The Role of Charged Amino Acids in the Origin of UV-Visible Electronic Absorption in Proteins

A Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

By

Ms. Saumya Prasad

Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Guwahati, Assam- 781039, India

December 2016

INDIAN INSTITUTE OF TECHNOLOGY
GUWAHATI, Assam, India

Department of Biosciences and Bioengineering

STATEMENT

I do hereby declare that the matter embodied in this thesis is the result of investigations carried out by me in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, India, under the guidance of Prof. Rajaram Swaminathan.

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.

IIT Guwahati — Saumya Prasad

December, 2016

INDIAN INSTITUTE OF TECHNOLOGY
GUWAHATI, Assam, India

Department of Biosciences and Bioengineering

CERTIFICATE

It is certified that the work described in this thesis, entitled "***The Role of charged amino acids in the origin of UV-Visible electronic absorption in proteins***" done by **Ms. Saumya Prasad** 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 and Bioengineering, Indian Institute of Technology Guwahati, India, and this work has not been submitted elsewhere for a degree.

IIT Guwahati
December, 2016

Prof. Rajaram Swaminathan
Department of BSBE
IIT Guwahati

Dedicated to my

Family

ACKNOWLEDGEMENT

Pursuing a Ph.D. degree has been a roller coaster ride for me which was full of ups and downs both at emotional and professional front. Honestly, it would not have been possible for me to make it till the very end of this ride had the following people not supported and encouraged me.

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. R. Swaminathan for his continuous support and guidance throughout the duration of my work. I would like to thank him for his patience, motivation, and his belief in me. His guidance helped me throughout my research journey and writing of this thesis. He has been a great mentor, from whom I have learnt a lot of things. He encouraged me not only to grow as a researcher but also as a person. I am indebted to him for making me learn to handle countless difficult situations. Without his support and ideas I would not have been able to complete my research work.

I would like to extend my sincere thanks to my Doctoral Committee members, Prof. P. Goswami, Dr. A. Limaye and Dr. D. Manna for their valuable suggestions which helped me to widen my research from various perspectives. I express my sincere thanks to Dr. B. Mandal for allowing me to synthesize peptides in his Lab. I take this opportunity to acknowledge Ashim Paul for helping me with the synthesis and characterization of all the peptides. Without his help, the work with peptides would not have been possible.

I thank Prof. J. B. Udgaonkar, NCBS, Bangalore and Prof. A. K. Sau, NII, New Delhi for allowing me to work in their respective labs during initial phase of my Ph.D.

Our collaborator Dr. R. Venkatramani, from Tata Institute of Fundamental Research, Mumbai deserves a special mention. His support, insightful ideas and encouragement throughout our collaboration has been instrumental. We have had some very exhaustive discussions over numerous teleconferences which helped us to carry the story forward. I would also like to thank Imon for her kind help and intensive theoretical calculations. Extensive calculations done by her helped us to arrive at a conclusion. She has truly been very supportive and helping to make me understand the nuances of MD simulation and theoretical calculations.

I would like to convey my heartfelt thanks to my lab mates and friends Shruti, Zia, Amrendra, Dileep, Anurag, Ekramul, Aditya, Garima, Adejoke, Sathvika and Vijya for their constant support and timely help. I thank them for making my stay at IIT Guwahati a memorable one. I would also like to thank my friends from college days Shilpi, Manika and Priyanka for motivating me whenever I was low.

My parents who are my pillars have been very supportive and understanding throughout this tough journey. I am very grateful to them for always believing in me and being patient with me to complete my work. My younger sister Surabhi and brother Shivam have always been my side whenever I needed them. I would like to thank them from the bottom of my heart for always being there for me and motivating me. I express my sincere gratitude to my in-laws who have been extremely supportive to carry forward my studies. I indeed consider myself very lucky and blessed to have such encouraging parents.

There were many other people who helped me both personally and professionally. Although it is not possible to pen down each of their names, I would like to thank each one of them for helping me in some or the other way during different stages of my Ph.D.

Finally, I would like to thank my husband Anil who is my backbone. His support, motivation and help during this phase of my life cannot be expressed in words. His understanding and unconditional love encouraged me to carry forward my work and not give up. He has truly been an immense support as a life partner. I owe every bit of my achievement to my family.

Last but not the least; I thank the almighty who in some form or the other has always taken care of me and showed me correct path whenever I was lost.

Saumya Prasad *December, 2016*

LIST OF ABBREVIATIONS

α_3C	Alpha 3C protein
Asp.K	Aspartate potassium salt
BSA	Bovine serum albumin
CD	Circular Dichroism
CT	Charge Transfer
FAD	Flavin adenine dinucleotide
Glu.Na	Glutamate monosodium salt
HEWL	Hen Egg-White Lysozyme
HSA	Human serum albumin
Lys.HCl	Lysine monohydrochloride
MALDI	Matrix-assisted laser desorption ionization
NAD	Nicotinamide adenine dinucleotide
OD	Optical Density
PDB	Protein Data Bank
TDDFT	Time Dependent Density Functional Theory
UV-Vis	Ultraviolet-Visible

ABSTRACT

Proteins are the most abundant intracellular macromolecules which perform a diverse range of functions within the living cell. They are known to absorb in the UV region of the electromagnetic spectrum owing to the presence of aromatic amino acids (Trp, Tyr and Phe) in them. This absorption has been well characterized using UV-Visible spectroscopy. However, L-Lys.HCl (Lysine monohydrochloride), a non-aromatic amino acid was reported to display a unique absorption at 270 nm and luminescence feature at high concentrations (~0.5 M) in aqueous medium. These features could not be accounted for by any chromophore present in the Lys molecule and a possible role of the ε-NH_2 moiety in Lys was anticipated. Similar observations arising from interactions between two or more lysine residues present in close spatial vicinity in lysine rich proteins like Human Serum Albumin have also been reported. However the origin of these novel spectra in the absence of any aromatic moiety has remained unanswered till date.

The work reported in this thesis is an attempt to understand the nature of the chromophore and underlying mechanism involved behind these unusual spectral signatures. Initial part of the thesis deals with studies of aliphatic compounds and short peptides devoid of any aromatic amino acids. In this study we show that the NH_2 moiety is actually crucial in order to see any significant absorption in the near UV region (270 nm) as compounds lacking the NH_2 moiety remained silent with no significant absorption in this region. To explore the effect of molecular interactions among the ε-NH_3^+ groups of two lysine residues behind the unusual spectral signatures which were observed in lysine rich proteins like HSA, investigations with short peptides (4-7 amino acids) containing pair of lysine residues placed at different positions in the sequence were carried out. Experiments on single lysine amino acids and peptides without any lysine residue served as controls. Compared to Lys.HCl all the peptides show ~100 fold increase in the absorptivity thereby hinting towards the possible role of peptide backbone and interaction among the Lys residues towards the unusual spectral features. It was further found that peptides terminating in -COOH group had absorption intensities much stronger than the peptides terminating in -$CONH_2$ at 270 nm. They also showed a broad tail extending up to 500 nm, a feature which was absent in peptides that did not contain the –COOH group. Besides Lys all other non-

aromatic amino acids were also studied which revealed that the charged amino acids namely Lys, Glu, Arg, His and Asp show unique absorption signatures above 250 nm, which extend up to 400 nm unlike their uncharged counterparts. Also Lys.HCl shows absorption intensities ~6 times lower than pure Lys which supports the role of participation of charged head group behind the observed spectral signatures. These features were insensitive to pH and also to the D_2O exchange, which reveal that proton transfer does not play a role in the absorption spectra. All these studies gave us initial clues to understand the phenomena involved. However, in order to characterize the transitions involved we required a protein which was devoid of aromatic amino acids and rich in charged amino acids. In the hunt for such a protein we devised a new methodology wherein unique prime numbers were allotted to each amino acid based on their hydrophobicities. This was then used to generate a unique numerical score for the protein (ProtID and PS-Score) which reflect the amino acid composition of a given sequence. This methodology was applied to search and hunt down a synthetic protein (Alpha 3C) from the PDB database. The protein is made up of 54 % of charged amino acids and was devoid of any aromatic amino acids. The last part of the thesis deals with employing α_3C protein to gain further insights into the mechanism of unusual absorption spectra. Despite lacking the conventional aromatic chromophores, α_3C exhibits moderate absorption features around 270 nm with a broad tail extending well into the visible spectrum. The role of the three dimensional fold of the protein behind the observed UV-Vis spectral features in the protein was examined. Elaborate theoretical studies revealed that interactions between spatially proximal Lys-Lys, Glu-Lys and Glu-Glu head-groups modulate the spectral transitions above 300 nm in the protein. The excited state TDDFT calculations on monomer and dimer forms of amino acids of the α_3C protein revealed that the unique spectral signatures of Lys and Glu amino acids arise from charge transfer transitions involving the amino (NH_3^+)/carboxylate (COO^-) head groups of Lys/Glu residues and the peptide backbone. The strength of absorption in the simulated spectra of Glu was more intense than that of Lys. This was attributed to the presence of a much stronger and potent electron donating head group in Glu and also to the shorter bridge between the electron donor state and the electron acceptor state, which makes the photo induced charge transfer. In summary, work from this thesis could propose an explanation for the unusual absorption spectra in Lys rich proteins observed in the past.

Table of Contents

Chapter 5: Investigations on α_3C Protein: Experimental and Theoretical Aspects

CHAPTER 1

INTRODUCTION AND REVIEW OF LITERATURE

Human biology relies primarily on biomolecules such as carbohydrates, proteins, lipids and nucleic acids. They are an extensive set of macromolecular machines which lay the foundation of all living systems[1]. All these biological macromolecules are important cellular components and perform well defined tasks necessary for the survival and growth of living organisms. Amongst the vast repertoire of biological macromolecules, proteins are one of the most versatile macromolecules present in the living systems. They carry out crucial functions in essentially all biological processes including catalyzing biochemical reactions[2]. Our understanding of living systems is therefore intimately connected to the characterization of structure and function of proteins. Understanding their structure and function is important in order to gain deeper insights about the various processes involved in a living body.

1.1 Spectroscopy

Spectroscopy has been employed for decades to answer the questions pertaining to protein structure and dynamics[3]. It occupies a very special position in Chemistry, Physics and in science in general. It is capable of providing accurate answers to some of the most difficult questions, particularly those concerning atomic and molecular structure. It is therefore a very handy tool to investigate the relevant structure and function of macromolecules.

Spectroscopy deals with interaction of electromagnetic radiation[4] with matter and we can harvest wealth of information, about the matter, from these interaction[5]. If matter is exposed to light or more precisely electromagnetic radiation, a number of processes can occur, including reflection, scattering, absorption followed by fluorescence/phosphorescence and/or photochemical reactions[6,7]. These processes can in turn provide important physical information like rotation, charge localization, molecular structure, symmetry and vibration (about a given molecule) which helps develop a better understanding of the system under investigation[8,9].

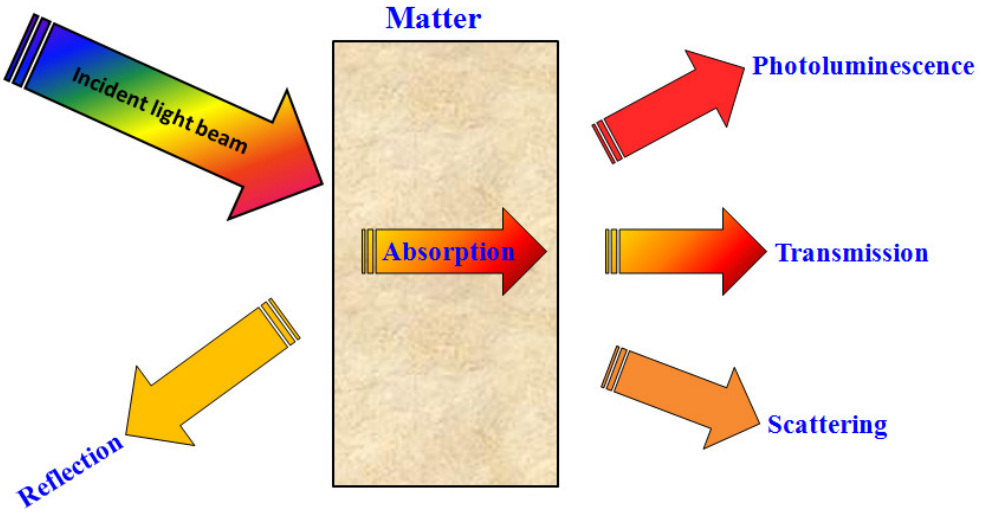


Figure 1.1: Schematic showing interaction of light with matter

Among all these processes, absorption has been widely employed to characterize various physical and chemical properties of biomolecules[10]. Absorption and emission of light by matter are important processes not only for the study of biological systems, but also for the function of life as a whole[11]. Without the interaction of light with chromophores, there would be no visual perception[12] and plants would not be able to perform photosynthesis[13,14]. These processes involve the interaction of matter with visible portion of the electromagnetic radiation[15]. However, in the study of biological structure, function and dynamics, the interaction between radiation and matter is not limited to the visible region of the electromagnetic spectrum.

1.2 Absorption Spectroscopy

Absorption spectroscopy is a general method applicable to almost all samples[16]. When a molecule absorbs electromagnetic radiation (more precisely a photon) the change which occurs depends on which part of the electromagnetic spectrum is absorbed. It is an extremely fast process[17], occurring in timescales of about 10^{-15} s. Absorption of energy causes an electron in the molecule to go from an initial lower energy state (ground state) to a final higher energy state (excited state).

A photon can only be absorbed, if the photon energy corresponds to the difference energy between the initial and final state of the sample[18] i.e.

$$\Delta E\ (h\nu) = E_{final} - E_{initial} \qquad (1.1)$$

where, h = Planck's constant (6.63 X 10^{-34} Js), ν = Frequency of the radiation, E = Energy of the radiation

Studying this change in energy state can yield information about various properties of the molecules.

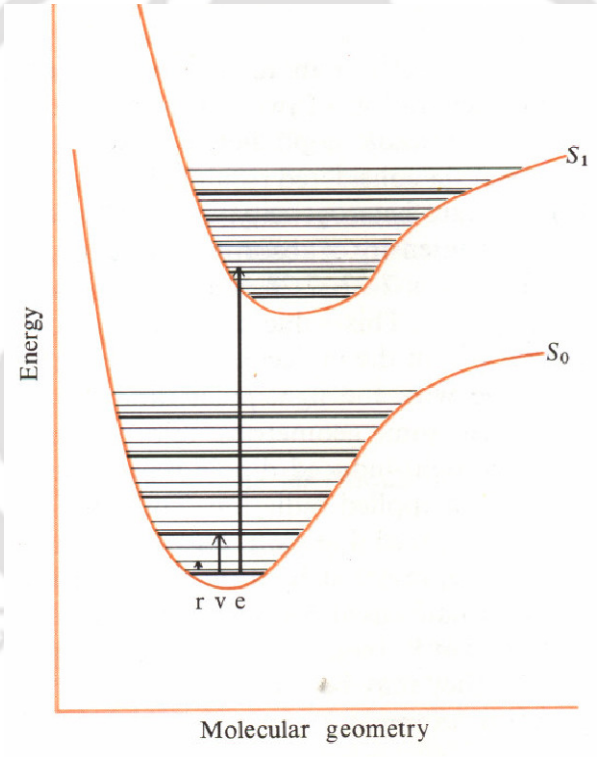


Figure 1.2: Different energy levels associated with a molecule and corresponding transition of an electron from ground state to excited state upon absorption of a photon (*Adapted from Biophysical Chemistry, Part II by Cantor and Schimmel*)[19]

Light is a rapidly oscillating electromagnetic field[20]. Molecules contain distribution of charges and spins that have electrical and magnetic properties. These distributions are altered when a molecule is exposed to light. The rate at which a molecule responds to this perturbation is often dealt with in absorption spectroscopy.

Figure 1.2 shows the potential energy surfaces of the two lowest electronic states of a simple molecule.

The energy between the electronic states (~80 kcal/mol) is larger than that between the vibrational states (~10 kcal/mol) which in turn is larger than energy between the rotational levels (~1 kcal/mol)[21] i.e.,

$$E_{elect} \gg E_{vib} > E_{rot}$$

The total potential energy of a molecule generally is represented as the sum of its electronic, vibrational, and rotational energies[22]:

$$E_{Total} = E_{elect} + E_{vib} + E_{rot} \quad (1.2)$$

The energy gap between the electronic states is much greater than the thermal energies at room temperature, thus in absence of any radiation that can excite a transition, all molecules in a solution are in the lowest electronic state[19]. When light of correct frequency is absorbed, the molecule can be excited to one of the many rotation-vibration levels of electronic state S_1.

1.2.1 Quantum mechanical treatment of absorption spectroscopy

Quantum mechanics has been an indispensible tool which has allowed us to understand how molecules respond to light and how they change from one state to another. Quantum mechanics embodies the principles that govern electrons and atoms[23]. It permits us to predict the characteristics of a transition observed by any type of spectroscopy applied to macromolecules[24]. Most of the theories described here have been extracted from Textbook on Biophysical Chemistry by Cantor and Schimmel[19].

In quantum mechanics, all the properties of a system (an atom, a molecule) that depend on position and time are described by a wavefunction ψ. In general, ψ is a function of the

positions and spins of all electrons and nuclei in the system. It is a complex number and is time dependent.

Every physical observable of a system (e.g., energy, position, dipole moment, etc) is governed by a Hamiltonian operator. A transition between two states of a system can be induced by a perturbation, which is measured by an operator. Thus, for predicting the properties of a molecule, we need its wavefunction which can be determined from the Schrodinger equation

$$\hat{H}\psi = i\hbar\frac{d\psi}{dt} \tag{1.3}$$

where, $\hat{H}$ is the Hamiltonian operator which describes the total energy (potential energy + kinetic energy) of a system and $\hbar$ is h/2π

1.2.1.1 Interaction of light with molecules

The ability of light to induce transitions within a molecule is calculated by its ability to induce dipole moments in the system. The induction of dipoles within the system can be represented classically by the following equation

$$\mu_{ind} = \tilde{\alpha}\,.E \tag{1.4}$$

where, μ_{ind} is the induced dipole moment, $\tilde{\alpha}$ is the polarizability of the molecule and E is the electric field of light. Since E fluctuates with time, μ_{ind} also fluctuates with time.

Suppose a molecule is initially in a state A denoted by ψ_a and that it is perturbed by light to state B denoted by ψ_b. The wavefunction in the presence of light may be represented as

$$\psi\,(t) = C_a(t)\psi_a e^{-iE_a t/\hbar} + C_b(t)\psi_b e^{-iE_b t/\hbar} \tag{1.5}$$

where, C_a and $C_b$are coefficients which relate to the probabilities that a system will be found in state A or in state B and E_a and $E_b$correspond to energies of states A and B respectively.

The Hamiltonian in the presence of light can be written as

$$\hat{\mathrm{H}}' = \hat{\mathrm{H}} + V(t) \tag{1.6}$$

The term $V(t)$ represents the effect of light on the system. This interaction energy between a molecule and light is given by:

$$V(t) = \tilde{\mu}.E_0 e^{i\omega t} \tag{1.7}$$

where, $\tilde{\mu}$ is the induced dipole moment, E_0 is the maximum amplitude (describes the polarization direction of the light) of electric field felt by the molecule and ω is the circular frequency ($2\pi\nu$) of light incident on the system.

The probability of transition from state A to state B after interaction with light is determined by the transition dipole integral (also denoted as μ_{ba}) which can be calculated by performing the integration $\int \psi_b^* \tilde{\mu}\, \psi_a d\tau$ which is also written as $\langle\psi_b|\tilde{\mu}|\psi_a\rangle$. The value $|\langle\psi_b|\tilde{\mu}|\psi_a\rangle|^2$ provides information about the electronic distribution within a molecule.

The probability P_b that the system is in state B at time *t* is given by

$$P_b = |C_b(t)|^2 \tag{1.8}$$

$$|C_b(t)|^2 = \frac{|\langle\psi_b|\tilde{\mu}|\psi_a\rangle.E_0|^2}{\hbar^2} \frac{t^2 Sin^2[(E_b/\hbar - E_a/\hbar - \omega)t/2]}{2[(E_b/\hbar - E_a/\hbar - \omega)t/2]^2} \tag{1.9}$$

Thus $|C_b(t)|^2$ is large when the denominator $(E_b/\hbar - E_a/\hbar - \omega)$ is small. Since $\hbar\omega$ is the energy of the light, the transitions from state A to state B will be induced only when the light energy is equal to the energy separation between the two states $(E_b - E_a)$:

$$h\nu = \hbar\omega = (E_b - E_a) \tag{1.10}$$

The rate at which energy is taken up by a molecule from the incident photon enables us to determine the absorption intensity.

