S, F17S, A20E, N89Y, G111D, L114P, A132D, I157M, N166Y	H-bonds and salt bridge formation, loss of intrinsic flexibility
3qmm	78	12	A15S,F17S,A20E,N89Y,G111D,L114P, A132D,M134E,M137P,I157M,S163P,N166Y	Loop stabilization, hydrogen bonds, salt bridge

All these available mutated thermostable structures of *Bacillus subtilis* lipase were observed to be ranked higher than the wild type structure (>0.5) by RankProt in Chapter IV. Furthermore the rank value was observed to increase with the increase in temperature stability of the mutants. Thus it can be concluded that, RankProt can successfully identify thermostabilizing mutations. Therefore it can be utilized to predict thermostabilizing mutations in *Bacillus subtilis* lipase. Physico-chemical feature analysis of 4 thermostable and mesostable structure (1i6w) of *Bacillus subtilis* lipase uncovered that this protocol can identify thermostabilizing features accurately as the deduced features corroborate earlier reports (Acharya et al. 2004; Srivastava et al. 2014).

Interestingly other than the mentioned features that led to thermostabilization, γ-turns were observed to increase in the mutated *Bacillus subtilis* lipases. Therefore

enhancement of γ-turns can be attempted for by performing mutations to increase thermostability of the wild type lipase.

5.3.2. *In silico* mutagenesis

Homology modeling and docking studies

Bacillus subtilis 168 lipase A is a 181 amino acid long protein with S77, D133 and H156 forming the catalytic triad residues. More than 60 stabilizing mutations, predicted by servers mentioned in Table 5.2, were carried out in wild type *Bacillus subtilis* 168 lipase A sequence to increase thermostability. Out of these 60 mutations, 18 combinations of double mutations were ranked higher by RankProt. Finally 2 double mutations with highest ranks were chosen for *in silico* mutagenesis. The mutated proteins were modeled by homology using the wild type structure as a template. Molecular superimposition in PyMol showed low RMSD values of 0.278 (mut 1 and 1i6w) and 0.312 (mut 2 and 1i6w), indicating that the wild type and mutated structures were alike without any massive structural changes.

Table 5.2. Mutations carried out on *Bacillus subtilis* structures

Residue position	Hot Spot	Old Residue	New Residue	Cupsat	I-mutant	iPTREE-STAB	ERIS	pH	T°C
47	Average	T	N	S	I	S	S	8	30-90
121	Average	Q	N	S	I	S	S	8	30-90
47	Average	T	S	S	I	S	S	8	30-90

* S=stabilizing mutation; I=Increase in stability; DS=Destabilizing mutation; *T-Temperature

Ranking via RankProt

The mutation presented in Table 5.2 were further ranked through RankProt. All these structures were ranked through RankProt to analyze the effect of mutations on the rank value. The designed stabilizing and destabilizing mutations that were further analyzed have been presented in Table 5.2. The aforementioned two combinations of double mutations were modeled and ranking through RankProt was performed. Only two double mutations gave highest rank value with respect to the wild type (Table 5.3).

Table 5.3. The ranks given by RankProt for *Bacillus subtilis* lipase mutated structures

Mutations	Mutated	1i6w
T47S, Q121N (mut 1)	0.50	0.49
T47N, Q121N (mut 2)	0.54	0.45

Analysis of 17 physicochemical features that are calculated by RankProt to enumerate ranks of the mutations, have been represented in Table 5.4. The percentage of features that increased in mutants' w.r.t wild type has been colour coded in ash colour, the features which dereased in orange, the features having equaled weights in white.

Table 5.4. Physicochemical features of mutated and wild type structures of *Bacillus subtilis* lipase

Mutations	%BT	%GT	%IGT	%HB	%SB	%II	%HI	%PV	%NASA	%PASA	%CASA	%ASI	%CPI	%AAI	%MMH	%MSH	%SSH
1i6w (WT)	14.46	0.55	0.55	3.11	0.37	0.33	5.56	85.8	60	27	13	0.03	0.07	0.03	8.21	2.29	2.37
T47S,Q121N (mut 1)	12.8	1.11	1.11	2.31	0.86	0.43	4.75	84.06	65	23	12	0.03	0.09	0.03	8.21	2.50	2.93
T47N,Q121N (mut 2)	12.8	1.11	1.11	3.09	0.59	0.29	5.66	84.51	61	25	14	0.03	0	0.03	8.30	2.68	2.64

*WT: wild type. Refer to abbreviations for the full text of the headers. The percentage of features that increased in mutants w.r.t wild type have been represented in ash colour, the features which dereased in orange, the features having equal weights in white.

From Table 5.4 the analysis of physicochemical properties showed that there was increment in γ- turns, salt bridges, ionic interaction, cation-π interaction (CPI), non-polar accessible surface area (NASA), main-chain side-chain (MSH) and side-chain side-chain (SSH) hydrogen bonds in mut 1. Again salt bridge (SB), γ-turns (GT and IGT), charged accessible surface area (CASA), hydrophobic interaction (HI), main-chain main-chain (MMH), main-chain side-chain (MSH) and side-chain side-chain (SSH) hydrogen bonds increased in mut 2. These intramolecular interactions have been implicated to increase thermostability of proteins by numerous research publications (Kumar et al. 2000; Vogt et al 2001; Sadeghi et al. 2006; Srivastava et al. 2014). It was interesting to note that packing volume decrease in both the mutants. This suggests that the mutants were more compact than the wild type structure. Compactness has been earlier proposed to enhance temperature stabilily by increasing protein rigidity (Russell et al. 1997). Beta turn was observed to decrease in both the mutants. Unfortunately the rationale behind the same is not known and has to be researched upon.

Further analysis of the secondary structure of WT and mutants further uncovered that the new γ-turns formed in the mutants were inverse and formed near

helix 7. Thus such turns may be involved in stabilization of the helices by clipping through the formation off intra-molecular hydrogen bonds.

Conclusively it can be reported that the highest ranked features (Chapter IV) increased in mut 1 and mut 2. However as mut 1 shows decrease in main-chain to main- chain hydrogen bonds (MMH) and charged accessible surface area (CASA) it gets a lower rank than mut 2. Thus it can be predicted that mut 2 will be more thermostable than mut 1. The overall properties differed with different mutations. Earlier studies have shown that the gamma-turn structure is fairly common in proteins. In a protein γ-turn was first observed by (Matthews et al 1972) in thermolysin. However he could not conclusively report that this could be the factor for thermo stabilization of Thermolysin. It has also been postualted that inverse γ-turns can be involved in the folding process leading to protein stability (Milner-White et al. 1990). Recently, γ-turns have brought attention through studies of peptide mimetics which lead to stable, receptor ligands (Alkorta et al. 1996).

Thus increase stability is a result of the cumulative effect of all such factors. Ranking further confirmed that the designed positive mutations for the mesostable lipase were thermostabilizing and the negative mutations designed for the engineered thermostable structure was not thermostabilizing.

Table 5.5. Docking energy for wild type and mutants of *Bacillus subtilis* lipase

Lipases	**Binding Energy**	***kI**	**Unbound Extended Energy**
1i6w	-5.33	123.43uM	-0.42
mut 1	-4.48	516.29uM	-0.58
mut 2	-4.91	326.61uM	-0.48

The results obtained after molecular docking have been presented in Table 5.5. Molecular docking with (C8) substrate gave a binding energy of -4.48 (mut 1), -4.91 (mut 2) and -5.53 (1i6w) for the mutated and the native structures respectively. The binding pocket was intact. This shows that the mutations did not affect the activity of the lipase. Thus mutation did not disturb the catalytic properties.

5.3.4. Contact map analysis to enumerate the importance of predicted stabilizing mutations

Graphical representations of unique contacts in mut 1 and mut 2 in comparison to 1i6w have been illustrated in Fig.5.2. In Fig.5.2 black colour represents common contacts, pink for contacts unique to the 1i6w structure and green for contacts unique to the mutants. It can be clearly observed that number of unique contacts in mutants is much higher than the wild type structure. Furthermore superimposition of wild type and mutant structures with their unique contacts revealed that the unique contacts were more in the loop region of the 3D-structure. Fig 5.3 is the graphical representation about the analysis of total number of unique contacts for all the mutants. The unique contacts were observed to be greater than the wild type. Further, it can be observed that backbone unique contacts always increased in the mutants. Similar trends were observed for mut 1 and mut 2. 1t4m and mut 1 did not show any increase in unique contacts pertaining to their side chains. Also side chain unique contacts were lower for mut 2 and were more for 1t2n, 3d2c and 3qmm. The total number of contacts formed is higher than the number of contacts lost suggesting an increase in the compactness of the protein. Thus from Fig 5.2 and 5.3 it can be concluded that as the mutants generated through RankProt show similar trends of unique contacts with those of already reported thermostable mutants, they have high probability of being thermostable.

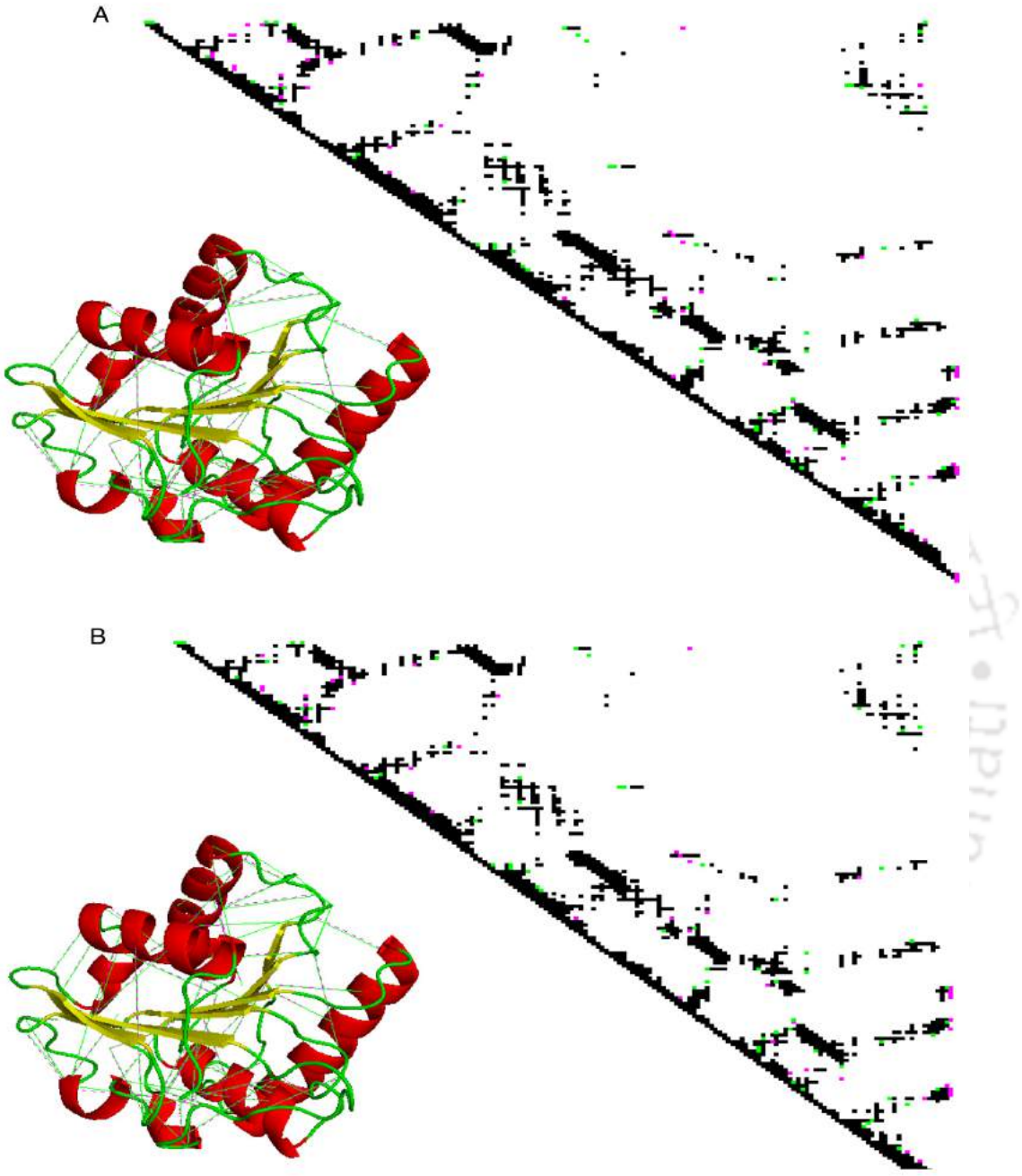


Fig.5.2. Graphical illustration of unique contacts formed in A) mut 1 vs. 1i6w along with the cartoon representation of mut 1 showing its unique contact. B) mut 2 vs. 1i6w along with the cartoon representation of mut 2 showing its unique contact.

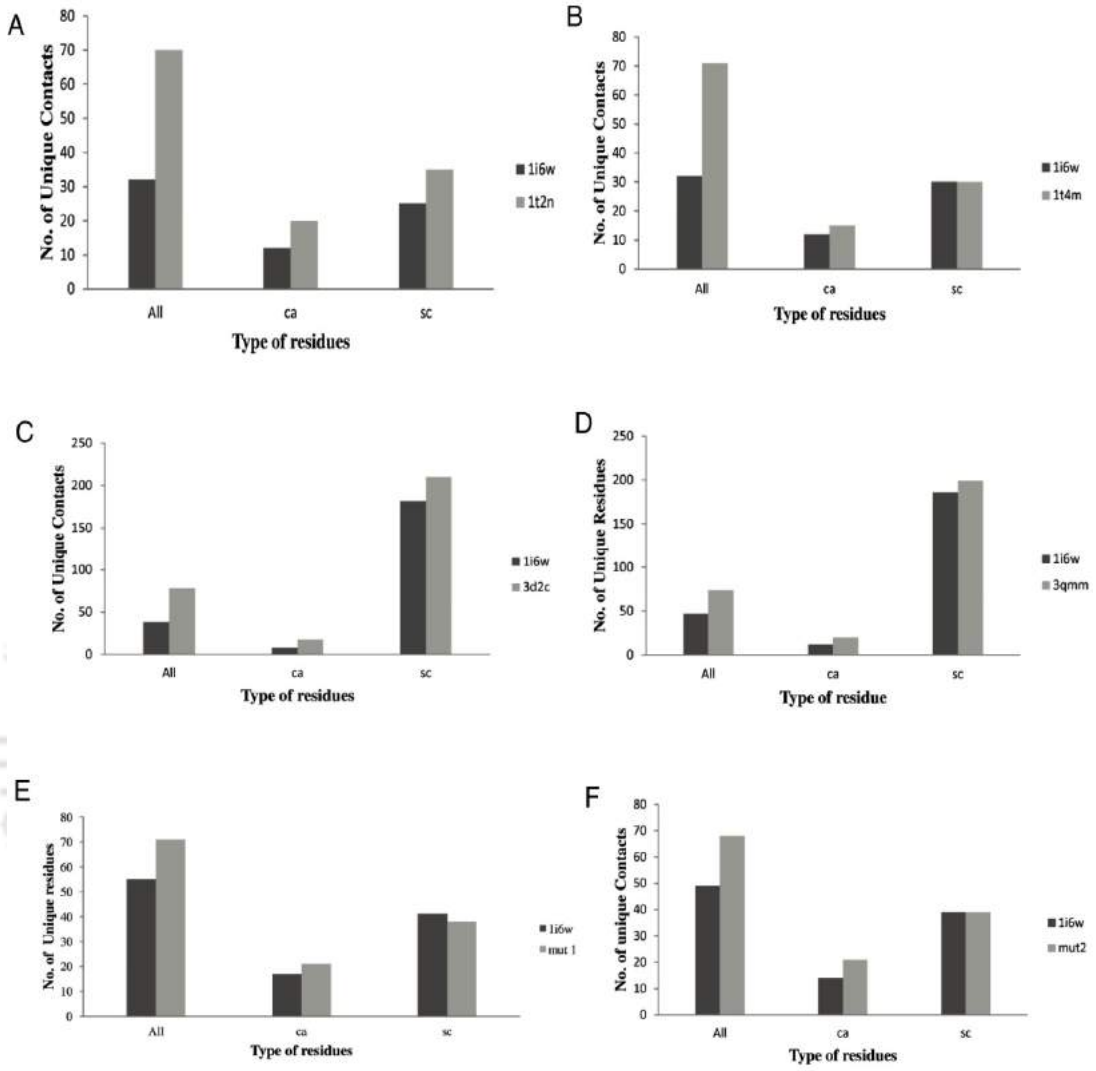


Fig 5.3. Comparative bar graphs of unique contacts in thermostable mutants of *Bacillus subtilis* lipase and wild type (1i6w). A) 1i6w vs. 1t2n; B) 1i6w vs. 1t4n; C)1i6w vs. 3d2c; D) 1i6w vs. 3qmm; E) 1i6w vs. mut 1; F) 1i6w vs. mut 2.

All represents all unique contacts in PDB structure.

ca represents backbone unique contacts and sc represents side chain unique contacts.

Hydrogen bond analysis of mutated and wild type structures

The HBplot analysis of hydrogen bond network in wild type and mutated structures of *Bacillus subtilis* lipase uncovered that main-chain main-chain hydrogen bonds increased in 4 mutated structures (3d2c, 1t4m, 3qmm, mut 1and mut 2). Main-chain side-chain and side-chain side chain hydrogen bonds increased in all the mutants (Fig 5.4). Along with increase in the number of hydrogen bonds it was observed that Hydrogen bonds <3Å were much greater for the mutants in comparison to the wild type structures. Therefore it can be concluded that a as the temperature stability increases, number of short distance hydrogen bond also increases (Fig 5.5).

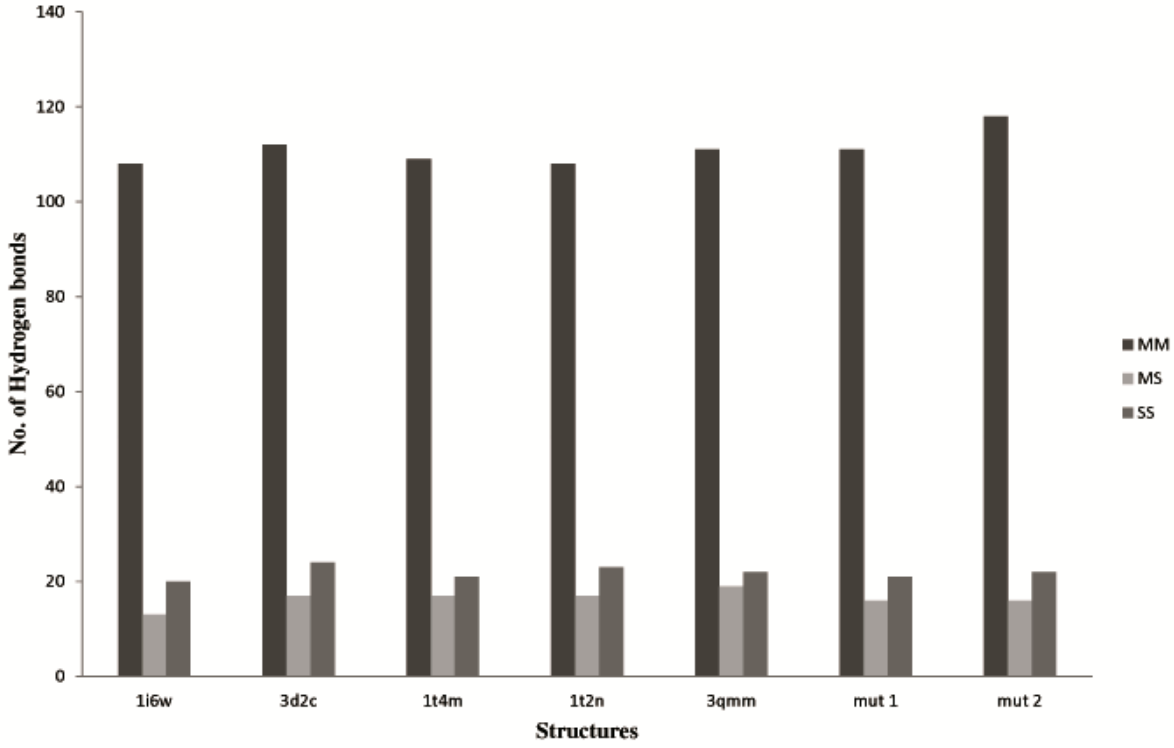


Fig 5.4. Number of hydrogen bond in wild type and mutated proteins.

MM: main chain-main chain; MS: main chain-side chain; SS: Side chain-side chain.

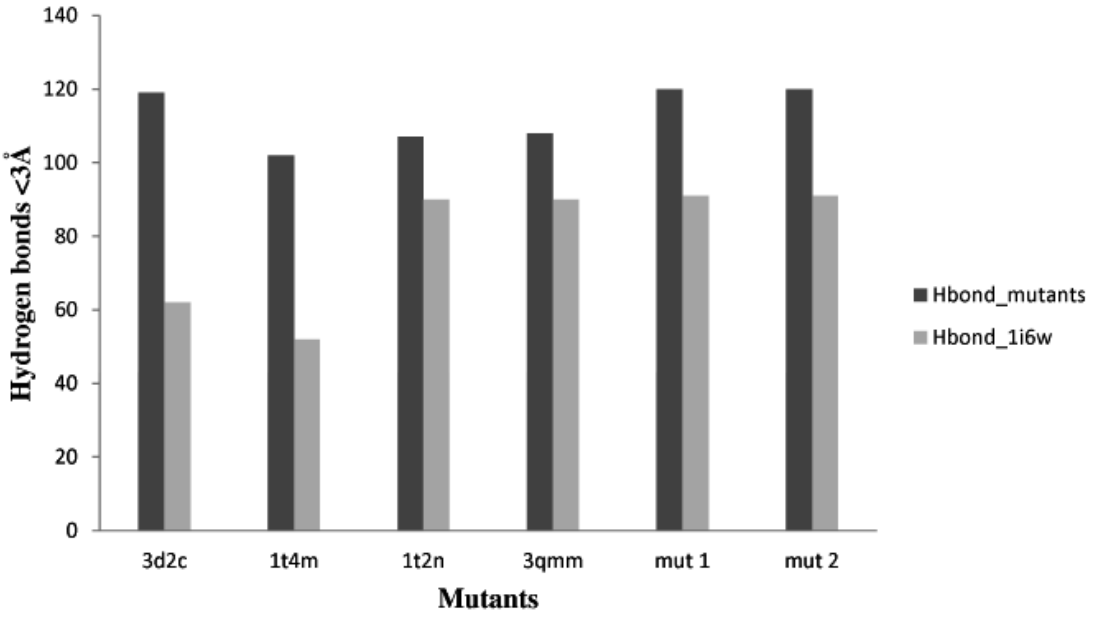


Fig. 5.5. Graphical illustration of the number of hydrogen bonds with distance <3Å in mutants and wild type.

Main chain and side chain hydrogen bonds have been previously linked to protein thermostability (Kumar et al. 2000; Sadeghi et al. 2006). Moreover it was observed that intramolecular hydrogen bonding networks increased near the β-strand and α-helices of the mesostable lipases. This can result in better packing due to pinning of helices and stands rendering them rigid to unfolding at elevated temperatures. These aforementioned results corroborate previous findings that increment in hydrogen bonds leads to thermostabilization of proteins. Increase in short distance hydrogen bonds can lead to better stability of proteins as it was reported that as distance becomes smaller, the charge-transfer contribution to the hydrogen-bond energy increases and the angle decreases (Kolman et al. 1972).

5.3.5. Molecular dynamics simulation analysis of the predicted thermostabilizing mutations of *Bacillus subtilis* lipase

MD simulation at higher temperatures for Wild type (1i6w: WT), mut 1 and mut 2 were performed because protein denaturation has been reported to occur in microsecond time scale (Duan et al. 1998). Therefore capturing protein unfolding at normal temperatures using molecular dynamic is rather difficult. This necessitates the use of much higher temperatures. It has been shown earlier that at higher temperature unfolding process is accelerated without alteration in unfolding pathway (Day et al. 2002). The difference in mut 1 and mut 2 was subtle at position T47 replaced by Ser in mut 1 and Asn in mut 2. Both the amino acids are hydrophilic in nature. The mutation at position Q121 was same for both mut 1 and mut 2 where it was replaced by Asn. Ranking provided by RankProt gave higher rank to mut 2 than mut 1 w.r.t. WT. Therefore it was assumed that mut 2 will be more stable than mut 1. RMSD, RMSF, Radius of gyration, hydrogen bonds and secondary structure analysis of the trajectory, were performed after 30 ns MD simulation at three different temperatures (320K, 330K and 350K). The average parameters of the analysis have been presented in Table 5.5 and discussed in the ensuing sections.

Table 5.5. Summarized parameters of MD simulations for wild type and mutants of *Bacillus subtilis* lipase

Variants	Rg (nm)	RMSD	RMSF
		320K	
1i6w (WT)	1.51±0.006	0.14±0.01	0.11±0.07
Mut 1	1.49±0.007	0.15±0.02	0.11±0.07
Mut 2	1.49±0.005	0.14±0.02	0.12±0.07
		330K	
1i6w (WT)	1.49±0.006	0.12±0.02	0.12±0.07
Mut 1	1.44±0.006	0.11±0.02	0.13±0.07
Mut 2	1.49±0.007	0.14±0.02	0.13±0.07
		350K	
1i6w (WT)	1.51±0.009	1.49±0.02	0.13±0.08
Mut 1	1.50±0.008	1.49±0.02	0.14±0.07
Mut 2	1.50±0.006	1.42±0.02	0.12±0.07

*Rg: Radius of gyration; RMSD: Root Mean Square Deviation; RMSF: Root Mean Squate Fluctuation; WT: wild type.

Root Mean Square Deviation (RMSD)

The degree of conformational changes of wild type lipase and its mutants identified is monitored by root mean square deviation of alpha carbon atom during the course of 30 ns of molecular dynamics simulation. In this work, the backbone RMSD of conformations from production run relative to its initial structure has been studied for wild type and the mutants and illustrated in Fig 5.6. The global average RMSD is similar for mut 1 and WT at all temperatures (Table 5.5). But at higher temperature of 350K the RMSD of mut 2 is much lower than mut 1 and wild type. This reflects that mut 2 is much stable having lower flexibility than mut 1 and WT at higher temperatures. Lower RMSD values indicates lower flexibility and thus stability of proteins (Singh et al. 2015). Kamal et al. (2012) reported low RMSD values for a thermostable mutant of *Bacillus subtilis* lipase A. Low RMSD value of thermostable mutant of Cocaine esterase was observed by Huang et al. in 2011. Singh et al. (2015) observed similar RMSD plots of wild type and *Bacillus subtilis* mutants (1t4n, 1t2n, 3d2a, 3d2b, 3d2c).

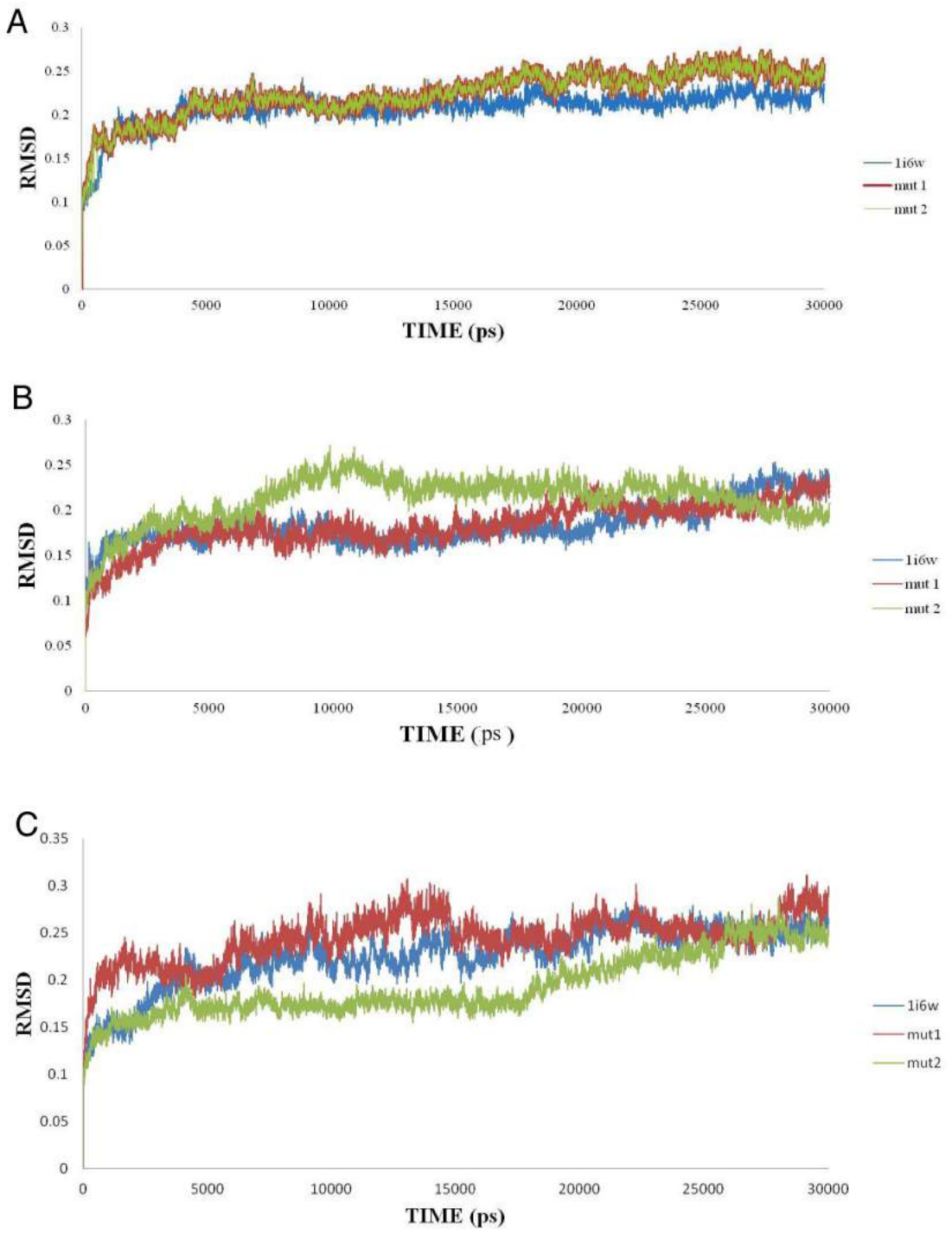


Fig.5.6. The Root-mean-square deviation (RMSD) to the starting structure as a function of time of WT and mutants during the 30ns time course of simulation is shown. (A) 320K. (B) 330K. (C) 350K.

Root Mean Square Fluctuation (RMSF)

The local deformability of the protein has been analyzed through root mean square fluctuations of alpha carbon atoms from MD simulations and illustrated in Fig. 5.7-5.9. Average RMSF plot shows global reduce in flexibility for mut 2 at 350K while mut 1 shows increase. Appreciable difference of RMSF between wild type and mutants is observed at the C- terminus. At 320K illustrated in Fig 5.7, the C-terminus is more stable in mut 1 and mut 2. At 330K, illustrated in Fig. 5.8, both N- and C-terminus are less flexible than 1i6w. This shows that mutations have resulted in lowering of flexibility of the mutants. At 350K illustrated in Fig 5.9, the flexibility of the WT and mutants can be arranged in descending order as mut 2> mut1> 1i6w (WT). The average value of RMSF shows that mut 2 has considerable reduction in the RMSF values globally. A balance in rigidity and flexibility at various region of a protein may lead to stability at high temperatures (Singh et al. 2015). Therefore it is important to deduce the difference in RMSF plots to understand the regions where flexibility or rigidity have ensued due to the mutations. It was also interesting to observe that the N-terminal and C-terminal of the mutant proteins have much more reduced flexibility than the wild type at elevated temperatures. It can also be seen that mut 2 shows much lower flexibility than mut 1 all throughout its structure and at elevated temperatures. Thus it can be comprehended that mut 2 will be more stable than mut 1. Kumar et al. (2000), Notomista et al. (2001) and Bhardwaj et al. (2012) reported the importance of protein termini on stability. Though Jacob et al. (2007) suggests that the terminal regions of a protein structure are more flexible and exposed to solvent and hence considered to have low influence on thermostability, further research carried out by Bhardwaj et al. (2010) reported that reduction in flexibility of N-terminal and C-terminal of proteins increases protein stability. Lower flexibility leads to rigidity of protein and thus stability. Therefore the obtained results corroborated previous work and suggest that reduced flexibility of N-terminal and C-terminal of mut 1 and mut 2 can play crucial role in their temperature stability.

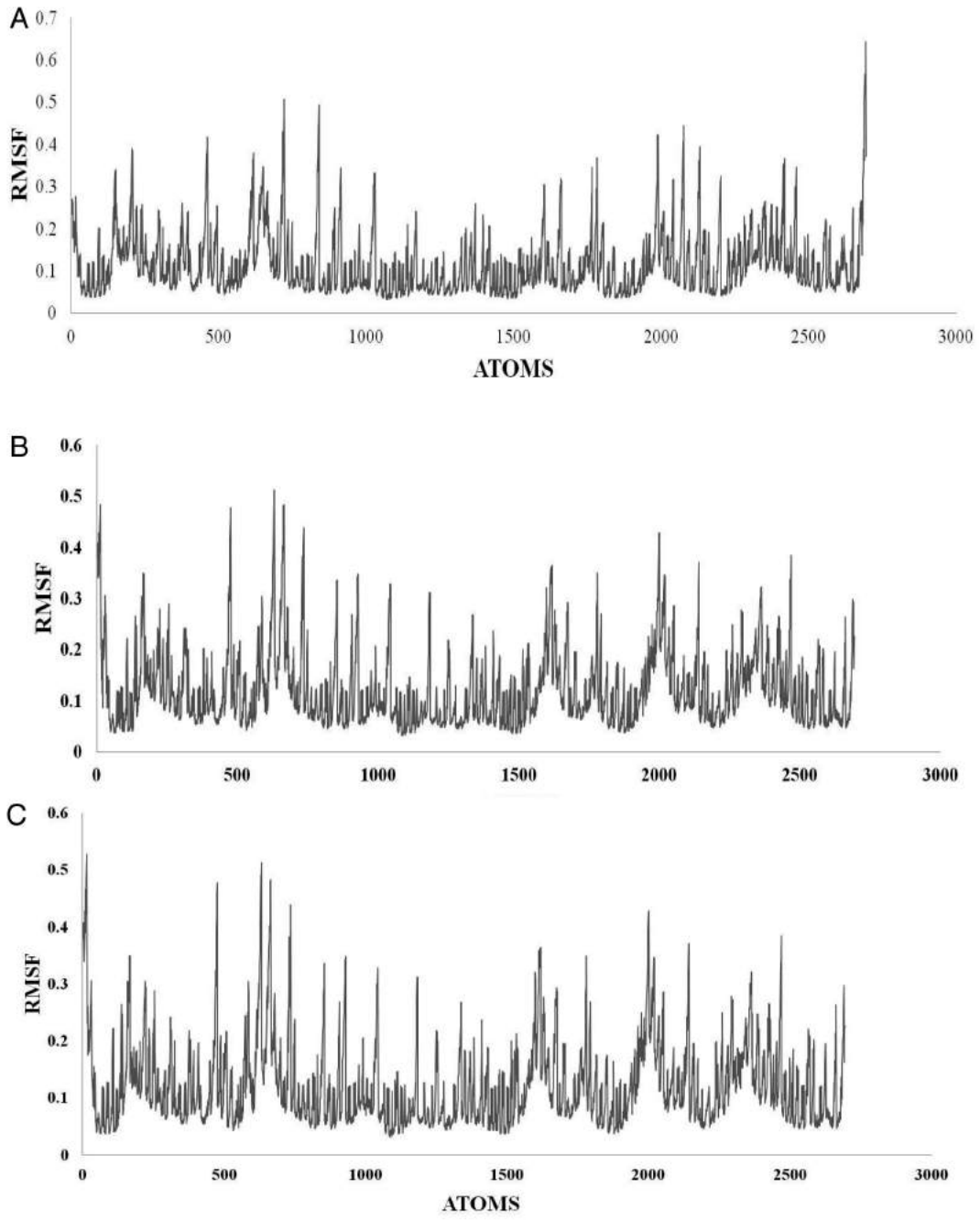


Fig.5.7. Root mean square fluctuation of α-carbon atoms as a function of atoms of 1i6w and its mutants from RMSF study at 320K during the 30ns simulation is shown. A) 1i6w at 320K, B) mut 1 at 320K, C) mut 2 at 320K.

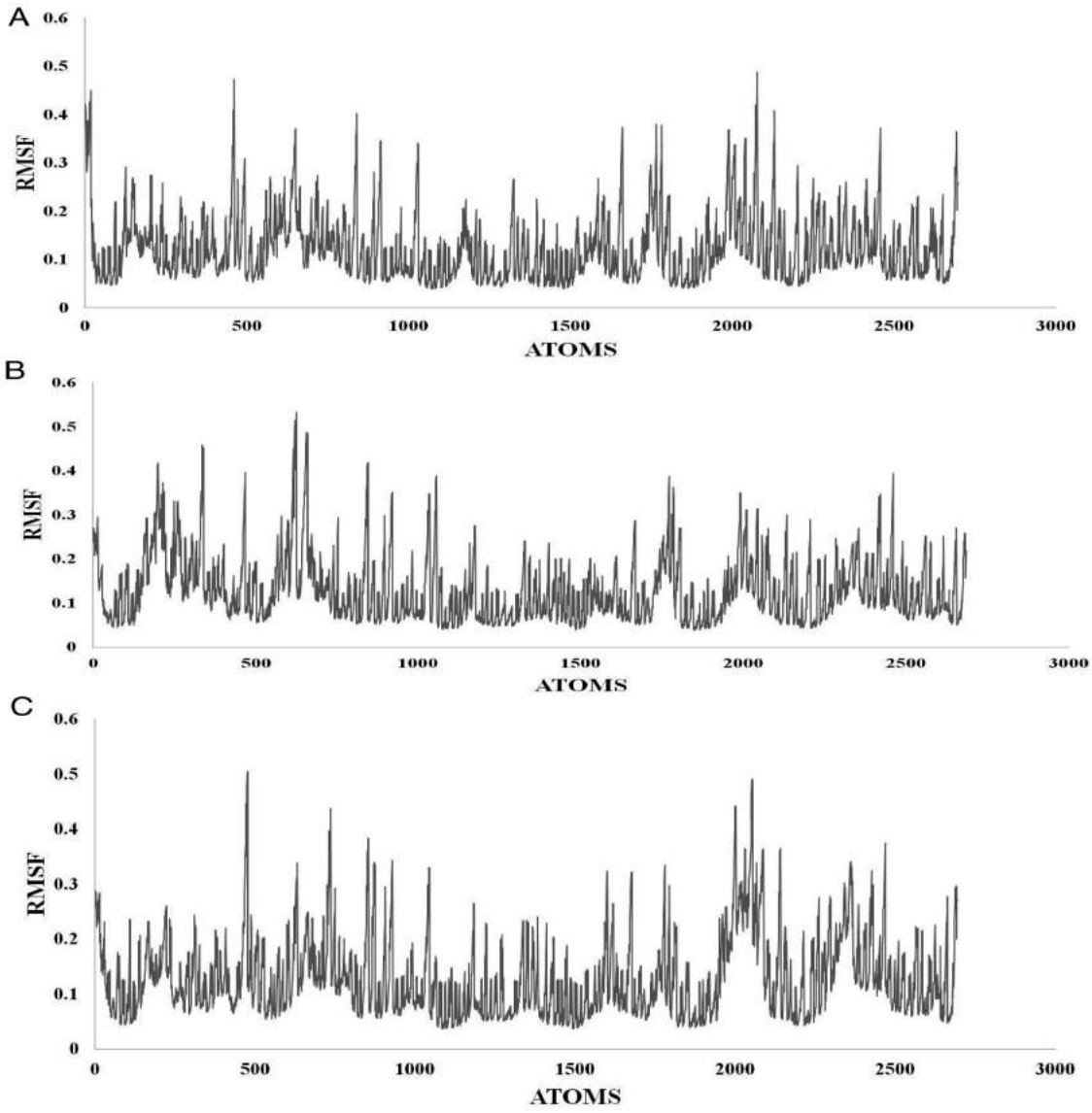


Fig.5.8. Root mean square fluctuation of α-carbon atoms as a function of atoms of 1i6w and its mutants from RMSF study at 330K during the 30ns simulation is shown. A) 1i6w at 330K, B) mut 1 at 330K, C) mut 2 at 330K.

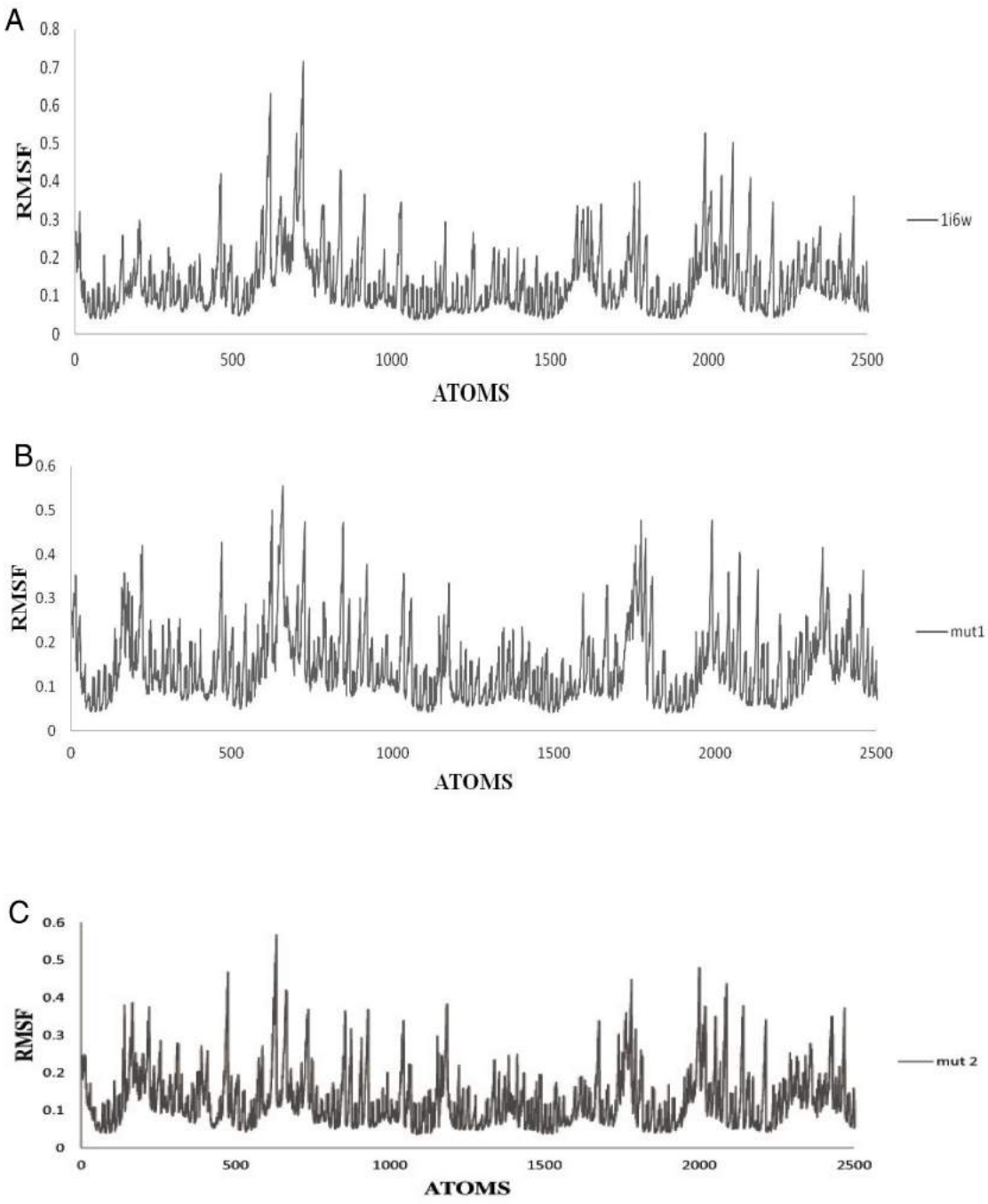


Fig.5.9. Root mean square fluctuation of α-carbon atoms as a function of atoms of 1i6w and its mutants from RMSF study at 350K during the 30ns simulation is shown. A) 1i6w at 350K, B) mut 1 at 350K, C) mut 2 at 350K.

The difference plots of RMSF of mut 1 (Red lines) and mut 2 (green lines) with wild type (blue lines) at 320K, 330K and 350K has been illustrated in Fig. 5.10. The comparison highlights that the reduction in the flexibility is specific to few regions. The regions where mutations were performed are atom numbers 689-702 for T47 and 1783 to 1796 for Q121 in mut 1 and mut 2. Both the mutants show lower flexibilily than the wild type at these regions at all the three temperatures. This observation show that the mutations have led to decrease in flexibility of the mutants w.r.t. the wild type.

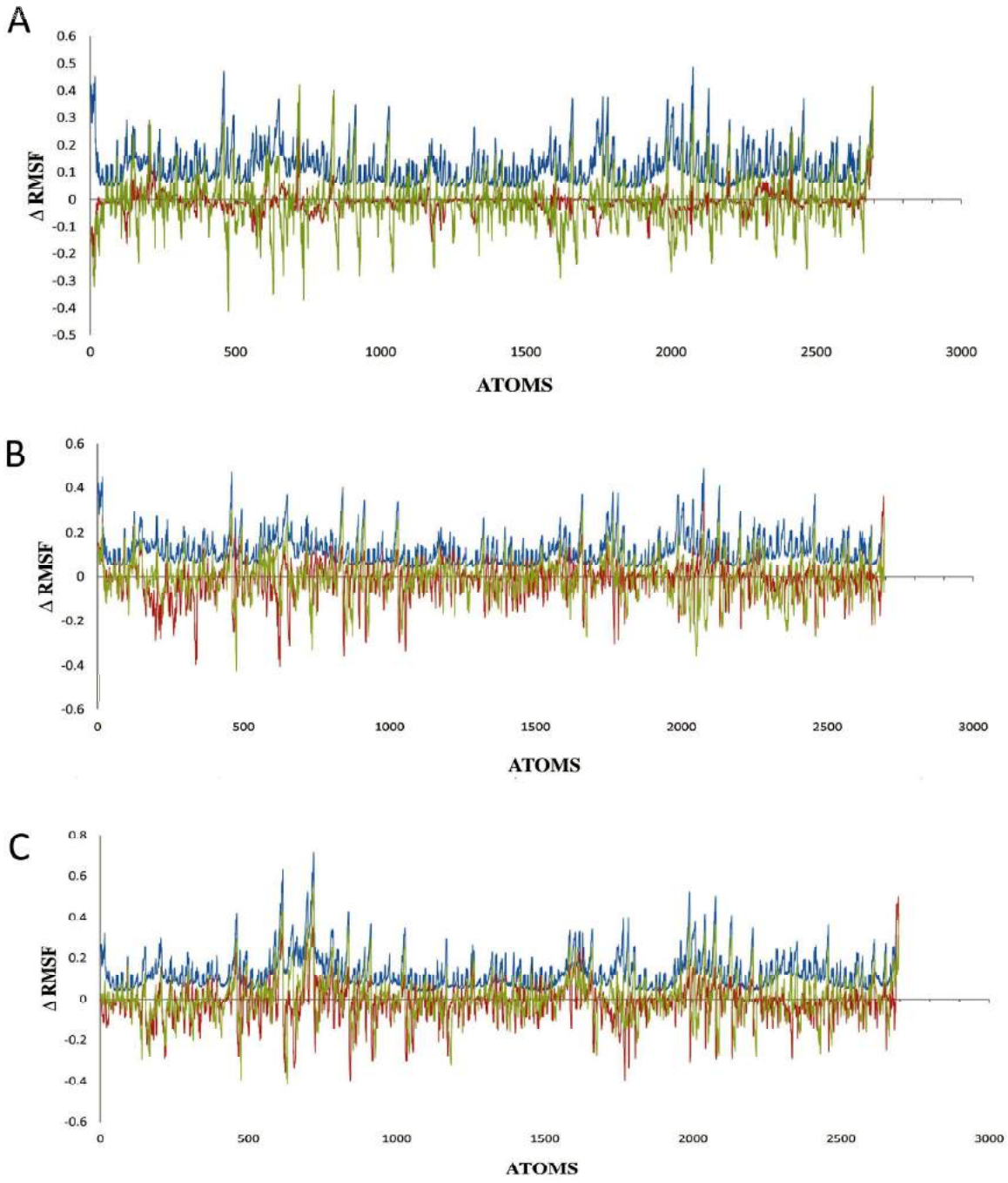


Fig. 5.10. Graph showing difference in RMSF between mut 1, mut 2 and wild type lipaseat A) 320K B) 330K C) 350K.

Radius of Gyration (Rg)

The radius of gyration has been analyzed in a time-dependent manner to investigate the compactness of wild type and mutants and illustrated in Fig.5.11. Lower average Rg of mutants at all the three temperatures reflects that the mutants are more compact than the wild type (Table 5.5). The radius of gyration reflects the packing of amino acids throughout the simulation thereby stability and folding rate. At 320K mut2 is more compact than mut 1 and 1i6w. However at 330K mut 1 becomes more compact. mut 2 compactness is still lower than 1i6w. Further at 350K mut 2 is more compact. Therefore mut 2 may be regarded as more stable than mut 1 and 1i6w.

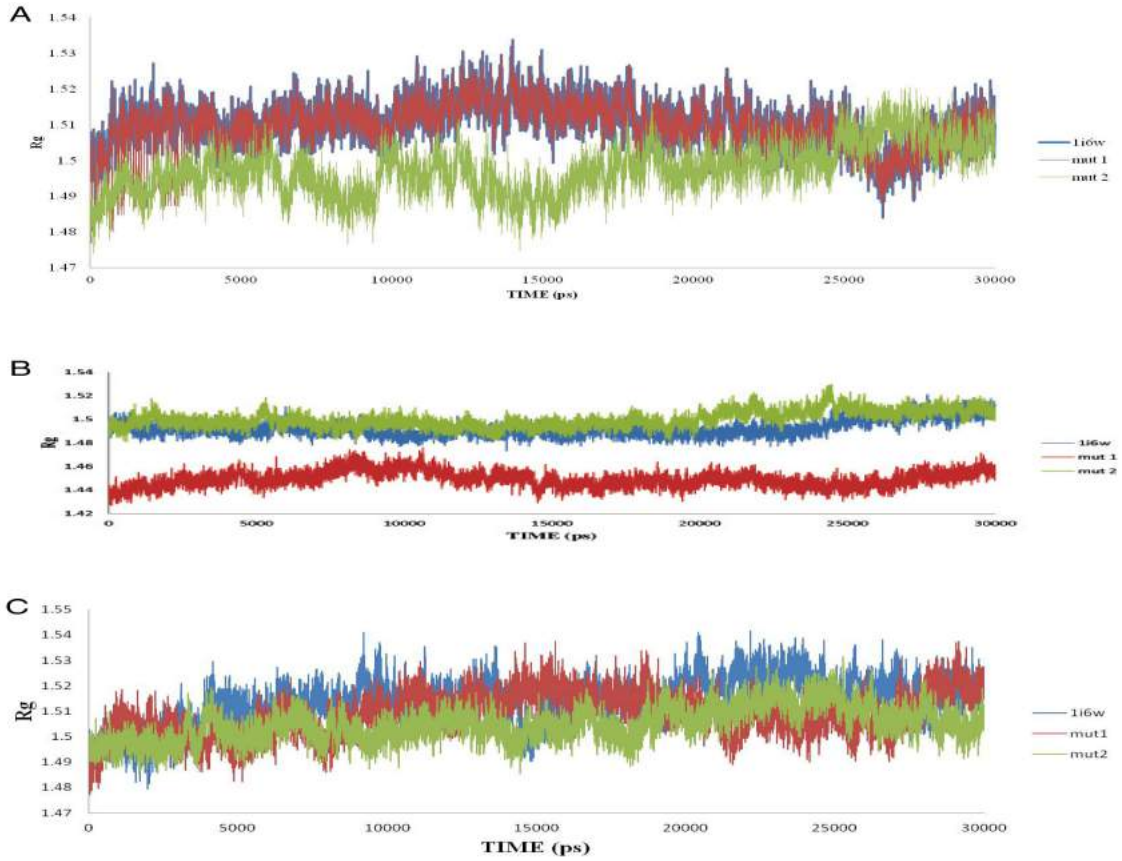


Fig. 5.11. Radius of gyration (nm) of α-carbon atoms as a function of time of 1i6w (WT) and its mutants mut 1 and mut 2 from RMSF study during the time course of simulation is shown. (A) 320K simulation. (B) 330K simulation (C) 350K simulation.

Hydrogen Bonds

The average number of hydrogen bonds per frame of the 30 ns MD simulations for WT, and thermostable proteins were calculated at 320K, 330K and 350K. The average number of hydrogen bonds has been illustrated in Fig. 5.12. The average number of intra-protein hydrogen bonds is much higher for mut 2 throughout the simulation at different temperatures. Higher number of hydrogen bonds has been previously related to the increased thermostability of thermophilic proteins (Vogt et al. 1997). Srivastava et al. (2014) also found hydrogen bonds to be greater in thermostable mutants of *Bacillus subtilis* lipase by performing MD simulation.

Interestingly average number of main chain hydrogen bonds is also much higher for mut 2 followed by WT and the least is observed for mut 1. The highest prioritized criteria by RankProt are main chain hydrogen bonds. Therefore proteins which show an increase in such bonds will be more thermostable. Then it may be assumed that mut 2 will be more stable than mut 1. Interestingly the simulation also shows the importance of hydrogen bonds in stabilizing proteins. Overall intra-protein side chain hydrogen bond decreases with increase in temperature. As temperature increases, unfolding ensues and hydrogen bonds are formed between the side chains of amino acid residues and solvent, thus lowering the intra-protein side chain hydrogen bonds. The reduction in side chain hydrogen bonds for the WT is more pronounced than the mut 1 and mut 2.

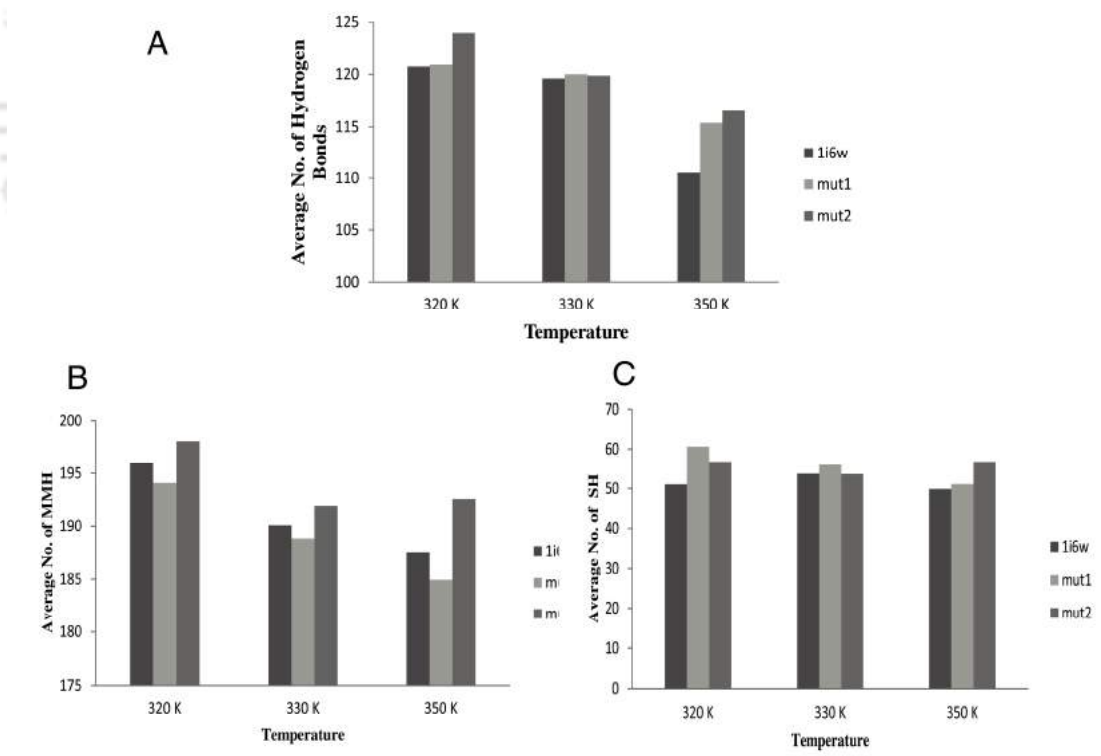


Fig. 5.12. The average number of hydrogen bonds per frame of the 30 ns MD simulations for A) All hydrogen bonds B) Main chain C) Side chain hydrogen bonds at 320, 330 and 350K.

Secondary structure analysis

Fig. 5.13-5.15, shows the average number of residues in regular secondary structure after MD simulations at 320, 330 and 350K for WT and mutants. At all three temperatures simulated here, it was observed that the percentage regular secondary structures and average number of residues are higher in mutants compared to WT. But the rise is more pronounced for the WT than the mutants. The mutations chosen were present in the turns and the average number of residues in turns was more for mutants than wild type at higher temperatures. It was also interesting to observe that residues forming the turns and beta sheets at higher temperature were more stable in mut 2 than mut 1 and the wild type. This shows stability of turns is important for thermostability. Moreover as it was observed through physico-chemical analysis available mutated thermostable proteins that thermostable mutants have higher number of γ- turns in the loop region than the WT. Therefore it can be correlated here that γ- turns may increase the stability of these mutants by stabilizing the turns. Total number of residues occurring in regular secondary structure is an indicator of the stability of the protein (Srivastava et al. 2014).