The transition rate dP_b/dt is given by:

$$dP_b/dt = B_{ab} I(\nu) \tag{1.11}$$

where, B_{ab} is the transition per unit energy density and $I(\nu)$ is the energy density incident on the molecule.

The term B_{ab} is given by

$$B_{ab} = (2/3)(\pi/\hbar^2)|\langle\psi_b|\tilde{\mu}|\psi_a\rangle|^2 \quad (1.12)$$

The rate at which energy is removed from the light will depend on the number of A to B absorption transitions stimulated by light, on the number of B to A emission transitions, and on the energy per transition $(E_b - E_a)$.

From equation (1.10), this rate is

$$-\frac{dI(\nu)}{dt} = h\nu(N_a B_{ab} - N_b B_{ba})I(\nu) \quad (1.13)$$

where, N_a and N_b are the number of molecules per cm^3 in states A and B respectively. Thus, light absorption (and other optical properties) depends on concentration of the samples. The quantities B_{ab} and B_{ba} are called the Einstein coefficients for stimulated absorption and emission respectively.

1.2.2 Types of absorption spectroscopy

Each portion of the electromagnetic spectrum consists of various levels of energy (different frequencies/wavelengths) appropriate for the excitation of certain types of physical processes. Depending on which frequency of light is absorbed by the sample we can have different classes of spectroscopy[25]. The sensitivity, specificity and resolution of all these spectroscopic methods allow one to gather a wealth of information for a given molecule[26,5]. These regions of the electromagnetic spectrum and their associated techniques are probably one of the most widely used techniques for analytical work and their applications to the solution of biological problems[27,28,29]. Table 1.2.2 shows various spectroscopic techniques employed to study different parameters of a molecule.

Spectroscopic methods are, in general, the method of choice (i) to investigate changes in the behavior of a molecule under different solvent conditions, and (ii) to compare the properties of related molecules, such as homologous or mutated forms of a protein[30]. In addition, they are widely used to measure protein stability and to follow structural transitions such as unfolding and refolding under a variety of conditions.

Table 1.2.2: Different regions of electromagnetic radiation and associated spectroscopic techniques (*Adapted from Biophysical Chemistry, Part II by Cantor and Schimmel*)[19]

Wavelength (cm)	**Approximate energy (kcal/mole)**	**Spectroscopic region**	**Molecular parameters**	**Techniques**
10^{-11}	3×10^{8}	γ-Ray	Transitions in nuclei	Mössbauer
10^{-8}	3×10^{5}	X-Ray	Transitions of electrons in inner orbitals	X-ray diffraction, Scattering
10^{-5}	3×10^{2}	Vacuum UV	Transitions of electrons in outer orbitals	Electronic or UV/Vis spectroscopy
3×10^{-5}	10^{2}	Near UV		
6×10^{-5}	5 x10	Visible		
10^{-3}	3×10^{0}	Infra Red (IR)	Change in molecular rotational and vibrational states	IR spectroscopy
10^{-2}	3×10^{-1}	Far IR		
10^{-1}	3×10^{-2}	Microwave	Change in rotational states of electron spin	ESR spectroscopy
10^{0}	3×10^{-3}	Microwave		
10	$3 \text{ X } 10^{-4}$	Radio frequency	Change in rotational states of nuclear spin	NMR spectroscopy

1.2.3 Ultraviolet-Visible Spectroscopy

It is one of the oldest methods in molecular spectroscopy. The development of quantum chemistry, has led to a better understanding of correlation between light absorption and the structure of matter[31]. Although UV-Vis occupies a very narrow portion in the entire electromagnetic spectrum, this range is of extreme importance, since the energy differences correspond to those of the electronic states of atoms and molecules. UV-Vis spectroscopy is hence also known as "**electronic spectroscopy**"[21]. If the energy of the electromagnetic radiation is sufficiently large, the electrons in the molecule are shifted to higher energy levels. The wavelength of absorption and the strength of absorbance of a molecule depend not only on the chemical nature but also on the molecular environment of its chromophores[32]. Hence, it can be used for purposes ranging from simple concentration determinations[33] to resolution of complex structural questions[34]. It is an excellent technique for following ligand-binding reactions, enzyme catalysis and conformational transitions in proteins and nucleic acids[35,36]. These measurements are nondestructive and very sensitive, and require only small amounts of samples for analysis. It is therefore a method of choice for studying of biopolymers like proteins and nucleic acids[37].

Conjugated and aromatic chromophores display bathochromic shift (shifting of absorption maxima to longer wavelengths) and an increase in intensity in the UV-Vis absorption because of a better overlap between the π orbitals which lowers the energy gap between adjacent orbitals[38]. The greater the degree of conjugation, the greater is the shift. The decrease in the energy between the HOMO and the LUMO as the conjugation increases can be explained on the basis of particle in a box model.

The particle-in-a-box is a simple quantum mechanical model[39] commonly used in to describe the behavior of a particle, such as an electron, that is under the influence of a constant potential in a given region of space. Outside of this region, the potential abruptly rises to infinity, meaning that it is impossible for the particle to leave this region.

When the Schrödinger equation is solved for the particle-in-a-box model, the resulting energy is given by the equation:

$$E_n = \frac{n^2h^2}{8mL^2} \tag{1.14}$$

where, n is a quantum number such that n = 1, 2, 3, … , h is Planck's constant, m is the mass of the particle, and L is the length of the box or region of space to which the particle is confined.

The energy difference for transition of an electron from HOMO to LUMO[40] can be calculated by the following expression:

$$\Delta E = E_{LUO} - E_{HFO} = \frac{n^2h^2}{8mL^2}(n^2_{LUMO} - n^2_{HOMO}) \tag{1.15}$$

We can easily calculate the wavelength of the main absorption band using the equation:

$$\lambda = \frac{hc}{\Delta E} \tag{1.16}$$

According to Eq. (1.14) the energy of a level varies inversely with the square of the length of the box. Thus the longer the conjugated chain, the closer the energy levels will be to each other and the less energy a photon will need to excite an electron. Naturally the lower the energy of a photon, the longer its wavelength will be. This model has been widely used to describe the effect of conjugation on the absorption spectrum of cyanine dyes[41,42].

1.2.3.1 Basis of UV-Vis Spectroscopy

One of the most important measures in light absorption is the molar extinction coefficient. For a sample of molecules incident with light intensity I_0 in a layer sufficiently thin (dl), the fraction of light absorbed is given by,

$$-dI/I = C\varepsilon' dl \tag{1.17}$$

where I is the intensity of light of wavelength (λ) transmitted through the sample, C is the concentration of absorbing molecules (in moles/L) , l is the path length the light travels in the sample (in cm) and ε' is a proportionality constant called the molar extinction coefficient ($M^{-1}cm^{-1}$)

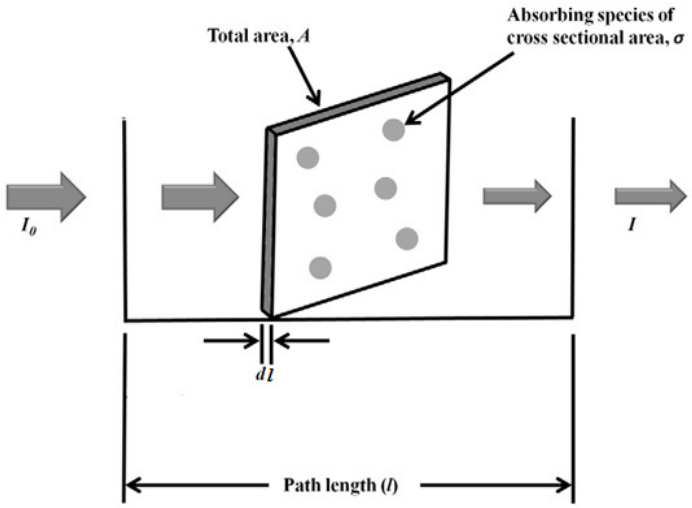


Figure 1.2.3.1: Diagrammatic representation of light absorption by molecules in an infinitesimal thin layer within a sample

Integrating equation (1.16) over the entire sample,

$$\int_{I_0}^{I}(-dI/I) = C\varepsilon' \int_0^l dl \tag{1.18}$$

we obtain,

$$\ln(I_0/I) = C\varepsilon' l \tag{1.19}$$

Converting this equation to log base 10, we have the common Beer-Lambert law:

$$A(\lambda) \equiv \log(I_0/I) = C\varepsilon\,(\lambda) l \tag{1.20}$$

where, $\varepsilon = \varepsilon'/2.303$

The Beer-Lambert law forms the basis of quantitative evaluation of light-absorption measurements. It states that the amount of light absorbed by a sample is proportional to the concentration of the substance[43], the path length of the cell containing the solution and the absorption coefficient[44]. The molar extinction coefficient defines the strength of absorption for a given molecule. It is a concentration independent property and is characteristic of the compound at a particular wavelength. The absorbance of a solute depends linearly on its concentration and therefore absorption spectroscopy is ideally suited for quantitative measurements.

It is sometimes useful to picture the extinction coefficient, ε, as the (hypothetical) cross-section area of the chromophore that is absorbing the light[1]. The absorbance cross-section (σ) is numerically related to ε as follows:

$$\sigma = 3.825 \text{ X } 10^{-21}\varepsilon \qquad (\text{cm}^2) \tag{1.21}$$

The molar extinction coefficient is related to the Einstein coefficient for stimulated absorption as

$$B_{ab} = (1000c/N_0h)\int(\varepsilon'/\nu)d\nu \tag{1.22}$$

where, N_0 is the Avogadro's number and c is velocity of light with frequency ν

Using equation (1.22) and equation (1.12) and converting ε' to ε we obtain the following relation

$$D_{ab} \equiv |\langle\psi_b|\tilde{\mu}|\psi_a\rangle|^2 = 9.180 \text{ X } 10^{-3} \int(\varepsilon/\nu)d\nu \qquad (\text{debye})^2 \tag{1.23}$$

D_{ab} is called the dipole strength which is a measure of strength of the transition dipole.

Another useful measure is oscillator strength, f_{ab} which compares the intensity of absorption to that expected from a three-dimensional harmonic oscillator. This can be shown as

$$f_{ab} = (8\pi^2 mc/3h\nu)D_{ab} = 4.315 \text{ X } 10^{-9} \int \varepsilon(\nu)d\nu \qquad (\text{dimensionless}) \tag{1.24}$$

where, m is the mass of the electron.

For a strongly allowed transition, f_{ab} may be in the range of 0.1 to 1. The quantities dipole strength and oscillator strength are very important in understanding some of the special optical effects observed in polymers.

1.2.3.2 Types of electronic spectra

Higher energy radiation in the UV (200-400 nm) and visible (400-800 nm) range of the electromagnetic spectrum causes molecules to undergo electronic transitions such that electrons are promoted to higher orbitals which results in electronic excitation. Different types of electronic transitions are possible[45]:

a) Transitions involving π, σ, and *n* electrons:

Valence electrons can generally be found in one of three types of molecular orbital: single, or σ, bonding orbitals; double or triple bonds (π bonding orbitals); and non-bonding orbitals (lone pair electrons). Hence, different electronic transitions can take place from a bonding orbital to anti bonding orbital[46].

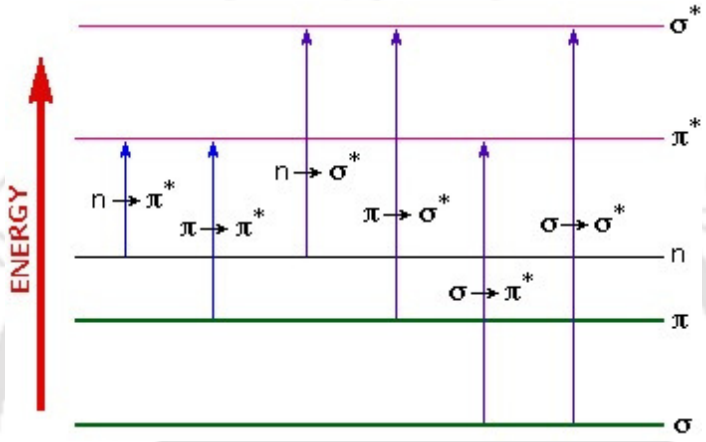


Figure 1.2.3.2: Different types of electronic transitions from different bonding orbitals to anti-bonding molecular orbitals

The most commonly observed transitions in biomolecules are the **π-π*** and **n-π*** transitions. Both σ-σ* and n-σ* transitions require a large amount of energy and therefore occur in the far ultraviolet region or weakly in the region 180-240 nm. Transitions of the n-π* and π-π* type occur in molecules with unsaturated centers; they require less energy and occur at longer wavelengths than transitions to σ* anti-bonding orbitals. Thus, only **π-π*** and **n-π*** transitions occur in the UV-Vis region[36].

The solvent in which the absorbing species is dissolved also has an effect on the spectrum of the species[47,48]. Peaks resulting from n- π* transitions are shifted to shorter wavelengths (blue shift) with increasing solvent polarity[49,50]. This arises from increased solvation of the lone pair, which lowers the energy of the n orbital. Often (but not always), the reverse (i.e. red shift) is seen for π - π* transitions[47]. This is caused by attractive polarization forces between the solvent and the absorber, which lower the energy levels of both the excited and ground states. This effect is greater for the excited state, and so the energy difference

between the excited and ground states is slightly reduced resulting in a small red shift. This effect also influences n- π* transitions but is overshadowed by the blue shift resulting from solvation of lone pairs[51].

b) Transitions involving *d* and *f* electrons:

These transitions involve *d* (First and Second transition metal series: Cr, Co, Cu and Ni) and *f* (ions of lanthanide and actinide elements)[52] electrons. The d-d transitions involve absorption between filled and unfilled d-orbitals[53]. They absorb broad bands of visible region[54]. These transitions appear as weakly intense on the spectrum because they are Laporte forbidden which states that if a molecule is centro-symmetric, transitions within a given set of p or d orbitals are forbidden[55].However, they are weakly allowed due to vibronic coupling[56,57] (Transitions that occur as a result of an asymmetrical vibration of a molecule, e.g. complexes of Manganese). Due to relatively low energy of transition, they can emit visible light upon relaxation which is why many transition metal complexes are brightly colored. The color of the transition metal complex solution is dependent on: the metal, the metal oxidation state, and the number of metal d-electrons[58]. For example Fe (II) complexes are green and Fe (III) complexes are orange/brown. The absorption spectra involving the *f*-electrons are narrow due to the screening of inner orbitals with well defined characteristic peaks[59,60].

c) Transitions involving charge-transfer electrons:

Many inorganic species form charge-transfer complexes which show charge-transfer (CT) transitions. For a complex to demonstrate charge-transfer behavior one of its components must have electron donating properties (donor) and another component must be able to accept electrons (acceptor). Absorption of radiation then involves the transfer of an electron from the donor to an orbital associated with the acceptor. The most common examples include the absorptions arising due charge transfer in metal-ligand complexes[61]. It can give rise to either ligand-to-metal charge-transfer (LMCT) bands (Eg. $KMnO_4$) or metal-to-ligand charge-transfer (MLCT) bands (For e.g. $[Ru(bpy)_3]Cl_2$)[62]. The position of charge transfer bands depend on the nature of the metal ion and ligand. It also depends on the relative ease of oxidation and reduction of the complex[63].

Charge transfer transitions are prevalent in biological systems as well[64,65,66,67]. Several metalloenzymes Ni (II), Cu (II) and Co (II) metal ions show prominent bands of charge transfer transitions[68]. One of the most popular examples of biological molecules having charge transfer transitions are blue copper proteins[69]. Color in these classes of proteins arises from ligand to metal charge transfer. Azurin[70], Plastocyanin and Stellacyanin[71,72] contain Copper ions surrounded by a distorted tetrahedron of two sulphur atoms and two nitrogen atoms, which originate from Cys and His residues in the protein. Strong spectral bands arising from a transition from π-bonding orbitals on sulphur atoms of Cys residues, to the empty t_2* orbitals on copper occur at wavelengths ranging from 470-830 nm[73].

Various photosynthetic reaction centers in bacteria and plants[74,75] also show characteristic charge transfer transitions. The best studied RCs are those isolated from two species of purple bacteria, *Rhodobacter sphaeroides* and *Rhodopseudornonavsi ridis*[76]. Molecular orbital calculations based on the crystal structure indicate that charge transfer transitions occur between two bacterial chlorophylls[77]. Cytochrome C has a broad CT band near 700 nm which is caused due to ligand to metal charge transfer. Movement of an electron to the heme Fe takes place from the methionine axial ligand[78]. Structural changes that separate the methionine from the heme cause the CT band to disappear. The photosynthetic reaction centers contain a metal ion (Mg^{2+}) which leads to MLCT bands which mix with local transitions. These transitions often get mixed with local transitions but have important influence on the position and line width of the absorption bands[79]. Charge transfer transitions are prevalent in enzymes (For e.g. 3-phosphoglyceraldehyde dehydrogenase) with coenzyme NAD. It has been reported that addition of NAD results in charge transfer complexing between the pyridinium ring of NAD and the enzyme side chain[80,81].

1.2.4 Chromophores in Proteins

The absorption of UV light by proteins has been analyzed in detail[82] and proposed as a structural probe from the very early days of molecular biology[83]. Measurements of biological macromolecules are limited to above 170 nm because they are studied mainly in aqueous solution and water itself absorbs strongly below 170 nm[84,85]. Basically there are three classes of chromophores in proteins.

1.2.4.1 Peptide bond

The electronic transitions in the peptide bond within a protein occur in the far UV region. The peptide bond contains π electrons which are delocalized over the N, C, and O atoms. An electron in a nonbonding, n-orbital is present near the O atom. Two distinct peaks owing to the electronic transitions within the peptide bond can be observed[86]. One strong peak at 190 nm (ε = 7000 $M^{-1}cm^{-1}$) occurs due to the π- π* transitions while another peak of a weaker intensity at about 210–220 nm (ε = 100 $M^{-1}cm^{-1}$) occurs due to n- π* transitions[87]. This transition is weak as it is symmetry forbidden and it forms a shoulder on the π- π* transition peak. A third transition can be observed at still higher energies of around 175 nm. This is tentatively assigned to an n-σ* transition[19].

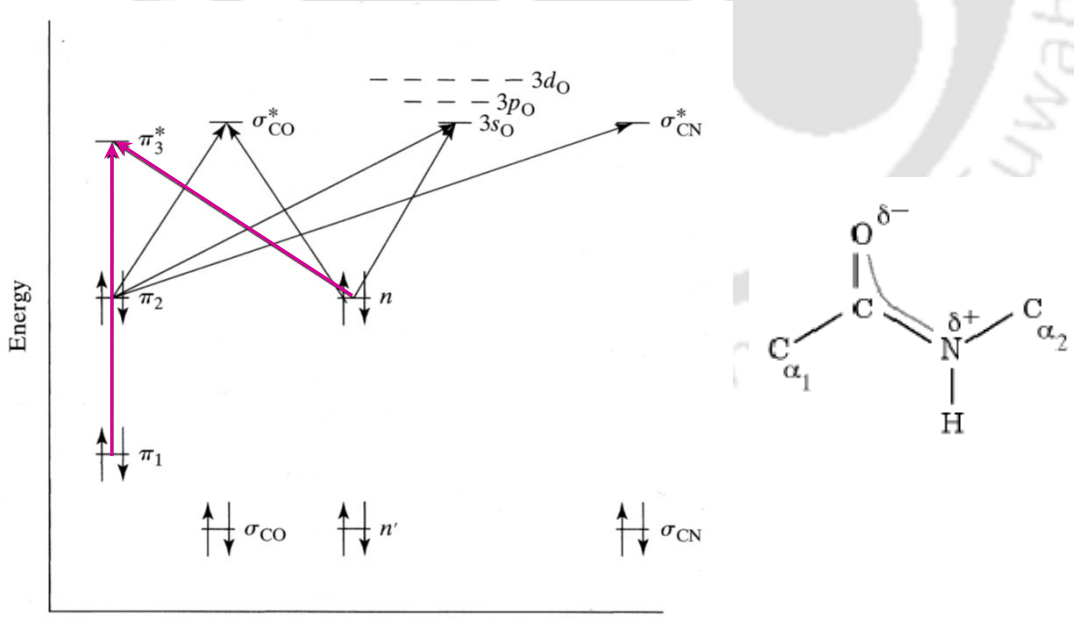


Figure 1.2.4.1: The occupied and unoccupied orbitals of the amide chromophore showing possible transitions from the filled orbitals to the unfilled orbitals. The amide chromophore is also shown. (*Adapted from Methods of Biochemical Analysis*)[88]

Peptide bond absorption is influenced by changes in secondary structures. It has been shown that for poly-L-glutamic acid[89] the peptide spectrum is conformation dependent. Poly-L-lysine also exhibits changes in absorption intensities with change in conformation[90]. The α-helical conformation of a peptide shows a decreased absorptivity in comparison to either random coil or β-conformation[91].