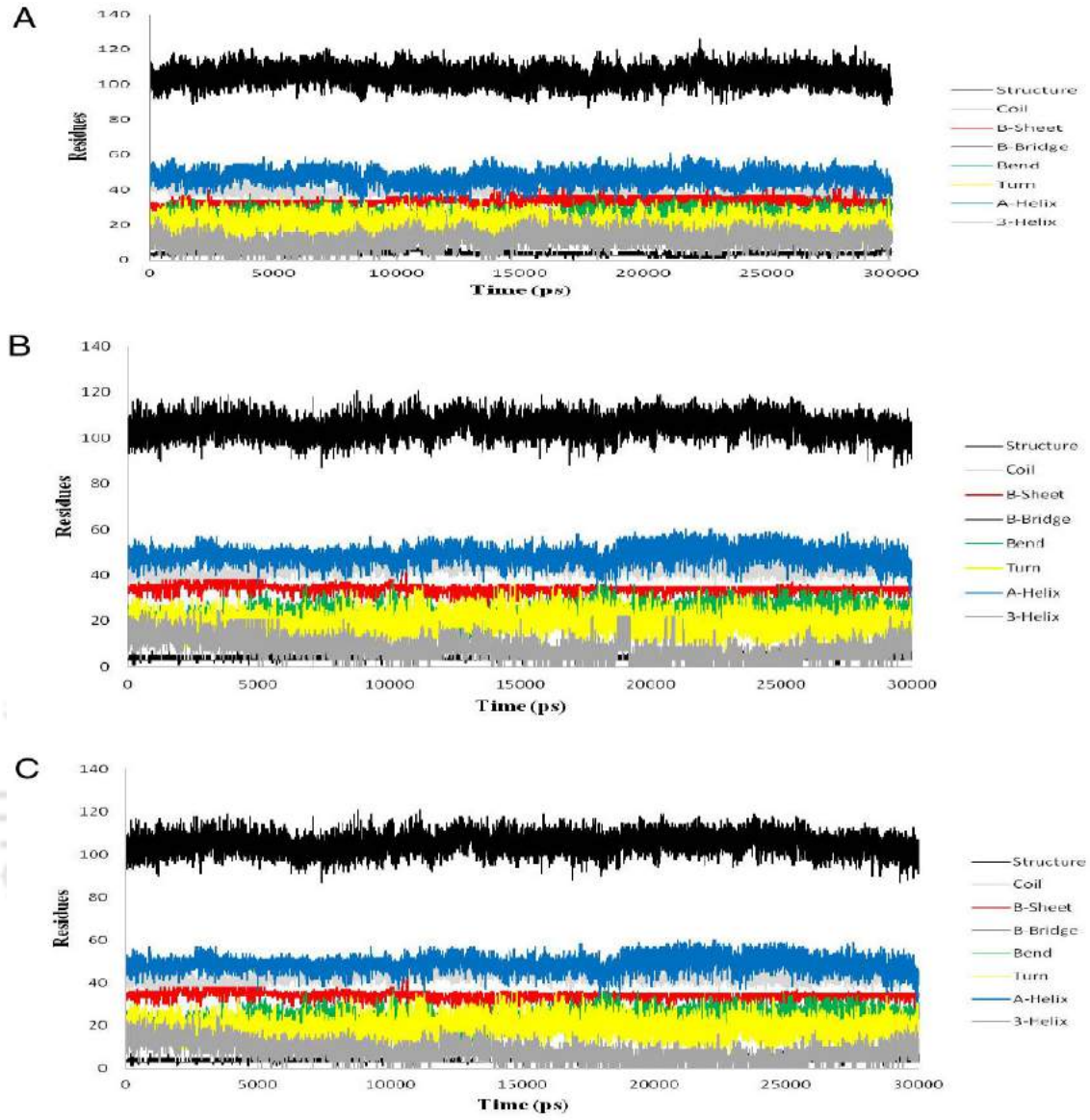


Fig. 5.13. Graphical representation of number of residues in secondary structure during 30ns simulation at 320K.

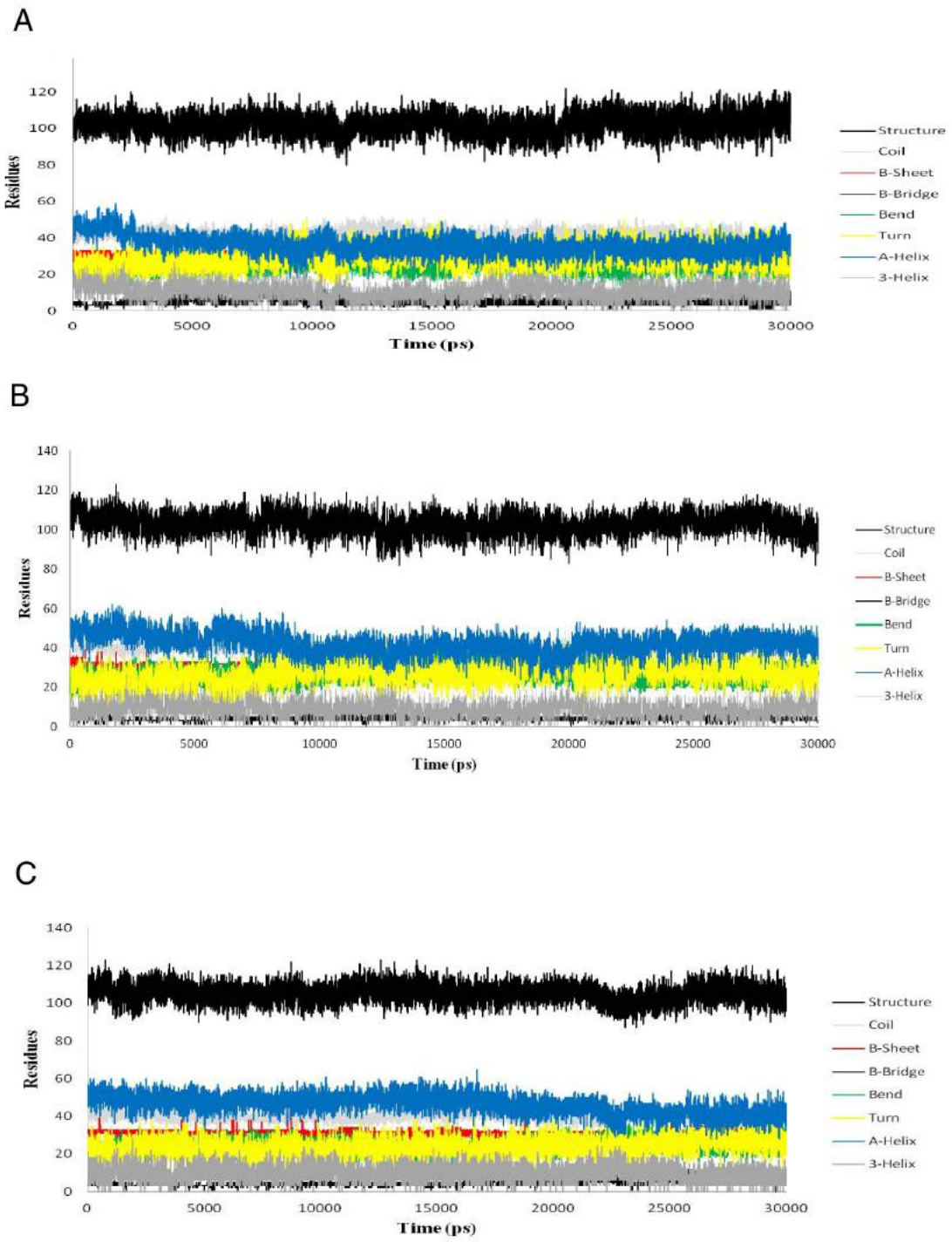


Fig. 5.14. Graphical representation of number of residues in secondary structure during 30ns simulation at 330K.

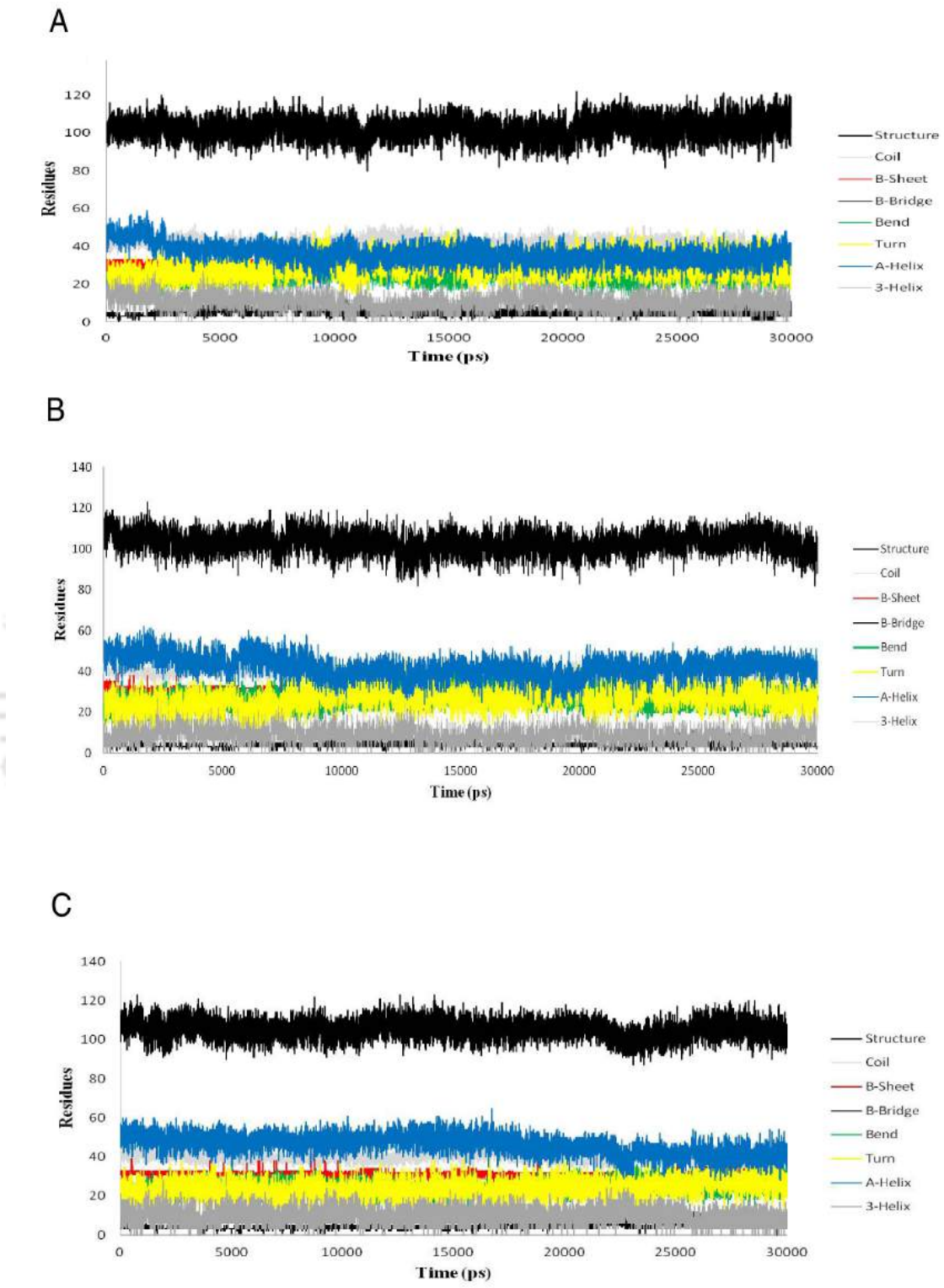


Fig. 5.15. Graphical representation of number of residues in secondary structure during 30ns simulation at 350K.

5.4. Conclusions

In this chapter we report a new *in silico* approach for enhancement of thermostability of lipases which can prove to be a guided path for their in vitro evolution, of thermostable proteins efficiently and conveniently. Herein, we discuss improver of thermal stability based on enthalpic contribution, in mesostable triacylglycerol lipase *Bacillus subtilis*.

Bacillus subtilis lipase A was chosen as the model enzyme to validate the developed tool RankProt as it is an industrially important enzyme. 60 single point mutations were predicted by servers like Cupsat, I-mutant, I-prestab and ERIS to be stabilizing. Out of these 60 mutations combination of 18 combinations of double mutations were ranked higher by RankProt than the wild type structure. Finally, 2 double mutations were predicted by RankProt to be the most thermostabilizing. The mutations were T47S, Q121N (mut 1) and T47N, Q121N (mut 2). The ranks obtained were 0.5 and 0.54 for mut 1 and mut 2 respectively. Therefore it was predicted that mut 2 will be more stable than mut 1. In silico mutagenesis was performed and the mutated structures were subjected to homology modeling. Molecular superimposition in PyMol showed low RMSD values of 0.278 (mut 1 and 1i6w) and 0.312 (mut 2 and 1i6w), indicating that the wild type and mutated structures were alike. Physicochemical characterization by RankProt showed that there was increment in γ-turns, salt bridges, ionic interaction, cation pi interaction (CPI), non polar accessible surface area (NASA), main-chain side-chain (MSH) and side-chain side-chain (SSH) hydrogen bonds in mut 1. Again salt bridge (SB), γ-turns (GT and IGT), charged accessible surface area (CASA), hydrophobic interaction (HI), main-chain main-chain (MMH), main-chain side-chain (MSH) and side-chain side-chain (SSH) hydrogen bonds increased in mut 2. It can be said that combinations of increment of different intra-molecular interactions and secondary structures were predicted to stabilize mut 1 and mut 2.

Furthermore molecular docking with (C8) substrate gave lower binding energy for mutants. The binding pocket was intact. This shows that the mutations did not affect the activity of the lipase. Thus mutation did not disturb the catalytic

properties. Contact map analysis of mut 1, mut 2 and the available thermostable structures of *Bacillus subtilis* lipase highlighted that number of unique contacts in mutants is much higher than the wild type structure. Such unique contacts were more in the loop region of the 3D-structure of the mutants. Furthermore The HBplot analysis of hydrogen bond network in wild type and mutated structures of *Bacillus subtilis* lipase uncovered that main-chain main-chain hydrogen bonds increased in 4 mutated structures. Main-chain side-chain and side-chain side chain hydrogen bonds increased in all the mutants. Hydrogen bonds <3Å were much greater for the mutants in comparison to the wild type structures. Intramolecular hydrogen bonding networks increased near the β-strand and α-helices of the mesostable lipases. This can result in better packing due to pinning of helices and stands rendering them rigid to unfolding at elevated temperatures.

Molecular dynamics simulation of the wild type and mutants were performed at 320K, 330K and 350K for 30 ns each. The results uncoverd many interesting factors supporting those mutants were more stable than the wild type. The analysis also predicted mut 2 have enhanced stability than mut 1. At higher temperature of 350K the RMSD of mut 2 was much lower than mut 1 and wild type. This reflects that mut 2 is much stable than mut 1 and WT at higher temperatures. Average RMSF plot showed global reduction in flexibility for mut 2 at 350K while mut 1 shows increase. Appreciable difference of RMSF between wild type and mutants is observed at the N- and C- terminus. The plots illustrating difference in RMSF value between mutants and wild type showed mutants to have lower flexibility at regions where mutations were performed. This observation shows that the mutations have led to decrease in flexibility of the mutants, w.r.t. the wild type. Furthermore lower average Radius of gyration of mutants at all the three temperatures reflects that the mutants are more compact than the wild type. Moreover average number of intra-protein hydrogen bonds was much higher for mut 2 throughout the simulation at different temperatures. This reflects that mut 2 to have enhanced stability than mut 1 and the wildtype. Again calculation of percentage regular secondary structures and average number of residues showed that they were higher in mutants compared to WT. The mutations

chosen were present in the turns and the average number of residues in turns was more for mutants than wild type at higher temperatures. This shows stability of turns is important for thermostability.

Existing technologies used to thermostabilize proteins rely on the principles of directed evolution and is random. A guided approach is lacking in the existing techniques. This is due to the interplay of various factors in thermostabilizing proteins. Therefore the aforementioned results and methodology reveals that this novel approach will aid in designing thermostabilizing mutations and the guided methodology provided will lead in its achievement with less time, effort, labour and capital. Thus it can revolutionize the method of engineering thermostable proteins *in vitro*.

Chapter VII
Conclusion and Future Perspective

7.1. Conclusions

This dissertation work has been successful to develop *a two-step hybrid design strategy to mutate amino acid residues which can render proteins stable at thermophilic range.* The interest was in developing such a strategy for thermostabilizing enzymes by protein engineering approaches because these enzymes hold high priority in industries. Though such enzymes are naturally obtainable from thermophiles that thrive in various extreme geothermal milieu of the Earth such as hot springs, such extreme environments are challenging to access. Thus an easy approach of engineering thermostable proteins will add to industrial economy.

In Chapter I of this current thesis, literature survey showed that though development of thermostable enzymes from their mesostable counterparts has become a trend through directed evolution. These strategies employed to engineer thermostable proteins is rather random and requires humongous amount of time and labour. A universal protocol for thermostabilizing protein ceases to exist. This is because each protein shows individual characteristics and patterns of intra molecular interactions (Kumar et al. 2000, Vogt and Argos 1997; Gromiha 2001; Trivedi et al. 2006). A guided and rationalized approach for thermostabilizing proteins is a need.

To achieve at a rationalized and predictable approach to thermostabilize proteins, the dissertation work began with data collection of thermostable proteins which was molded in the form of a thermostable protein database accessible to researchers through www.extreme-stabledb.in. Chapter II presents the details about the same. To summarize, it can be said that the database is a curated information repository for thermostable proteins and mutants. It is the only database on thermostable proteins created until date. The architecture is built on Apache, MySql, and PHP platform. The database hosts a total of 378 thermostable proteins from 132 hyperthermophilic and thermophilic organisms and 261 mutants. Around 14% of

proteins have temperature optimum of 70-80^{o}C. Most of the thermostable protein structures available were found to belong to the hydrolase class. The database unlike any other available database comprises of information regarding the percentage of intra-protein interaction for each protein structure. An intra-protein interaction calculator was developed in Python platform and is freely available for download. The collected data were filtered and homologous mesostable proteins chosen through BLAST search. A total of 25 protein stability features were generated using the developed python tool and other webservers and softwares like VADAR and Promotif. The features were filtered through statistical tests of correlation. Highly correlated features were discarded. The final dataset consisting 17 features of 127 proteins were subjected to machine learning and multicriteria decision making approach. The features were ranked or prioritized. *A new thermostabilizing feature- γ-turns was identified. It was observed that such turns were greater in thermostable proteins when compared to their mesostable counterparts.*

Data collection and database creation led to the observation that hydrolases were most populated by thermostable enzymes, majority of which were lipases. Therefore the dissertation work further presents in Chapter III, an in depth analysis of factors that lead to thermostabilization of lipases. By sequence and structure analysis of thermostable lipases with an added approach of tree based annotation it was concluded that each thermostable lipase adopts its own strategy in relation with the three dimensional arrangement of amino acid residues to increase its stability at elevated temperatures. Conclusively a set of strategies that can enhance thermostability of lipases was reported. The most important fact to be endorsed leading to enhancement of temperature stability is the increment in inverse γ-turn near the amino or carboxyl end of helices and strands of lipases. This corroborates previous observation of increment in γ-turn in other protein classes.

Following data filtration and feature generation it was observed that thermostability is a mutifactorial problem. This necessitated the development of rules

that can prioritize these features in accordance to their importance in rendering proteins thermostable. For achieving the aforementioned, in Chapter IV, various machine learning approaches were utilized to classify thermostable proteins by application of the same on a dataset of 127 thermostable-mesostable protein pairs with 17 statistically significant protein features. This was followed by ranking of the features through Analytical Hierarchical Process. This method has been utilized for the first time to prioritize protein thermostability features. This method was successful to come up with a scoring model which can differentiate thermophilic/mesophilic proteins. The problem of ranking thermostable proteins features was decomposed into hierarchies and the factors ranked with the aid of eigen vectors. Ionic interaction and main chain to main chain hydrogen bonding were given the highest priority for conferring thermal stability. This resulted in the development of a tool RankProt. Further the tool was validated through blind tests. A random set of 100 proteins were ranked and the thermostable proteins were assigned an average rank value of 0.54. The accuracy of the method was calculated to be 91%. Furthermore three case studies to check the efficiency of RankProt were performed. In all the three cases RankProt successfully recognized thermostabilizing mutants. Thus it can be conclusively said that this method can successfully identify thermostabilizing mutations. Moreover the edge of this method is that multiple combinations of mutations can be prioritized at a single go with higher rank assigned to the higher thermostabilizing ones.

Chapter V presents further validation of RankProt. To achieve at the same, thermostabilizing mutations have been identified for the commercially important *Bacillus subtilis* lipase. The mesostable lipase is stable at 35^{o}C and pH 8. The lipase was chosen as it is lid less and does not require lipid-water interphase for its functionality. *In silico* mutations were identified using the aforementioned thermostability protocol. Two sets of mutation were identified as thermostabilizing. The mutations were T47S, Q121N (mut 1) and T47N, Q121N (mut 2). They were

given the ranks 0.5 and 0.54 by RankProt. The mutations were designed to increase main-chain to main-chain hydrogen bonds, ionic interactions and γ-turns. These features were the highest prioritized features to attain thermostability by multi criteria decision making algorithm. Molecular superimposition in PyMol of the mutated proteins with the wild type structure, showed low RMSD values of 0.278 (mut 1 and 1i6w) and 0.312 (mut 2 and 1i6w), indicating that the wild type and mutated structures were alike. Physicochemical characterization by RankProt showed that there was increment in γ- turns, salt bridges, ionic interaction, cation pi interaction (CPI), non polar accessible surface area (NASA), main-chain side-chain (MSH) and side-chain side-chain (SSH) hydrogen bonds in mut 1. Again salt bridge (SB), γ-turns (GT and IGT), charged accessible surface area (CASA), hydrophobic interaction (HI), main-chain main-chain (MMH), main-chain side-chain (MSH) and side-chain side-chain (SSH) hydrogen bonds increased in mut 2. It can be said that combinations of increment of different intra-molecular interactions and secondary structures were predicted to stabilize mut 1 and mut 2. Furthermore, molecular docking with (C8) substrate gave a binding energy of -6.79 (mut 1), -6.91 (mut 2) and -6.53 (wild type) for the mutated and the native structures respectively. The binding pocket was intact. Contact map analysis of mut 1, mut 2 and the available thermostable structures of *Bacillus subtilis* lipase highlighted that number of unique contacts in mutants is much higher than the wild type structure. Such unique contacts were more in the loop region of the 3D-structure of the mutants. Furthermore, the HBplot analysis of hydrogen bond network in wild type and mutated structures of *Bacillus subtilis* lipase uncovered that main-chain main-chain hydrogen bonds increased in mut 1 and mut 2. Main-chain side-chain and side-chain side chain hydrogen bonds increased in all the mutants. Hydrogen bonds <3Å were much greater for the mutants in comparison to the wild type structures. intramolecular hydrogen bonding networks increased near the β-strand and α-helices of the mesostable lipases. This can result in better packing due to pinning of helices and strands rendering them

rigid to unfolding at elevated temperatures. Molecular dynamics simulation of the wild type and mutants were performed at 320K, 330K and 350K for 30 ns each. The results uncovered many interesting factors supporting those mutants were more stable than the wild type. The analysis also predicted mut 2 has enhanced stability than mut 1. At higher temperature of 350K the RMSD of mut 2 was much lower than mut 1 and wild type. This reflects that mut 2 is much stable than mut 1 and WT at higher temperatures. Average RMSF plot showed global reduction in flexibility for mut 2 at 350K while mut 1 shows increase. Appreciable difference of RMSF between wild type and mutants was observed at the C- terminus. The plots illustrating difference in RMSF value between mutants and wild type showed mutants to have lower flexibility at regions where mutations were performed. This observation shows that the mutations have led to decrease in flexibility of the mutants, w.r.t. the wild type. Furthermore, lower average radius of gyration of mutants at all the three temperatures reflects that the mutants are more compact than the wild type. Average number of intra-protein hydrogen bonds was much higher for mut 2 throughout the simulation at different temperatures. This reflects mut 2 to have enhanced stability than mut 1 and the wildtype. Calculation of percentage regular secondary structures and average number of residues showed that they were higher in mutants compared to wild type. The mutations chosen were present in the turns and the average number of residues in turns was more for mutants than wild type at higher temperatures. This shows stability of turns is important for thermostability.

Therefore all these aforementioned methodologies predicted mut 2 to have greater stability followed by mut 1 than the wild type lipase. Chapter VI presents further validations of these observations. The mesostable gene of *Bacillus subtilis* lipase, was cloned in pET28 a vector with N-terminal His tag. Multi site mutations have been performed by site directed mutagenesis. Protein was purified by affinity purification and single bands were obtained after coomassie blue staining of polyacrylamide gel. The wild type and mutants were observed to have an optimum

pH of 8. Kinetic analysis with pNP-octanoate as substrate revealed that the two mutants showed highest activity and stability at 40^{o}C and 55^{o}C respectively. There is an enhancement by 5^{o}C and 20^{o}C in its kinetic temperature stability. Thus mutant 2 was found to be more thermostable than mutant 1. This corroborates previous observation where the ranks obtained by RankProt for mut 2 was higher than mut 1. Therefore the protocol for engineering thermostable protein from mesostable souce was successfully validated. Furthermore, it was observed that two double mutations induced in mut 1 and mut 2 enhanced temperature stability comparable to those evolved through directed evolution. The melting temperature was calculated using thermal shift assay. The T_m of the mutants was calculated to be 63^{o}C and 66^{o}C with respect to the wild type which has a T_m of 59^{o}C. The catalytic efficiency of mut 2 was higher than mut 1 and the wild type lipase. This shows that along with enhancement of temperature stability the mutations also were able to enhance the catalytic efficiency of mut 2.

Thus it can be conclusively said that this dissertation work successfully developed a rationalized strategy to predict thermostabilizing mutations in any test protein of interest.

7.2. Commercial Viability

It can be mentioned here that the method of rationally predicting thermostabilizing mutations has been developed and proven to be successful. It is now ready to be implemented in thermostabilizing of other commercially important enzymes. The aforementioned work has the following attributes for commercial viability:

i. Preplanned, specific and direct approach.

ii. Incorporation of single or multiple mutations through site directed mutagenesis is based on the *in silico* ranks, success prediction has 91% accuracy.

iii. Proteins will be rendered thermostable by a single mutagenesis experiment, thus commercially cheaper requiring lesser time and capital.

iv. Minimizes initial investment as the method does not utilize sophisticated instrumentation.

7.3. Research Output

i. A thermostable protein database that can be accessed through www.extreme-stabledb.in.
ii. A python tool to calculate intra-protein interactions.
iii. RankProt tool to predict and rank thermostabilizing mutations of any protein of interest. It is available for download from the database.
iv. A thermostable mutant *Bacillus subtilis* lipase.

7.4. Future perspective

The goal of enhancing thermostability of proteins using this methodology was proven to be successful. The mutated thermostable *Bacillus subtilis* lipase using this methodology has potential to be used in detergent and organic synthesis industries. Thus we have a ready to use product to be utilized by industries for commercial benefits. Bulk production can be done and downstreamed further. Other factors that need to be considered in future, while mutating proteins for thermostability, is protein aggregation and codon bias while expressing eukaryotic proteins in prokaryotic hosts. As a well characterized bacterial lipase was chosen for validation of RankProt and expressed in prokaryotic host, the problem of aggregation was not observed and codon bias was ruled out. Further, the outline of the method can be tested on datasets of proteins possessing extremophilic properties like pH stability, halophilicity, and

solvent stability. RankProt can be extended to other extremophilic behavior of proteins.