1.2.4.2 Aromatic amino acids

A number of amino acids (Asp, Glu, Asn, Gln, Arg and His) have weak electronic transitions at around 210 nm. Usually, these cannot be observed in proteins because they are masked by the more intense peptide bond absorption[19]. Absorption in the region of 220 nm due to the peptide bond has been used in the quantification of proteins but many other compounds also absorb at this wavelength and as a result such methods suffer from a considerable degree of interference. Thus, the most useful side chain optical properties are those that occur at wavelengths longer than 230 nm, where peptide absorption is reduced to negligible values.

Among the 20 naturally occurring amino acids, only the aromatic amino acids namely Trp, Tyr and Phe have been reported to absorb significantly in the near UV region owing to the presence of aromatic moieties in them.

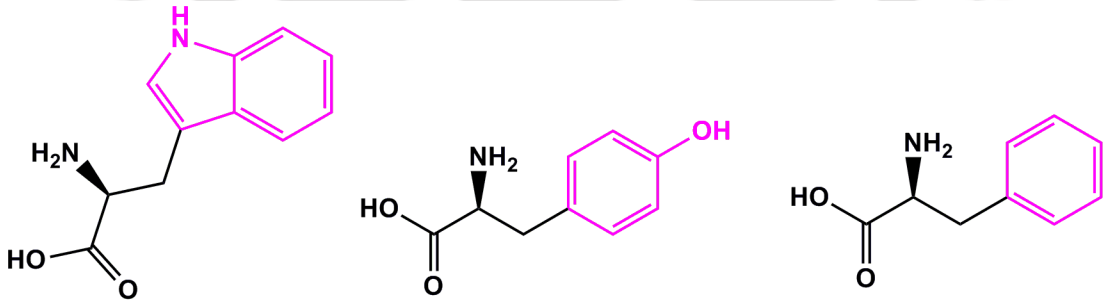


Figure 1.2.4.2A: Three aromatic amino acids (Trp, Tyr and Phe) with their side chains

Among the three aromatic amino acids, Trp has the strongest absorption in the near UV region (ε = 5,600 $M^{-1}cm^{-1}$ at 280 nm)[92]. The absorption spectrum of the indole side chain of Trp is complex. It consists of basically two major peaks; one near 220 nm (ε = 36,000 M^{-1}

cm^{-1}) and one at 280 nm (ε = 5,600 M^{-1} cm^{-1})[93]. At least two independent electronic transitions are responsible for the spectra of Trp in the 260-310 nm, with one of the transitions being π- π* transition.

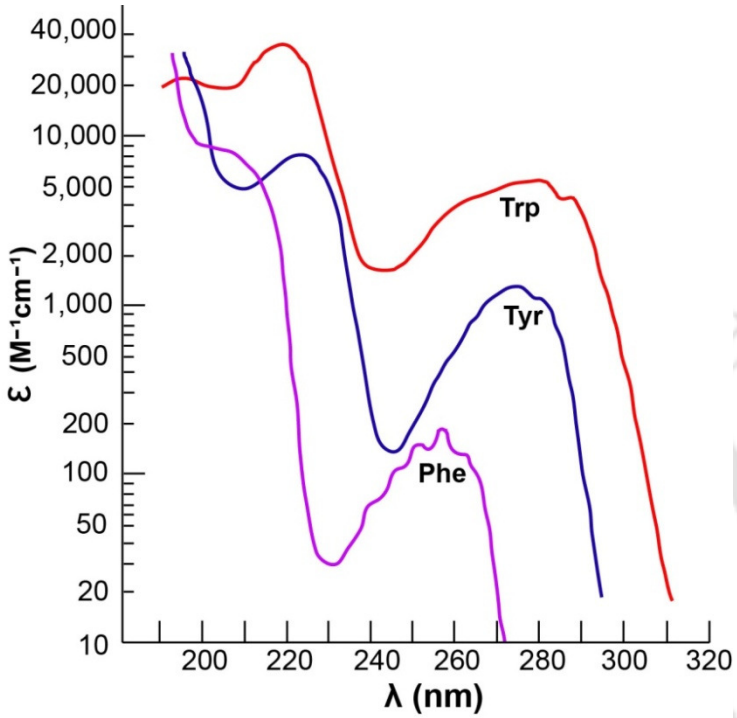


Figure 1.2.4.2B: Absorption spectra of aromatic amino acids (*Adapted from "Ultraviolet spectra of Proteins and Amino Acid" by D.B Wetlaufer)*[94]

Another aromatic residue with non-negligible absorption in the near-UV region is tyrosine (Tyr-OH)[95]. Electronic transitions occur at 275 nm (ε = 1400 $M^{-1}cm^{-1}$) and 222 nm (ε = 9000 $M^{-1}cm^{-1}$)[96]. The 275 nm absorption band corresponds to a π- π* transition[94]. At alkaline pH, the OH group of tyrosine side chain de-protonates (pK_a = 10.07)[97]. The resulting tyrosinate ion (Tyr-O^-) has a slightly red-shifted absorption compared to tyrosine, with maxima at 240 nm (ε = 1100 $M^{-1}cm^{-1}$) and 290 nm (ε = 2300 $M^{-1}cm^{-1}$)[98,99]. This sensitivity has been used in accurate titration of Tyr residues in proteins as well as in the separate determination of Tyr and Trp contributions to an observed absorption spectrum[100]. The absorbance of Trp is less sensitive to pH than that of Tyr. Also, the pK_a values of Trp lie outside the pH range in which most proteins can be handled safely[19].

Phe shows low intensity absorption[101] band around 257 nm (ε = 200 $M^{-1}cm^{-1}$) which corresponds to a symmetry forbidden π- π* transition[94] and another band is observed

around 205 nm (ε = 9600 $M^{-1}cm^{-1}$). Variation in pH gives rise to small changes in the spectrum of Phe[96]. Besides these three aromatic amino acids, the sulphur containing amino acids, namely Cys and Met have been reported to exhibit low absorption bands in 230-240 nm range[94]. However, these transitions are not easily measurable in proteins because their largest wavelength strong transition (~230 nm) is submerged in the peptide region[19]. Disulphides (cystine) have longer-wavelength transitions with λ_{max} values between 250-270 nm (ε = 300 $M^{-1}cm^{-1}$)[102]. Cystine occurs in high proportions in many proteins and thus along with the aromatic amino acids it is also considered in order to account for the near-UV absorption in proteins.

His which contains imidazole group in its side chain also absorbs appreciably between 185-220 nm (ε = 6000 M^{-1} cm^{-1} at 212 nm)[94].However, there are no extensive studies on this amino acid as other amino acids absorb much more strongly in this region.

The spectra of aromatic amino acids and peptide bonds in proteins are significantly influenced by their local environment. A denatured protein usually has a somewhat different ultraviolet spectrum than the native protein so that protein folding reactions can be monitored by measuring spectral changes[103].

1.2.4.3 Prosthetic groups and Co-Enzymes

Numerous proteins contain a tightly bound non-protein part known as prosthetic groups (For e.g. heme, flavin, carotenoid) which are important for the biological activity of the molecule. These proteins along with some metal–protein (For e.g. Azurin, Xanthin oxidase) complexes exhibit strong absorption in the UV-Vis region. Important coenzymes of proteins such as FAD, NADH and NAD^+ show spectra in the UV-Vis region. FAD absorbs at 450 nm (ε = 11,300 $M^{-1}cm^{-1}$)[104], NADH at 340 nm (ε = 6220 $M^{-1}cm^{-1}$) and NAD^+ at 259 nm (ε = 16,900 $M^{-1}cm^{-1}$)[105]. Heme shows a very intense absorption due to the presence of porphyrin ring at 404 nm (ε = 170,000 $M^{-1}cm^{-1}$)[106]. The spectra of these prosthetic groups are often in the visible region of the spectrum, easily separable from the absorption due to the protein, and can be monitored to follow the reactions undergone by the protein. The light absorption properties of a protein associated chromophore may be of direct biological relevance as in retinal in vision and chlorophyll in photosynthesis.

If we look carefully, the only players which have been reported to significantly contribute to the UV-Vis spectra of proteins are the peptide bond, aromatic amino acids, disulphides and the prosthetic groups present in some proteins. Therefore, the absorption spectrum of a protein is expected to remain optically silent beyond 315 nm if it lacks these chromophores and practically no absorption signals in the UV region are expected from a molecule lacking aromatic moiety.

1.2.4.4 Absorption beyond 250 nm arising from non-aromatic amino acids

1.2.4.4.1 Lysine

Several years ago, L-Lysine monohydrochloride (Lys.HCl) was reported to display new near UV absorption spectra with λ_{max} ~270 nm (ε = 0.34 $M^{-1}cm^{-1}$)[107]. These unique spectral features were observed only at high concentrations (0.5-1 M) in aqueous medium at pH 7 and were concentration dependent which were attributed to likely aggregates of Lys.HCl. It also displayed blue fluorescence (~435 nm) on excitation with 355 nm. Since Lys has an aliphatic side chain and does not contain any aromatic moiety, the observed spectral features could not be explained in terms of the known electronic transitions expected from the Lys side chain. Control studies with Gly showed that it was the Lys side chain and not the terminal amino or carboxyl groups of the amino acid, which contribute towards the unique spectral features at 270 nm. These features were deduced to arise from intermolecular interactions between adjacent Lys chains in the aggregates. However, the nature of the electronic transitions from this aggregate like species was not probed thoroughly. Similar observations of absorption at ~270 nm for Lys.HCl have been reported by other workers as well. Chai and co-workers attributed the anomalous UV-Vis peak at 270 nm to characteristic features of water structure which order around the charged moiety of Lys amino acid[108]. They interpreted the 270 nm absorption as a reflection of the overall quasicrystalline nature of water in the vicinity of charged or hydrophilic entities. Marti et al. have assigned the main absorption feature of structured water to the presence of two π-stacked ground-state water molecules. They postulated the corresponding fluorescence emission to arise from two relaxed side-hydrated π-stacked water molecules (a relaxed tetramer) which constitutes a unique excimer-type fluorescent moiety[109]. The spectral features of Lys.HCl and Gly have also been studied by Degtyareva et al. They observed

unusual spectral and optical features from both Lys.HCl and Gly at high concentrations (0.3-0.5 M). They reported multiple excitation and emission peaks from both the amino acids and suggested formation of multiple states and /or multiple species of these amino acids in aqueous medium as plausible reasons behind different peaks[110].

Proteins rich in Lys residues such as HSA, Calf thymus Histone and poly-L-Lysine.HCl were also reported to exhibit absorption features which could not be explained on the basis of conventional chromophores present within a protein[111]. Poly-L-Lys was shown to display similar absorption features as Lys, but at a concentration which was about 250 fold less than that of Lys amino acid. The absorption features observed in these proteins were deduced to arise from intramolecular interactions between two or more Lys residues present in close vicinity in these proteins. Since HSA and Calf thymus Histone contain aromatic amino acids as well, the absorption spectra below 300 nm could not be investigated as the Lys spectra at 270 nm were masked by the presence of much stronger absorbing aromatic amino acids. However, these proteins showed a broad tail beyond 315 nm where contribution from conventional chromophores such as Trp, Tyr, Phe, Cys and Cystine are negligible[112]. HSA was reported to have a molar extinction coefficient of about 1546 $M^{-1}cm^{-1}$ at 325 nm. The unique absorption features in Lys rich proteins were attributed to the close proximity (separated by 7-8 Å) of the Lys head groups as proteins which lacked proximal Lys such as HEWL, Barstar and Subtilisin Carlsberg where optically silent beyond 320 nm. The three dimensional architecture of the protein was thought to enable a multitude of intramolecular interactions between the Lys residues in HSA. Proteins unfolding studies with calf thymus histone clearly demonstrated a fourfold decrease in absorption intensity beyond 300 nm upon unfolding. This study suggested that the population of spatially proximal Lys residues was significantly reduced upon unfolding.

Detection of very low concentrations of lead ions in aqueous solutions (picomolar concentrations) by fluorescence change at 430 nm in HSA has been reported. Interaction among the Lys residues within HSA has been thought of as possible reasons for this fluorescence change[113].

How can two Lys residues be in close proximity of each other?

Since Lys is a charged amino acid with a positively charged ε-NH_3^+ group in its side chain, the interaction and close proximity among the Lys residues cannot be explained in terms of electrostatic forces between them as repulsive forces are expected to dominate beyond certain proximity. It was hypothesized that a common water molecule acts as a bridge between two Lys residues to bring them together[111]. The two lone pairs in the O atom of the bridging water molecule can act as hydrogen bond acceptor for one H atom each of the two nitrogens facing each other.

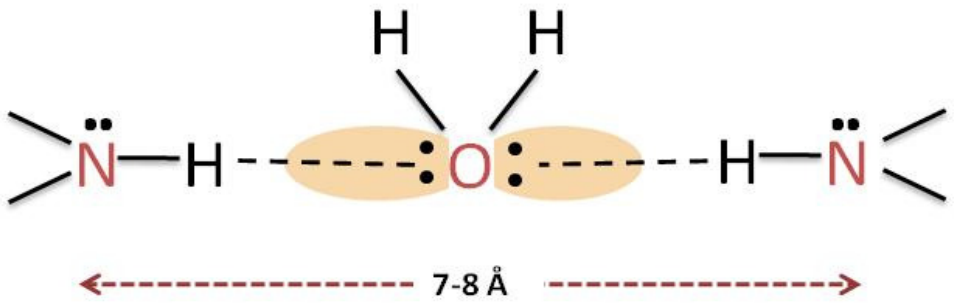


Figure 1.2.4.4.1: Illustration of proposed bridging water molecule between two Lys residues

The above results are intriguing because Lys has an aliphatic side chain ending with a primary amine and possesses no identifiable chromophore which can absorb in near-UV region. However, the nature of the chromophore and electronic transitions involved were not probed. Unusual absorption/fluorescence from protein fibrils and aggregates lacking any aromatic moiety has been reported by several other groups.

1.2.4.4.2 Protein aggregates lacking aromatic amino acids

Recently, it has been reported that the amyloid fibrils formed from protein as well as small peptides exhibit luminescence in blue and green region, although the peptide or protein is devoid of chromophores which can emit in the visible region. In the study of Aβ (16–24) peptide, Anand and Mukherjee have ascribed the visible fluorescence of amyloid fibril obtained from peptide to the excitonic transitions[114]. Studies on elastin-related polypeptide, poly (Val-Gly-Gly-Leu-Gly) have suggested that direct excitation of the fibrils may induce electronic transitions in peptides (e.g., in amide groups)[115]. Studies on peptide nanotubes (PNTs), where the self-assembly process of the dipeptides NH_2-Phe-Phe-COOH (FF) were

studied were reported to show step like absorption features at 245-264 and 300-370 nm. Strong photoluminescence of PNTs in the blue and UV spectra of exciton origin has been reported[116]. Studies on proteins and large peptides such as Gamma-II crystalline and a peptide fragment of this protein which lacks any aromatic amino acid have reported novel excitation that peaks at 340 nm and yields visible violet-blue radiation with apparent band maxima at 425, 445, 470, and 500 nm. They propose the delocalization of the peptide electrons through intramolecular or intermolecular hydrogen bond formation to be behind the long-wavelength electronic transitions[117,118].

Smith et al. suggested that the blue emission in Fmoc-di-phenylalanine (Fmoc-FF) is related to the formation of extensive J-aggregates that arise from π-stacking of aromatic moieties[119]. However, Chan et al. have proposed that the intrinsic fluorescence observed in peptides such as Aβ (1-42), lysozyme and Tau is independent of the presence of aromatic side chain residues within the polypeptides. Electronic delocalization via hydrogen bonds in β-sheet structure has been postulated to cause this phenomenon[120]. Studies on Elastin-Derived Peptide GVGVAGVG by Sharpe et al. also support the theory of hydrogen-bonded water molecules within the cross-β structure, which results in fluorescence[121]. A very recent report by Pinotsi and co-workers on human Aβ (1–42) and Aβ (33–42) peptides shows proton transfer across hydrogen bonds in fibrils which lack aromatic amino acids as one of the probable mechanisms behind the emergence of blue fluorescence[122]. The origin of this 'unusual luminescence' in the visible region is highly debated and accordingly various theories have been put forward to explain the phenomena. Although there are several reports on unusual fluorescence from different types of protein aggregates which lack aromatic moiety, there are no reports which explain the mechanism involved in unusual absorption features of amino acids lacking aromatic moiety.

Preliminary reports by Homchaudhuri and Swaminathan point towards the interaction among Lys residues as one of the possible reasons behind the observed unusual spectral features. However, a concrete mechanism behind the underlying phenomenon and the nature of the chromophore involved still remains to be answered. This is rather an interesting observation which could open an altogether new aspect in protein absorption studies which could find applications in protein dynamics studies. Another aspect to be

noted is that none of these studies have reported investigations on a whole protein which is completely devoid of any of the aromatic amino acids. Spectra beyond 300 nm have not been reported for solvated proteins in monomeric form in any of these reports. Studies on such a model protein would definitely aid the present studies and also provide better insights into the mechanism involved as the interference from aromatic amino acids in the ~270 nm region can be clearly ruled out. Additionally, a protein devoid of any aromatic amino acid and yet rich in Lys residues would be an ideal candidate for such studies. However finding such a protein in such a vast repertoire of proteins is a big challenge in itself. Protein databases such as PDB, Uniprot, etc. in present date account for over millions of proteins and the number keeps getting bigger and bigger every single day. At present there are no tools available which can help us segregate proteins on the basis of their amino acid content. Current approaches which store the sequences as one letter alphabetical code make the above task tedious and computationally intensive. Novel techniques which can give quick information about the amino acid composition of a protein are therefore required.

Being motivated by the intriguing spectral features from Lys amino acid and reports of absorption beyond 300 nm in proteins rich in Lys residues, we decided to carry forward the work in order to gain better insights into the phenomena involved. This thesis work is an attempt to understand the molecular mechanisms involved behind unique spectral signatures arising from Lys residues. The aim of our studies is to understand the nature of the chromophore and transitions involved behind these unique spectral features.

1.3 Objectives for the thesis work

1. To identify the chromophore behind the unique spectra arising from Lys residues and understand the nature of electronic transitions involved.

2. To investigate the contribution of other aliphatic amines, non-aromatic amino acids and their peptides to the observed spectra.

3. To devise a new approach to hunt for proteins rich in specific amino acids in the proteome and develop graphically informative tools to display protein sequences.

4. To investigate the spectra from a Lys rich protein devoid of aromatic amino acids.

Towards this end, we have investigated the absorption spectra of amine containing non-aromatic compounds and Lys containing short peptides devoid of aromatic amino acids in aqueous medium. We have also studied the absorption spectra of all non-aromatic amino acids extensively. Further to gain better insights into the mechanism behind the unusual absorption of proteins involved, it is desirable to study a protein devoid of any aromatic amino acids yet rich in Lys residues that are spatially located in close proximity within the protein fold. Using a new tool developed by us a synthetic protein α_3C (67 residues) which is rich in Lys residues and is devoid of any aromatic amino acids was identified.