BIBLIOGRAPHY

1. Acharya, P., Rajakumara, E., Sankaranarayanan, R., Rao, N.M. (2004). Structural basis of selection and thermostability of laboratory evolved Bacillus subtilis lipase. *Journal of Molecular Biology*, 341(5), 1271-81.
2. Ahern, T.J., & Klibanov, A.M. (1988). Analysis of processes causing thermal inactivation of enzymes. *Methods of Biochemical Analysis*, 33, 91-127.
3. Ahmad, S., Kamal, M.Z., Sankaranarayanan, R., Rao, N.M. (2008). Thermostable *Bacillus subtilis* lipases: in vitro evolution and structural insight. *Journal of Molecular Biology*, 381(2), 324-340.
4. Ahn, J. O., Choi, E. S., Lee, W., Hwang, S. H., Kim, C. S., Jang, H. W., Haam, S. J., Jung, J. K. (2004). Enhanced secretion of Bacillus stearothermophilus L1 lipase in Saccharomyces cerevisiae by translational fusion to cellulose-binding domain. *Applied Microbiology and Biotechnology*, 64, 833–839.
5. Allali-Hassani, A., Wasney, G.A., Chau, I., Hong, B.S., Senisterra, G., Loppnau, P., Shi, Z., Moult, J., Edwards, A.M., Arrowsmith, C.H., Park, H.W., Schapira, M., Vedadi, M. (2009). A survey of proteins encoded by non-synonymous single nucleotide polymorphisms reveals a significant fraction with altered stability and activity. *The Biochemical Journal*, 424, 15–26.
6. Alonso, A.J., Lamata, M.T. (2006). Consistency in the analytic hierarchy process: a new approach. *International Journal of Uncertainty Fuzziness and Knowledge*, 14, 445-459.
7. Aparna, V., Rambabu, G., Panigrahi, S. K., Sarma, J. A. R. P., Desiraju, G. R. (2005). Virtual screening of 4-anilinoquinazoline analogs as EGFR kinase inhibitors: importance of hydrogen bonds in the evaluation of poses and scoring functions. *Journal of Chemical Information and Modeling*, 45, 725–738.
8. Argos, P., Rossman, M.G., Grau, U.M., Zuber, H. (1979). Thermal stability and protein structure. *Biochemistry*, 18, 5698-5703.
9. Arpigny, J. L., & Jaeger, K. E. (1999). Bacterial lipolytic enzymes: classification and properties. *Biochemistry*, 343, 177–183.
10. Bagler, G., Sinha, S. (2007). Assortative Mixing in Protein Contact Networks and Protein Folding Kinetics. *Structural Bioinformatics*, 23(14), 1760-1767.
11. Baharum, S.N., Sulong, B.M.R., Rahmanm R., Salleh, A., Basri, M. (2006). Organic solvent tolerant lipases. In: Salleh AB, Rahman R, Basri M (eds) New Lipases and Proteases. Nova Science, New York, pp 63-76.
12. Bailey, T.L., Bodén, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, WS. (2009). MEME SUITE: tools for motif discovery and searching, *Nucleic Acids Research*, 37, W202-W208,
13. Bao, Q., Tian, Y., Li, W., Xu, Z., Xuan, Z., Hu, S., Dong, W., Yang, J., Chen, Y., Xue, Y., Xu, Y., Lai, X., Huang, L., Dong, X., Ma, Y., Ling, L., Tan, H., Chen, R., Wang, J., Yu, J., Yang, H. (2002). A complete sequence of the *T. tengcongensis* genome. *Genome Research*, 12, 689–700.

14. Barah, P., Sinha, S. (2008). Analysis of protein folds using protein contact networks. *Pramana*, 71, 369-378.
15. Bava, K.A., Gromiha, M.M., Uedaira, H., Kitajima, K., Sarai, A. (2004). ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res*, 32(Database issue), D120-1.
16. Ben Hur, A., Horn, D., Siegelman, H., & Vapnik, V. (2000). A support vector method for clustering. Advances in Neural Information Processing Systems 13, Cambridge, MA: MIT Press.
17. Benedix, A., Becker, C.M., de Groot, B.L., Caflisch, A., Böckmann, R.A. (2009). Predicting free energy changes using structural ensembles. *Nature Methods*, 6(1), 3-4.
18. Berendsen, H.J.C., van der Spoel, D., van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3), 43-46.
19. Betz, SF. (1993). Disulfide bonds and the stability of globular proteins. *Protein Science*, 2(10), 1551-1558.
20. Bhardwaj, A., Leelavathi, S., Mazumdar-Leighton, S., Ghosh, A., Ramakumar, S., Reddy, V.S. (2010). The Critical Role of N- and C-Terminal Contact in Protein Stability and Folding of a Family 10 Xylanase under Extreme Conditions. *PLoS ONE*, 5(6), e11347.
21. Bhardwaj, A., Mahanta, P., Ramakumar, S., Ghosh, A., Leelavathi, S., & Reddy, V. S. (2012). Emerging role of N- and C-terminal interactions in stabilizing (β/α)8 fold with special emphasis on Family 10 xylanases. *Computational and Structural Biotechnology Journal*, 2, e201209014.
22. Bhavani, D.S. Suvarnavani, K., Sinha, S. (2011). Mining of protein contact maps for protein fold prediction. *WIREs: Data Mining and Knowledge Discovery*, 1(4), 362–368.
23. Bikadi, Z., Demko, L., Hazai, E. (2007). Functional and structural characterization of a protein based on analysis of its hydrogen bonding network by hydrogen bonding plot. *Archives of Biochemistry and Biophysics*, 461(2), 225-34.
24. Bogin, O., Peretz, M., Hacham, Y., Korkhin, Y., Frolow, F., Kalb(Gilboa), A.J., Burstein,Y. (1998). Enhanced thermal stability of Clostridium beijerinckii alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic Thermoanaerobacter brockii alcohol dehydrogenase. *Protein Science*, 7 (5), 1156–1163.
25. Bolen, D.W., Rose, G.D. (2008). Structure and energetics of the hydrogen-bonded backbone in protein folding. *Annual Review of Biochemistry*, 77, 339-62
26. Bowers, K. J., Chow, E., Xu, H., Dror, R.O., Eastwood, M.P., Gregersen, B.A., Klepeis, J.L., Kolossvary, I., Moraes, M.A., Sacerdoti, F.D., Salmon, J.K., Shan, Y., Shaw, D.E. (2006). Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. Proceedings of the ACM/IEEE Conference on Supercomputing (SC06), Tampa, Florida, November 11-17.
27. Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of quantities of protein utilising the principle of protein–dye binding. *Analytical Biochemistry*, 72, 248–254

28. Branden, C., Tooze, J. (1999). Introduction to protein structure, Garland, New York.
29. Braxton, S. (1996). In Protein Engineering: Principles and Practice, Cleland J. & Craik C. Eds.; Wiley-Liss, New York, NY, Chapter 11.
30. Brinda, K. V., Vishveshwara, S. (2005). A network representation of protein structures: Implications for protein stability. *Biophysical Journal*, 89, 4159–4170.
31. Brock, T.D. (1985). Life at high temperatures. *Science*, 230 (4722), 132-138.
32. Brown, D.P., Krishnamurthy, N., Sjölander, K. (2007). Automated Protein Subfamily Identification and Classification. *PLoS Computational Biology*, 3 (8), e160.
33. Burley, S.K., Petsko, G.A. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*, 229 (4708), 23-28.
34. Bussi, G., Donadio, D., Parrinello, M. (2007). Canonical Sampling through Velocity Rescaling. *The Journal of Chemical Physics*, 126, 014101.
35. Cambillau, C. & Claverie, J. M. (2000). Structural and genomic correlates of hyperthermostability. *Journal of Biological Chemistry*, 275, 32383-32386.
36. Capriotti, E., Fariselli, Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, 33, W306-10.
37. Castagnoli, L., Vetriani, C., Cesareni, G.(1994). Linking an easily detectable phenotype to the folding of a common structural motif. Selection of rare turn mutations that prevent the folding of Rop. *Journal of Molecular Biology*, 237(4), 378-87.
38. Ceroni, A., Passerini, A., Vullo, A., Frasconi, P. (2006). DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Research*, 34 (suppl 2), W177-W181.
39. Chakravarty, S., & Varadarajan, R. (2000). Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Letters.* 470, 65-69.
40. Chakravarty, S., Varadarajan, R. (2002). Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. *Biochemistry*, 41, 8152-8161.
41. Chan, C.H., Yu, T.H., Wong, K.B. (2011). Stabilizing salt-bridge enhances protein thermostability by reducing the heat capacity change of unfolding. *PLoS One*, 6(6), e21624.
42. Chang, C., Park, B.C., Lee, D.S., Suh, S.W. (1999). Crystal structures of thermostable xylose isomerases from Thermus caldophilus and Thermus thermophilus: possible structural determinants of thermostability. *Journal of Molecular Biology*, 288, 623–634.
43. Chen, C-W., Lin, J., Chu, Y-W. (2013). iStable: Off-the-shelf Predictor Integration for Predicting Protein Stability Changes. *BMC Bioinformatics*, 14(suppl 2), S5.
44. Cheng, J., Randall, A., Baldi, P. (2006). Prediction of Protein Stability Changes for Single Site Mutations Using Support Vector Machines. *Proteins*, 62(4), 1125-32.
45. Cherukuvada, S.L., Seshasayee, A.S.N., Raghunathan, K., Anishetty, S., Pennathur, G. (2005). Evidence of a Double-Lid Movement in Pseudomonas aeruginosa Lipase: Insights from Molecular Dynamics Simulations. *PLoS Computational Biology*, 1(3), e28.

46. Chung, G.H., Lee, Y.P., Jeohn, G.H., Yoo, O.J., Rhee, J.S. (1991). Cloning and nucleotide sequence of thermostable lipase gene from *Pseudomonas fluorescens SIK W1*. *Agricultural and Biological Chemistry*, 55(9), 2359–2365.
47. Costantini, S., Colonna, G., Facchiano, A. M. (2008). ESBRI: A web server for evaluating salt bridges in proteins. *Bioinformation*, 3(3), 137–138.
48. Covalt, J.C. Jr., Roy, M., Jennings, P.A. (2001). Core and surface mutations affect folding kinetics, stability and cooperativity in IL-1 beta: does alteration in buried water play a role? *Journal of Molecular Biology*, 307(2), 657-69.
49. Creighton, T.E. (1994). The energetic ups and downs of protein folding. *Nature Structural Biology*, 1, 135 – 138.
50. D'Amico, S., Collins, T., Marx, J.-C., Feller, G., & Gerday, C. (2006). Psychrophilic microorganisms: challenges for life. *EMBO Reports*, *7*(4), 385–389.
51. Dancey, D., Bandar, Z.A., McLean, D. (2007). Logistic model tree extraction from artificial neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics focuses on cybernetics*, 37, 794-802.
52. Darby, N. & Creighton, T.E. (1995). Disulfide bonds in protein folding and stability. *Methods in Molecular* Biology, *40* 219–252.
53. Darby, N., Creighton, T.E. (1997). Probing protein folding and stability using disulfide bonds. *Molecular Biotechnology*, 7(1), 57-77.
54. Dartois, V., Baulard, A., Schanck, K., Colson, C. (1992). Cloning, nucleotide sequence and expression in *Escherichia coli* of a lipase gene from *Bacillus subtilis* 168. *Biochimica et Biophysica Acta*, 1131(3), 253-60.
55. Das, S., Chew, M.Y.L. (2008). Building Grading Systems:A Review of the State-of-the-Art. *Architectural Science Reviews*, 50, 2-12.
56. Day, R., Bennion, B.J., Ham, S., Daggett, V. (2002). Increasing temperature accelerates protein unfolding without changing the pathway of unfolding. *Journal of Molecular Biology*, 322(1), 189-203.
57. De Farias, S.T. & Bonato, M.C. (2002). Preferred codons and amino acid couples in hyperthermophiles. *Genome Biology*, 3 (online publication).
58. Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., Rooman, M. (2009). Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, 25(19), 2537-43.
59. DeLano, W.L. (2002). The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org.
60. Delboni, L. F., Mande, S. C., Rentier-Delrue, F., Mainfroid, V., Turley, S., Vellieux, F. M., Martial, J.A., Hol, W. G. (1995). 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 Science : A Publication of the Protein Society*, 4(12), 2594–2604.
61. Denisov, V.P., Jonsson, B.H., Halle, B. (1999). Hydration of denatured and molten globule proteins. *Nature Structure Biology*, 6(3), 253–260.
62. Desjarlais, J. R. & Handel, T. M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Science*, 4, 2006-2018.

63. Ding, Y., Cai, Y., Zhang, G., Xu, W. (2004). The influence of dipeptide composition on protein thermostability. *FEBS Letters*, 569(1-3), 284-8.
64. Do, C.B., Batzoglou, S. (2008).What is the expectation maximization algorithm? *Nature Biotechnology*, 26(8), 897-9.
65. Dosztányi, Z., Magyar, C., Tusnády, G., Simon, I. (2003). SCide: identification of stabilization centers in proteins. *Bioinformatics*, 19: 899–900.
66. Ebrahimi, M., Lakizadeh, A., Agha-Golzadeh, P., Ebrahimie, E., Ebrahimi, M. (2011). Prediction of Thermostability from Amino Acid Attributes by Combination of Clustering with Attribute Weighting: A New Vista in Engineering Enzymes. *PLoS One,* 6(8), e23146.
67. Eijsink, V.G., Bjørk, A., Gåseidnes, S., Sirevåg, R., Synstad, B., van den Burg, B., Vriend, G. (2004). Rational engineering of enzyme stability. *Journal of Biotechnology*, 113(1-3), 105-20.
68. Eijsink, V.G., Vriend, G., van den Burg, B., van der Zee, J.R., Venema, G. (1992). Increasing the thermostability of a neutral protease by replacing positively charged amino acids in the N-terminal turn of a-helices. *Protein Engineering*, 5(2), 165–170.
69. Elcock, A. H. (1998). The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. *Journal of Molecular Biology*, 284(2), 89-502.
70. Eltaweel, M.A., Rahman, R.A., Salleh, A.B., Basri, M. (2005). An organic solvent stable lipase from *Bacillus sp. strain 42*. *Annals of Microbiology*, 55(3), 187-192.
71. Emond, S., André, I., Jaziri, K., Potocki-Véronèse, G., Mondon, P., Bouayadi, K., Kharrat, H., Monsan,P., Remaud-Simeon, M. (2008). Combinatorial engineering to enhance thermostability of amylosucrase. *Protein Science : A Publication of the Protein Society*, 17(6), 967–976.
72. Essman U, Perera, L., Berkowitz, M.L., Darden, T., Lee, H., Pedersen, L.G., (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103, 8577–8592.
73. Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, 226–231.
74. Feller, G., Arpigny, L., Narinx, E., Gerday, Ch. (1997). Molecular adaptation of enzymes from psychrophilic organisms. *Comparative Biochemistry and Physiology*, 118(3), 495-499.
75. Ferre, F. & Clote, P. (2005). DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Research*, 33(Web Server issue), W230-2.
76. Fitter, J., & Heberle, J. (2000). Structural equilibrium fluctuations in mesophilic and thermophilic alpha-amylase. *Biophysical Journal*, 79(3), 1629–1636.
77. Fogliatto, F.S., Albin, S.L. (2001). A hierarchical method for evaluating products with quantitative and sensory characteristics. *IEEE Trans*, 33, 1081-1092.
78. Folcarelli, S., Battistoni, A., Carrì, M.T., Polticelli, F., Falconi, M., Nicolini, L., Stella, L., Rosato, N., Rotilio, G., Desideri, A. (1996). Effect of Lys-->Arg mutation on the thermal stability of Cu,Zn superoxide dismutase: influence on the monomer-dimer equilibrium. *Protein Engineering*, 9(4), 323-5.

79. Frank, A.C., Lobry, J.R. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, 238(1), 65-77.
80. Frappier, V., Najmanovich, R.J. (2014). 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 Computational Biology,* 10(4):e1003569.
81. Fuji, T., Hata, Y., Wakagi, T., Tanaka, N., Oshima, T. (1996). Novel zinc-binding centre in thermoacidophilic archaeal ferredoxins. *Nature Structural Biology*, 3, 34–837.
82. Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M. (2003). Unique amino acid composition of proteins in halophilic bacteria. *Journal of Molecular Biology*, 327, 347-357
83. Gong, H., Rose, G.D. (2008). Assessing the solvent-dependent surface area of unfolded proteins using an ensemble model. PNAS, 105 (9), 3321-3326.
84. Grantcharova, V.P., Riddle, D.S., Baker, D. (2000). Long-range order in the src SH3 folding transition state. *Proceedings of the National Academy of Sciences of the United States of America.* 97(13), 7084–7089.
85. Greaves RB, Warwicker J. Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles. BMC Struct Biol. 2007;7:18.
86. Grimsley, G.R., Shaw, K.L., Fee, L.R., Alston, R.W., Huyghues-Despointes, B.M.P., Thurlkill, R.L., Scholtz, J.M., Pace, C. (1999). Increasing protein stability by altering long-range coulombic interactions. *Protein Science*, 8, 1843–1849.
87. Gromiha, M. M., Oobatake, M., Sarai, A. (1999). Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophysical Chemistry*, 82, 51–67.
88. Gromiha, M.M. (2001). Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. *Biophysical Chemistry*, 91, 71-77.
89. Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedaira, H., Sarai, A. (1999). ProTherm: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Research*, 27, 286-288.
90. Gromiha, M.M., Pujadas, G., Magyar, C., Selvaraj, S., Simon, I. (2004). Locating the stabilizing residues in (a/b)8 barrel proteins based on hydrophobicity, long-range interactions, and sequence conservation. *Proteins*, 55, 316–329.
91. Gromiha, M.M., & Suresh, M.X. (2008). Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *PROTEINS: Proteins: Structure, Function, and Bioinformatics*, 70, 1274-1279.
92. Gromiha, M.M., Thomas, S., Santhosh, C. (2002). Role of cation -pi interactions to the stability of thermophilic proteins. *Preparative Biochemistry and Biotechnology*, 32(4), 355-62.
93. Grottesi, A., Ceruso, M.A., Colosimo, A., Di Nola, A. (2002). Molecular dynamics study of a hyperthermophilic and a mesophilic rubredoxin. *Proteins*, 46(3), 287-94.
94. Guerois, R., Nielsen, J. E., Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology*, 320, 369–387.

95. Gunasekaran, V., & Das, D. (2005). Lipase fermentation progress and prospect. Indian Journal of Biotechnology, 4, 437-445.
96. Haki, G.D., Rakshit, S.K. (2003). Developments in industrially important thermostable enzymes: a review. *Bioresource Technology*, 89(1), 17-34.
97. Han, J., Kamber, M., Tung, A. K. H. (2001). Spatial Clustering Methods in Data Mining: A Survey, H. Miller and J. Han(eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis.
98. Haney, P. J., Badger, J. H., Buldak, G. L., Reich, C. I., Woese, C. R. and Olsen, G. J. (1999). Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species. *Proceedings of the National Academy of Sciences USA,* 96, 3578-3583.
99. Haney, P., Konisky, J., Koretke, K. K., Luthey-Schulten, Z., Wolynes, P. G. (1997). Structural basis for thermostability and identification of potential active site residues for adenylate kinases from the archaeal genus *Methanococcus*. *Proteins*, 28, 117–130.
100. Hartigan, J.A., Wong, M.A. (1979). A k-means clustering algorithm. *Applied Statistics*, 28, 100-108.
101. Hayashi, C.O.M., Nagai, S. (1987). Purification and some properties of a thermostable lipase from Humicola lanuginosa no. 3. *Agricultural and Biological Chemistry*, 51(1), 37–45.
102. Hellinga, H.W., Richards, FM. (1994). Optimal sequence selection in proteins of known structure by simulated evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 91(13), 5803-7.
103. Hess, B., Bekker, H., Berendsen, H. J. C., Fraaije, J. G. E. M. (1997). LINCS: A linear constraint solver for molecular simulations, *Journal of Computational Chemistry*, 18, 1463-1472.
104. Horchani, H., Mosbah, H., Salem, N.B., Gargouri, Y., Sayari, A. (2008). Biochemical and molecular characterisation of a thermoactive, alkaline and detergent-stable lipase from a newly isolated *Staphylococcus aureus* strain. *Journal of Molecular Catalysis B: Enzymatic*, 56(4), 237-245.
105. Huang, G., Ying, T, Huo, P. and Jiang, Y.Z. (2006). Purification and characterization of a protease from thermophilic Bacillus strain HS08. *African Journal of Biotechnology*, 5, 2433-2438.
106. Huang, J., Lu, J., Ling, C.X. (2003).Comparing Naïve Bayes, Decision Trees, and SVM with AUC and Accuracy, Proceedings of the Third IEEE International Conference on Data Mining.
107. Huang, L.T., Gromiha, M.M., Ho, S.Y. (2007). iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics*, 23(10), 1292-3.
108. Huang, L-T., Gromiha, M.M. (2009). Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. *Bioinformatics*, 25(17), 2181-2187.
109. Huang, X., Gao, D., Zhan, C.G. (2011). Computational design of a thermostable mutant of cocaine esterase via molecular dynamics simulations. Org Biomol Chem, 9(11), 4138-43.

110. Hutchinson, E.G., Thornton, J.M. (1996). PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Science*, 5(2), 212-20.
111. Iizumi, T., Nakamura, K., Fukase, T. (1990.) Purification and Characterization of a Thermo stable Lipase from Newly Isolated Pseudomonas sp. KWI-56. *Agricultural and Biological Chemistry*, 54 (5), 1253-1258.
112. Invernizzi, G., Papaleo, E., Grandori, R., Gioia, L., Lott, M. (2009). Relevance of metal ions for lipase stability: Structural rearrangements induced in the *Burkholderia glumae* lipase by calcium depletion. *Journal of Structural Biology*, 168(3), 562-570
113. Jacob, E. & Unger, R. (2007). A tale of two tails: why are terminal residues of proteins exposed? *Bioinformatics*, 23, e225–30.
114. Jaeger, K.E., Reetz, M.T. (1998). Microbial lipases form versatile tools for biotechnology. *Trends Biotechnology*, 16(9), 396–403.
115. Jaenicke, R. (2000). Do ultrastable proteins from hyperthermophiles have high or low conformational rigidity? PNAS USA, 97, 2962–2964.
116. Jeong, S.T., Kim, H.K., Kim, S.J., Chi, S.W. (2002). Novel zincbinding center and a temperature switch in the *Bacillus stearothermophilus* L1 lipase. *Journal of Biological Chemistry*, 277, 17041–17047.
117. Joel, D.A., Supachok, S., Linda, A.F.G. (2002). Crystal structure of a thermostable lipase from *Bacillus stearothermophilus P1*. *Journal of Molecular Biology*, 323(5), 859–869.
118. Jurgen, P., Markus, F., Marcus, P., Claudia, T. (2000). Lipase engineering database understanding and exploiting sequence structure function relationships. *Journal of Molecular Catalysis A: Chemical*, 10, 491-508.
119. Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22, 2577–2637.
120. Kamal, M.Z., Ahmad, S., Molugu, T.R., Vijayalakshmi, A., Deshmukh, M.V., Sankaranarayanan, R., Rao, N.M. (2011). In vitro evolved non-aggregating and thermostable lipase: structural and thermodynamic investigation. *Journal of Molecular Biology*, 413(3), 726-41.
121. Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238(1), 143-55.
122. Kannan, N., Vishveshwara, S. (2000). Aromatic clusters: a determinant of thermal stability of thermophilic proteins. *PEDS*, 13(11), 753-761.
123. Karshikoff, A., & Ladenstein, R. (1998).Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. *Protein Engineering*, 11(10), 867-72.
124. Karshinkoff, A. & Landenstien, R. (2001). Ion pairs and the thermotolerance of proteins from hyperthermophiles: A traffic rule for hot roads. *Trends in Biochemistry Science*, 26, 550–556.
125. Kaur, H. & Raghava, G.P.S. (2006). Prediction of Cα-H...O and Cα-H...π interactions in proteins using recurrent neural network. *In-Silico Biology*, 6, 0011.

126. Kelil, A., Wang, S., Brzezinski, R., Fleury, A. (2007). CLUSS: Clustering of protein sequences based on a new similarity. *BMC Bioinformatics*, 8, 1-19.
127. Khayat, R., Tang, L., Larson, E.T., Lawrence, C.M., Young, M., Johnson, J.E. (2005). Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52), 18944-9.
128. Khechinashvili, N.N., Fedorov, M.V., Kabanov, A.V., Monti, S., Ghio, C., Soda, K. (2006). Side chain dynamics and alternative hydrogen bonding in the mechanism of protein thermostabilization. *Journal of Biomolecular Structure and Dynamics*, 24(3), 255-62.
129. Khyami-Horani, H. (1996). Thermotolerant strain of *Bacillus licheniformis* producing lipase. Wor Journal of Microbiol Biotechnology, 12, 399-401.
130. Kleiger, G., Grothe, R., Mallick, P., Eisenberg, D. (2002). GXXXG and AXXXA: Common α-Helical Interaction Motifs in Proteins, Particularly in Extremophiles. *Biochemistry*, 41(19), 5990–5997.
131. Klein,C., Georges,G., Kunkele,K.-P., Huber,R., Engh,R. and Hansen,S. (2001). High Thermostability and Lack of Cooperative DNA Binding Distinguish the p63 Core Domain from the Homologous Tumor Suppressor p53. *Journal of Biological Chemistry*, 276, 37390–37401.
132. Knoche, T.R., Hennig, M., Merz, A., Darimont, B., Kirschner, K., Jansonius, J.N. (1996). The crystal structure of indole-3-glycerol phosphate synthase from the hyperthermophilic archaeon *Sulfolobus solfataricus* in three different crystal forms: effects of ionic strength. *Journal of Molecular Biology*, 262, 502-515.
133. Kohonen, J., Talikota, S., Corander, J., Auvinen, P., Arjas, E. (2008). A naïve bayes classifier for protein function prediction. *In Silico Biology*, 9, 3.
134. Kollman, P.A., Allen, L.C. (1972). Theory of the hydrogen bond. *Chemical Reviews*, 72 (3), 283–303.
135. Kumar, M., Thakur, V., Raghava, G.P. (2008). COPid: composition based protein identification. *In Silico Biology*, 8(2), 121-8.
136. Kumar, S. & Nussinov, R. (2001). How do thermophiles deal with heat? *Cellular and Molecular Life Sciences*, 58. 1216–1233.
137. Kumar, S., Chung-Jung, T., Ruth, N. (2000).. Factors enhancing protein thermostability. *Protein Engineering*, 13 (3), 179-191.
138. Kumar, S., Ma, B., Tsai, C.J., Nussinov, R. (2000). Electrostatic strengths of salt bridges in thermophilic and mesophilic glutamate dehydrogenase monomers. *Proteins*, 38(4), 368–383.
139. Kyte, J., & Doolittle, R.F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, 157, 105-132.
140. Ladenstein, R., Antranikian, G. (1998). Proteins from hyperthermophiles: stability and enzymatic catalysis close to the boiling point of water. *Advances in Biochemical Engineering / Biotechnology*, 61, 37-85.
141. Laemmli, U.K. (1970). Cleavage of structural protein during the assembly of the head of bacteriophage T4. *Nature*, 227, 680–685.
142. Lee, D., Kok, Y., Kim, K., Kim, B., Choi, H., Kim, D., Suhartono, M.T., Pyun, Y. (1999). Isolation and characterization of a thermophilic lipase from *Bacillus thermoleovorans ID-1*. *FEMS Microbiol Letters*, 179(2), 393–400.

143. Lehmann, M., Wyss, M. (2001). Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Current Opinion in Biotechnology*, 12(4), 371-5.
144. Leonov, H., Arkin, I.T. (2005). A periodicity analysis of transmembrane helices. *Bioinformatics*, 21(11), 2604–2610.
145. Leow, T.C., Rahman, R.N., Basri, M., Salleh, A.B. (2004). High level expression of thermostable lipase from *Geobacillus sp. strain T1*. *Bioscience, Biotechnology, and Biochemistry*, 68(1), 96-103.
146. Lesuisse, E., Schanck, K., Colson, C. (1993). Purification and preliminary characterization of the extracellular lipase of *Bacillus subtilis* 168, an extremely basic pH-tolerant enzyme. *European Journal of Biochemistry*, 216(1), 155-60.
147. Li, H., Zhang, X. (2005). Characterization of thermostable lipase from thermophilic *Geobacillus sp. TW1*. *Protein Expression and Purification*, 42(1), 153-159.
148. Li, W.F., Zhou, X.X., Lu, P. (2005). Structural features of thermozymes. *Biotechnogy Advances*, 23(4), 271-281.
149. Li, Y., Zhang, J., Tai, D., Middaugh, C.R., Zhang, Y., Fang, J. (2012). PROTS: A fragment based protein thermo-stability potential. *Proteins*, 80(1), 81–92.
150. Liang, H.K., Huang, C.M., Ko, M.T., Hwang, J.K. (2005). Amino acid coupling patterns in thermophilic proteins. *Proteins*, 59, 58–63.
151. Lobry, J.R., Chessel, D. (2003). Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *Journal of Applied Genetics*, 44, 235-261.
152. Lu, J.L., Hu, X.H., Hu, D.G. (2012). A new hybrid fractal algorithm for predicting thermophilic nucleotide sequences. *Journal of Theoretical Biology*, 293, 74-81.
153. Luke, K.A., Higgins, C.L., Wittung-Stafshede, P. (2007).Thermodynamic stability and folding of proteins from hyperthermophilic organisms. *FEBS Journal*, 274(16), 4023-33.
154. Lynn, D.J., Singer, G.A., Hickey, D.A. (2002). Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Research*, 30(19), 4272-7.
155. Ma, J., Zhang, Z., Wang, B., Kong, X., Wang, Y., Cao, S., Feng, Y. (2006). Overexpression and characterization of a lipase from *Bacillus subtilis*. *Protein Expression and Purification*, 45(1), 22-9.
156. Macrae, A. R., Hammond, R.C. (1985). Present and future applications of lipases. *Biotechnology & Genetic Engineering Reviews*, 3, 193–219.
157. Madigan, M. T., Marrs, B. L. (1997). Extremophiles. *Scientific American*, 276(4), 82-87.
158. Madigan, M.T., Martinko, J.M., Parker J. (2000). Brock Biology of Microorganisms, 9th ed. Prentice Hall: USA.
159. Magyar, C., Gromiha, M.M., Pujadas, G., Tusnady, G.E., Simon, I. (2005). SRide: a server for identifying stabilizing residues in proteins. *Nucleic Acids Research*, 33, W303–W305.
160. Malcolm, B.A., Wilson, K.P., Matthews, B.W., Kirsch, J.F., Wilson, A.C. (1990). Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature*, 34, 536 – 89.