The protein α_3C does not contain any aromatic amino acids which can contribute to the spectra in the near UV region. Further, the protein has 14 pairs of Lys residues which were within 10 Å of each other. We carried out systematic experimental investigations on α_3C using optical absorption and Circular Dichroism. Further, in collaboration with Dr. R. Venkatramani's research group at Tata Institute of Fundamental Research, Mumbai, we initiated computational approaches to understand the spectra. I carried out classical Molecular Dynamics (MD) simulations and associated trajectory analysis for α_3C in consultation with Dr. Venkatramani. Additionally, Ms. Imon Mandal (Graduate student from Dr. Venkatramani's research group) carried out excited state Time Dependent Density Functional Theory (TDDFT) electronic structure calculations on amino acid fragments extracted from MD snapshots.

Finally our hunt for Lys rich protein sequences led us to develop a new tool to quantify amino acid composition among different proteins in the proteome. We devised a novel way of doing so by assigning a unique prime number to each of the 20 amino acids in order of their increasing hydrophobicity. This has then been used to assign unique scores to protein sequences which can quickly give information about the amino acid composition in a given protein sequence. Additionally, new visual tools to represent protein sequences were developed.

CHAPTER 2

STUDIES ON AMINE CONTAINING ALIPHATIC COMPOUNDS AND SHORT PEPTIDES

2.1 Introduction:

Unusual absorption ~270 nm from Lys.HCl has been studied by Homchaudhuri and Swaminathan in which they predicted the role of Lys side chain as one of the probable reasons behind the unique spectral features of Lys in the near UV region. Initial reports had hinted towards the involvement of ε-NH_2 group of the Lys side chain behind the unique spectral features of Lys (chapter 1), therefore we decided to investigate few aliphatic compounds which contained amine group in order to probe the role of NH_2 moiety behind the observed optical signatures.

Far UV absorption spectra from 159-256 nm of some simple alkyl amines such as ammonia (λ_{max}: 194 nm, ε = 5600 $M^{-1}cm^{-1}$); methyl amine (λ_{max}: 215 nm, ε = 600 $M^{-1}cm^{-1}$); dimethyl amine (λ_{max}: 222 nm, ε = 100 $M^{-1}cm^{-1}$); trimethyl amine (λ_{max}: 227 nm, ε = 900 $M^{-1}cm^{-1}$) have been reported by Harrison and co-workers[123]. UV absorption values for ethylamine (λ_{max}: 210 nm, ε = 800 $M^{-1}cm^{-1}$); diethyl amine (λ_{max}: 193 nm, ε = 3000 $M^{-1}cm^{-1}$) and trimethylamine (λ_{max}: 213 nm, ε = 6000 $M^{-1}cm^{-1}$) have also been reported[124]. However, all these reports were restricted to below 260 nm. We investigated the absorption from some amine containing compounds in the UV-Vis region.

Also, studies with proteins rich in Lys residues such HSA speculated the close proximity between the side chains of Lys residues as one of the probable reasons behind the observed spectral features (chapter 1). Based on the analysis of available three-dimensional structures from PDB and earlier work on the Lys amino acid, an intra-molecular interaction between Lys side chains in close spatial proximity was deduced to be the origin for the above spectra[111]. To understand this feature more clearly, absorption studies on short peptides (4-7 amino acids) were carried out. These peptides lacked any aromatic amino acid, and contained pair of Lys residues placed at different positions. The objective was to understand the effect of intramolecular distance between two Lys residues within a peptide sequence on the absorption spectra of the peptides. Experiments on single Lys amino acids and peptides without any Lys residue served as controls.

2.2 Materials and methods:

2.2.1 Materials:

The aliphatic compounds, *viz.* trans-4-Cyclohexene-1, 2-diamine dihydrochloride, MW: 185.09 g/mol (726184); (1R, 2R)-(−)-1, 2-Diaminocyclohexane, MW: 114.19 g/mol, (346721); Piperazine, MW: 86.14 g/mol (P45907); trans-1, 4-Diaminocyclohexane, MW: 114.19 g/mol (32851); (1S, 2S)-trans-1, 2-Cyclopentanediamine dihydrochloride MW: 178.08 g/mol (670219), Cyclopropanemethylamine hydrochloride, MW: 107.58 g/mol (A63805); Acetylcholine chloride, MW: 181.66 g/mol (A6625); and Gly-Gly hydrochloride, MW: 181.66 g/mol (G1127) were purchased from Sigma Aldrich. Rink amide MBHA resin was purchased from Fluka (Loading 1.1 mmol/g). BOP [(Benzotriazole-1-yloxy) tris (dimethylamino) phosphoniumhexafluorophosphate] (MW: 442 g/mol), PyBOP [(Benzotriazole-1-yl-oxy-tris-pyrolidine-phosphonium hexafluorophosphate], Diisopropyl ethylamine (DIPEA) (MW: 129 g/mol), were purchased from Sigma. Dimethylformamide (DMF, extra pure grade), dichloromethane (extra pure grade) and Acetonitrile of HPLC grade were obtained from Merck (India). Acetic anhydride (synthesis grade), N-methyl imidazole (extrapure), Trifluoroacetic acid (TFA) of extra pure grade were purchased from SRL (India). Deionised water at 18.2 MΩ was used. All Fmoc (Fluorenylmethyloxycarbonyl) amino acids were purchased from GL Biochem (Shanghai) with Boc (butyloxycarbonyl) protected side chain for Lys.

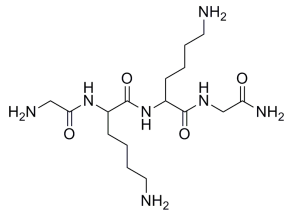
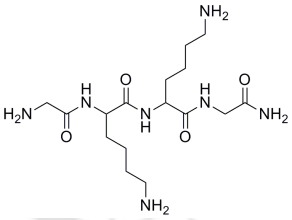
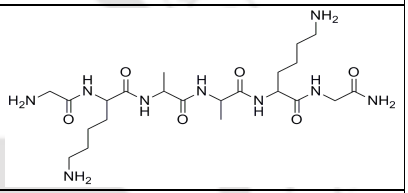
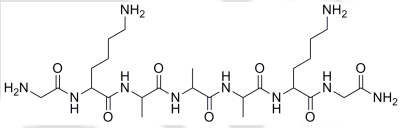
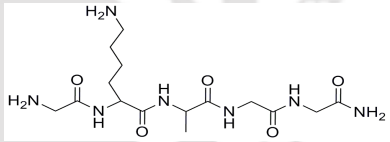
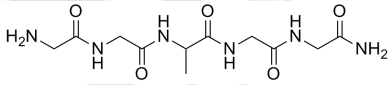
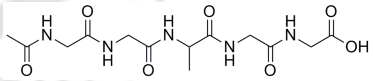
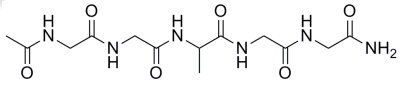
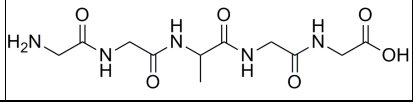
2.2.2 Methods:

2.2.2.1 Solid Phase Peptide synthesis

Peptides with varying distance between the Lys residues were synthesized by standard Fmoc/tertiary-Butyl orthogonal protection strategy using solid phase peptide synthesis. The syntheses were performed manually on a Stuart blood tube rotator.

Table 2.2.2.1 shows the target peptides along with their molecular weight and structure.

Table 2.2.2.1: Sequence, calculated molecular weight and structure of target peptides

Sl.No.	Peptide sequence	MW	Structure
1	NH_2-G-**K**-**K**-G-$CONH_2$	387.48	
2	NH_2-G-**K**-A-**K**-G-$CONH_2$	458.56	
3	NH_2-G-**K**-A-A-**K**-G-$CONH_2$	529.63	
4	NH_2-G-**K**-A-A-A-**K**-G-$CONH_2$	600.71	
5	NH_2-G-**K**-A-G-G-$CONH_2$	387.43	
6	NH_2-G-G-A-G-G-$CONH_2$	316.31	
7	Ac-G-G-A-G-G-COOH	359.34	
8	Ac-G-G-A-G-G-$CONH_2$	358.35	
9	NH_2-G-G-A-G-G-COOH	317.30	

Peptides 1-4 were synthesized such that each peptide had two Lys residues while the distance between them in the sequence kept on increasing. For example peptide 1 has two consecutive Lys residues and peptide 4 contained two Lys residues separated by three Ala residues. Peptide 5 contained only single Lys residue while peptide 6 was devoid of any Lys residue. All these peptides however contained an amidated C-terminal. Since earlier studies predicted the plausible role of ε-NH_2 moiety of Lys in the unusual spectral features, we synthesized few peptides (7-9) which did not contain any Lys residue and also had modified C- and N-terminal. The C-terminal was carboxylated for peptides 7 and 9, while the N-terminal was acetylated for peptides 7 and 8. Each step for peptide synthesis is described in detail below.

Unless stated specifically all reactions were carried out at room temperature.

STEP I: Swelling of Resin

a) 100 mg of Rink amide resin (loading 1.1 mmol/g) was soaked in 2 mL of Dichloromethane (DCM) for swelling.
b) DCM was then replaced with dimethylformamide (DMF) and the resin was further allowed to swell for another 1 hour.

STEP II: Deprotection of Fmoc group from the resin

a) 1.5 mL of 20% piperidine in DMF was added to the resin and the resin was washed 3 times for 7 minutes each.
b) After the deprotection was complete (about 21 minutes), the reaction column was drained and the resin was washed with (5 x 1.5 mL) portions of DMF for 1 minute each to remove piperidine.

STEP III: Amino Acid Coupling

a) In a small vial, 3 equivalents (98 mg) of Fmoc-AA (AA: Amino acid of interest) was pre-activated by combining it with 3.5 equivalents (170 mg) of BOP, 6 equivalents (85 mg) of DIPEA and 3 mL of DMF.
b) The contents were fully dissolved and then added to the activated resin.
c) The coupling was allowed to occur for 3 hours at room temperature.
d) Since the coupling of the first amino acid is often difficult, the above steps were repeated to ensure proper coupling of the first amino acid.

STEP IV: Washing

a) The resin containing the peptide was washed thoroughly with (5 x 1.5 mL) portions of DMF for 1 minute each to remove unbound amino acid.

STEP V: Kaiser Test

The Kaiser test[125] is a qualitative test performed monitor completeness of amino acid coupling in Solid Phase Peptide Synthesis. The test is based on the reaction of ninhydrin with primary amines, which gives a characteristic dark blue color[126]. The test is used to monitor the presence of free amine after deprotection (dark blue color) and the completeness of the amino acid coupling step (yellow color).

a) Few resin beads were taken in a fusion tube.
b) Kaiser A solution (5% Ninhydrin in ethanol) and Kaiser B solution (Pinch of KCN in Pyridine: Ethanol solution, 80:20) was added in the tube.
c) The contents were heated for 5 minutes up to 80-85 ˚C on a sand bath.
d) The Kaiser test was negative as there was no blue color rather yellowish color was observed.

STEP VI: Capping

Apparently there was no free amine present in the resin, still capping of the resin was performed to make sure there were no free amine was present. Before performing this step the resin was washed with DCM (3 x 1.5 mL) for 1 minute each. Then,

a) Three fold molar excess of each; acetic anhydride and N-methyl imidazole were dissolved in DCM and added to the resin.
b) The reaction was kept for 40 minutes at room temperature.
c) The solution was filtered and the peptide resin was washed alternately with DCM followed by DMF.

STEP VII: Deprotection of amino acid (Fmoc cleavage from the amino acid)

a) 1.5 mL of 20% piperidine in DMF was added to the resin and the resin was washed 3 times for 7 minute each.
b) After the deprotection was complete (about 21 minutes), the reaction column was drained and the resin was washed with (5 x 1.5 mL) portions of DMF for 1 minute each to remove piperidine

Steps III through VII were repeated until the desired peptide sequence was synthesized on the resin. The last step was to cleave the peptide from the resin and collect it.

STEP VIII: Final peptide cleavage from the resin

a) After final washing with DMF, the resin was washed with (6 x 1.5 mL) portions of DCM for 1 minute each.
b) 2 mL of cleavage cocktail (TFA: DCM; 8.5: 1.5) was added to resin and the final cleavage was allowed to occur for 3 hours.

STEP IX: Peptide Precipitation

a) The contents were transferred to a tube containing 10 mL of chilled diethyl ether.
b) The contents were then centrifuged at 4000 rpm for about 10 minutes.
c) Supernatant was discarded and the pellet was stored at -20 ˚C

STEP X: Peptide purification and characterization

The synthesized crude peptides were dissolved in water/acetonitrile (2:1) and purified by Waters 600E RP-HPLC with a flow rate of 4 mL/ minute. Binary solvent system was used: solvent A (0.1 % TFA in water) and solvent B (0.1 % TFA in acetonitrile). C18-µ Bondapak column and a Waters 2489 UV detector were used with the detection at 214 nm. A total run time of 20 minutes and linear gradient was employed, as mentioned in the HPLC profiles. The purified peptides were characterized by mass spectrometry. A very small amount each peptide was dissolved in water/acetonitrile (2:1) solvent and then filtered with 0.2 µm filter. The mass of each peptide was then recorded in a Mass Spectrometer. (Make: Agilent, Q-TOF 6500) in ESI positive mode. All the peptides were later lyophilized and stored at -20 ˚C for further studies.

2.2.2.2 Estimation of peptide concentration

Since the synthesized peptides did not contain any aromatic amino acid, determination of their concentration was not possible by using the conventional method which uses the extinction coefficient of aromatic amino acids for concentration calculations. We therefore employed a dipeptide (Gly-Gly hydrochloride) to calculate the extinction coefficient of a peptide bond. Absorbance at 220 nm was recorded for different concentrations (1-5 mM) of Gly-Gly hydrochloride. The slope of the absorbance at 220 nm vs. concentration was used as the extinction coefficient of a single peptide bond (Figure 2.2.2.2). This value was then later used for determination of extinction coefficients of each peptide (Table 2.2.2.2) based on the number of peptide bonds in each peptide. This was further used to calculate peptide concentration using equation 2.1

$$C = \frac{A_{220}}{\varepsilon . l} \qquad (2.1)$$

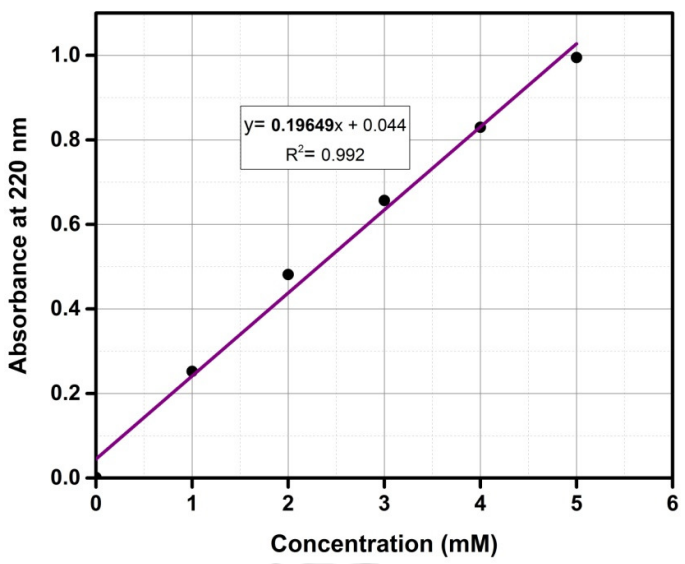


Figure 2.2.2.2: Standard plot (Absorbance at 220 nm vs. Concentration) for Gly-Gly hydrochloride in deionised water

Table 2.2.2.2: Calculated molar extinction coefficients (at 220 nm) for all peptides

Sl.No.	Number of peptide bonds	Calculated extinction coefficient ($M^{-1}cm^{-1}$) at 220 nm using Gly-Gly standard
Peptide 1	3	589.5
Peptide 2	4	786.0
Peptide 3	5	982.4
Peptide 4	6	1178.9
Peptide 5	4	786.0
Peptide 6	4	786.0
Peptide 7	4	786.0
Peptide 8	4	786.0
Peptide 9	4	786.0

2.2.2.3 Absorption measurements

The absorption spectra for different small aliphatic compounds and peptides were recorded at room temperature using a double beam Lamda-25 UV-VIS Spectrophotometer (Make: Perkin Elmer). The scans were recorded from 250-500 nm with fixed 1 nm bandwidth and a scan speed of 420 nm/minute. Quartz Cuvette with 1 cm path length (Make: Opti glass-UK) with transmission range up to 200 nm was used for recording all the measurements. Water was kept as blank for samples dissolved in water while 0.1 N HCl served as blank for samples in 0.1 N HCl. A concentration of 4 mM was used for all absorption measurements of the peptides. Different concentrations for compounds were used for recording the absorption spectra of the compounds. Absorption spectrum of Lys.HCl (1M) was also recorded in order to compare its absorption with the peptides and compounds. For the determination of concentration of the peptides, a scan between 200-300 nm was recorded for Gly-Gly hydrochloride and the peptides.

2.3 Results and discussion:

2.3.1 Characterization of peptides

Peptide 1: NH_2-G-**K**-**K**-G-$CONH_2$

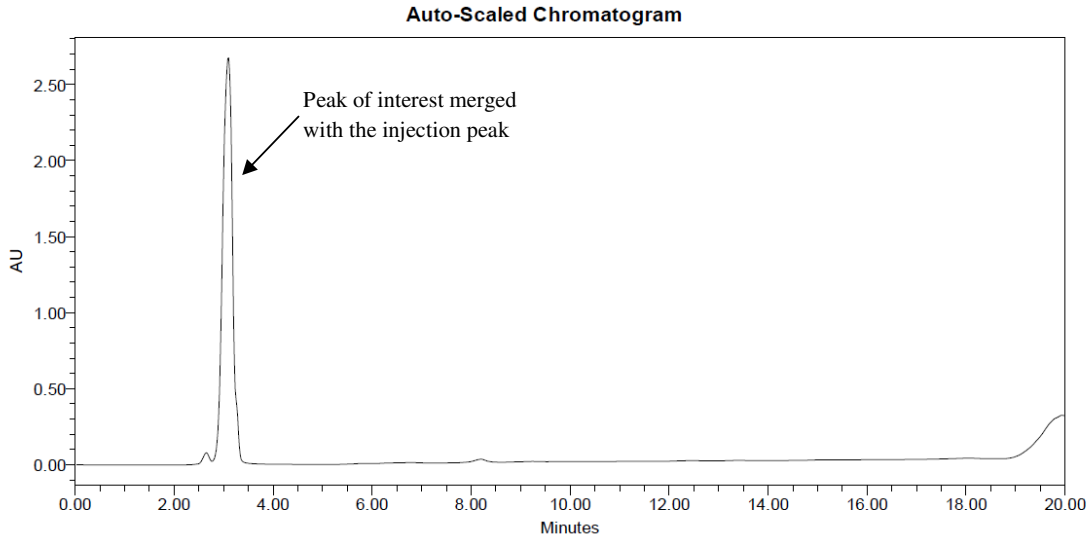


Figure 2.3.1A: HPLC profile of Peptide 1.Gradient: 0-15 min 0-10 % CH_3CN, 15-18 min 10-100% CH_3CN and 18-20 min 100% CH_3CN. Retention time: 3.1 minutes