161. Manjunath, K., Sekar, K. (2013). Molecular Dynamics Perspective on the Protein Thermal Stability: A Case Study Using SAICAR Synthetase *Journal of Chemical Information and Modeling*, 53 (9), 2448–2461.
162. Maria, P., Carboni-Oerlamans, C., Tuin, B., Bargeman, G., van der Meer, A., van Gemert, R. (2005). Biotechnological applications of *Candida antarctica* lipase A: state-of-the-art. *Journal of Molecular Catalysis B: Enzymatic*, 37(1-6), 36–46.
163. Masso, M. & Vaisman, I.I. (2011). A structure-based computational mutagenesis elucidates the spectrum of stability-activity relationships in proteins. *Proceedings 33rd IEEE EMBC*, 3225-3228.
164. Matsumura, M., Matthews, BW. (1991). Stabilization of functional proteins by introduction of multiple disulfide bonds. *Methods in Enzymology*, 202, 336–356.
165. Mattos, C. (2002). Protein–water interactions in dynamic world. *TRENDS in Biochemical Sciences*, 27(4), 203–208.
166. Mattos, C., Ringe, D. (1996). Locating and characterizing binding sites on proteins. *Nature Biotechnology*, 14, 595–599.
167. McDonald, I.K., Thornton, J.M. (1994). Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology*, 238(5), 777-93.
168. MCQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281–297.
169. Milner-White, E.J. (1990). Situations of Gamma-turns in Proteins Their Relation to Alpha-helices, Beta-sheets and Ligand Binding Sites. *Journal of Molecular Biology*, 216(2), 385-397
170. Milner-White, E.J. & Poet, R. (1986). Four classes of beta-hairpins in proteins. *Biochemical Journal*, 240, 289–292.
171. Mirchevska, V., Luštrek, M., & Gams, M. (2009). Combining Machine Learning and Expert Knowledge for Classifying Human Posture. In Proceedings of ERK 2009, 183–186.
172. Miyazaki, K., Takenouchi, M., Kondo, H., Noro, N., Suzuki, M., Tsuda, S. (2006). Thermal stabilization of Bacillus subtilis family-11 xylanase by directed evolution. *Journal of Biological Chemistry*, 281, 10236-10242.
173. Montanucci, L., Fariselli, P., Martelli, P. L., & Casadio, R. (2008). Predicting protein thermostability changes from sequence upon multiple mutations. *Bioinformatics*, 24(13), i190–i195.
174. Nagendra, H.G., Sukumar, N., Vijayan, M. (1998). Role of water in plasticity, stability, and action of proteins: the crystal structures of lysozyme at very low levels of hydration. *Proteins,* 32(2), 229-40.
175. Narlikar, L., Gôrdan, R., Hartemink, A.J. (2007). A nucleosome-guided map of transcription factor binding sites in yeast. *Plos Computational Biology*, 3(11), e215.
176. Nicholson, E.M., Mo, H., Prusiner, S.B., Cohen, F.E., Marqusee, S. (2002). Differences between the prion protein and its homolog doppel: a partially structured state with implications for scrapie formation. *Journal of Molecular Biology*, 316, 807-815.

177. Niesen, F.H., Berglund, H., Vedadi, M. (2007).The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols*, 2(9), 2212-21.
178. Norris, P. R., Burton, N. P., Foulis, N. A. M. (2000). Acidophiles in bioreactor mineral processing. *Extremophiles,* 4(2), 71-76.
179. Notomista, E. Catanzano, F., Graziano, G., Di Gaetano, S., Barone, G., Di, Donato, A. (2001). Contribution of chain termini to the conformational stability and biological activity of onconase. *Biochemistry*, 40, 9097–103.
180. Ohnishi, K., Yoshida, Y., Sekiguchi, J. (1994). Lipase production of *Aspergillus oryzae*. *Journal of Fermentation and Bioengineering*, 77, 490 – 5.
181. Okada, J., Okamoto, T., Mukaiyama, A., Tadokoro, T., You, D.-J., Chon, H., Koga, Y., Takano, K., Kanaya, S. (2010). Evolution and thermodynamics of the slow unfolding of hyperstable monomeric proteins. *BMC Evolutionary Biology*, 10, 207.
182. Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M. (1992). The alpha/beta hydrolase fold. *Protein Engineering*, 5(3), 197–211.
183. Overington,J.P., Johnson,M.S., Sali,A. and Blundell,T.L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proceedings of the Royal Society B: Biological Sciences*, 241, 132–145.
184. Pace, C. N., Fu, H., Fryar, K. L., Landua, J., Trevino, S. R., Shirley, B. A., Hendricks, M.M., Iimura, S., Gajiwala, K., Scholtz, J.M. Grimsley, G. R. (2011). Contribution of Hydrophobic Interactions to Protein Stability. *Journal of Molecular Biology*, 408(3), 514–528.
185. Panigrahi, P., Sule, M., Ghanate, A., Ramasamy, S., Suresh, C.G. (2015).Engineering Proteins for Thermostability with iRDP Web Server. *PLoS One*, 10(10), e0139486.
186. Panigrahi, S.K. (2008). Strong and weak hydrogen bonds in protein-ligand complexes of kinases: a comparative study. *Amino Acids*. 34(4), 617-33.
187. Panja, A. S., Bandopadhyay, B., & Maiti, S. (2015). 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(7), e0131495.
188. Park, H.S., Lee, J.S., Jun, C.H. (2009). A K-MeansLike Algorithm for K-Medoids Clustering and Its Performance, Department of Industrial and Management Engineering, POSTECH, South Korea.
189. Parrinello, M., Rahman, A. (1981). Polymorphic transitions in single-crystals: a new molecular-dynamics method. *Journal of Applied Physics*, 52, 7182–7190.
190. Parthiban, V. (2006). Prediction of Factors Determining Changes in Stability in Protein Mutants. PhD thesis, Universität zu Köln.
191. Parthiban, V., Gromiha, M.M., Schomburg, D. (2006), CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Research*, 34, W239-W242.
192. Paul, M., Hazra, M., Barman, A., Hazra, S. (2014). Comparative molecular dynamics simulation studies for determining factors contributing to the thermostability of chemotaxis protein "CheY". *Journal of Biomolecular Structure and Dynamics*, 32(6), 928-49.

193. Pauptit, R.A., Karlsson, R., Picot, D., Jenkins, J.A., Niklaus-Reimer, A.S., Jansonius, J.N. (1988). Crystal structure of neutral protease from *Bacillus cereus* refined at 3.0 A resolution and comparison with the homologous but more thermostable enzyme thermolysin. *Journal of Molecular Biology*, 199(3), 525-37.
194. Pavelka, A., Chovancova, E., Damborsky, J. (2009). HotSpot Wizard: a Web Server for Identification of Hot Spots in Protein Engineering. *Nucleic Acids Research*, 37:W376-W383.
195. Paz, A., Mester, D., Baca, I., Nevo, E. (2004). Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. *PNAS*, 101, 2951-2956.
196. Perutz, M. F. (1978). Electrostatic effects in proteins. *Science*, 201, 1187–1191.
197. Perutz, M. F., & Raidt, H. (1975). Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. *Nature*, 25, 256–259.
198. Petersen, M.T.N., Fojan, P., Petersen, S.B. (2001). How do lipases and esterases work: the electrostatic contribution. *Journal of Biotechnology*, 85(2), 115-147.
199. Petsko, G. A. (2001). Structural basis of thermostability in hyperthermophilic proteins, or "there's more than one way to skin a cat". *Methods in Enzymology*, 334, 469–478.
200. Petsko, G.A., Ringe D. (2004). Bonds that stabilize folded proteins, in "Protein Structure and Function", New Science Press Ltd., London, pp.10-11.
201. Philip, J., Bell, L., Sunna, A., Gibbs, M.D., Curach, N.C., Nevalainen, H., Bergquist, P.L. (2002). Prospecting for novel lipase genes using PCR. *Microbiology*, 148, 2283–2291.
202. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26, 1781-1802.
203. Pleiss, J., Fischer, M., Peiker, M., Thiele, C., Schmid, R.D. (2000). Lipase engineering database understanding and exploiting sequence structure function relationships. *Journal of Molecular Catalysis B: Enzymatic*, 10, 91–508.
204. Poklar, N. Lah, J. Salobir, M. Maček, P. Vesnaver G. (1997). 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.
205. Potapov, V., Cohen, M., Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Engineering, Design & Selection*, 22(9), 553-60.
206. Privalov, P.L., Gill, S.J. (1988). Stability of protein structure and hydrophobic interaction. *Advances in Protein Chemistry*, 39, 191-234.
207. Quyen, D.T., Schmidt-Dannert, C., Schmid, R.D. (2003). High-level expression of a lipase from *Bacillus thermocatenulatus* BTL2 in *Pichia pastoris* and some properties of the recombinant lipase. *Protein Expression and Purification*, 28, 102-110.
208. Rathi, P., Bradoo, S., Saxena, R.K., Gupta, R. (2000). A hyper-thermostable, alkaline lipase from Pseudomonas sp. with the property of thermal activation. *Biotechnology Letters*, 22, 495–498.

209. Rathi, P.C., Jaeger, K.E., Gohlke, H. (2015). Structural Rigidity and Protein Thermostability in Variants of Lipase A from *Bacillus subtilis*. *PLoS One*, 10(7), e0130289.
210. Reed, C.J., Lewis, H., Trejo, E., Winston, V., Evilia, C. (2013). Protein Adaptations in Archaeal Extremophiles. *Archaea*, 2013, 1-14.
211. Reetz, M.T., Carballeira, J.D., Vogel, A. (2006). Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angewandte Chemie International Edition*, 45, 7745–7751.
212. Reid, K.S.C., Lindley, P.F., Thornton, J.M. (1985). Sulphur-aromatic interactions in proteins, *FEBS Letters*, 190 (2), 14 209-213.
213. Reysenbach, A. L., Shock, E. (2002). Merging Genomes with Geochemistry in Hydrothermal Ecosystems. *Science*, 296, 1077.
214. Rohl, C.A., Fiori, W., Baldwin, R.L. (1999) Alanine is helix-stabilizing in both template-nucleated and standard peptide helices. *PNAS*, 96(7), 3682-3687.
215. Rohl, C.A., Strauss, C.E., Misura, K.M., Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymology*, 383: 66-93.
216. Rokach, L. (2010). Chapter 14: A Survey of Clustering Algorithms. In Rokach L., Maimon O. (Eds.), Data Mining and Knowledge Discovery Handbook 2nd Edition, Part III, 269–298, Springer.
217. Rosato, V., Pucello, N., Giuliano, G. (2002). Evidence for cysteine clustering in thermophilic proteomes. Trends in Genetics, 18, 278–281.
218. Rose, G.D., Gierasch, L.M., Smith, J.A. (1985). Turns in peptides and proteins. *Advances in Protein Chemistry*, 37, 1–109.
219. Rosen, G.L., Reichenberger, E.R., Rosenfeld, A.M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, 27(1):127-9.
220. Rosenstein, R., Götz, F. (2000). Staphylococcal lipases, biochemical and molecular characterization. *Biochimie Journal*, 82(11), 1005-1014
221. Royter, M., Schmidt, M., Elend, C., Höbenreich, H., Schäfer, T., Bornscheuer, U.T., Antranikian, G. (2009). Thermostable lipases from the extreme thermophilic anaerobic bacteria *Thermoanaerobacter thermohydrosulfuricus SOL1* and *Caldanaerobacter subterraneus subsp. Tengcongensis*. *Extremophiles*, 13(5), 769–783.
222. Russell, R.J., Gerike, U., Danson, M.J., Hough, D.W., Taylor, G.L. (1998). Structural adaptations of the cold-active citrate synthase from an Antarctic bacterium. *Structure*, 6(3), 351-61.
223. Russell, R.J., Ferguson, J.M., Hough, D.W., Danson, M.J., Taylor, G.L. (1997). The crystal structure of citrate synthase from the hyperthermophilic archaeon *Pyrococcus furiosus* at 1.9 A resolution. *Biochemistry*, 36(33), 9983-94.
224. Russell, R.J., Ferguson, J.M., Hough, D.W., Danson, M.J., Taylor, G.L. (1997). The crystal structure of citrate synthase from the hyperthermophilic archaeon *Pyrococcus furiosus* at 1.9A° resolution. *Biochemistry*, 36(33), 9983–9994.
225. Russell, R.J.M., Ferguson, J.M.C., Haugh, D.W., Danson, M.J., Taylor, G.L. (1997). The Crystal Structure of Citrate Synthase from the Hyperthermophilic Archaeon *Pyrococcus furiosus* at 1.9 Å resolution. *Biochemistry*, 36, 9983–9994.

226. Saaty, T.L. (1982). Decision Making for Leaders; The Analytical Hierarchy Process for Decisions in a Complex World, Belmont, CA: Wadsworth.
227. Saaty, T.L. (2008). Decision making with the analytic hierarchy process. International Journal of Services Sciences, 1, 83-98.
228. Sadeghi, M., Naderi-Manesh, H., Zarrabi, M., Ranjbar, B. (2006). Effective factors in thermostability of thermophilic proteins. *Biophysical Chemistry*, 119(3), 256-70.
229. Sælensminde, G., Jr Ø.H., Jonassen, I. (2009). Amino acid contacts in proteins adapted to different temperatures: hydrophobic interactions and surface charges play a key role. *Extremophiles*, 13(1), 11-20.
230. Santarossa, G., Lafranconi, PG., Alquati, C., DeGioia, L., Alberghina, L., Fantucci, P., Lott, M. (2005). Mutations in the "lid" region affect chain length specificity and thermostability of a *Pseudomonas fragi* lipase. *Febs letters*, 579(11), 2383-2386.
231. Saraboji, K., Gromiha, M.M., Ponnuswamy, M.N. (2000). Importance of main-chain hydrophobic free energy to the stability of thermophilic proteins. *International Journal of Biological Macromolecules*, 2005(35), 211-20.
232. Satyapriya, R., &Vishveshwara, S. (2004). Interaction of DNA with clusters of amino acids in proteins. *Nucleic Acids Research*, 32, 4109–411.
233. Saunders, N.F.W., Thomas, T., Curmi, P.M., Mattick, J.S., Kuczek, E., Slade, R., Davis, J., Franzmann, P.D., Boone, D., Rusterholtz, K., Feldman, R., Gates, C., Bench, S., Sowers, K., Kadner, K., Aerts, A., Dehal, P., Detter, C., Glavina, T., Lucas, S., Richardson, P., Larimer, F., Hauser, L., Land, M., Cavicchioli, R. (2003). Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii. Genome Research*, 13, 1580–1588.
234. Scandurra, R., Consalvi, V., Chiaraluce, R., Politi, L., Engel, P.C. (1998). Protein thermostability in extremophiles. *Biochimie*, 80, 933–941.
235. Schrag, J.D., Cygler, M. (1997). Lipases and alpha/beta hydrolase fold. *Methods in Enzymology*, 284, 85-107.
236. Schumann, J., Bohm, G., Schumacher, G., Rudolph, R., Jaenicke, R. (1993). Stabilization of creatinase from *Pseudomonas putida* by random mutagenesis. *Protein Science*, 2(10), 1612-20.
237. Seeliger, D., de Groot, B.L. (2010). Protein thermostability calculations using alchemical free energy simulations. *Biophysical Journal*, 98(10), 2309-16.
238. Serrano, L., Bycroft, M., Fersht, A.R. (1991). Aromatic-aromatic interactions and protein stability. Investigation by double-mutant cycles. *Journal of Molecular Biology*, 218(2), 465-75.
239. Shalongo, W., Dugad, L., Stellwagen, E. (1994). Analysis of the thermal transitions of a model helical peptide using 13C NMR. *Journal of the American Chemical Society*, 116(6), 2500–2507.
240. Shin, D. S., Pratt, A. J., Tainer, J. A. (2014). Archaeal Genome Guardians Give Insights into Eukaryotic DNA Replication and Damage Response Proteins. *Archaea*, 2014, 206735.
241. Shindyalov, IN, Bourne, P.E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9), 739-747.

242. Shirley, B.A. (1995). Protein Stability and Folding: Theory and Practice. Humana Press, pp. 387.
243. Siddiqui, K.S., Cavicchioli, R. (2006). Cold-adapted enzymes. Annual Review of Biochemistry, 75, 403–433.
244. Siddiqui, K.S., Thomas, T. (2008). Protein Adaptation in Extremophiles: New York: Nova Biomedical Books.
245. Silver, A.M., Livesay, D.R. (2003). Optimized electrostatic surfaces parallel increased thermostability: a structural bioinformatic analysis. *Protein Engineering*, 16 (12), 871–874.
246. Singer, G.A.C., Hickey, D.A. (2003). Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene*, 317: 39–47.
247. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15, 327-332.
248. Soliman, N.A., Knoll, M., Abdel-Fattah, Y., Schmid, R.D., Lange, S. (2007). Molecular cloning and characterization of thermostable esterase and lipase from *Geobacillus thermoleovorans* YN isolated from desert soil in Egypt. *Protein Biochemistry*, 42(7), 1090-1100.
249. Spassov, V.Z., Karshikoff, A.D., Ladenstein, R. (1995). The optimization of protein-solvent interactions: thermostability and the role of hydrophobic and electrostatic interactions. *Protein Science,* 4, 1516-1527.
250. Spector, S., Wang, M., Carp, S.A., Robblee, J., Hendsch, Z.S., Fairman, R., Tidor, B., and Raleigh, D.P. (2000). Rational modification of protein stability by the mutation of charged surface residues. *Biochemistry*, 39, 872–879.
251. Srivastava, A, Sinha, S. (2014). Thermostability of *in vitro* evolved *Bacillus subtilis* lipase A: a network and dynamics perspective. *PLoS One*. 9(8), e102856.
252. Stanger, H.E., Syud, F.A., Espinosa, J.F., Giriat, I., Muir, T., Gellman, S.H. (2001). Length-dependent stability and strand length limits in antiparallel β-sheet secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 12015–12020.
253. Stark, W., Pauptit, R.A., Wilson, K.S., Jansonius, J.N. (1992). The structure of neutral protease from Bacillus cereus at 0.2-nm resolution. European Journal of Biochemistry, 207(2), 781 -791.
254. Stenico, M., Lloyd, A.T., Sharp, P.M. (1994). Codon usage in *Caenorhabditis elegans*, delineation of translational selection and mutational biases. *Nucleic Acids Research*, 22, 2437–2446.
255. Stetter, K. O. (1999). Extremophiles and their adaptation to hot environments. *FEBS Letters*, 452, 22-25.
256. Suen, W.C., Zhang, N.Y., Xiao, L., Madison, V., Zaks, A. (2004). Improved activity and thermostability of *Candida antarctica* lipase B by DNA family shuffling. *Protein Engineering Design & Selection*, 17(2), 133–140.
257. Sueoka, N., Kawanishi,Y. (2000). DNA G+C content of the third codon position and codon usage biases of human genes. *Gene*, 261, 53–62.

258. Sugihara, A., Ueshima, M., Shimada, Y., Tsunasawa, S., Tominaga, Y. (1992). Purification and Characterization of a Novel Thermostable Lipase from *Pseudomonas cepacia*. *Journal of Biochemistry*, 112(5), 598-603.
259. Sujak, A., Sanghamitra, N.J., Maneg, O., Ludwig, B., Mazumdar, S. (2007). Thermostability of proteins: role of metal binding and pH on the stability of the dinuclear CuA site of *Thermus thermophilus*. *Biophysical Journal*, 93(8), 2845-51.
260. Sun, S.Y., Xu, Y., Wang, D. (2009). Novel minor lipase from *Rhizopus chinensis* during solid-state fermentation: Biochemical characterization and its esterification potential for ester synthesis. *Bioresource Technology*, 100(9), 2607-2612.
261. Suvd, D., Fujimoto, Z., Takase, K., Matsumura, M., Mizuno, H. (2001). Crystal structure of *Bacillus stearothermophilus* aamylase: possible factors determining the thermostability. *Journal of Biochemistry*, 129, 461– 468.
262. Swaim, M.W., & Pizzo, S.V. (1988). Methionine sulfoxide and the oxidative regulation of plasma proteinase inhibitors. *Journal of Leukocyte Biology*, 43, 365-379.
263. Szilágyi, A. and Závodszky, P. (2000). Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure Folding and Design*, 8, 493-504.
264. Takagi, H., Takahashi, T., Momose, H., Inouye, M., Maeda, Y., Matsuzawa, H., and Ohta, T. 1990. 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. *The Journal of Biological Chemistry*, 265, 6874–6878.
265. Takano, K., Yamagata, Y., Fujii, S., Yutani, K. (1997). Contribution of the hydrophobic effect to the stability of human lysozyme: calorimetric studies and X-ray structural analyses of the nine valine to alanine mutants. *Biochemistry*, 36(4), 688-98.
266. Tanaka, Y., Tsumoto, K., Yasutake, Y., Umetsu, M., Yao, M., Fukada, H., Tanaka, I., Kumagai, I. (2004). How oligomerization contributes to the thermostability of an archaeon protein. Protein L-isoaspartyl-O-methyltransferase from *Sulfolobus tokodaii*. *The Journal of Biological Chemistry*, 279(31), 32957-67.
267. Tanner, J., Hecht, R. M., Krause, K. L. (1996). 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.
268. Tansey, M. R., Brock, T. D. (1971). Isolation of thermophilic and thermotolerant fungi from hot spring effluents and thermal soils of Yellowstone National Park. *Bacteriological Proceedings*, 36.
269. Teng, S., Srivastava, A.K. and Wang, L. (2010). Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC* Genomics, 11(Suppl 2), S5.
270. Thompson, M. J. & Eisenberg, D. (1999). Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *Journal of Molecular Biology*, 290, 595-604.

271. Tian, J., Wu, N., Guo, J., Fan, Y. (2009). Prediction of amyloid fibril-forming segments based on a support vector machine. *BMC Bioinformatics,* 10(Suppl. 1), S45.
272. Timasheff, S.N. (1995). Solvent stabilization of protein structure. In: Shirley BA. editor. Protein Stability and Folding: Theory and Practice. New Jersey: Humana Press, pp 253–269.
273. Tina, K. G., Bhadra, R., Srinivasan, N. (2007). PIC: Protein Interactions Calculator, *Nucleic Acids Research*, 35, W473–W476.
274. Tiwari, A., Panigrahi, S.K. (2007). HBAT: a complete package for analysing strong and weak hydrogen bonds in macromolecular crystal structures. *In Silico Biology*, 7(6), 651-61.
275. Trevino, S.R., Schaefer, S., Scholtz, J.M., Pace, C.N. (2007). Increasing Protein Conformational Stability by Optimizing β-turn Sequence. *Journal of Molecular Biology*, 373(1): 211-218.
276. Trivedi, S., Gehlot, H.S., Rao, S.R. (2006). Protein thermostability in Archaea and Eubacteria. *Genetic and Molecular Research*, 5(4), 816-827.
277. Tsai, P. C. (2009). Directed Evolution of Phosphotriesterase for Stereoselective Detoxification of Organophosphate Nerve Agents (Doctoral dissertation, Texas A&M University).
278. Turner, P., Mamo, G., Karlsson, E.N. (2007). Potential and utilization of thermophiles and thermostable enzymes in biorefining. *Microbial Cell Factories.* 6, 9.
279. Tyndall, J., Sinchaikul, S., Fothergill-Gilmore, L., Taylor, P., Walkinshaw, M.D. (2002). Crystal Structure of a Thermostable Lipase from *Bacillus stearothermophilus* P1. *Journal of Molecular Biology*, 323, 859-869.
280. Van den Burg ,B., Vriend, G., Veltman, O. R., Venema, G., Eijsink, V. G. H. (1998). Engineering an enzyme to resist boiling. *PNAS USA,* 95, 2056–2060.
281. Van den Burg, B., Dijkstra, B.W., Vriend, G., Van der Vinne, B., Venema, G., Eijsink, V.G. (1994). Protein stabilization by hydrophobic interactions at the surface. *European Journal of Biochemistry*, 220(3), 981-5.
282. van Pouderoyen, G., Eggert, T., Jaeger, K.E., Dijkstra, B.W. (2001). The crystal structure of Bacillus subtilis lipase: a minimal alpha/beta hydrolase fold enzyme. *Journal of Molecular Biology*, 309, 216–226.
283. Vedadi, M., Arrowsmith, C. H., Allali-Hassani, A., Senisterra, G., & Wasney, G. A. (2010). Biophysical characterization of recombinant proteins: A key to higher structural genomics success. *Journal of Structural Biology*, 172(1-2), 107–119.
284. Vehlow, C., Stehr, H., Winkelmann, M., Duarte, J.M., Petzold, L., Dinse, J., Lappe, M. (2011). CMView: interactive contact map visualization and analysis. *Bioinformatics*, 27(11), 1573-4.
285. Vendruscolo, M., Kussell, E., Domany, E. (1997). Recovery of protein structure from contact maps. *Folding Design*, 2(5), 295-306.
286. Ventura, S., Vega, M.C., Lacroix, E., Angrand, I., Spagnolo, L., and Serrano, L. (2002). Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nature Structural & Molecular Biology*, 9485–493.
287. Vetriani, C., Maeder, D. L., Tolliday, N., Yip, K. S. P., Stillman, T. J., Britton, K. L., Rice, D. W., Klump, H. H. and Robb, F. T. (1998). Protein thermostability

above 100°C: A key role for ionic interactions. *Proceedings of the National Academy of Sciences USA*, 95, 12300-12305.

288. Vieille, C., Zeikus, G.J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiology and Molecular Biology Reviews*, 65(1), 1-43.
289. Villeret, V., Clantin, B., Tricot, C., Legrain, C., Roovers, M., Stalon, V., Van Beeumen, J. (1998). The crystal structure of Pyrococcus furiosus ornithine carbamoyltransferase reveals a key role for oligomerization in enzyme stability at extremely high temperatures. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6), 2801–2806.
290. Vogt, G., Woell S, Argos P. (1997). Protein thermal stability: hydrogen bonds or internal packing? *Journal of Molecular Biology*, 269, 631–643.
291. Volkin, D. B., Middaugh, C. R. (1992). The effect of temperature on protein structure. In: Ahern T J, Manning M C, editors. Stability of protein pharmaceuticals. A Chemical and physical pathways of protein degradation. New York, N.Y: Plenum Press, pp. 215–247.
292. Wang, L.K., Shuman, S. (2001). Domain structure and mutational analysis of T4 polynucleotide kinase. *Journal of Biological Chemistry*, 276, 26868–26874.
293. Wang, L.K., Shuman, S. (2002). Mutational analysis defines the 5′ kinase and 3′ phosphatase active sites of T4 polynucleotide kinase. *Nucleic Acids Research*, 30, 1073–1080.
294. Wang, X., He, X., Yang, S., An, X., Chang, W., Liang, D. (2003). Structural Basis for Thermostability of ß-Glycosidase from the Thermophilic Eubacterium *Thermus nonproteolyticus* HG102. *Journal of Bacteriology*, 185, 4248-4255.
295. Wang, X.Y., Meng, F .G . & Zhou, H.M. (2004). Unfolding and inactivation during thermal denaturation of an enzyme that exhibits phytase and acid phosphatase activities. *The International Journal of Biochemistry & Cell Biology*, 36, 447-459.
296. Wang, Y., Srivastava, K.C., Shen, G.J., Wang, H.Y. (1995). Thermostable alkaline lipase from a newly isolated thermophilic *Bacillus* strain, A30-1 (ATCC 53841). Journal of Fermentation and Bioengineering, 79(5), 433–438.
297. Watson, J. D., Baker, T.A., Bell, S.P., Levine, M., Losick, R., CSHLP, I. (2008). Alexander Gann Eds. *Molecular Biology of the Gene*, Pearson; 6 edition.
298. Westers, H., Braun, P.G., Westers, L., Antelmann, H., Hecker, M., Jongbloed, J.D., Yoshikawa, H., Tanaka, T., van Dijl, J.M., Quax, W.J. (2005). Genes involved in SkfA killing factor production protect a *Bacillus subtilis* lipase against proteolysis. *Applied Environmental Microbiology*, 71(4), 1899-908.
299. Whitaker, J.R., & Feeney, R.E. (1983). Chemical and physical modification of proteins by the hydroxide ion. *Critical Reviews in Food Science and Nutrition*, 19, 173-212.
300. Wicker, N., Perrin, G.R., Thierry, J.C., Poch, O. (2001). Secator: A Program for Inferring Protein Subfamilies from Phylogenetic Trees. *Molecular Biology and Evolution*, 18(32), 1435-1441.
301. Wijma, H.J., Floor, R.J., Janssen, D.B. (2013). Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Current Opinion in Structural Biology*, 23(4), 588-94.

302. Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R. F., Sykes, B. D., & Wishart, D. S. (2003). VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Research*, 31(13), 3316–3319.
303. Worth, C.L., Preissner, R., Blundell, T.L. (2011). SDM--a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Research*, 39(Web Server issue), W215-22.
304. Wright, H.T. (1991). Sequence and structure determinants of the nonenzymatic deamidation of asparagine and glutamine residues in proteins. *Protein Engineering*, 4(3), 283-94.
305. Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L-S., Natale, D.A., Vinayaka, C.R., Hu, Z.Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R.S., Suzek, B.E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J.L., Chung, S., Castro-Alvear, J., Dinkov, G., Barker, W.C. (2004). PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Research*, 1, D112–D114.
306. Wu, L., Liu, B., Hong, Y., Sheng, D., Shen, Y., Ni, J. (2009). Residue Tyr224 is critical for the thermostability of *Geobacillus sp.* RD-2 lipase. *Biotechnology Letters*, 32, 107-112.
307. Wu, X. & Kumar, V. (2009). The Top Ten Algoritms in Data Mining, Chapman and Hall, Boca Raton.
308. Xiao, L., & B. Honig. (1999). Electrostatic contributions to the stability of hyperthermophilic proteins. *Journal of Molecular Biology*, 289, 1435–1444.
309. Xu, J., Baase, W.A., Baldwin, E., and Matthews, B.W. (1998).The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Protein Science*, 7, 158-177.
310. Yadav, R.P., Saxena, R.K., Gupta, R., Davidson, W.S. (1998). Purification and characterization of a regiospecific lipase from Aspergillus terreus. *Biotechnology and Applied Biochemistry*, 28(3), 243-249.
311. Yano, J.K., Poulos, T.L. (2003). New understandings of thermostable and peizostable enzymes. *Current Opinion in Biotechnology*, 14(4), 360-5.
312. Yaseen, A., & Li, Y. (2013). Dinosolve: a protein disulfide bonding prediction server using context-based features to enhance prediction accuracy. *BMC Bioinformatics*, 14(Suppl 13), S9.
313. Yin, S., Ding, F., Dokholyan, N.V. (2007). Modeling backbone flexibility improves protein stability estimation. *Structure*, 15(12):1567-76.
314. Yokota, K., Satou, K., Ohki, S. (2006). Comparative analysis of protein thermostability: Differences in amino acid content and substitution at the surfaces and in the core regions of thermophilic and mesophilic proteins. *Science and Technology of Advanced Materials*, 7(3), 255-262.
315. Zavodszky, M., Chen, C.-W., Huang, J.-K., Zolkiewski, M., Wen, L., & Krishnamoorthi, R. (2001). Disulfide bond effects on protein stability: Designed variants of Cucurbita maxima trypsin inhibitor-V. *Protein Science* , 10(1), 149–160.
316. Zeldovich, K.B., Berezovsky, I.N., Shakhnovich, E.I. (2007). Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Computational Biology*, 3(1), e5.

317. Zhang, M., Stauffacher, C.V., Linand, D., Etten, R.L. (1998). Crystal structure of a human low molecular weight phosphotyrosyl phosphatase implications for substrate specificity. *Journal of Biological Chemistry*, 273, 21714-21720.
318. Zhou, H.-X. (2002). A Gaussian-chain model for treating residual charge-charge interactions in the unfolded state of proteins. *PNAS USA*, 99, 3569–3574.
319. Zhou, X.X., Wang, Y.B., Pan, Y.J., Li, W.F. (2008). Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids*, 34, 25–33.
320. Zhu, K., Jutila, A., Tuominen, E.K.J., Patkar, S.A., Svendsen, A., Kinnunen, P.K.J. (2001). Impact of the tryptophan residues of Humicola lanuginosa lipase on its thermal stability. *Biochim et Biophys Acta*, 1547 (2), 329-338.
321. Zou, D., Ma, L., Yu, J., & Zhang, Z. (2015). Biological Databases for Human Research. *Genomics, Proteomics & Bioinformatics*, 13(1), 55–63.

PUBLICATIONS

Thesis Publications

Journal Papers

1. Debamitra C, Saravanan P, Dubey VK and Sanjukta Patra. *In Silico* Characterization of Thermostable Lipases. Extremophiles (2010) 15, 89-103.
2. Debamitra Chakravorty, Faheem Khan, Sanjukta Patra. Thermostability of proteins revisited through machine learning methodologies. Current Biotechnology (2015) Volume 5 (E-pub ahead of print).
3. Debamitra Chakravorty, Faheem Khan, Sanjukta Patra. The creation of Extremostable Protein Database **(Manuscript under preparation)**.
4. Debamitra Chakravorty, Sanjukta Patra. An insight into attaining thermostable proteins: process and challenge **(Manuscript under preparation)**.
5. Debamitra Chakravorty, Sanjukta Patra. Prioritizing thermostabilizing features and development of a two-step hybrid approach to attain thermostabilizing mutations **(Manuscript under preparation).**

Book Chapters

1. Debamitra Chakravorty, Ashwini Shreshtha, V.R.Sarath Babu and Sanjukta Patra. Molecular evolution of extremophiles. Chapter 1, in Exremophiles: Sustainable resources and Biotechnological Implications. Ed. 2. Singh OV; Wiley-Blackwell; USA.2012.
2. Debamitra Chakravorty and Sanjukta Patra. Attaining extremophiles and extremolytes: methodologies and limitations. Chapter 2, in Exremophiles: Sustainable resources and Biotechnological Implications. Ed. 4. Singh OV; Wiley-Blackwell; USA.2012.
3. Debamitra Chakravorty, Sanjukta Patra. Advance Techniques in Enzyme Research In Advances in Enzyme Biotechnology. Ed. P. Shukla; Springer; India. 2013. pp 89-109.

Award

1. 4th prize cash award at GBP`s Talent Search contest on Innovative Research Ideas Leading To Entrepreneurial Venture In Biotechnology and Allied Areas. For, oral presentation of “A two-step method to generate protein thermostable mutants for industrial applications”.

Conference/Workshop presentation

1. Thermostable lipases and *in silico* characterization. Chakravorty D, Saravanan P, Alpana Thorat, Ayan Sadhukhan, Sanjukta Patra. 48th Annual conference AMI National Chemical Laboratory, Pune, December 15-19, 2009.
2. International conference on "Biomolecular forms and functions". 2013. Enhancing thermostability of *Bacillus* lipases by increment of γ-turns. Charavorty D, Patra S. Indian Institute of Science, Bangalore.
3. Analytical Hierarchical Process Ranking Method to Distinguish Thermostable and Mesostable Proteins. Debamitra Chakravorty, Sanjukta Patra. Symposium cum Workshop on Advances in Computational Biology and Computer Aided Drug Design, .Bioinformatics Infrastructure Facility, IIT Guwahati, Guwahati 2015. **(Oral presentation)**.
4. Debamitra Chakravorty. A two-step method to generate protein thermostable mutants for industrial applications. GBP`s Talent Search contest on Innovative Research Ideas Leading To Entrepreneurial Venture In Biotechnology and Allied Areas. Guwahati, 20th-21st November, 2015. **(Oral presentation)**.
5. 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, Singapore, 25th -27th November, 2015.

Journal (Other than Thesis work)

1. Design of lead peptide drugs from mushroom targeting cysteine proteases. D Chakravorty, S Singh, P Saravanan, S Patra. Medicinal Chemistry Research, 2013. 22 (4), 2038-2049.
2. Unraveling the rationale behind organic solvent stability of lipases. D Chakravorty, S Parameswaran, VK Dubey, S Patra. Applied biochemistry and biotechnology, 2012. 167 (3), 439-461.
3. An insight into plant lipase research–challenges encountered. S Seth, D Chakravorty, VK Dubey, S Patra. Protein expression and purification, 2014. 95, 13-21.
4. Debamitra Chakravorty, Sanjukta Patra. Significance of Lipolytic Enzymes in Pathogenesis and Treatment of Neglected Diseases. Current Protein and Peptide Science, 2016, 17, 000-000. **(Accepted).**

Conference (Other than Thesis work)

1. Debamitra Chakravorty, Sumit Singh, Sanjukta Patra. Design of peptide inhibitors for cysteine proteases. Society of Biological Chemists 12-15th Nov 2011. CIMAP Lucknow, India.
2. Sonali Seth, Debamitra Chakravorty, Sanjukta Patra. An insight into pH stability of

enzymes.Society of Biological Chemists 12-15th Nov 2011. CIMAP Lucknow, India.

3. Debamitra Chakravorty, Sanjukta Patra. The rationale behind organic solvent stability of lipases: an in silico approach. Society of Biological Chemists 12-15th Nov 2011. CIMAP Lucknow, India.
4. Chakravorty D, Singh SK, Sanjukta Patra. In silico characterization of mechanism of action of cysteine protease inhibitors from mushroom. Annual conference AMI, Birla Institute of Technology, Ranchi, 14-17 Dec, 2010.
5. Raveendran A, Chakravorty D, Sanjukta Patra. Analysis of bactericidal activity from mushroom extract of North East India. Annual conference AMI, Birla Institute of Technology, Ranchi, 14-17 Dec, 2010.
6. Saravanan Parameswaran, Alpana Thorat, Debamitra Chakravorty, Sanjukta Patra. In Silico Characterization and Structural Modeling of Thermoactive and Alkaline Staphylococcus Lipase. Asia-Pacific Bioinformatics Conference, NCBS, Jan 18-21, 2010.

Appendix I

Table A1: The 127 thermostable and mesostable protein pairs forming the final dataset

PDB ID Thermo	Source	*T °C	Macromolecule name	PDB ID Meso	*T °C	Macromolecule name
1uz5	*Pyrococcus horikoshii*	98	Molybdopterin biosynthesis moea protein	1bev1	38	Bovine enterovirus coat proteins vp1 to vp4
1t1g	*Bacillus sp. Mn-32*	60	Kumamolisin	1bh6	37	Subtilisin
1thm	*Thermoactinomyces vulgaris*	50	Thermitase	1bh6	37	Subtilisin
2ajr	*Thermotoga maritima*	80	Sugar kinase, pfkb family	1bx4	37	Protein (adenosine kinase)
1nee	*Methanothermobacter thermautotrophicus*	65	Probable translation initiation factor 2 beta subunit	1cf5	20	Protein (beta-momorcharin)
1rfk	*Mastigocladus laminosus*	75	Ferredoxin	1czp	20	Ferredoxin i
1j6r	*Thermotoga maritima*	80	Methionine synthase	1d4m	37	Protein (coxsackievirus a9)
1nz0	*Thermotoga maritima*	80	Ribonuclease P protein component	1d6t	37	Ribonuclease p
1ayg	*Hydrogenobacter thermophilus*	70	Cytochrome C-552	1dvv	30	Cytochrome c551
1clc	*Clostridium thermocellum*	60	Endoglucanase celd; ec: 3.2.1.4	1e1f	20	Beta-glucosidase
1ixk	*Pyrococcus horikoshii*	98	Methyltransferase	1ej0	37	Ftsj
1tzv	*Thermotoga maritima*	80	N utilization substance protein B homolog	1ey1	37	Antitermination factor nusb
1io9	*Sulfolobus solfataricus*	85	Cytochrome p450 cyp119	1f20	37	Nitric-oxide synthase
1lfp	*Aquifex aeolicus*	85	Hypothetical protein aq_1575	1f5n	37	Interferon-induced guanylate-binding protein 1
1mpp	*Rhizomucor pusillus*	50	Pepsin	1fmx	25	Saccharopepsin
1v37	*Thermus thermophilus*	75	Phosphoglycerate mutase	1fzt	24	Phosphoglycerate mutase

PDB ID Thermo	Source	T °C	Macromolecule name	PDB ID Meso	T°C	Macromolecule name
1pzx	*Geobacillus stearothermophilus*	55	Hypothetical protein apc36103	1g7n	37	Adipocyte lipid-binding protein
1vkz	*Thermotoga maritima*	80	Phosphoribosylamine--glycine ligase	1gso	37	Protein (glycinamide ribonucleotide synthetase)
1gw0	*Melanocarpus albomyces*	50	Laccase-1	1gyc	20	Laccase 2
1vp5	*Thermotoga maritima*	80	2,5-diketo-d-gluconic acid reductase	1hw6	30	2,5-diketo-d-gluconic acid reductase
1vku	*Thermotoga maritima*	80	Acyl carrier protein	1hy8	20	Acyl carrier protein
1wr2	*Pyrococcus horikoshii*	98	Hypothetical protein ph1789	1j0n	30	Xanthan lyase
1o5z	*Thermotoga maritima*	80	Folylpolyglutamate synthase/dihydrofolate synthase	1jbw	37	Folylpolyglutamate synthase
1jdq	*Thermotoga maritima*	80	Hypothetical protein tm0983	1je3	37	Hypothetical 8.6 kda protein in amya-flie intergenic region
1zar	*Archaeoglobus fulgidus*	82	Rio2 kinase	1jkk	37	Death-associated protein kinase
1x0t	*Pyrococcus horikoshii*	98	Ribonuclease p protein component 4	1jln	37	Protein tyrosine phosphatase, receptor type, r
1xqo	*Pyrobaculum aerophilum*	100	8-oxoguanine dna glycosylase	1jxo	37	Postsynaptic density protein
1lab	*Geobacillus stearothermophilus*	55	Dihydrolipoamide acetyltransferase	1k8o	37	E2 component of branched-chain alpha-ketoacid dehydrogenase
1v7r	*Pyrococcus horikoshii*	98	Hypothetical protein ph1917	1kfq	28	Phosphoglucomutase 1
1vr8	*Thermotoga maritima*	80	Gtp binding regulator	1kfr	25	Hemoglobin
1jcu	*Methanothermobacter thermautotrophicus*	65	Conserved protein mth1692	1kk9	37	Probable translation factor ycio
1jg1	*Pyrococcus furiosus*	100	Protein-l-isoaspartate o-methyltransferase	1kr5	37	Protein-l-isoaspartate o-methyltransferase
1olr	*Humicola grisea*	45	Endo-beta-1,4-glucanase	1ks4	30	Endoglucanase a
1vcd	*Thermus thermophilus*	75	Ndx1	1ktg	25	Diadenosine hydrolase

PDB ID Thermo	Source	T °C	Macromolecule name	PDB ID Meso	T°C	Macromolecule name
2c4x	*Clostridium thermocellum*	60	Endoglucanase	1l0q	37	Surface layer protein
1xjk	*Thermotoga maritima*	80	Ribonucleotide reductase, b12-dependent	1l11	37	Ribonucleoside triphosphate reductase
1ldn	*Geobacillus stearothermophilus*	80	L-lactate dehydrogenase	1ldg	37	L-lactate dehydrogenase
1bqc	*Thermobifida fusca*	45	Protein (beta-mannanase)	1lf1	20	Cel5
1xhc	*Pyrococcus furiosus*	100	Nadh oxidase /nitrite reductase	1m6i	37	Programmed cell death protein 8
1wl7	*Geobacillus thermodenitrificans*	60	Arabinanase-ts	1mdw	37	Calpain ii, catalytic subunit
1oi0	*Archaeoglobus fulgidus*	82	Hypothetical protein af2198	1mqa	37	Integrin alpha-l
2d2e	*Thermus thermophilus*	75	Sufc protein	1mt0	37	Hemolysin secretion atp-binding protein
1kgs	*Thermotoga maritima*	80	Dna binding response regulator d	1mvo	20	Phop response regulator
1i1w	*Thermoascus aurantiacus*	45	Endo-1,4-beta-xylanase	1mzd	37	Pro-granzyme k
1jrm	*Methanothermobacter thermautotrophicus*	65	Conserved hypothetical protein mth637	1n91	37	Orf, hypothetical protein
1iv0	*Thermus thermophilus*	75	Hypothetical protein	1nmn	37	Hypothetical protein yqgf
1urs	*Alicyclobacillus acidocaldarius*	60	Maltose-binding protein	1nnf	37	Iron-utilization periplasmic protein
1l7m	*Methanocaldococcus jannaschii*	85	Phosphoserine phosphatase	1nnl	37	L-3-phosphoserine phosphatase
1wot	*Thermus thermophilus*	75	Putative minimal nucleotidyltransferase	1no5	37	Hypothetical protein hi0073
2etd	*Thermotoga maritima*	80	Lema protein	1nwk	37	Actin, alpha skeletal muscle
1pz3	*Geobacillus stearothermophilus*	55	Alpha-l-arabinofuranosidase	1nxc	37	Mannosyl-oligosaccharide 1,2-alpha-mannosidase ia

PDB ID Thermo	Source	T °C	Macromolecule name	PDB ID Meso	T °C	Macromolecule name
1n75	*Thermus thermophilus*	75	Glutamyl-trna synthetase	1nyl	37	Glutaminyl-trna synthetase
1t6c	*Aquifex aeolicus*	85	Exopolyphosphatase	1nyn	25	Hypothetical 12.0 kda protein in nam8-gar1 intergenic region
1iq0	*Thermus thermophilus*	75	Arginyl-trna synthetase	1nzj	37	Hypothetical protein yadb
1t4y	*Synechococcus elongatus*	60	Adaptive-response sensory-kinase sasa	1o2f	37	Pts system, glucose-specific iia component
1zko	*Thermotoga maritima*	80	Glycine cleavage system h protein	1o78	30	Biotin carboxyl carrier protein of methylmalonyl-coa carboxyl-transferase
1z5z	*Sulfolobus solfataricus*	85	Helicase of the snf2/rad54 family	1oyy	37	Atp-dependent dna helicase
1pmh	*Caldicellulosiruptor saccharolyticus*	65	Beta-1,4-mannanase	1oyz	37	Hypothetical protein yiba
1t6t	*Aquifex aeolicus*	85	Putative protein	1pui	37	Probable gtp-binding protein engb
1vbl	*Bacillus sp. Ts-47*	60	Pectate lyase 47	1pxz	20	Major pollen allergen jun a 1
1uet	*Archaeoglobus fulgidus*	82	Trna nucleotidyltransferase	1q6d	20	Beta-amylase
1o98	*Bacillus stearothermophilus*	55	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	1q8k	37	Eukaryotic translation initiation factor 2 subunit 1
1mrz	*Thermotoga maritima*	80	Riboflavin kinase/fmn adenylyltransferase	1q9s	37	Hypothetical protein flj11149
1mgt	*Thermococcus kodakarensis*	95	Protein (o6-methylguanine-dna methyltransferase)	1qnt	37	Methylated-dna--protein-cysteine methyltransferase
2bm3	*Clostridium thermocellum*	60	Scaffolding dockerin binding protein a	1qzn	37	Cellulosomal scaffoldin adaptor protein b
1wf3	*Thermus thermophilus*	75	Gtp-binding protein	1rfl	37	Probable trna modification gtpase trme
1vrx	*Acidothermus cellulolyticus*	60	Endocellulase e1	1rh9	25	Endo-beta-mannanase

PDB ID Thermo	Source	T °C	Macromolecule name	PDB ID Meso	T°C	Macromolecule name
1ud9	*Sulfolobus tokodaii*	80	Dna polymerase sliding clamp a	1ri6	37	Putative isomerase ybhe
1ye8	*Aquifex aeolicus*	85	Hypothetical upf0334 kinase-like protein aq_1292	1rkb	37	Protein ad-004
1vjr	*Thermotoga maritima*	80	4-nitrophenylphosphatase	1rkq	37	Hypothetical protein yida
1mxg	*Pyrococcus woesei*	100	Alpha amylase	1rpa	20	Prostatic acid phosphatase
1q0u	*Geobacillus stearothermophilus*	55	Bstdead	1s2m	25	Putative atp-dependent rna helicase dhh1
1lva	*Moorella thermoacetica*	55	Selenocysteine-specific elongation factor	1sjx	38	Immunoglobulin vh domain
1c3p	*Aquifex aeolicus*	85	Protein (hdlp (histone deacetylase-like protein))	1sy1	20	Nitrophorin 4
1y8a	*Archaeoglobus fulgidus*	82	Hypothetical protein af1437	1tf1	37	Negative regulator of allantoin and glyoxylate utilization operons
1tty	*Thermotoga maritima*	80	Rna polymerase sigma factor rpod	1tlh	37	10 kda anti-sigma factor
1t7l	*Thermotoga maritima*	80	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase	1u22	18	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase
1pe5	*Bacillus thermoproteolyticus*	60	Thermolysin	1u4g	30	Elastase
1vkc	*Pyrococcus furiosus*	100	Putative acetyl transferase	1u6m	37	Acetyltransferase, gnat family
1zy9	*Thermotoga maritima*	80	Alpha-galactosidase	1uas	20	Alpha-galactosidase
1wj9	*Thermus thermophilus*	75	Crispr-associated protein	1ued	28	P450 monooxygenase
1lf6	*Thermoanaerobacterium thermosaccharolyticum*	60	Glucoamylase	1ulv	30	Glucodextranase
1u4h	*Thermoanaerobacter tengcongensis*	75	Heme-based methyl-accepting chemotaxis protein	1upw	37	Oxysterols receptor lxr-beta

PDB ID Thermo	Source	T °C	Macromolecule name	PDB ID Meso	T °C	Macromolecule name
1dq3	*Pyrococcus furiosus*	100	Endonuclease	1v5d	20	Chitosanase
1gh8	*Methanothermobacter thermautotrophicus*	65	Translation elongation factor 1beta	1va9	37	Down syndrome cell adhesion molecule like-protein 1b
2ars	*Thermoplasma acidophilum*	59	Lipoate-protein ligase a	1vqz	37	Lipoate-protein ligase, putative
1qo2	*Thermotoga maritima*	80		1vzw	28	Phosphoribosyl isomerase a
1t95	*Archaeoglobus fulgidus*	82	Hypothetical protein af0491	1w45	37	Annexin a8
1o2d	*Thermotoga maritima*	80	Alcohol dehydrogenase, iron-containing	1wik	37	Thioredoxin-like protein 2
1o7i	*Sulfolobus solfataricus*	85	Single stranded dna binding protein	1wjj	18	Hypothetical protein f20o9.120
1mtz	*Thermoplasma acidophilum*	59	Proline iminopeptidase	1wm1	30	Proline iminopeptidase
1ujp	*Thermus thermophilus*	75	Tryptophan synthase alpha chain	1wq5	37	Tryptophan synthase alpha chain
1jl2	*Escherichia coli, Thermus thermophilus*	75	Chimeric rnase h	1wsh	37	Ribonuclease hi
1oz9	*Aquifex aeolicus*	85	Hypothetical protein aq_1354	1xax	37	Hypothetical upf0054 protein hi0004
1z0w	*Archaeoglobus fulgidus*	82	Putative protease la homolog type	1xmj	37	Cystic fibrosis transmembrane conductance regulator
1cz4	*Thermoplasma acidophilum*	59	Vcp-like atpase	1xmv	37	Reca protein
1ryj	*Methanothermococcus thermolithotrophicus*	65	Unknown	1xs3	37	Hypothetical protein xc975
1in4	*Thermotoga maritima*	80	Holliday junction dna helicase ruvb	1xwi	37	Skd1 protein
1vlm	*Thermotoga maritima*	80	Sam-dependent methyltransferase	1xxl	20	Ycgj protein
1ix5	*Methanothermococcus thermolithotrophicus*	65	Fkbp	1y0o	18	Fkbp-type peptidyl-prolyl cis-trans isomerase 3

PDB ID Thermo	Source	T °C	Macromolecule name	PDB ID Meso	T °C	Macromolecule name
2bog	*Thermomonospora fusca*	45	Endoglucanase e-2	1y7m	20	Hypothetical protein
1pbt	*Thermotoga maritima*	80	6-phosphogluconolactonase	1y89	28	Devb protein
1vhu	*Archaeoglobus fulgidus*	82	Hypothetical protein af1521	1yd9	37	Core histone macro-h2a.1
1zdr	*Geobacillus stearothermophilus*	55	Dihydrofolate reductase	1yho	37	Dihydrofolate reductase
1ku0	*Geobacillus stearothermophilus*	55	L1 lipase	1ys2	28	Lipase
1sau	*Archaeoglobus fulgidus*	82	Sulfite reductase, desulfoviridin-type subunit gamma	1yx3	25	Hypothetical protein
1u04	*Pyrococcus furiosus*	100	Hypothetical protein pf0537	1z6t	37	Apoptotic protease activating factor 1
1ilo	*Methanothermobacter thermautotrophicus*	65	Conserved hypothetical protein mth895	1z8f	37	Guanylate kinase
1ytl	*Archaeoglobus fulgidus*	82	Acetyl-coa decarboxylase/synthase complex epsilon subunit 2	1zbe2	38	Coat protein vp1
1wg8	*Thermus thermophilus*	75	Predicted s-adenosylmethionine-dependent methyltransferase	1zq9	37	Probable dimethyladenosine transferase
1z8s	*Geobacillus stearothermophilus*	55	Dna primase	1zrh	37	Heparan sulfate glucosamine 3-o-sulfotransferase 1
1sfs	*Geobacillus stearothermophilus*	55	Hypothetical protein	1zsw	30	Glyoxalase family protein
1ui9	*Thermus thermophilus*	75	Chorismate mutase	2a22	20	Vacuolar protein sorting 29
1kkh	*Methanocaldococcus jannaschii*	85	Mevalonate kinase	2a2d	37	N-acetylgalactosamine kinase
1ihn	*Methanothermobacter thermautotrophicus*	65	Hypothetical protein mth938	2ab1	37	Hypothetical protein
1v3y	*Thermus thermophilus*	75	Peptide deformylase	2ai9	37	Peptide deformylase

PDB ID Thermo	Source	T °C	Macromolecule name	PDB ID Meso	T °C	Macromolecule name
1xbi	*Methanocaldococcus jannaschii*	85	50s ribosomal protein l7ae	2aif	25	Ribosomal protein l7a
1ryq	*Pyrococcus furiosus*	100	Dna-directed rna polymerase, subunit e"	2aou	37	Histamine n-methyltransferase
1vhn	*Thermotoga maritima*	80	Putative flavin oxidoreductase	2b0m	37	Dihydroorotate dehydrogenase, mitochondrial
1o0x	*Thermotoga maritima*	80	Methionine aminopeptidase	2b3l	37	Methionine aminopeptidase 1
1nv8	*Thermotoga maritima*	80	Hemk protein	2b3t	37	Protein methyltransferase hemk
1ufk	*Thermus thermophilus*	75	Tt0836 protein	2b3t	37	Protein methyltransferase hemk
1j6o	*Thermotoga maritima*	80	Tatd-related deoxyribonuclease	2b75	37	Lysozyme
1d1n	*Geobacillus stearothermophilus*	55	Initiation factor 2	2crv	37	Translation initiation factor if-2
1v98	*Thermus thermophilus*	75	Thioredoxin	2fch	37	Thioredoxin 1
1brf	*Pyrococcus furiosus*	100	Protein (rubredoxin)	2rdv	30	Rubredoxin
1caa	*Pyrococcus furiosus*	80	Rubredoxin	8rxn	37	Rubredoxin

*****T -Temperature,
Thermo- Thermostable Proteins;
Meso- Mesostable Proteins.

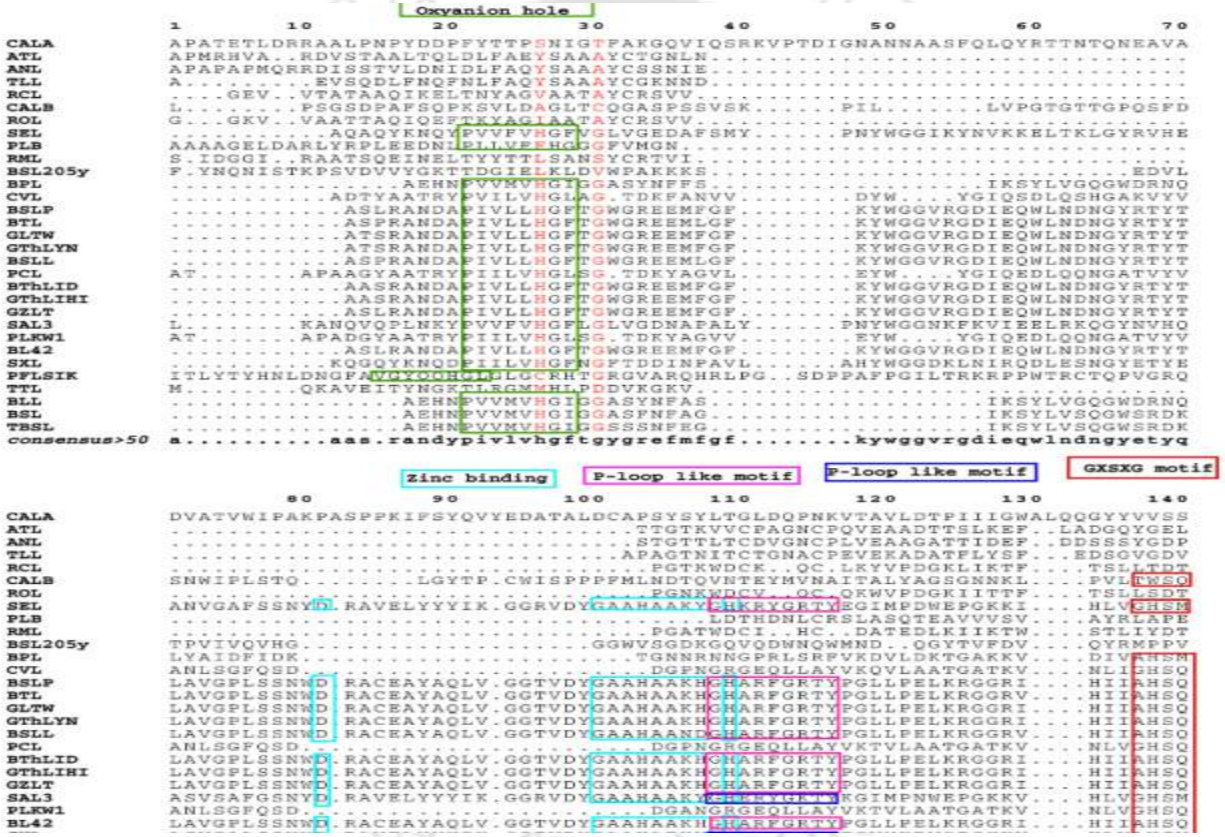
Table A2: Dipeptide composition of 127 thermostable proteins