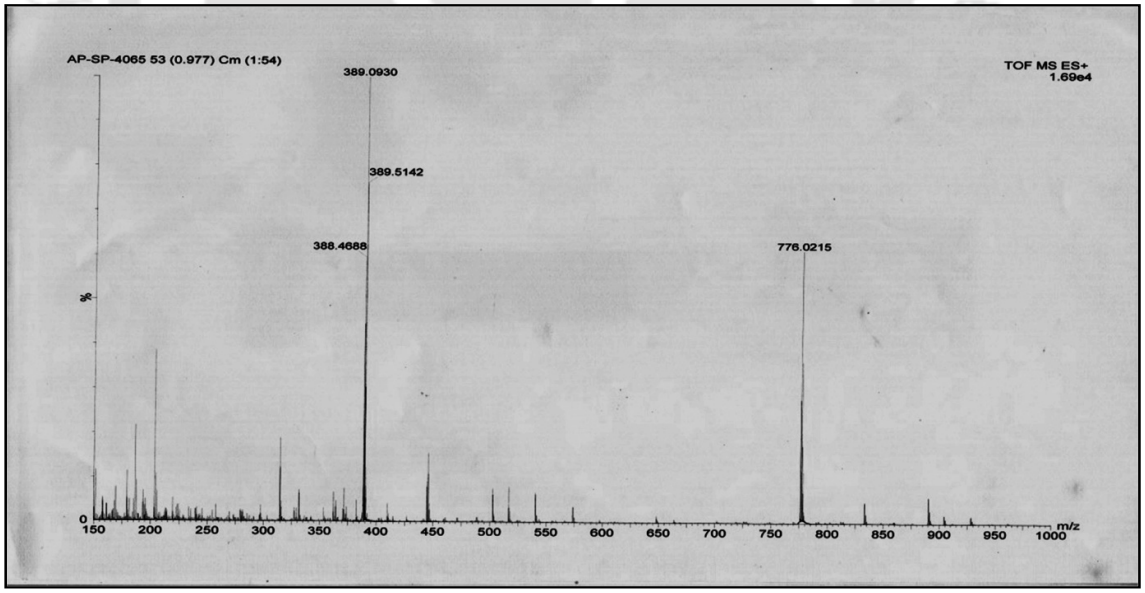


Figure 2.3.1B: Mass spectrum (ESI-MS) of Peptide 1. Calculated mass for $C_{16}H_{34}N_7O_4$ is 388.48 Da $[M+H]^+$, observed 389.09 Da $[M+H]^+$

Peptide 2: NH_2-G-**K**-A-**K**-G-$CONH_2$

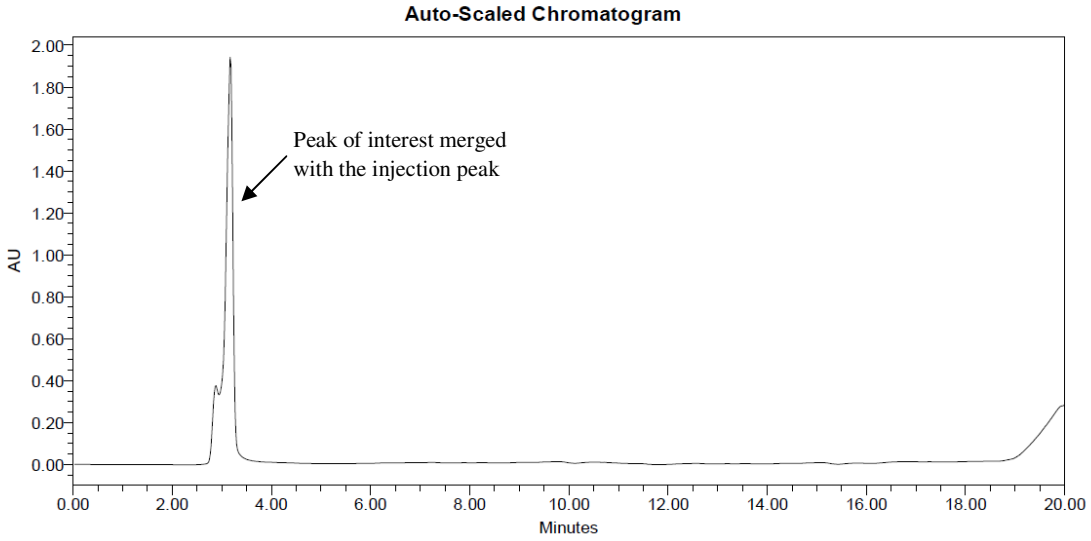


Figure 2.3.1C: HPLC profile of Peptide 2.Gradient: 0-15 min 0-10 % CH_3CN, 15-18 min 10-100% CH_3CN and 18-20 min 100% CH_3CN. Retention time: 3.1 minutes

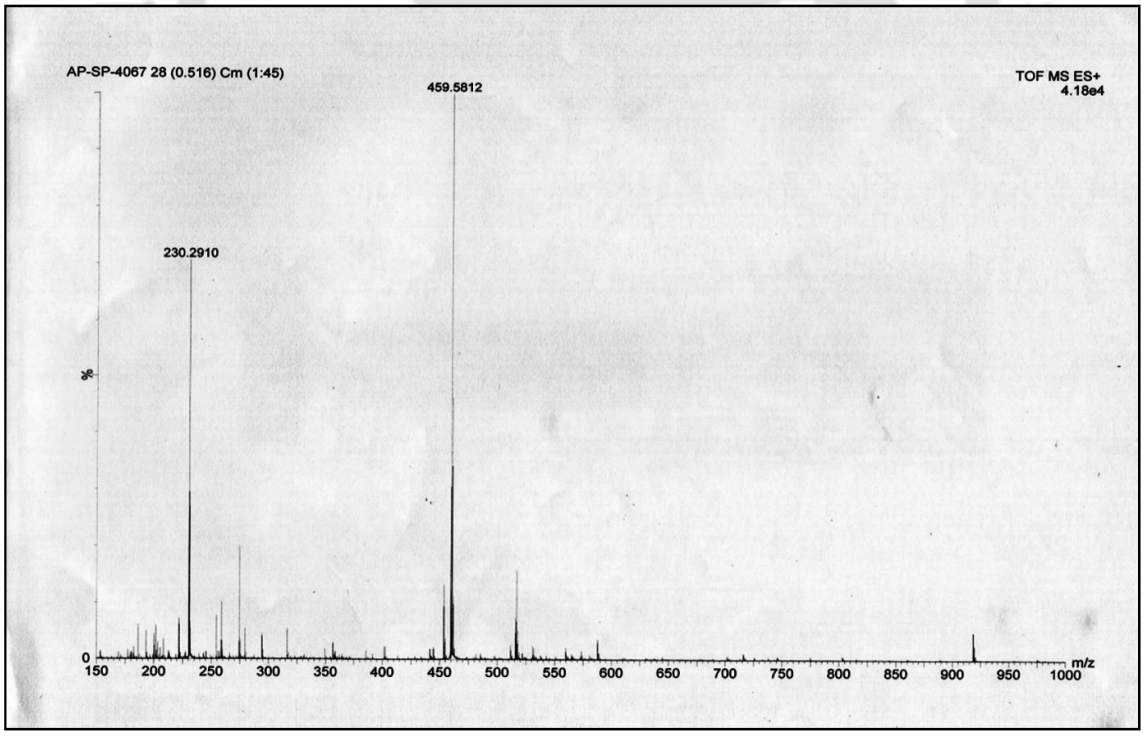


Figure 2.3.1D: Mass spectrum (ESI-MS) of Peptide 2. Calculated mass for $C_{19}H_{39}N_8O_5$ is 459.56 Da $[M+H]^+$, observed 459.58 Da $[M+H]^+$

Peptide 3: NH_2-G-**K**-A-A-**K**-G-$CONH_2$

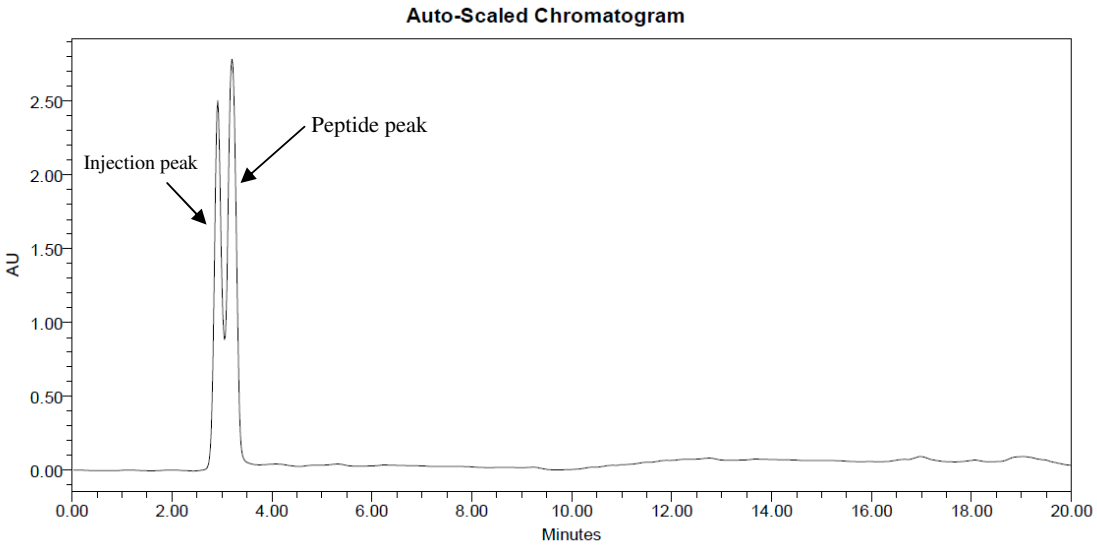


Figure 2.3.1E: HPLC profile of Peptide 3. Gradient: 0-15 min 0-10 % CH_3CN, 15-18 min 10-100% CH_3CN and 18-20 min 100% CH_3CN. Retention time: 3.3 minutes

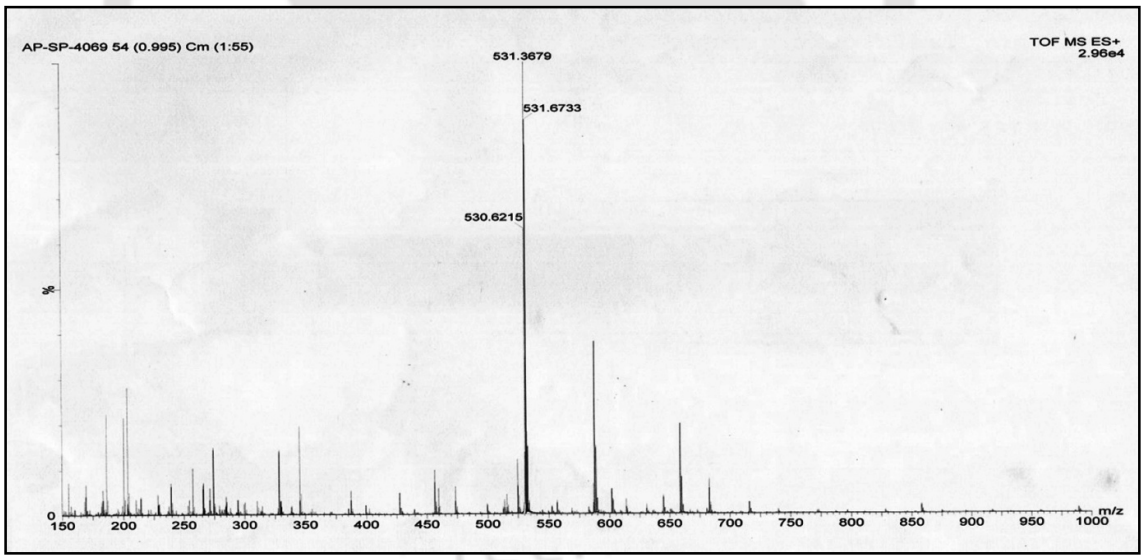


Figure 2.3.1F: Mass spectrum (ESI-MS) of Peptide 3. Calculated mass for $C_{22}H_{44}N_9O_6$ is 530.63 Da $[M+H]^+$, observed 531.36 Da $[M+H]^+$

Peptide 4: NH_2-G-**K**-A-A-A-**K**-G-$CONH_2$

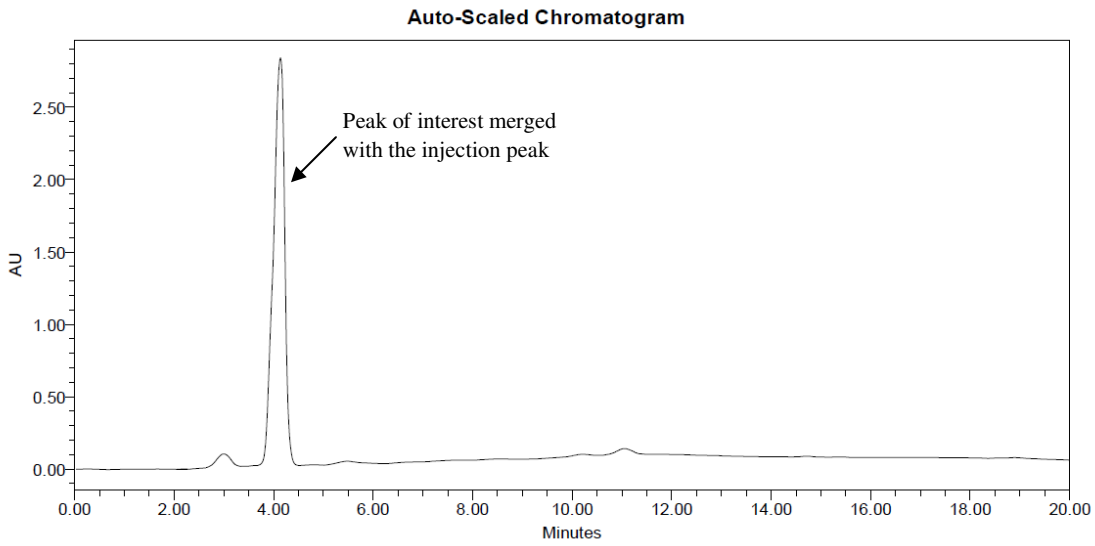


Figure 2.3.1G: HPLC profile of Peptide 4. Gradient: 0-15 min 0-10 % CH_3CN, 15-18min 10-100% CH_3CN and 18-20 min 100% CH_3CN. Retention time: 4.0 minutes

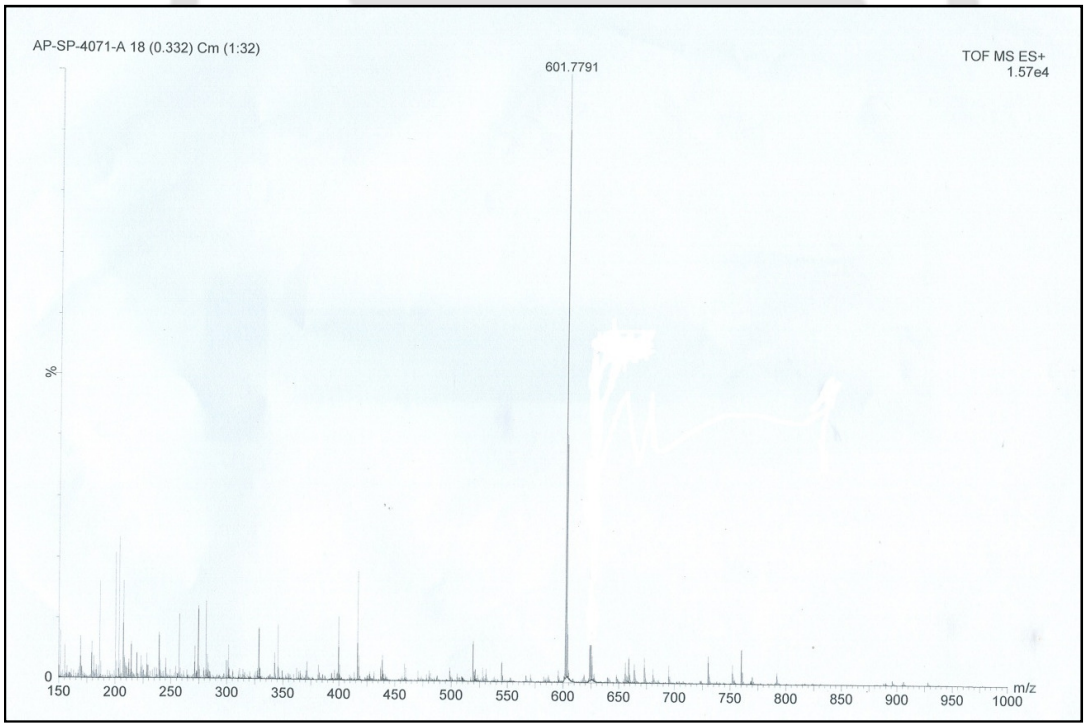


Figure 2.3.1H: Mass spectrum (ESI-MS) of Peptide 4. Calculated mass for $C_{25}H_{49}N_{10}O_7$ is 601.71 Da $[M+H]^+$, observed 601.77 Da $[M+H]^+$

Peptide 5: NH_2-G-**K**-A-G-G-$CONH_2$

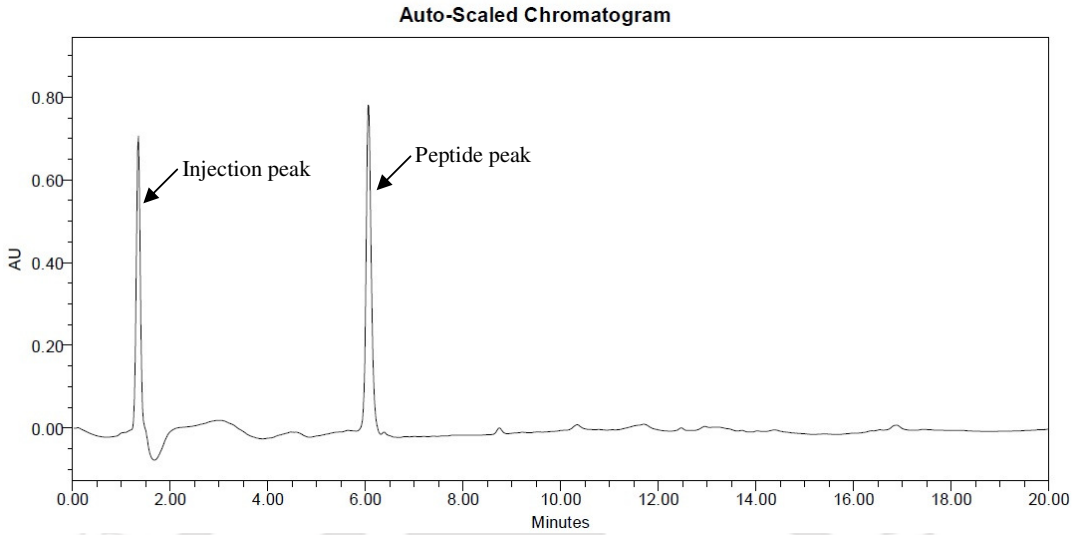


Figure 2.3.1I: HPLC profile of Peptide 5. Gradient: 0-15 min 0-10 % CH_3CN, 15-18min 10-100% CH_3CN and 18-20 min 100% CH_3CN. Retention time: 6.0 minutes

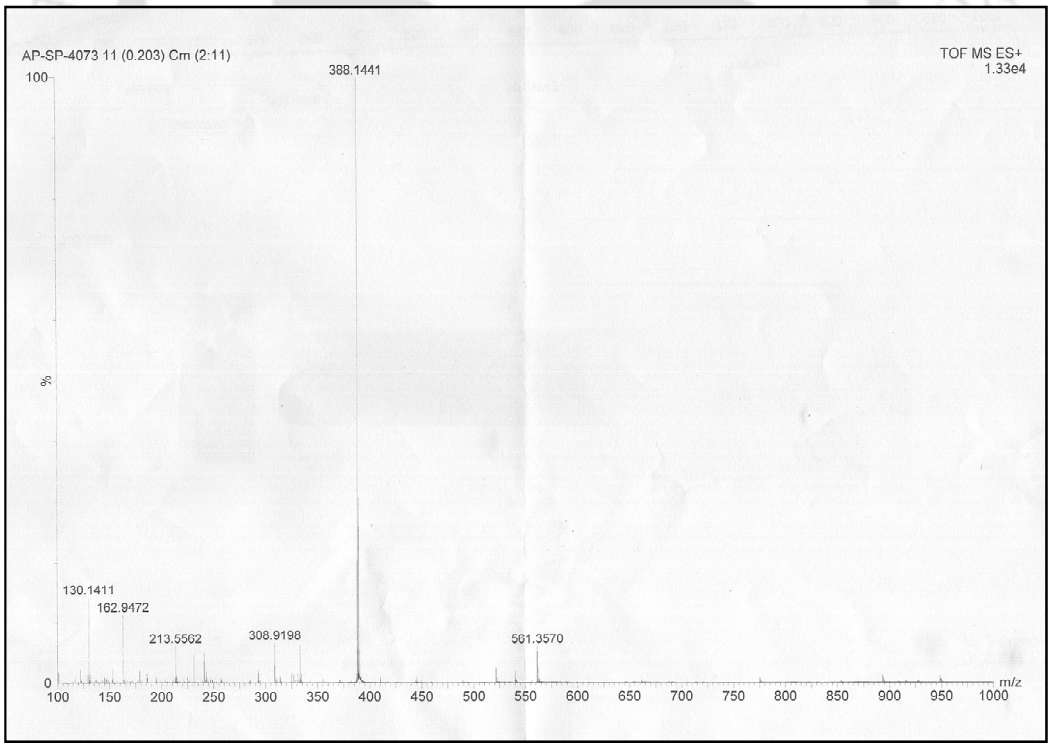


Figure 2.3.1J: Mass spectrum (ESI-MS) of Peptide 5. Calculated mass for $C_{15}H_{29}N_7O_5$ is 388.43 Da $[M+H]^+$, observed 388.14 Da $[M+H]^+$