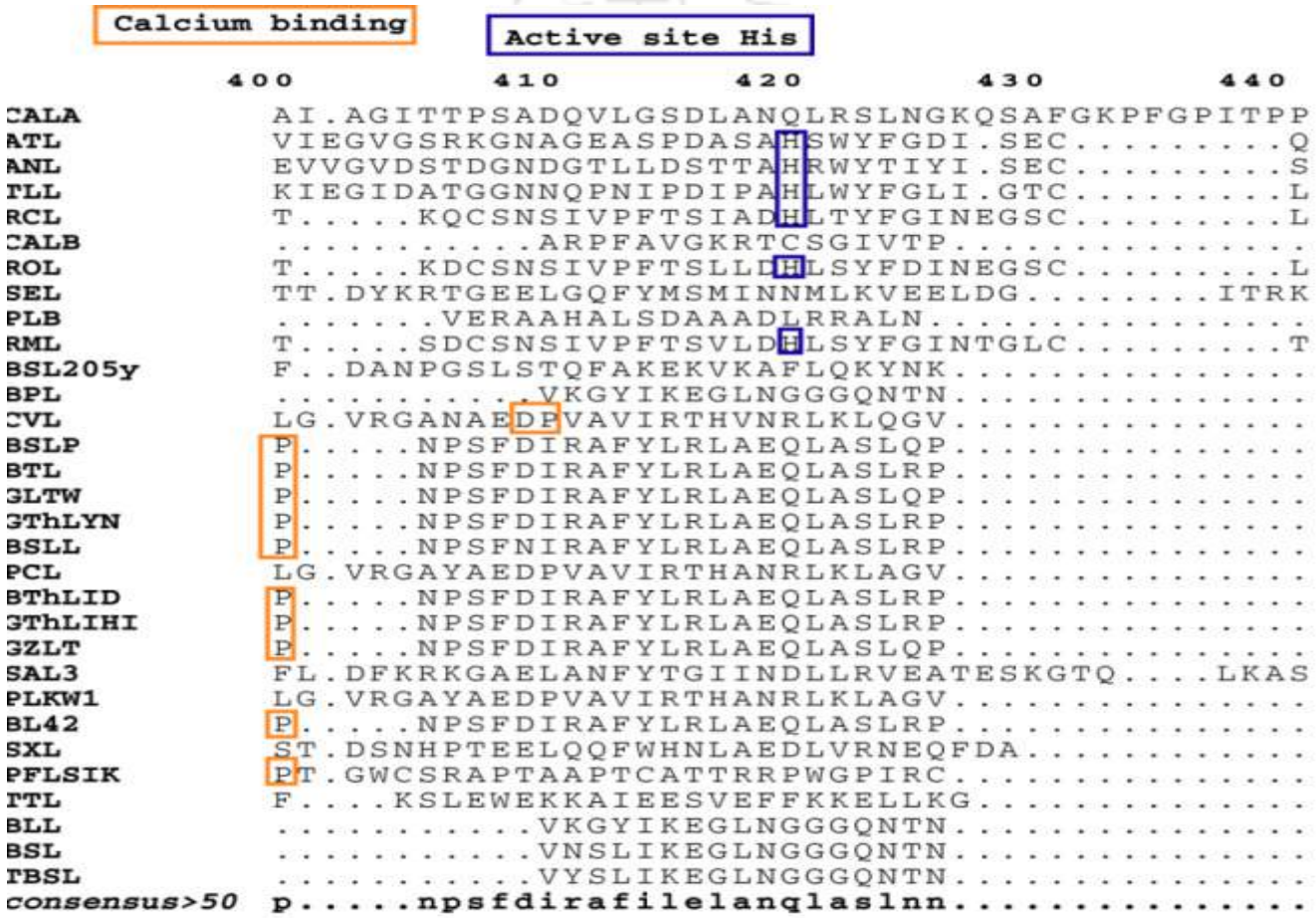
Amino acid	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	0.56	0.09	0.29	0.55	0.30	0.58	0.11	0.50	0.62	0.80	0.14	0.26	0.36	0.21	0.54	0.36	0.34	0.64	0.08	0.27
Cys	0.08	0.00	0.03	0.05	0.02	0.11	0.04	0.03	0.05	0.08	0.02	0.03	0.07	0.02	0.09	0.05	0.03	0.07	0.00	0.03
Asp	0.39	0.06	0.25	0.53	0.27	0.37	0.09	0.34	0.19	0.65	0.11	0.17	0.28	0.17	0.25	0.19	0.26	0.49	0.11	0.26
Glu	0.75	0.03	0.47	1.15	0.29	0.72	0.12	0.71	0.84	0.87	0.15	0.29	0.24	0.17	0.55	0.26	0.38	0.85	0.15	0.24
Phe	0.21	0.03	0.28	0.35	0.15	0.31	0.08	0.19	0.24	0.42	0.05	0.11	0.16	0.11	0.25	0.26	0.20	0.31	0.04	0.12
Gly	0.52	0.09	0.37	0.63	0.34	0.56	0.12	0.53	0.66	0.57	0.12	0.23	0.25	0.18	0.47	0.40	0.40	0.60	0.12	0.33
His	0.15	0.04	0.10	0.14	0.07	0.16	0.06	0.12	0.07	0.16	0.02	0.02	0.13	0.03	0.07	0.08	0.09	0.13	0.04	0.05
Ile	0.50	0.08	0.46	0.62	0.22	0.44	0.14	0.38	0.49	0.56	0.09	0.20	0.35	0.19	0.34	0.38	0.33	0.50	0.05	0.21
Lys	0.61	0.03	0.37	0.74	0.34	0.52	0.09	0.61	0.70	0.61	0.11	0.35	0.28	0.15	0.45	0.21	0.29	0.58	0.10	0.25
Leu	0.75	0.05	0.57	1.04	0.33	0.59	0.12	0.44	0.77	0.83	0.17	0.30	0.47	0.20	0.59	0.59	0.35	0.65	0.12	0.21
Met	0.19	0.00	0.13	0.18	0.08	0.16	0.04	0.10	0.21	0.20	0.04	0.08	0.10	0.02	0.15	0.11	0.08	0.16	0.02	0.04
Asn	0.28	0.03	0.12	0.30	0.16	0.31	0.08	0.21	0.18	0.29	0.08	0.13	0.22	0.06	0.24	0.15	0.16	0.31	0.04	0.11
Pro	0.28	0.04	0.33	0.50	0.15	0.41	0.09	0.26	0.26	0.34	0.07	0.16	0.21	0.17	0.24	0.27	0.22	0.34	0.06	0.17
Gln	0.25	0.01	0.10	0.21	0.06	0.18	0.04	0.22	0.20	0.19	0.05	0.11	0.09	0.06	0.08	0.12	0.12	0.16	0.04	0.09
Arg	0.45	0.04	0.32	0.59	0.23	0.45	0.09	0.40	0.47	0.56	0.08	0.23	0.19	0.13	0.41	0.23	0.20	0.52	0.06	0.22
Ser	0.34	0.04	0.24	0.40	0.18	0.47	0.09	0.26	0.32	0.47	0.07	0.14	0.26	0.16	0.23	0.28	0.21	0.37	0.07	0.17
Thr	0.30	0.06	0.25	0.31	0.18	0.37	0.07	0.31	0.28	0.46	0.08	0.13	0.30	0.07	0.20	0.24	0.16	0.39	0.06	0.17
Val	0.67	0.10	0.51	0.59	0.31	0.48	0.19	0.61	0.62	0.73	0.13	0.31	0.41	0.18	0.48	0.40	0.37	0.73	0.09	0.26
Trp	0.08	0.02	0.08	0.11	0.06	0.09	0.04	0.10	0.09	0.11	0.01	0.08	0.02	0.04	0.08	0.07	0.08	0.12	0.03	0.03
Tyr	0.28	0.03	0.19	0.27	0.13	0.28	0.07	0.24	0.16	0.30	0.04	0.14	0.20	0.09	0.21	0.16	0.13	0.25	0.03	0.11

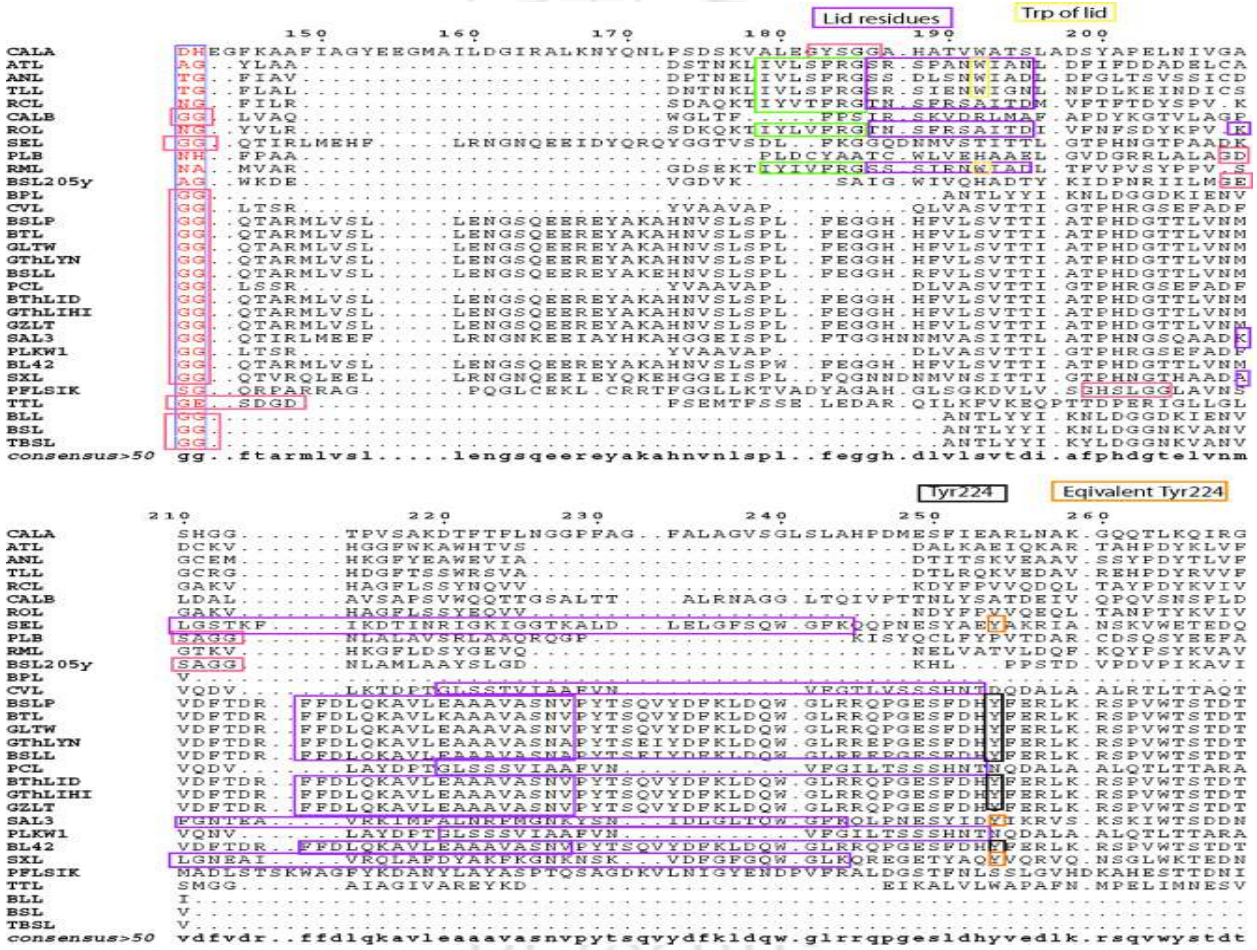
Table A3: Dipeptide composition of 127 mesostable proteins

Amino acid	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
Ala	0.88	0.11	0.46	0.59	0.30	0.64	0.17	0.50	0.47	0.75	0.16	0.27	0.39	0.31	0.39	0.49	0.51	0.61	0.08	0.26
Cys	0.09	0.01	0.06	0.08	0.05	0.15	0.04	0.07	0.07	0.10	0.01	0.06	0.11	0.04	0.10	0.08	0.09	0.08	0.01	0.03
Asp	0.39	0.08	0.39	0.43	0.25	0.51	0.12	0.36	0.35	0.54	0.12	0.26	0.36	0.20	0.17	0.31	0.32	0.52	0.11	0.17
Glu	0.65	0.08	0.39	0.48	0.26	0.43	0.15	0.46	0.47	0.66	0.15	0.23	0.22	0.32	0.28	0.32	0.33	0.47	0.11	0.23
Phe	0.25	0.05	0.24	0.24	0.11	0.26	0.08	0.17	0.23	0.29	0.06	0.16	0.17	0.14	0.18	0.33	0.20	0.27	0.04	0.13
Gly	0.66	0.12	0.49	0.46	0.27	0.55	0.18	0.43	0.55	0.58	0.11	0.28	0.25	0.29	0.34	0.45	0.56	0.60	0.09	0.31
His	0.16	0.03	0.06	0.15	0.15	0.21	0.07	0.15	0.14	0.23	0.05	0.07	0.10	0.07	0.09	0.13	0.11	0.13	0.06	0.08
Ile	0.49	0.08	0.36	0.41	0.17	0.40	0.17	0.37	0.39	0.52	0.10	0.30	0.27	0.24	0.30	0.33	0.31	0.46	0.05	0.14
Lys	0.50	0.06	0.34	0.49	0.19	0.45	0.11	0.39	0.42	0.61	0.12	0.32	0.32	0.24	0.26	0.40	0.28	0.49	0.09	0.23
Leu	0.78	0.08	0.54	0.63	0.33	0.62	0.20	0.48	0.62	0.83	0.18	0.39	0.45	0.39	0.48	0.61	0.51	0.62	0.06	0.26
Met	0.23	0.04	0.10	0.14	0.05	0.14	0.03	0.12	0.25	0.22	0.02	0.07	0.10	0.06	0.10	0.18	0.14	0.15	0.01	0.07
Asn	0.27	0.06	0.18	0.24	0.16	0.36	0.10	0.26	0.26	0.35	0.09	0.19	0.25	0.16	0.19	0.26	0.20	0.29	0.07	0.13
Pro	0.48	0.05	0.31	0.34	0.13	0.40	0.10	0.22	0.24	0.40	0.08	0.21	0.27	0.20	0.15	0.30	0.32	0.32	0.05	0.18
Gln	0.31	0.05	0.24	0.21	0.13	0.34	0.06	0.27	0.22	0.39	0.08	0.18	0.20	0.25	0.19	0.19	0.19	0.25	0.06	0.11
Arg	0.33	0.05	0.26	0.36	0.17	0.27	0.12	0.32	0.31	0.47	0.09	0.18	0.18	0.17	0.24	0.28	0.16	0.32	0.04	0.17
Ser	0.41	0.07	0.35	0.43	0.27	0.54	0.16	0.35	0.31	0.58	0.08	0.24	0.29	0.24	0.31	0.45	0.37	0.38	0.09	0.21
Thr	0.51	0.09	0.27	0.30	0.19	0.48	0.11	0.30	0.35	0.55	0.08	0.22	0.29	0.17	0.20	0.34	0.29	0.54	0.08	0.15
Val	0.60	0.18	0.53	0.47	0.26	0.53	0.17	0.43	0.48	0.67	0.16	0.27	0.33	0.26	0.34	0.46	0.41	0.55	0.08	0.19
Trp	0.12	0.02	0.10	0.05	0.04	0.08	0.03	0.07	0.08	0.10	0.02	0.06	0.05	0.08	0.05	0.06	0.04	0.10	0.02	0.06
Tyr	0.24	0.04	0.27	0.20	0.13	0.23	0.07	0.17	0.19	0.26	0.04	0.14	0.15	0.14	0.15	0.17	0.18	0.25	0.05	0.12

Appendix II







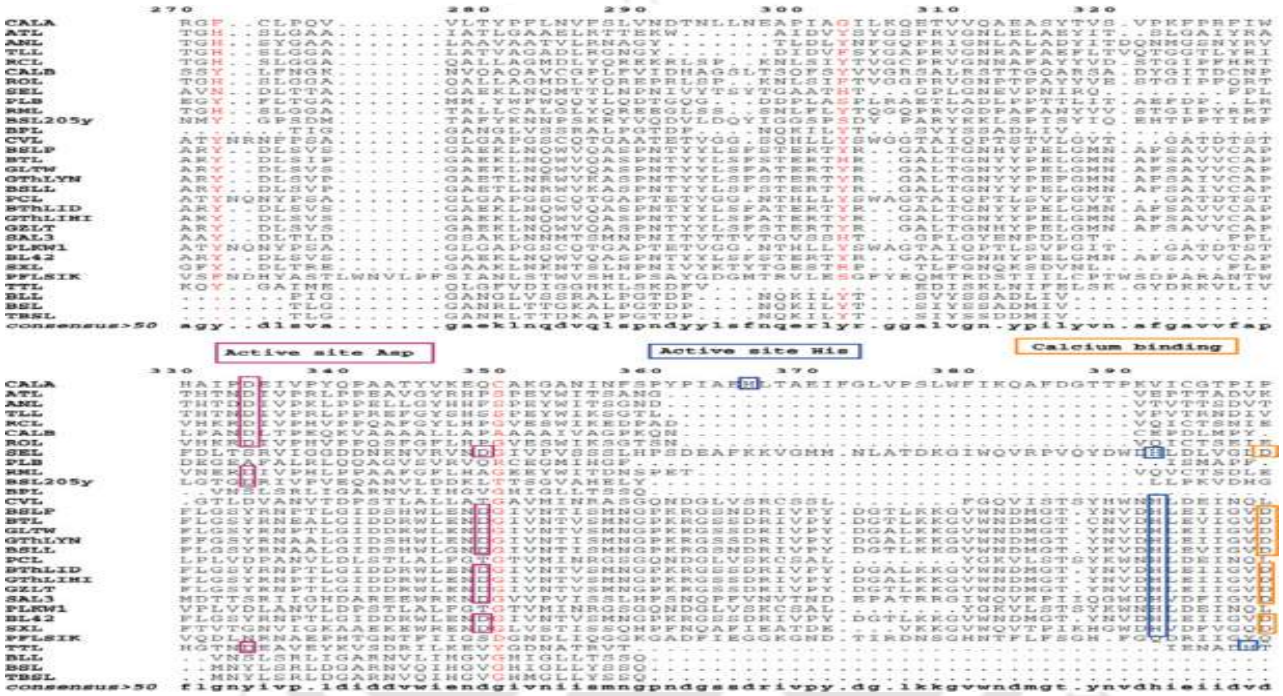


Fig. A2.1. Multiple sequence alignment of thermostable and mesostable lipases showing conserved motifs as generated by ESPript 2.2 (Gaut et al. 1999) showing the concensus sequence at greater than 50% sequence similarity. The different boxes are coloured according to the following, green–oxyanion hole sequence, cyan- Zinc binding residues and conserved sequences, pink- P-loop like motif, dark blue-P-loop motif, red-active site motif, purple- lid residues, yellow-Trp residue of lid domain, black-Tyr224 in *Bacillus* lipases, orange-Equivalent position of Tyr residue in *Staphylococcus* lipases, brown-active site Asp, black-Active site His, and yellow-Calcium binding residues. The numbers represent the positions of the amino acid residues that start from the first amino acid residue for each mature peptide.

Table A2.1: Signature sequences of lipases showing conserved sequences

Lipases	GXSXG motif residues	Oxyanion hole residues	Lid domain residues	P-loop motif	Zinc domain	No. GxxxG	No. AxxxA
BSLP1	AHSQG	PIVLLHGFTG	FFDLQKAVLEAAAVASNV	GHKRYGRT	GAAHAAKH	3	5
BTL	AHSQG	PIVLLHGFTG	FFDLQKAVLKAAAVASNV	GHKRYGRT	GAAHAAKH	3	5
GLTW1	AHSQG	PIVLLHGFTG	FFDLQKAVLEAAAVASNV	GHKRYGRT	GAAHAAKH	3	5
GThLYN	AHSQG	PIVLLHGFTG	FFDLQKAVLEAAAVASNA	GHKRYGRT	GAAHAAKH	3	6
BSLL1	AHSQG	PIVLLHGFTG	FFDLQKAVLEAAAVASNA	GHKRYGRT	GAAHAAND	3	5
BLL	AHSMG	PVVMVHGIGG	x	x	x	2	0
BL42	AHSQG	PIVLLHGFTG	FFDLQKAVLEAAAVASNV	GHARFGRT	GAAHAAKH	3	6
BThlID-1	AHSQG	PIVLLHGFTG	FFDLQKAVLEAAAVASNV	GHKRYGRT	GAAHAAKH	3	6
GThLIH1	AHSQG	PIVLLHGFTG	FFDLQKAVLEAAAVASNV	GHKRYGRT	GAAHAAKH	3	5
GZLT1	AHSQG	PIVLLHGFTG	FFDLQKAVLEAAAVASNV	GHKRYGRT	GAAHAAKH	3	6
SXL	AHSMG	PIVLLHGFTG	ALGNEAIVRQLAFDYAKFKGNKNSKVDFGFG QWGLK	GHKRYGRT	GAAHAAKY	1	3
SAL	GHSMG	PVVFVHGFLG	KFGNTEAVRKIMFALNRFMGNKYSNIDLGLT QWGFK	GHKRYGRT	GAAHAAKY	1	3
ATL	GHSLG	IVLSFRG	SRSPANWIANLDFIF	x	x	3	4
ANL	GHSYG	IVLSFRG	SSDLSNWIADLDFGL	x	x	2	3
PCL	GHSQG	PIILVHGLSG	GLSSSVIAAFVNVFGILTSSSHNT	x	x	1	4
PLKW1-56	GHSQG	PIILVHGLSG	GLSSSVIAAFVNVFGILTSSSHNT	x	x	1	4
CALA	GYSGG	-	-	x	x	4	5
TLL	GHSLG	IVLSFRG	SRSIENWIGNLNFDL	x	x	1	1
RCL	GHSLG	IYVTFRG	TNSFRSAITDMVFTF	x	x	1	2
CALB	TWSQG	x	x	x	x	0	5
TTL	GESDG	PMVIMFHGF		x	x	3	1
ROL	GHSLG	IYLVFRG	TNSFRSAITDIVFNF	x	x	3	2
PLB	GHSMG	PLLVFFHGGGF	_	x	x	2	4
BSL205y	GESAG	PVIVQVHGGG	_	x	x	2	1

Lipases	GXSXG motif residues	Oxyanion hole residues	Lid domain residues	P-loop motif	Zinc domain	No. GxxxG	No. AxxxA
SEL	GHSMG	PVVFVHGFVG	KIGGTKALDLELGFSQWGFK	GHKRYGRT	GAAHAAKY	3	3
RML	GHSLG	IYLVFRG	SSSIRNWIADLTFVP	x	x	2	1
BPL	AHSMG	PVVMVHG	_	x	x	1	0
CVL	GHSQG	PVILVHGLAG	GLSSTVIAAFVNVFGTLVSSSHNT	x	x	2	3
RML	GDSLG	IVFRG	SSSIRNWIAD	x	x	2	1
BSL	AHSMGG	PVVMVHGIGG	x	x	x	1	0
TBSL	AHSMGG	PVVMVHGIGG	x	x	x	1	0

x represents absence of the particular motif
– represents not found.

Appendix III

Table A3.1: The Difference matrix represented by 1 and 0 vectors for 127 thermostable proteins.

Sl No.	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	CASA	ASI	CPI	AAI	MMH	MSH	SSH
1	1	0	0	1	1	1	1	1	0	0	1	1	1	0	1	1	1
2	0	1	1	1	1	0	1	0	0	1	1	0	0	0	1	1	1
3	0	1	1	1	1	1	1	0	0	1	0	1	1	0	1	1	1
4	1	1	1	1	0	1	0	0	0	0	1	1	1	0	1	1	0
5	1	1	1	0	1	0	0	1	0	1	0	0	0	0	1	1	1
6	1	0	0	0	1	0	1	1	0	1	0	1	1	0	1	1	1
7	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
8	0	0	0	0	1	1	1	0	1	0	0	0	0	1	0	1	1
9	1	1	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0
10	1	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1
11	0	1	1	0	1	1	1	1	0	0	1	1	1	0	1	1	1
12	0	1	1	1	1	1	1	0	1	0	1	1	0	1	1	1	1

Sl No.	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	CASA	ASI	CPI	AAI	MMH	MSH	SSH
13	0	0	0	1	1	1	1	0	1	0	1	1	1	1	1	1	1
14	0	1	1	1	1	0	1	0	0	0	1	0	1	1	1	1	1
15	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	0
16	1	1	1	1	0	0	1	0	0	1	1	0	1	1	1	1	0
17	0	1	1	1	1	0	1	1	0	1	1	0	0	0	1	1	0
18	0	0	0	1	0	1	1	0	0	0	1	0	1	1	1	0	0
19	0	1	1	1	0	1	1	0	0	1	1	0	1	0	1	0	0
20	1	1	1	0	1	0	0	1	0	0	1	0	0	0	1	1	0
21	0	1	1	1	1	1	0	1	0	0	1	0	1	0	1	1	0
22	1	1	1	0	0	1	0	0	1	1	0	0	1	1	1	1	1
23	1	1	1	0	0	0	0	1	1	0	1	1	0	0	1	0	1
24	1	1	1	0	0	0	1	1	0	1	0	1	1	0	1	1	0
25	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1
26	0	1	1	1	1	1	1	1	0	1	1	0	1	0	1	1	0

Sl No.	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	CASA	ASI	CPI	AAI	MMH	MSH	SSH
27	1	1	1	0	1	1	1	1	0	1	0	0	0	0	1	1	1
28	0	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1
29	1	0	0	0	1	1	0	1	0	1	0	0	0	0	1	0	1
30	0	1	1	1	1	1	1	0	1	0	3	0	0	1	1	1	1
31	0	0	0	1	0	1	1	0	0	0	1	0	0	1	0	0	1
32	0	0	0	1	0	0	1	1	0	0	0	1	0	1	1	0	0
33	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0
34	1	1	1	0	0	0	1	0	0	1	0	1	1	0	1	1	1
35	0	1	1	0	0	1	1	0	0	0	1	0	0	0	1	1	1
36	1	0	0	0	1	1	0	1	1	0	1	0	1	1	0	1	1
37	0	1	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0
38	0	0	0	1	1	1	0	0	1	0	0	1	1	0	0	1	1
39	0	1	1	0	1	1	1	0	0	0	1	0	0	0	1	0	1
40	0	1	1	1	0	0	0	1	1	0	1	0	0	0	0	0	0

Sl No.	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	CASA	ASI	CPI	AAI	MMH	MSH	SSH
41	0	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1
42	0	0	0	1	0	1	1	1	1	0	0	0	1	1	0	0	0
43	1	1	1	1	1	1	0	1	0	0	0	0	1	1	1	1	0
44	1	0	0	0	0	0	1	0	1	0	0	0	1	1	1	1	0
45	0	0	0	1	0	0	1	0	1	0	1	0	0	0	1	1	0
46	1	0	0	1	1	1	1	0	1	0	1	0	0	0	1	0	0
47	0	0	0	1	1	1	1	0	0	0	1	0	1	0	1	1	1
48	0	0	0	1	1	1	1	0	0	0	1	1	1	0	1	0	0
49	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0	0
50	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0
51	0	1	1	0	1	0	1	0	1	0	1	0	0	0	1	1	1
52	0	0	0	1	1	1	1	0	1	0	1	1	1	1	1	0	1
53	0	0	0	1	1	1	1	1	0	0	0	0	0	1	1	0	0
54	0	1	1	1	0	0	0	0	1	0	1	0	1	0	1	0	0

Sl No.	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	CASA	ASI	CPI	AAI	MMH	MSH	SSH
55	0	1	1	0	1	1	1	1	1	1	1	0	1	1	1	0	1
56	0	1	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
57	0	0	0	1	0	1	0	1	1	0	1	0	0	0	0	0	0
58	0	0	0	1	1	1	1	0	1	0	0	1	1	1	1	0	0
59	0	1	1	1	0	1	0	0	0	1	0	0	1	0	0	0	0
60	0	0	0	0	1	1	0	1	0	0	1	0	1	1	0	0	1
61	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0
62	1	0	0	0	0	1	1	0	1	0	0	0	1	1	1	1	1
63	1	1	1	0	1	1	1	1	0	0	0	0	1	0	1	0	0
64	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	0	0
65	1	1	1	1	0	1	1	0	0	1	0	0	1	1	1	1	0
66	1	1	1	1	1	1	1	1	0	0	0	0	1	0	1	0	0
67	1	0	0	0	0	1	1	1	1	0	1	0	0	0	1	0	0
68	0	1	1	0	0	0	0	0	0	0	1	1	1	0	1	1	0

Sl No.	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	CASA	ASI	CPI	AAI	MMH	MSH	SSH
69	1	1	1	1	1	1	1	0	1	0	1	0	0	0	1	0	0
70	1	1	1	0	0	1	0	1	0	0	1	0	0	0	1	1	1
71	1	0	0	0	1	1	1	1	0	0	1	0	0	0	1	1	0
72	0	1	1	1	1	1	1	0	1	0	1	1	0	0	1	1	1
73	0	0	0	1	1	1	1	0	0	1	1	1	1	1	0	1	1
74	0	1	1	1	1	1	1	0	1	0	0	0	1	1	1	0	0
75	0	0	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1
76	0	0	0	1	0	1	1	0	1	0	1	1	1	1	0	0	0
77	0	1	1	1	1	1	0	0	0	0	1	0	1	0	1	0	0
78	0	0	1	0	1	1	1	0	0	0	1	0	1	1	0	0	1
79	0	0	0	0	1	1	0	1	0	0	1	0	1	1	0	0	0
80	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	1	0
81	0	1	1	1	1	1	0	1	1	0	1	0	0	0	0	0	1
82	0	1	1	1	0	1	0	1	1	0	1	0	1	1	1	0	1

Sl No.	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	CASA	ASI	CPI	AAI	MMH	MSH	SSH
83	0	1	1	1	1	1	0	1	1	0	0	0	1	1	0	1	1
84	0	0	0	1	1	1	1	0	0	1	1	0	0	0	1	0	1
85	0	1	1	1	1	1	0	1	0	1	1	0	0	1	0	0	0
86	0	0	0	1	1	1	0	0	0	0	1	0	1	1	1	1	1
87	0	0	0	1	0	0	0	0	1	0	0	1	1	0	1	0	0
88	0	1	1	0	1	1	1	0	0	0	1	1	0	1	1	1	1
89	0	0	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0
90	0	0	0	1	0	1	1	1	1	0	1	0	0	1	1	0	0
91	0	1	1	1	0	0	0	1	0	1	1	0	0	0	1	0	0
92	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
93	1	1	1	1	0	1	1	0	0	0	0	0	1	1	1	1	0
94	0	1	1	0	0	1	0	0	1	0	1	0	0	0	1	0	0
95	0	1	1	1	1	0	1	0	0	0	1	0	0	0	1	0	1
96	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1	1	1

Sl No.	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	CASA	ASI	CPI	AAI	MMH	MSH	SSH
97	1	1	1	1	1	1	1	1	0	0	1	0	1	1	0	1	1
98	0	0	0	1	1	1	1	0	0	0	1	1	1	1	1	0	1
99	0	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	0
100	0	1	1	1	1	1	1	0	0	0	1	0	1	0	0	0	0
101	1	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0
102	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
103	0	0	0	1	1	1	1	0	0	0	1	1	1	0	1	1	1
104	1	1	1	1	1	1	0	0	1	0	0	0	1	0	1	0	0
105	0	0	1	1	1	1	1	0	0	0	1	0	1	0	1	1	1
106	0	1	1	1	1	1	0	1	1	0	1	0	1	1	1	1	0
107	0	1	1	1	1	1	1	1	0	0	1	0	1	0	1	1	1
108	0	1	1	1	1	1	1	0	1	0	0	0	1	0	1	0	1
109	0	0	0	1	1	1	1	1	1	0	0	0	0	1	1	1	1
110	0	0	0	1	1	1	1	0	1	0	1	0	1	1	0	0	1

Sl No.	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	CASA	ASI	CPI	AAI	MMH	MSH	SSH
111	0	1	1	1	1	1	0	0	0	0	1	1	0	1	1	0	0
112	0	1	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0
113	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1
114	1	1	1	1	1	0	0	1	0	1	0	0	0	1	0	1	1
115	1	0	1	1	0	1	0	0	0	0	1	0	1	0	1	0	0
116	0	1	1	1	1	1	1	0	1	0	9	1	0	0	1	0	0
117	0	0	0	0	1	0	1	0	0	0	1	1	1	1	1	0	0
118	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	0	0
119	1	1	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0
120	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
121	0	0	0	0	1	1	1	1	1	0	0	0	1	0	1	1	1
122	1	1	0	0	0	0	1	1	0	0	1	0	1	0	1	0	0
123	0	0	0	0	1	1	0	1	0	0	1	0	1	0	1	0	0
124	0	0	0	1	0	0	1	1	1	0	0	0	1	0	0	0	0

Sl No.	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	CASA	ASI	CPI	AAI	MMH	MSH	SSH
125	1	1	1	6	0	1	1	1	0	0	1	1	1	0	1	1	1
126	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0
127	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0
Total	44	77	79	73	79	91	77	51	48	26	81	34	73	50	89	59	57

Vector 1 represents increase in the value of the respective factor for the thermostable set (TP).
0 for decrease or equality w.r.t. the mesostable set (MP).
Total represents the number of proteins in the TP set showing increase in factor or having the vector 1.

Table A3.2: The pairwise comparison matrix for thermostability

Interactions	BT	GT	IGT	HB	SB	II	HI	PV	NASA	PASA	C ASA	ASI	CPI	AAI	MMH	MSH	SSH
BT	1.00	0.40	0.40	0.40	0.40	0.30	0.50	0.80	0.80	3.00	0.30	1.70	0.40	0.80	0.30	0.60	0.60
GT	2.30	1.00	0.90	1.00	0.90	0.78	1.00	1.80	1.80	7.00	0.90	3.50	1.00	1.80	0.80	1.40	1.40
IGT	2.67	1.14	1.00	1.14	1.00	0.89	1.14	2.00	2.00	8.00	1.00	4.00	1.14	2.00	0.90	1.60	1.60
HB	2.33	1.00	0.88	1.00	0.88	0.78	1.00	1.75	1.75	7.00	0.88	3.50	1.00	1.75	0.80	1.40	1.40
SB	2.67	1.14	1.00	1.14	1.00	0.89	1.14	2.00	2.00	8.00	1.00	4.00	1.14	2.00	0.90	1.60	1.60
II	3.00	1.29	1.13	1.29	1.13	1.00	1.29	2.25	2.25	9.00	1.13	4.50	1.29	2.25	1.00	1.80	1.80
HI	2.33	1.00	0.88	1.00	0.88	0.78	1.00	1.75	1.75	7.00	0.88	3.50	1.00	1.75	0.80	1.40	1.40
PV	1.33	0.57	0.50	0.57	0.50	0.44	0.57	1.00	1.00	4.00	0.50	2.00	0.57	1.00	0.44	0.80	0.80
NASA	1.33	0.57	0.50	0.57	0.50	0.44	0.57	1.00	1.00	4.00	0.50	2.00	0.57	1.00	0.44	0.80	0.80
PASA	0.33	0.14	0.13	0.14	0.13	0.11	0.14	0.25	0.25	1.00	0.13	0.50	0.14	0.25	0.11	0.20	0.20
CASA	2.70	1.14	1.00	1.14	1.00	0.89	1.14	2.00	2.00	8.00	1.00	4.00	1.14	2.00	0.90	1.60	1.60
ASI	0.67	0.29	0.25	0.29	0.25	0.22	0.29	0.50	0.50	2.00	0.25	1.00	0.29	0.50	0.22	0.40	0.40
CPI	2.33	1.00	0.90	1.00	0.88	0.78	1.00	1.75	1.75	7.00	0.88	3.50	1.00	1.75	0.80	1.40	1.40
AAI	1.33	0.57	0.50	0.57	0.50	0.44	0.57	1.00	1.00	4.00	0.50	2.00	0.57	1.00	0.44	0.80	0.80
MMH	3.00	1.29	1.13	1.29	1.13	1.00	1.29	2.25	2.25	9.00	1.13	4.50	1.29	2.25	1.00	1.80	1.80
MSH	1.67	0.71	0.63	0.71	0.63	0.56	0.71	1.25	1.25	5.00	0.63	2.50	0.71	1.25	0.60	1.00	1.00
SSH	1.67	0.71	0.63	0.71	0.63	0.56	0.71	1.25	1.25	5.00	0.63	2.50	0.71	1.25	0.60	1.00	1.00

* The pairwise comparison matrix was normalized and priority values or eigen vectors for each feature was compute

Table A3.3: Ranks obtained for the thermostable protein dataset set through RankProt

Sl	TP	Temperature (°C)	TP Protein	TP Source	RANK TP	MP	RANK MP
1	1vp9	70	DNA ligase	*Thermus filiformis*	0.53	1owo	0.47
2	3d2c	65	Lipase	*Bacillus subtilis*	0.58	1i6w	0.42
3	2vul	75	Gh11 xylanase	*E. Coli*	0.59	2vug	0.41
4	2ak9	50-90	Subtilisin BPN'	*Bacillus amyloliquefaciens*	0.60	1sbt	0.40
5	1a5z	50-90	Lactate dehydrogenase	*Thermotoga maritima*	0.50	1ldn	0.50
6	1b26	75	Glutamate dehydrogenase	*Thermotoga maritima*	0.52	1bgv	0.48
7	1je0	120	5'-methylthioadenosine phosphorylase	*Sulfolobus solfataricus*	0.54	1eu8	0.46
8	2zf5	105.4	Glycerol kinase	*Thermococcus kodakarensis kod1*	0.54	3pnk	0.46
9	1t2n	50	Lipase	*Bacillus subtilis*	0.54	1i6w	0.50
10	1bxc	75	Xylose isomerase	*Thermus caldophilus*	0.53	1bhw	0.47
11	1bxz	94	Alcohol dehydrogenase	*Thermoanaerobacter brockii*	0.51	1kev	0.49
12	1tmy	80	Chey	*Thermotoga maritima*	0.55	3chy	0.45
13	1bxb	90	Xylose isomerase	*Thermus thermophilus*	0.59	1qti	0.41
14	1xyz	60	Xylanhydrolase	*Clostridium thermocellum*	0.53	2exo	0.47
15	1thl	80	Neutral protease	*Bacillus thermoproteolyticus*	0.52	1npc	0.48
16	2prd	73	Hydrolase	*Thermus thermophilus*	0.45	1ino	0.55
17	1gtm	75-100	Glutamate dehydrogenase	*Pyrococcus furiosus*	0.59	1hrd	0.41
18	3pfk	53	Phosphofructose kinase	*Geobacillus stearothermophilus*	0.56	2pfk	0.44
19	3mds	60	Maganese superoxide dismutase	*Thermus thermophilus*	0.46	1qnm	0.54

Sl	TP	Temperature (°C)	TP Protein	TP Source	RANK TP	MP	RANK MP
20	1lnf	86	Thermolysin	*Bacillus thermoproteolyticus*	0.48	1npc	0.52
21	1bdm	90	Malate dehydrogenase	*Thermus thermophilus*	0.51	4mdh	0.49
22	1a2z	83	Pyrrolidone carboxyl peptidase	*Thermococcus litoralis*	0.51	1-aug	0.49
23	1a53	70-80	Indole-3-glycerolphosphate synthase	*Sulfolobus solfataricus*	0.42	1pii	0.58
24	1a5z	80	Lactate dehydrogenase	*Thermotoga maritima*	0.53	9ldt	0.47
25	1a8h	75	Methionyl-trna synthetase	*Thermus thermophilus*	0.56	1qtt	0.44
26	1b69	83	Histone hmfa	*Methanothermus fervidus*	0.41	1aoi	0.59
27	1bjw	75	Aspartate aminotransferase	*Thermus thermophilus*	0.53	1bw0	0.47
28	1bmd	70	Malate dehydrogenase	*Thermus aquaticus flavus*	0.50	1b8p	0.50
29	1bvu	83	Glutamate dehydrogenase	*Thermococcus litoralis*	0.55	1hrd	0.45
30	1bxb	75	Xylose isomerase	*Thermus thermophilus*	0.55	1xif	0.45
31	1c3u	80	Adenylosuccinate lyase	*Thermotoga maritima*	0.51	1auw	0.49
32	1coj	85	Superoxide dismutase	*Aquifex pyrophilus*	0.53	1var	0.47
33	1dv7	65	Orotidine 5V-phosphate decarboxylase	*Methanothermobacter thermoautot*	0.59	1eix	0.41
34	1ffh	70	Gtpase domains of the signal sequence	*Thermus aquaticus*	0.51	1fts	0.49
35	1gln	75	Glutamyl-trna synthetase	*Thermus thermophilus*	0.53	1euq	0.47
36	1gtm	97-100	Glutamate dehydrogenase	*Pyrococcus furiosus*	0.47	1aup	0.53
37	1sss	70-85	Iron superoxide dismutase	*Sulfolobus solfataricus*	0.53	1qnn	0.47
38	1xgs	97-100	Methionine aminopeptidase	*Pyrococcus furiosus*	0.55	1mat	0.45
39	1ykf	65	NADP-dependent alcohol dehydrogenase	*Thermoanaerobium brockii*	0.53	1kev	0.47
40	1yna	48	Endo-1,4-beta-xylanase	*Thermomyces lanuginosus*	0.48	1xnd	0.37
41	2btm	60	Triosephosphate isomerase	*Geobacillus stearothermophilus*	0.45	1tpf	0.47

Sl	TP	Temperature (°C)	TP Protein	TP Source	RANK TP	MP	RANK MP
42	4pfk	55	Phosphofructokinase	*Geobacillus stearothermophilus*	0.51	1pfk	0.49
43	1iqz	52	Ferredoxin	*Bacillus thermoproteolyticus*	0.70	1fca	0.30
44	1hjs	50	Fungal beta-1,4-galactanases	*Thielavia heterothallica*	0.53	1zqd	0.47
45	1uek	75	4-(cytidine 5'-diphospho)-2c-methyl-d-erythritol kinase	*Thermus thermophilus*	0.55	1dqj	0.45
46	1wfr	75	Conserved hypothetical protein tt1886, possibly sterol carrier protein, from thermus thermophilus hb8	*Thermus thermophilus hb8*	0.60	1x4v	0.40
47	1yz7	103	C-terminal segment of alpha subunit of aif2 from pyrococcus abyssi	*Pyrococcus abyssi*	0.51	1a1a	0.49
48	1vrm	80	Hypothetical protein (tm1553) from thermotoga maritima	*Thermotoga maritima msb8*	0.51	1n7j	0.49
49	1a76	85	5'-3' exo/endo nuclease	*Methanococcus jannaschii*	0.55	1ut8	0.45
50	1v6s	75	Phosphoglycerate kinase	*Thermus thermophilus hb8*	0.60	1puz	0.40
51	1jji	82	Carboxylesterase	*Archaeon archaeoglobus*	0.53	1k4y	0.47
52	1pjr	55	Dna helicase	*Bacillus stearothermophilus*	0.61	1zrr	0.39
53	1lf6	60	Glucoamylase	*Thermoanaerobacterium thermosaccharolyticum*	0.53	1ulv	0.45
54	1js4	45	Endo/exocellulase:cellobiose	*Thermomonospora fusca*	0.51	1kfq	0.47
55	1woy	75	Methionyl trna synthetase y225f mutant	*Thermus thermophilus*	0.48	1l1l	0.50
56	1vmb	80	30s ribosomal protein s6 (tm0603)	*Thermotoga maritima*	0.59	1wjh	0.41
57	1yqe	82	Conserved hypothetical protein af0625	*Archaeoglobus fulgidus*	0.53	1hdh	0.47
58	1esw	70	Amylomaltase	*Thermus aquaticus*	0.53	1yht	0.47
59	1im5	98	Alpha-glucosidase (tm0752)	*Thermotoga maritima*	0.64	1frr	0.36
60	1vjt	80	Alpha-glucosidase (tm0752)	*Thermotoga maritima*	49.88	2bw0	48.38

Sl	TP	Temperature (°C)	TP Protein	TP Source	RANK TP	MP	RANK MP
61	1bqc	45	Beta-mannanase	*Thermomonospora fusca*	0.61	1lf1	0.39
62	2bog	45	Catalytic domain of endo-1,4-glucanase cel6a mutant y73s	*Thermobifida fusca*	0.56	1y7m	0.44
63	1i1w	45	Thermostable xylanase	*Thermoascus aurantiacus*	0.59	1mzd	0.41
64	1olr	45	Humicola grisea cel12a enzyme structure	*Humicola grisea*	0.62	1ks4	0.38
65	1d1n	55	Fmet-trnafmet binding domain of becillus stearothermophillus translation initiation factor if2	*Bacillus stearothermophillus*	0.62	2crv	0.38
66	2ars	59	Lipoate-protein ligase a	*Thermoplasma acidophilum*	0.53	1vqz	0.47
67	1wl7	60	Thermostable arabinanase	*Bacillus thermodenitrificans*	0.59	1mdw	0.41
68	1vbl	60	Thermostable pectate lyase pl 47	*Bacillus sp. Ts-47*	0.57	1pxz	0.43
69	1t1g	60	Mutant e23a of kumamolisin, a sedolisin type proteinase (previously called kumamolysin or kscp)	*Bacillus sp. Mn-32*	0.52	1bh6	0.48
70	1clc	60	Endoglucanase D	*Clostridium Thermocellum 1clc*	0.54	1eif	0.46
71	1v3y	75	Peptide deformylase from thermus thermophilus hb8	*Thermus thermophilus*	0.56	2ai9	0.44
72	1ujp	75	Tryptophan synthase a-subunit from thermus thermophilus hb8	*Thermus thermophilus*	0.54	1wq5	0.46
73	1n75	75	Glutamyl-trna synthetase	*Thermus thermophilus*	0.53	1nyl	0.47
74	1v37	75	Phosphoglycerate mutase	*Thermus thermophilus hb8*	0.60	1fzt	0.40
75	1v35	75	Crystal Structure of Eoyl-ACP Reductase with NADH	*Plasmodium falciparum*	0.56	2ai9	0.44
76	1vjr	80	4-nitrophenylphosphatase (tm1742)	*Thermotoga maritima*	0.53	1rkq	0.47
77	1o2d	80	Alcohol dehydrogenase, iron-containing (tm0920)	*Thermotoga maritima*	0.55	1wik	0.45

Sl No.	TP	Temperature (°C)	TP Protein	TP Source	RANK TP	MP	RANK MP
78	1vkz	80	Phosphoribosylamine--glycine ligase (tm1250)	*Thermotoga maritima*	0.53	1gso	0.47
79	1o5z	80	Folylpolyglutamate synthase (tm0166)	*Thermotoga maritima*	0.54	1jbw	0.46
80	1qo2	80	Isomerase	*Thermotoga maritima*	0.53	1vzw	0.47
81	1jdq	80	Tm006 protein	*Thermotoga maritima*	0.58	1je3	0.42
82	1vku	80	Acyl carrier protein (TM0175)	*Thermotoga maritima*	0.61	1hy8	0.39
83	1z0w	82	Lon proteolytic domain	*Archaeoglobus fulgidus*	0.53	1xmj	0.47
84	1io9	85	Oxidoreductase	*Sulfolobus solfataricus*	0.59	1f20	0.41
85	1c3p	85	Hdac homolog	*Aquifex aeolicus*	0.61	1sy1	0.39
86	1t6t	85	Putative protein	*Aquifex aeolicus*	0.51	1puib	0.49
87	1t6c	85	Putative protein	*Aquifex aeolicus*	0.63	1nyn	0.37
88	1z5z	85	Swi2/snf2 atpase c-terminal domain	*Sulfolobus solfataricus p2*	0.54	1oyy	0.46
89	1l7m	85	Phosphoserine phosphatase (pi complex)	*Methanococcus jannaschii*	0.52	1nnl	0.48
90	1wr2	98	Ph1788	*Pyrococcus horikoshii ot3*	0.50	1j0n	0.50
91	1v7r	98	Nucleotide triphosphate pyrophosphatase	*Pyrococcus horikoshii ot3*	0.51	1kfq	0.49
92	1mxg	100	(Ca,zn)-dependent alpha-amylase	*Pyrococcus woesei*	0.53	1rpa	0.47
93	1u04	100	Full length argonaute	*Pyrococcus furiosus dsm 3638*	0.51	1z6t	0.49
94	1dq3	100	An archaeal intein-encoded homing endonuclease pi-pfui	*Pyrococcus furiosus*	0.51	1v5d	0.49
95	1xqo	100	Pa-agog, 8-oxoguanine dna glycosylase	*Pyrobaculum aerophilum*	0.63	1jxo	0.37
96	1brf	100	Rubredoxin (wild type)	*Pyrococcus furiosus*	0.58	2rdv	0.42
97	1sfs	55	Uncharacterized	*Bacillus Stearothermophilus*	0.50	1zsw	0.50
98	1lab	55	Lipoyl domain pyruvate dehydrogenase multienzyme complex.	*Bacillus stearothermophilus*	0.54	1k8o	0.46

Sl No.	TP	Temperature (°C)	TP Protein	TP Source	RANK TP	MP	RANK MP
99	1mgt	95	O6-methylguanine-dna methyltransferase	*Pyrococcus kodakaraensis*	0.59	1qnt	0.41
100	1ixk	98	Methyltransferase homolog protein	*Pyrococcus horikoshii*	0.57	1ej0	0.43

Table A3.4: Mutant ranks of bacteriophage T4 lysozyme obtained by RankProt

Mutant	Temperature (^{o}C)	Rank_Mutant	Rank_wild type
1DYA	53.08	0.52	0.43
1DYB	53.08	0.5	0.45
1DYC	68.3	0.52	0.43
1DYD	64.7	0.5	0.45
1DYE	53.08	0.51	0.44
1DYF	53.08	0.5	0.44
1DYG	53.08	0.51	0.44
1L00	66.3	0.5	0.45
1L02	53.6	0.51	0.44
1L03	42	0.5	0.42
1L04	42	0.52	0.43
1L06	53.6	0.5	0.42
1L07	42	0.51	0.44
1L08	42	0.52	0.43
1L09	42	0.51	0.43
1L10	64.5	0.51	0.44
1L11	53.6	0.52	0.43
1L12	53.6	0.5	0.42
1L13	53.6	0.5	0.42
1L14	42	0.52	0.43
1L15	53.6	0.51	0.44
1L16	64.7	0.51	0.44
1L17	64.7	0.51	0.44

Mutant	Temperature (^{o}C)	Rank_Mutant	Rank_wild type
1L18	64.7	0.52	0.43
1L19	66.51	0.52	0.43
1L20	66.51	0.51	0.44
1L21	64.7	0.51	0.43
1L22	64.7	0.5	0.42
1L23	64.7	0.5	0.42
1L24	66.51	0.51	0.43
1L33	66.51	0.5	0.42
1L34	70	0.51	0.44
1L35	NR	0.52	0.43
1L36	53.4	0.51	0.43
1L37	66.7	0.51	0.43
1L38	66.7	0.51	0.44
1L39	NR	0.52	0.43
1L40	66.7	0.5	0.42
1L41	64.6	0.5	0.44
1L42	0.5 kcal/mol more stable than wild-type	0.52	0.43
1L42	66.7	0.5	0.44
1L44	66.7	0.52	0.43
1L45	66.7	0.51	0.43
1L46	66.7	0.5	0.42
1L47	66.7	0.51	0.43
1L48	40	0.52	0.43
1L49	40	0.51	0.44

Mutant	Temperature (^{o}C)	Rank_Mutant	Rank_wild type
1L50	40	0.51	0.43
1L51	40	0.52	0.43
1L52	40	0.51	0.43
1L53	40	0.5	0.42
1L54	65.3	0.51	0.44
1L55	62.4	0.52	0.43
1L56	64.7	0.5	0.42
1L57	66.51	0.51	0.44
1L59	62.4	0.51	0.43
1L60	64.7	0.52	0.43
1L61	62.4	0.5	0.42
1L62	62.4	0.51	0.44
1L63	53.6	0.5	0.42
1L64	increased by 3.1^{o}C	0.51	0.43
1L65	62.2	0.51	0.43
1L66	62.2	0.5	0.42
1L67	62.2	0.5	0.44
1L68	62.2	0.51	0.43
1L69	51.8	0.5	0.44
1L70	40.75	0.52	0.43
1L71	40.75	0.5	0.44
1L72	40.75	0.5	0.42
1L73	53.4	0.52	0.43
1L74	53.4	0.5	0.44
1L75	53.4	0.51	0.43

Mutant	Temperature (^{o}C)	Rank_Mutant	Rank_wild type
1L76	63.4	0.51	0.44
1L77	64.89	0.51	0.43
1L85	65.3	0.5	0.42
1L86	64.88	0.51	0.43
1L87	65.15	0.52	0.43
1L88	65.3	0.51	0.43
1L89	64.88	0.52	0.43
1L90	64.88	0.52	0.43
1L91	64.88	0.52	0.43
1L92	64.88	0.52	0.43
1L93	64.88	0.5	0.42
1L94	64.88	0.51	0.43
1L95	64.88	0.51	0.43
1L96	53.56	0.52	0.43
1L98	66.3	0.51	0.43
1L99	66.3	0.52	0.43
1LAV	52.5	0.47	0.48
1LAW	52	0.47	0.48
1LHH	63.5	0.55	0.4
1LHI	63.5	0.41	0.54
1LHJ	66.2	0.4	0.4
1LHK	61.6	0.4	0.41
1LHL	67.6	0.4	0.41
1LHM	49	0.54	0.41

Mutant	Temperature (^{o}C)	Rank_Mutant	Rank_wild type
1LRA	48.6	0.54	0.4
1LSN	74	0.48	0.47
1LYE	63	0.51	0.43
1LYF	63	0.5	0.45
1LYG	63	0.51	0.44
1LYH	63	0.51	0.44
1LYI	63	0.52	0.43
1LYJ	63	0.5	0.45

NR- Not reported

Table A3.5: Mutant ranks of human lysozyme obtained by RankProt

Mutant	T°C	Rank_Mutant	Rank_wild type
2BQB	46	0.39	0.41
2BQC	47	0.38	0.41
2BQD	42.9	0.39	0.41
2BQE	44.3	0.39	0.41
2BQF	46	0.39	0.41
2BQG	46	0.39	0.41
2BQH	49.5	0.39	0.41
2BQI	39.9	0.39	0.41
2BQJ	42.2	0.39	0.41
2BQK	44.4	0.39	0.41
2BQM	47	0.39	0.41
2BQN	44	0.39	0.41
2BQO	44.8	0.39	0.41
2HEA	61.9	0.38	0.42
2HEB	56.8	0.38	0.41
2HEC	52.3	0.48	0.31
2HED	59.7	0.54	0.41
2HEE	52.2	0.39	0.41
2HEF	56.2	0.39	0.41
1WQM	63.8	0.38	0.41
1WQN	61.1	0.53	0.42
1WQO	64.3	0.54	0.41

Mutant	T°C	Rank_Mutant	Rank_wild type
1WQP	65.1	0.39	0.41
1YAM	58.3	0.39	0.41
1YAN	59.8	0.39	0.41
1YAO	57.9	0.39	0.41
1YAP	57.7	0.39	0.41
1YAQ	59.4	0.39	0.41
1OUB	64.1	0.54	0.41
1OUC	66.4	0.39	0.41
1OUD	60.4	0.39	0.41
1OUE	60.9	0.39	0.41
1OUF	62.3	0.39	0.41
1OUG	60.3	0.39	0.41
1OUH	63.8	0.39	0.41
1OUI	62.6	0.39	0.41
1OUJ	61.9	0.39	0.41
1LHH	63.5	0.55	0.4
1LHI	63.5	0.41	0.54
1LHJ	66.2	0.4	0.4
1LHK	61.6	0.4	0.41
1LHL	67.6	0.4	0.41
1LHM	49	0.54	0.41
1GAY	57.4	0.4	0.41
1GAZ	68.3	0.4	0.41

Mutant	T°C	Rank_Mutant	Rank_wild type
1GB2	63.9	0.39	0.41
1GB3	62.2	0.54	0.41