Peptide 6: NH_2-G-G-A-G-G-$CONH_2$

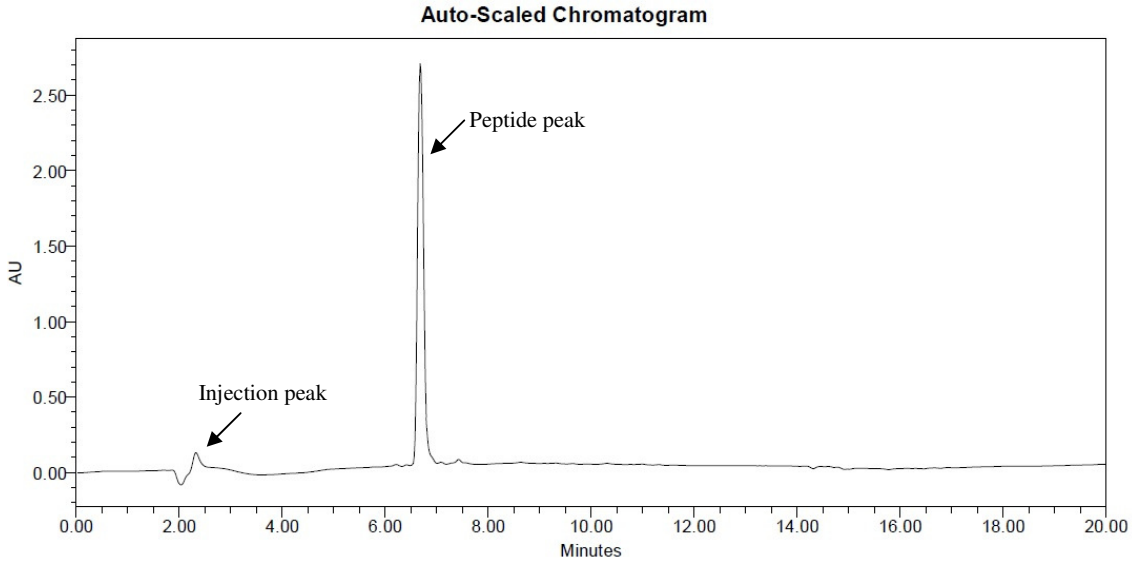


Figure 2.3.1K: HPLC profile of Peptide 6. Gradient: 0-15 min 0-10 % CH_3CN, 15-18 min 10-100% CH_3CN and 18-20 min 100% CH_3CN. Retention time: 6.5 minutes

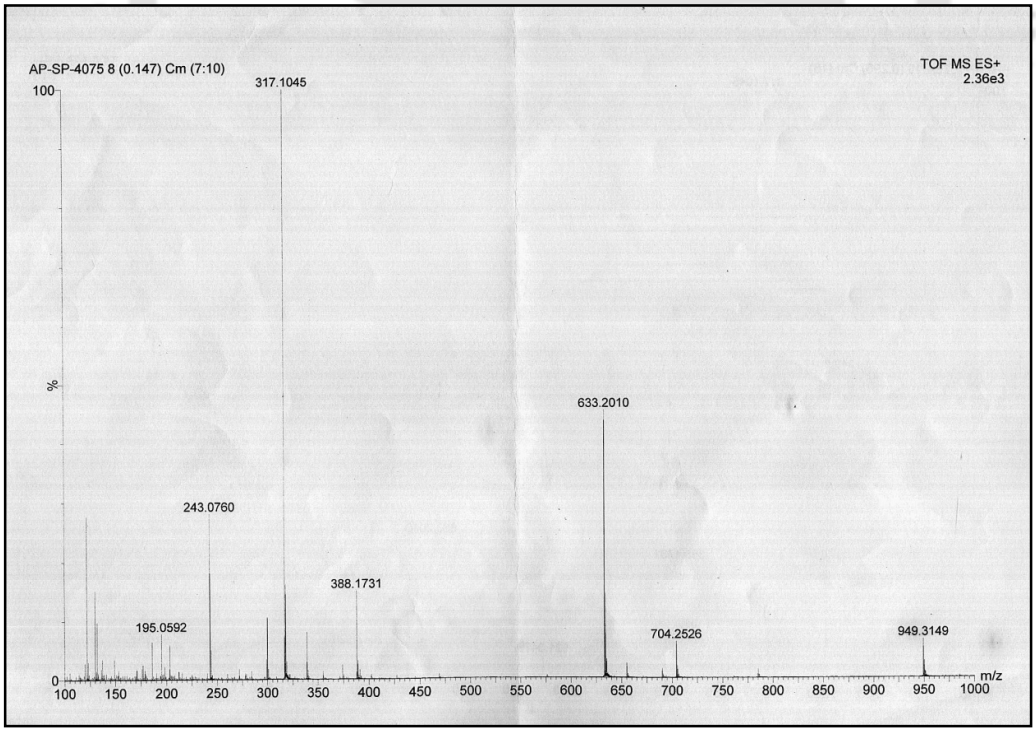


Figure 2.3.1L: Mass spectrum (ESI-MS) of Peptide 6. Calculated mass for $C_{11}H_{21}N_6O_5$ is 317.31 Da $[M+H]^+$, observed 317.10 Da $[M+H]^+$

Peptide 7: Ac-G-G-A-G-G-COOH

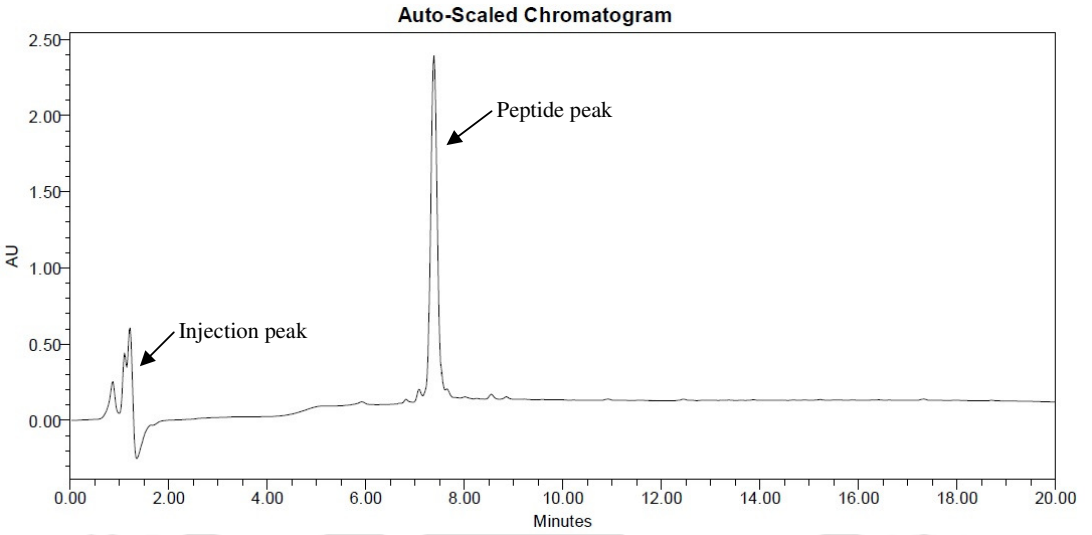


Figure 2.3.1M: HPLC profile of Peptide 7. Gradient: 0-15 min 0-10 % CH_3CN, 15-18 min 10-100% CH_3CN and 18-20 min 100% CH_3CN. Retention time: 7.6 minutes

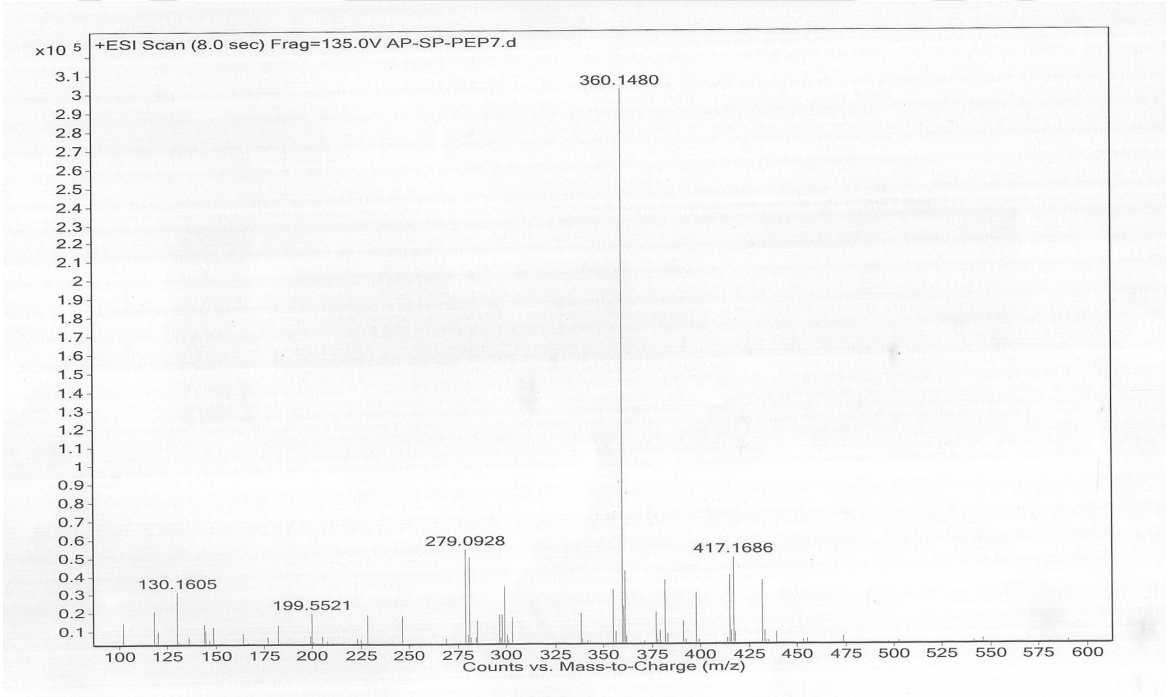


Figure 2.3.1N: Mass spectrum (ESI-MS) of Peptide 7. Calculated mass for $C_{13}H_{22}N_5O_7$ is 360.34 Da $[M+H]^+$, observed 360.14 Da $[M+H]^+$

Peptide 8:

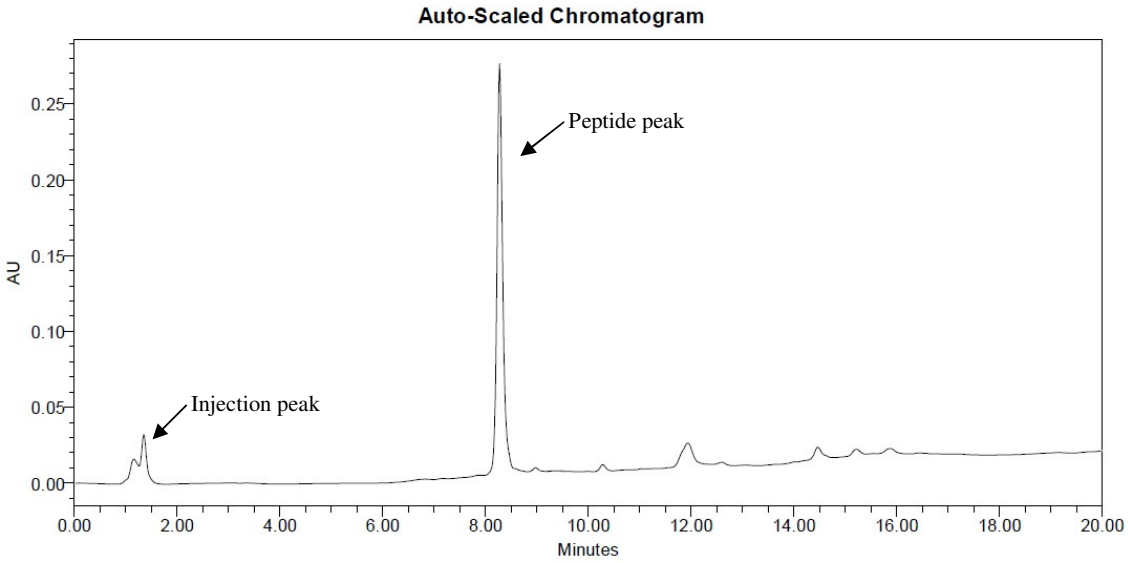


Figure 2.3.1O: HPLC profile of Peptide 8. Gradient: 0-15 min 0-10 % CH_3CN, 15-18 min 10-100% CH_3CN and 18-20 min 100% CH_3CN. Retention time: 8.2 minutes

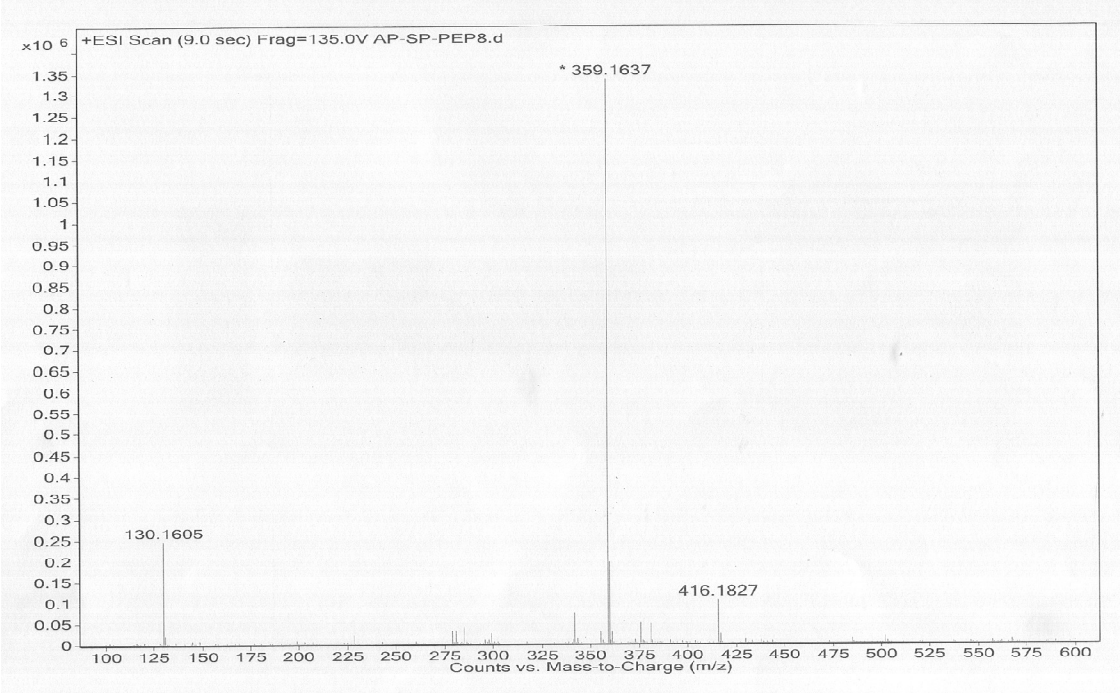


Figure 2.3.1P: Mass spectrum (ESI-MS) of Peptide 8. Calculated mass for $C_{13}H_{23}N_6O_6$ is 359.35 Da $[M+H]^+$, observed 359.16 Da $[M+H]^+$

Peptide 9: NH_2-G-G-A-G-G-COOH

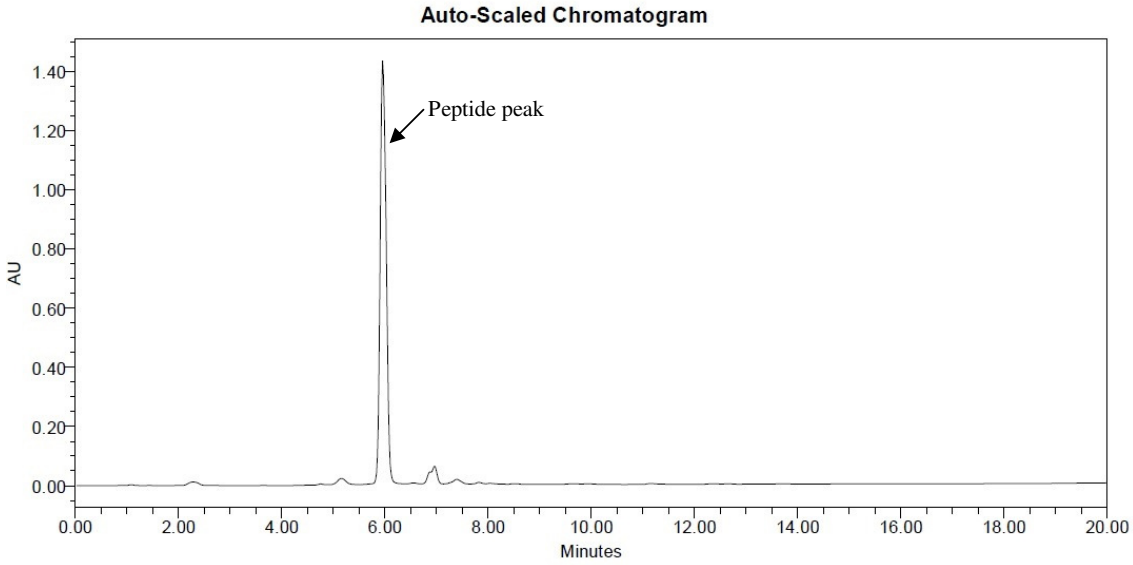


Figure 2.3.1Q: HPLC profile of Peptide 9. Gradient: 0-15 min 0-10 % CH_3CN, 15-18 min 10-100% CH_3CN and 18-20 min 100% CH_3CN. Retention time: 6.0 minutes

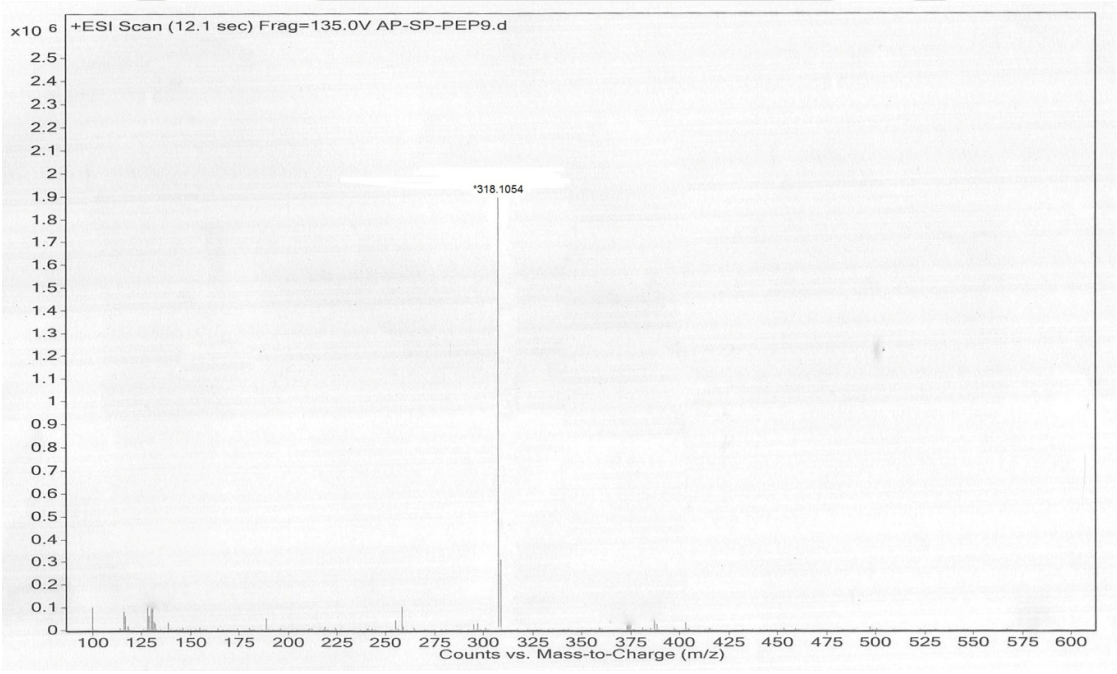
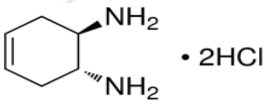
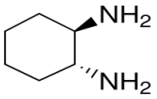
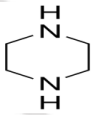
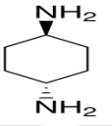
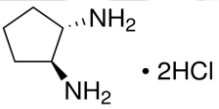
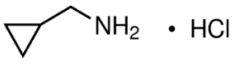
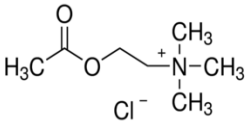


Figure 2.3.1R: Mass spectrum (ESI-MS) of Peptide 9. Calculated mass for $C_{11}H_{18}N_5O_6$ is 318.30 Da $[M+H]^+$, observed 318.1054 Da $[M+H]^+$

2.3.2 Absorption spectra of aliphatic compounds

To assess the role of primary amine, compounds containing either two or one primary amine (NH_2) moieties were chosen. Compounds with secondary amine (NH-) and tertiary amine (N-) were also studied. Following aliphatic compounds containing different forms of amine group were studied.

Table 2.3.2: Name and structure of the aliphatic compounds studied

Sl.No.	Name	Structure
1	trans-4-Cyclohexene-1,2-diamine dihydrochloride	
2	(1R,2R)-(−)-1,2-Diaminocyclohexane	
3	Piperazine	
4	trans-1,4-Diaminocyclohexane	
5	(1S,2S)-trans-1,2-Cyclopentanediamine dihydrochloride	
6	Cyclopropanemethylamine hydrochloride	
7	Acetylcholine chloride	

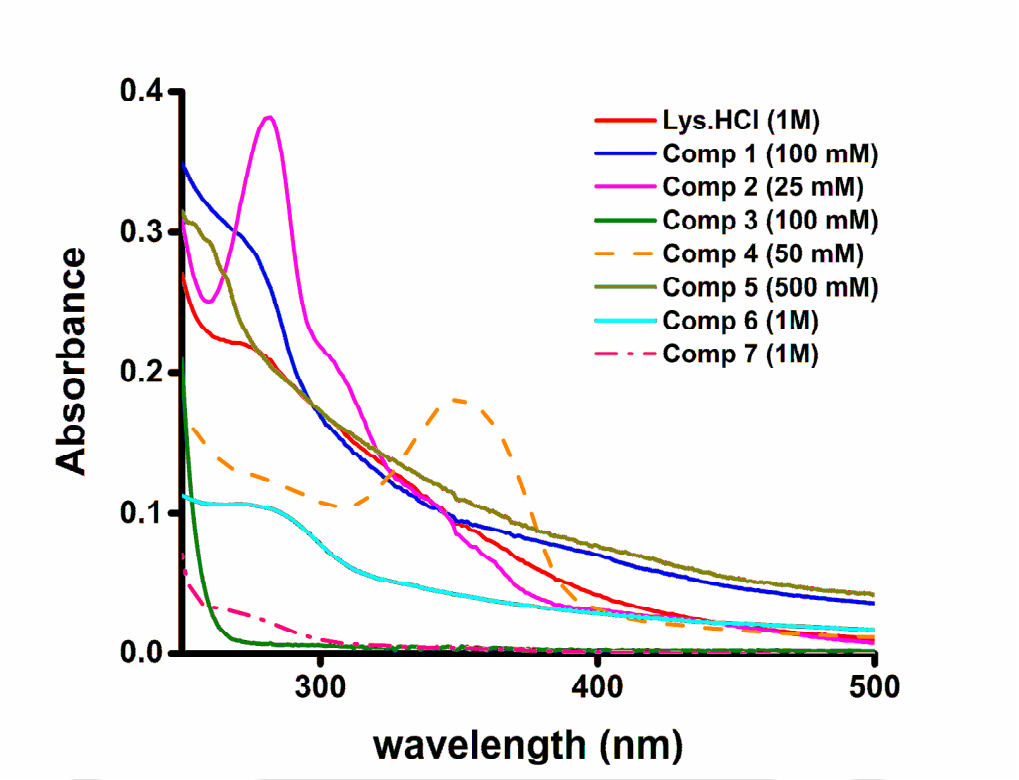


Figure 2.3.2A: Absorption spectra of small aliphatic compounds in deionised water

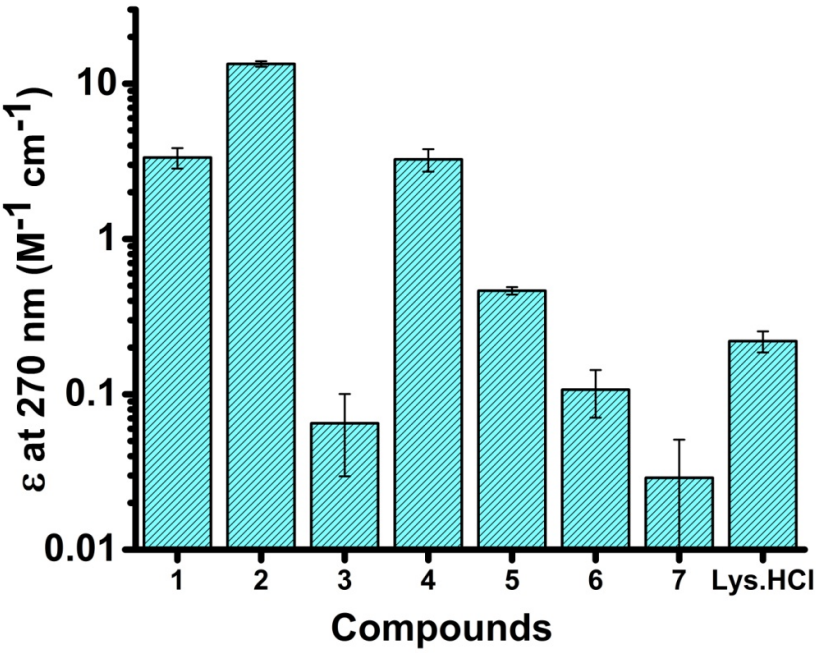


Figure 2.3.2B: Comparison of molar extinction coefficients at 270 nm for all the compounds

The absorption spectra from 250-500 nm for all the compounds are shown in Figure 2.3.2A. The absorption spectrum of Lys.HCl was also recorded for comparison. The absorption spectra for compounds were recorded at different concentrations as each of them had different molar absorptivities. The first thing to note is that most of the compounds display absorption in the UV-VIS region despite lacking any aromatic moiety. Compound 1 and compound 5 which contain two primary amine moieties have similar absorption profile as Lys.HCl. Compound 2 and 4 however showed quite different absorption profiles. Compound 2 had a sharp peak at 270 nm while compound 4 displayed an unusual peak at around 350 nm which was not observed for any other samples. Compound 6 which has a single primary amine also shows similar spectral features like Lys.HCl, but has almost half the intensity when compared to Lys.HCl. The compounds (1, 5 and 6) also display a long tail beyond 300 nm. An important aspect to be noted is that the compounds which lack primary amine (compounds 3 and 7) display negligible absorption features beyond 300 nm unlike the rest.

Figure 2.3.2B depicts a comparison of molar extinction coefficients at 270 nm for all the compounds. It gives a much clear picture as the effects of different concentrations used for recording the absorption spectra have been taken care of. It shows that compounds with two primary amines (1, 2, 4 and 5) have significantly higher molar absorptivities than compounds with either single primary amine (6), secondary amine (3) or tertiary amine (7). Also the compounds with two primary amine moieties (1, 2, 4 and 5) showed intensities which were almost 15 times more than Lys.HCl at 270 nm. Compound 6 which has a single primary amine moiety has absorption intensity which is almost half to that of Lys.HCl (Figure 2.3.2B). Another clear feature to be noted is that the compounds which lack any primary amine (Compounds 3 and 7) have absolutely no absorption at 270 nm. Interestingly, comparison of isomeric compound 2 and 4 shows that pair of proximal primary amines (in ortho position in compound 2) compared to distinct primary amines (in para position in compound 4) absorbs three fold better.

Overall this study has revealed that the NH_2 moiety is needed in order to observe any significant absorption in the near UV region (270 nm). Also, the absorption intensity at 270 nm drops significantly when the number of primary amine moieties in a compound decreases from two to one or when a pair of primary amines in the same molecule are placed far apart. Taken together, these results hint towards the importance of intramolecular interactions among the NH_2 moieties.

2.3.3 Absorption spectra of peptides

Peptide	Sequence
1	NH_2-G-**K**-**K**-G-$CONH_2$
2	NH_2-G-**K**-A-**K**-G-$CONH_2$
3	NH_2-G-**K**-A-A-**K**-G-$CONH_2$
4	NH_2-G-**K**-A-A-A-**K**-G-$CONH_2$
5	NH_2-G-**K**-A-G-G-$CONH_2$
6	*NH_2-G-G-A-G-G-$CONH_2$*
7	*Ac-G-G-A-G-G-COOH*
8	*Ac-G-G-A-G-G-$CONH_2$*
9	*NH_2-G-G-A-G-G-COOH*

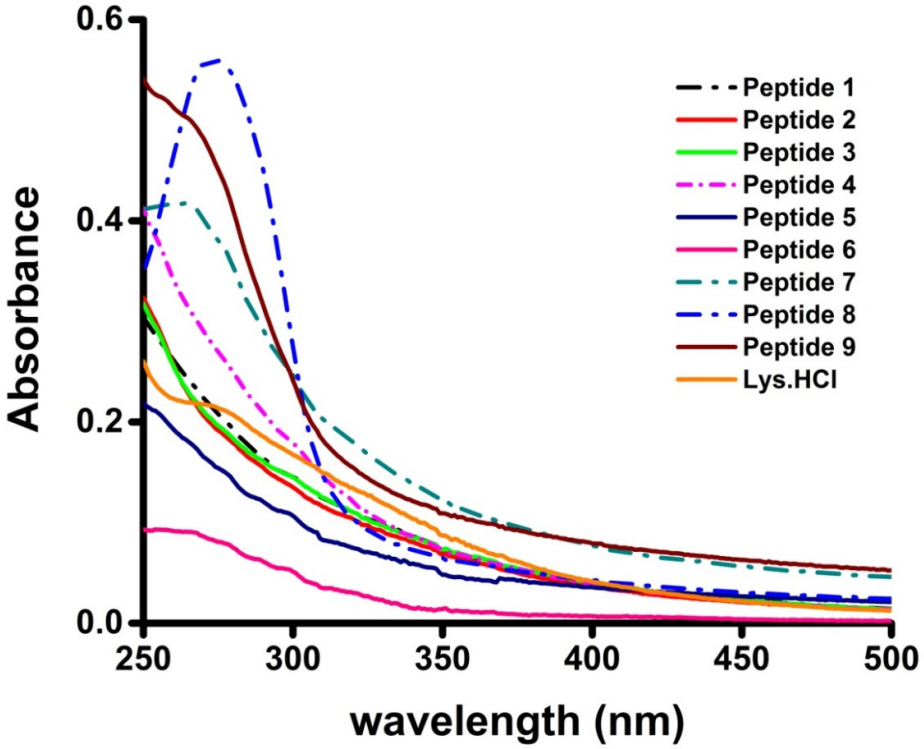


Figure 2.3.3A: Absorption spectra of peptides (4 mM) and Lys.HCl (1 M) in deionised water. The table above shows the sequence of peptides. Sequences shown in *italics* do not contain Lys residues

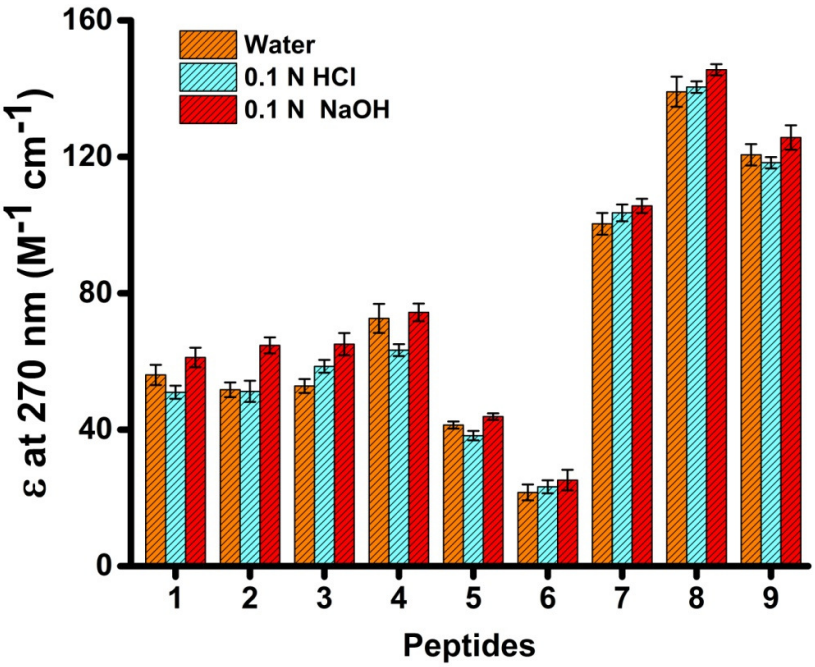


Figure 2.3.3B: Comparison of absorption spectra at 270 nm of all the peptides in various conditions

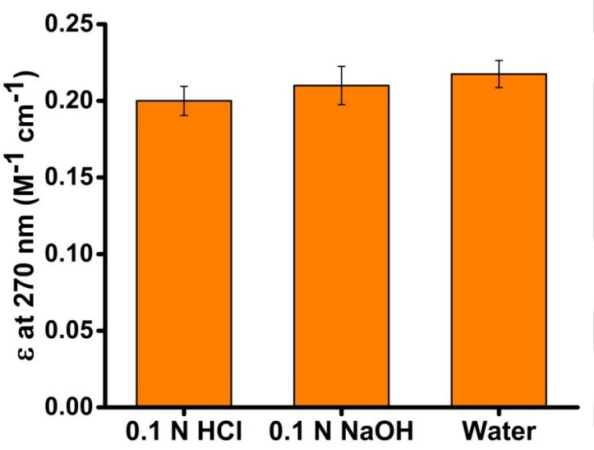


Figure 2.3.3C: Comparison of absorption spectra at 270 nm for Lys.HCl in various conditions

Figure 2.3.3A depicts the absorption spectra (250–500 nm) of all the peptides in deionised water. Absorption spectrum of Lys.HCl was recorded for comparison. All the peptides which contained two Lys residues (Peptide 1-4) display absorption features similar to Lys.HCl. These peptides also display a long tail up to 500 nm as shown by Lys.HCl as well. Based on previous work, we expected the absorption to drop as the distance between Lys residues increased. However, if we compare the peptides which have two Lys residues, peptide 4 which had the maximum distance between the Lys

residues showed a slightly higher absorption when compared to peptides 1-3. Probably peptide 4 might be forming a helical structure thereby bringing the two Lys residues close enough to interact, thus giving rise to higher intensities of peptide 4. The slight decrease in absorption intensities of peptides 1-3 might be due to steric hindrance between the bulky side chains of Lys residues which might be preventing a close interaction among the ε-NH_2 of Lys residues.

Peptide 5 which had one Lys residue had absorption intensities which were much lower (reduced almost by half) than all the peptides which contained two Lys residues. Peptide 6 which was devoid of any Lys residue also showed significant absorption although the intensity was much lower than peptide 5. This was believed to arise from amidated C-terminal and N-terminal of the peptide. To understand this better we studied three peptides (Peptide 7-9) which did not contain any Lys residue and had modified C- and N-terminal. However, to our surprise all the three peptides displayed much greater absorption intensities when compared with any other peptide. Also each of them had a much greater tail extending all the way up to 500 nm. Among peptides 7, 8 and 9, peptide 8 showed the greatest intensity. However, we were not in a position to explain this alluring result at that point of time. The effect of acetylation and carboxylation of N- and C-terminals respectively will be explained in subsequent chapter 5.

Comparison of absorption intensities (Figure 2.3.3B) at 270 nm depicts that all the peptides have molar absorptivities which are much greater (~100 fold higher) than Lys.HCl. This is in agreement with previous studies where poly L-Lys was shown to have 250 fold greater absorption intensities than Lys.HCl[111]. Our study thereby also hints towards the role of intramolecular interaction among the Lys residues behind the unusual spectral features. To understand the effect of protonation/deprotonation of the -NH_2 and –COOH group, on the absorption spectra, we recorded the absorption spectra of all the peptides in 0.1 N HCl (pH ~2) and 0.1 N NaOH (pH ~12). However, none of the peptides showed any major change in the absorption intensity with change in pH (Figure 2.3.3B). This holds true for Lys.HCl as well (Figure 2.3.3C).

2.4 Conclusions:

a) Investigating small aliphatic compounds revealed that -NH_2 moiety plays an important role in the origin of spectral signatures in the near UV region. Intramolecular interactions among a pair of –NH_2 groups enhanced the spectra.

b) Compared to Lys.HCl, all peptides (irrespective of the number of Lys residues) show ~100 fold more intense molar absorptivities. Probably the peptide backbone has a role in this absorption.

c) The absorption intensities for both peptides and Lys.HCl were insensitive to change in pH.

2.5 Implications of the work:

The results obtained in this chapter clearly highlight the role of -NH_2 moiety behind the unusual spectral signatures as speculated in earlier studies. Studies with short peptides reveal the importance of intra-molecular interactions between two Lys residues within a peptide. We have unambiguously shown that peptides lacking aromatic amino acids posses absorption features in the UV-Visible region with maxima around 270 nm. These observations are in direct conflict with standard textbooks which deal with protein spectroscopy and therefore have significant implications in the field of protein biophysics as -NH_2 moiety is present in all proteins (as N-terminal and in Lys residues). Therefore, these results can be employed towards better understanding of spectroscopic properties of proteins.

CHAPTER 3

STUDIES ON NON-AROMATIC AMINO ACIDS AND THEIR DERIVATIVES

3.1 Introduction:

Unusual absorption ~270 nm from Lys.HCl has been reported by Homchaudhuri and Swaminathan in which they predicted the role of Lys side chain as one of the probable reasons behind the unique spectral features of Lys in the near UV region (chapter 1). They studied Gly as a control which did not show any noticeable absorption features. Lys.HCl was also reported to show blue fluorescence (~435 nm) on excitation with 350 nm. Apart from Lys.HCl and Gly, they also studied emission from few randomly selected amino acids (such as Arg, Ser, Glu and Ile) as well. However, excluding Lys.HCl no other amino acid showed any significant luminescence[107]. In order to understand if any other amino acid has similar unusual absorption features like Lys.HCl, we decided to carry out systematic absorption studies on all non-aromatic amino acids in the near UV and visible region. The objective was identify other amino acids (if possible) which show similar unique absorption features as seen for Lys.HCl. Other factors such as pH and ionic strength which might affect the spectral features were also studied.

An extensive study in the far UV region (200-230 nm) on all the 20 amino acids has been reported by Saidel et al[127]. However, their study is restricted to the far UV region and there are no reports on absorption spectra of non-aromatic amino acids above 250 nm.

We investigated the absorption spectra above 250 nm for each non-aromatic amino acid in detail. Along with the non-aromatic amino acids, we also studied some derivatives of these amino acids in order to see if they show these absorption features as well. Factors such as pH, presence of D_2O instead of H_2O which could play a role in perturbing the spectral features were also studied.

3.2 Materials and methods:

3.2.1 Materials:

The L-form of all the non-aromatic amino acids *viz.* Alanine (A7469), Arginine (A5006), Asparagine (A4159), Cysteine hydrochloride (C1276), Glutamic acid monosodium salt (G1626), Aspartic acid potassium salt (A6558), Glycine (G7126), Histidine (53319) , Isoleucine (I7403), Leucine (L8912), Lysine (5501) , Lysine monohydorchloride (L5626), Methionine (M5308), Proline (P0380), Serine (S4311), Threonine (T8441), Valine (VO513) and amino acid derivatives *viz.* L-Ornithine Monohydrochloride (O2375), L-Glutamic acid dimethyl ester hydrochloride (49560), N-Acetyl-L-proline (A0783), N-Acetyl-L-alanine (A4625), Nε-Acetyl-L-lysine (A4021), Nα-Acetyl-L-ornithine (A3626), L-Glutamic acid amide (G3521) were purchased from Sigma Aldrich. Poly-L-Lys.HBr (P0879, MW: 1000-5000 Da), 2-(N-morpholino)ethanesulfonic acid (MES) hydrate (M8250) and D_2O (364312) were also purchased from Sigma Aldrich. KCl (104936), tris(hydroxymethyl)aminomethane (Tris buffer) (17714), and Sodium dihydrogen phosphate (17845) were procured from Merck.

3.2.2 Methods:

3.2.2.1 Absorption measurements

The absorption spectra for different non-aromatic amino acids and various amino acid derivatives were recorded at room temperature from 250-800 nm according to the procedure described in chapter 2. The amino acids namely Ala, Arg, Glutamic acid monosodium salt (Glu.Na), Aspartic acid potassium salt (Asp.K), Gly, Lys, Lys monohydorchloride (Lys.HCl), Pro, Ser and Cysteine hydrochloride were dissolved in deionised water (18.2 MΩ) while Asn, His, Ile, Leu, Thr and Val were dissolved in 0.1 N HCl as they were insoluble in pure water. N-Acetyl-L-proline was dissolved in in 0.1 N HCl, while the remaining derivatives were dissolved in deionised water. 1 M concentration was used for all the readings unless otherwise stated. Water was kept as blank for samples dissolved in water while 0.1 N HCl served as blank for samples in 0.1 N HCl.

3.2.2.2 pH titration for charged amino acids and poly-L-Lys.HBr

1 M solution for each of the charged amino acid was prepared (20 mL). The solution was continuously stirred using a magnetic stirrer (Make: REMI) and pH was monitored with a pH meter (Make: Sartorius). The pH was changed gradually by adding either 0.1 N HCl or 0.1 N NaOH in small aliquots. The change in the concentration caused by the addition of acid/base was later corrected in the final calculations.

Poly-L-Lys.HBr was dissolved in deionised water (10 mg/mL). The pH was gradually changed by adding 0.1 N NaOH/HCl under constant stirring. The change in pH was monitored by using a pH strip. Absorption spectra were then recorded for each sample as stated earlier.

3.2.2.3 Effect of KCl on Lys absorption

1 M stock solution of KCl was prepared in deionised water. Solutions of lower molarities were prepared from this stock. Lys (weight equivalent to make 1 M solution of Lys) was then dissolved in KCl solution of different molarities. Absorption spectra were then recorded as stated earlier.

3.2.2.4 D_2O exchange

1 M of Lys solution was prepared both in H_2O and D_2O separately. Both the samples were then incubated for 24 hours. The absorption spectra were recorded at zero hour as well as after incubation of the samples at 24 hours. Pure D_2O and H_2O were kept as blank for the respective samples during measurement of absorption.

3.2.2.5 Absorption spectra of Lys.HCl and Glu.Na mixture

1 M stock solution of both Lys.HCl and Glu.Na were prepared separately. They were then mixed in different ratios and absorption spectra for various samples were recorded as stated earlier. Pure deionised water was kept as blank for absorption measurements.

3.3 Results and discussion

3.3.1 Absorption spectra of all non-aromatic amino acids

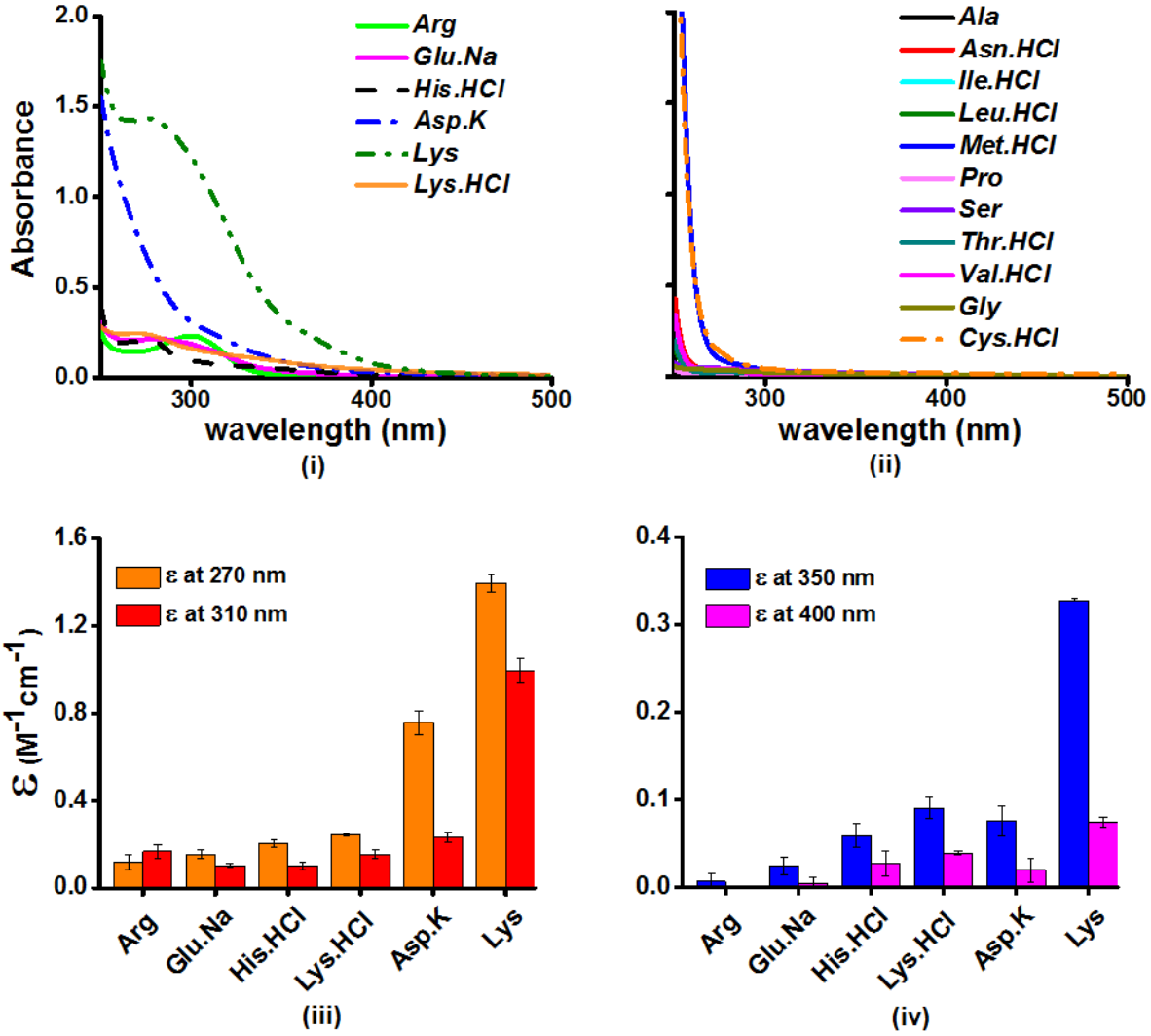


Figure 3.3.1A: Absorption spectra of (i) Charged amino acids; (ii) Uncharged amino acids in deionised water; Comparison of molar extinction coefficients (iii) at 270 nm and 310 nm and (iv) at 350 nm and 400 nm for charged amino acids. The aqueous concentration of amino acids used was 1M

Previous studies by Homchaudhuri and Swaminathan reported unusual absorption spectral features of Lys.HCl (chapter 1). To explore if other non-aromatic amino acids also display unusual absorption features like Lys.HCl in the near UV region absorption studies on all non-aromatic amino acids were carried out. In this context we recorded absorption spectra in high concentration solutions of all the non-aromatic amino acids (Figure 3.3.1A). Among the 17 non-aromatic amino acids, we found significant absorption between 250-400 nm for charged amino acid solutions of Lys, Glu monosodium salt (Glu.Na), Arg, Asp potassium salt (Asp.K) and His (Figure 3.3.1A i). Among all the charged amino acids, Lys showed the maximum absorption intensity (ε~1.2 at 270 nm) with a tail extending all the way to 400 nm. In contrast, uncharged amino acid solutions of Ala, Asn, Ile, Leu, Pro, Ser, Thr, and Val had negligible absorption beyond 270 nm (Figure 3.3.1A ii). Only the sulphur containing amino acids, i.e., Met and Cys showed some absorption below 270 nm. This clearly suggests that charged amino acids are special and have the capability to display novel spectral features in the near UV region which can extend up to the visible portion of the electromagnetic spectrum.

Also to notice was that Lys.HCl had a molar absorptivity ~6 times smaller than that for pure Lys solutions lacking the hydrochloride ion (Figure 3.3.1A iii). The modulation of absorption by ions supports participation of the charged amino acid side chains in the initial photo-excited transitions. The hydrochloride ion may screen the side chain charge to reduce the net absorption of the sample. A similar reasoning suggests that pure Glu (insoluble in aqueous medium) may have a higher molar absorptivity than that measured for its monosodium salt solution. Since Lys.HCl and Glu.Na have very similar absorption intensities (0.23 and 0.20 respectively, at 270 nm), the molar absorptivity of pure Glu may match that of pure Lys.

Since Lys showed maximum absorption among all the charged amino acids, we carried out concentration dependent studies on it. Studies on different concentrations of Lys revealed that the absorption features are concentration dependent which increases with increase in concentration (Figure 3.3.1B).

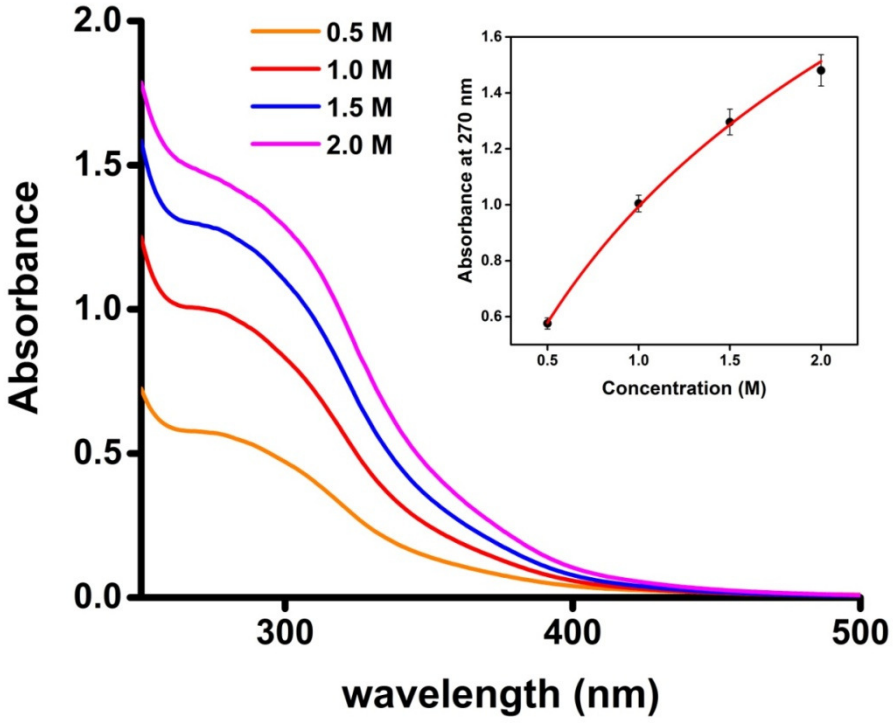
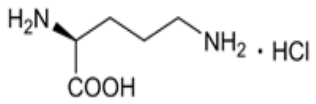
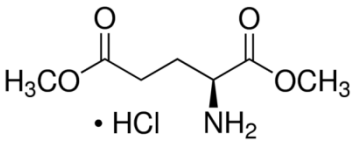
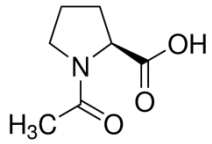
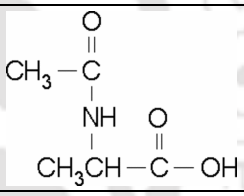
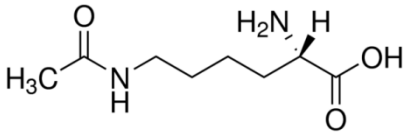
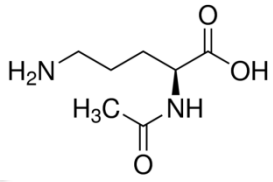
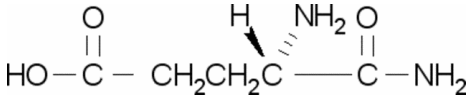


Figure 3.3.1B: Absorption spectra of Lys in deionised water at different concentrations. *Inset* shows increase in absorbance at 270 nm with increasing concentration

The non-linear increase in absorption intensity with increase in concentration suggests that intermolecular interactions between Lys molecules may play a role in the absorption intensity.

3.3.2 Absorption spectra of non-aromatic amino acid derivatives

Table 3.3.2: Name and structure of non-aromatic amino acid derivatives studied

Sl.No.	Name	Structure
1	L-Ornithine Monohydrochloride	
2	L-Glutamic acid dimethyl ester hydrochloride	
3	N-Acetyl-L-proline	
4	N-Acetyl-L-alanine	
5	N_{ε}-Acetyl-L-lysine	
6	N_{α}-Acetyl-L-ornithine	
7	L-Glutamic acid amide	

After having an indication that charged amino acids are special with a capability to absorb in the near UV region, we studied few derivatives of amino acids (Figure 3.3.2). Among the derivatives studied, N_{ε}-Acetyl-L-lysine (Sample 5) had a very distinct profile with a long tail extending up till ~550 nm when compared to rest of the compounds studied. It also showed the maximum molar extinction coefficient at 270 nm (ε~0.8). The derivatives of Glutamic acid (Samples 2 and 7) also showed moderate absorption at 270 nm.

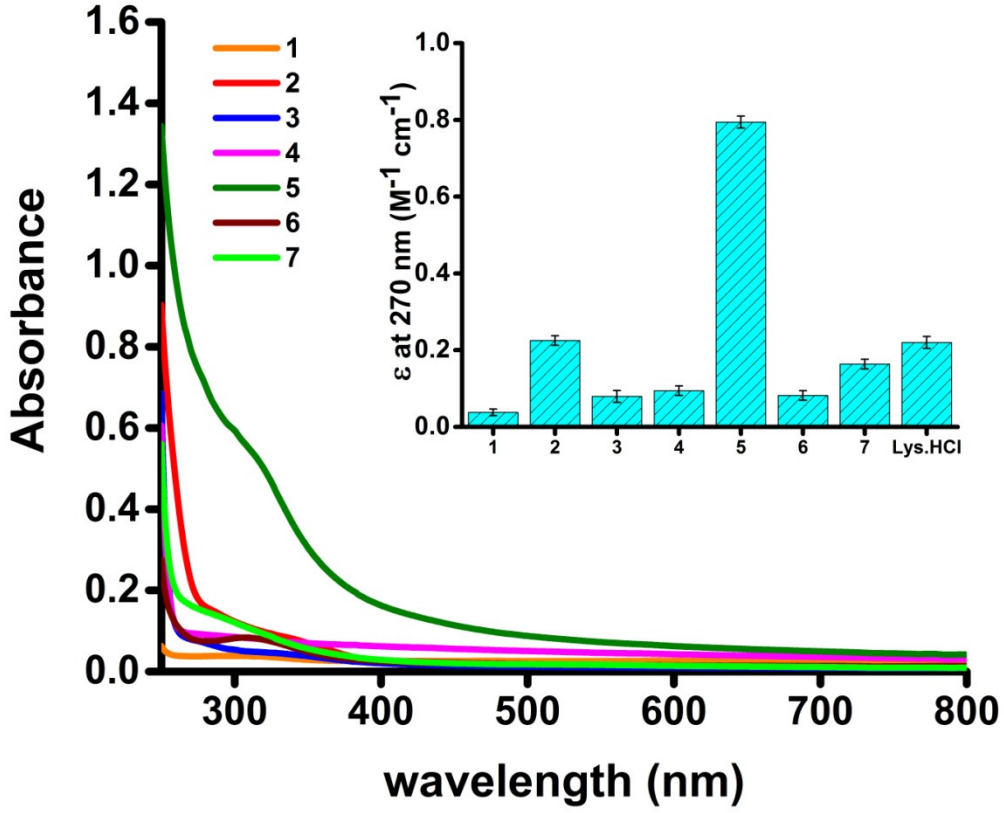


Figure 3.3.2: Absorption spectra of some amino acid derivatives in deionised water (1M). *Inset* shows comparison of absorbance at 270 nm across different samples

The above data reveal interesting facts. Decrease of one carbon in the side chain of Lys (compound 1) appears to drastically reduce its absorptivity in comparison to Lys.HCl. This fact is also evident in the absorption spectra of compounds 5 and 6. It is clear that Glu and Lys derivatives are the dominant players in absorption as derivatives of other amino acids (3 and 4) do not show any significant absorption.

3.3.3 pH dependent studies of charged amino acids

Charged amino acids are special when compared to their uncharged counterparts. They show unique absorption signatures above 250 nm, which extend up to 400 nm unlike their uncharged counterparts. Since these amino acids contain charged headgroups the next step was to study the effect of change in pH on the absorption spectra of each charged amino acid. The idea was to see the effect of protonation/deprotonation of the headgroups of each of the amino acid on the absorption spectra. Each charged amino acid has a specific pK_a for its side chain as shown in Figure 3.3.3A

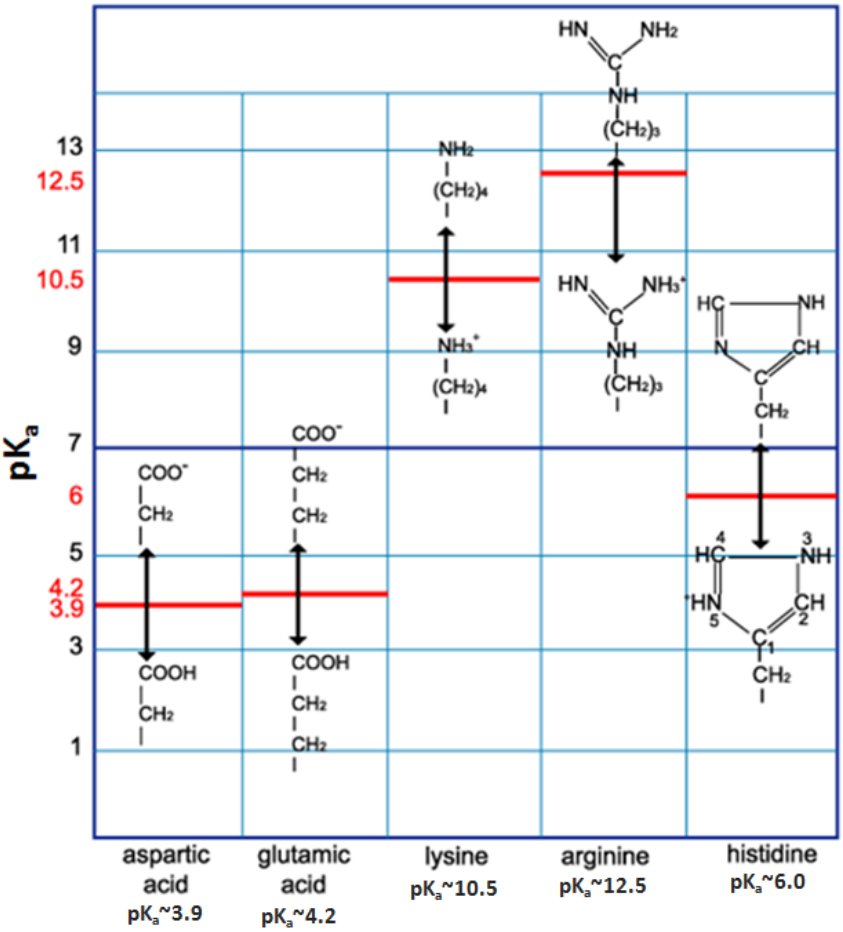


Figure 3.3.3A: pK_a of head groups for charged amino acids. (Adapted from: "IMGT®, the international ImMunoGeneTics information system® http://www.imgt.org)

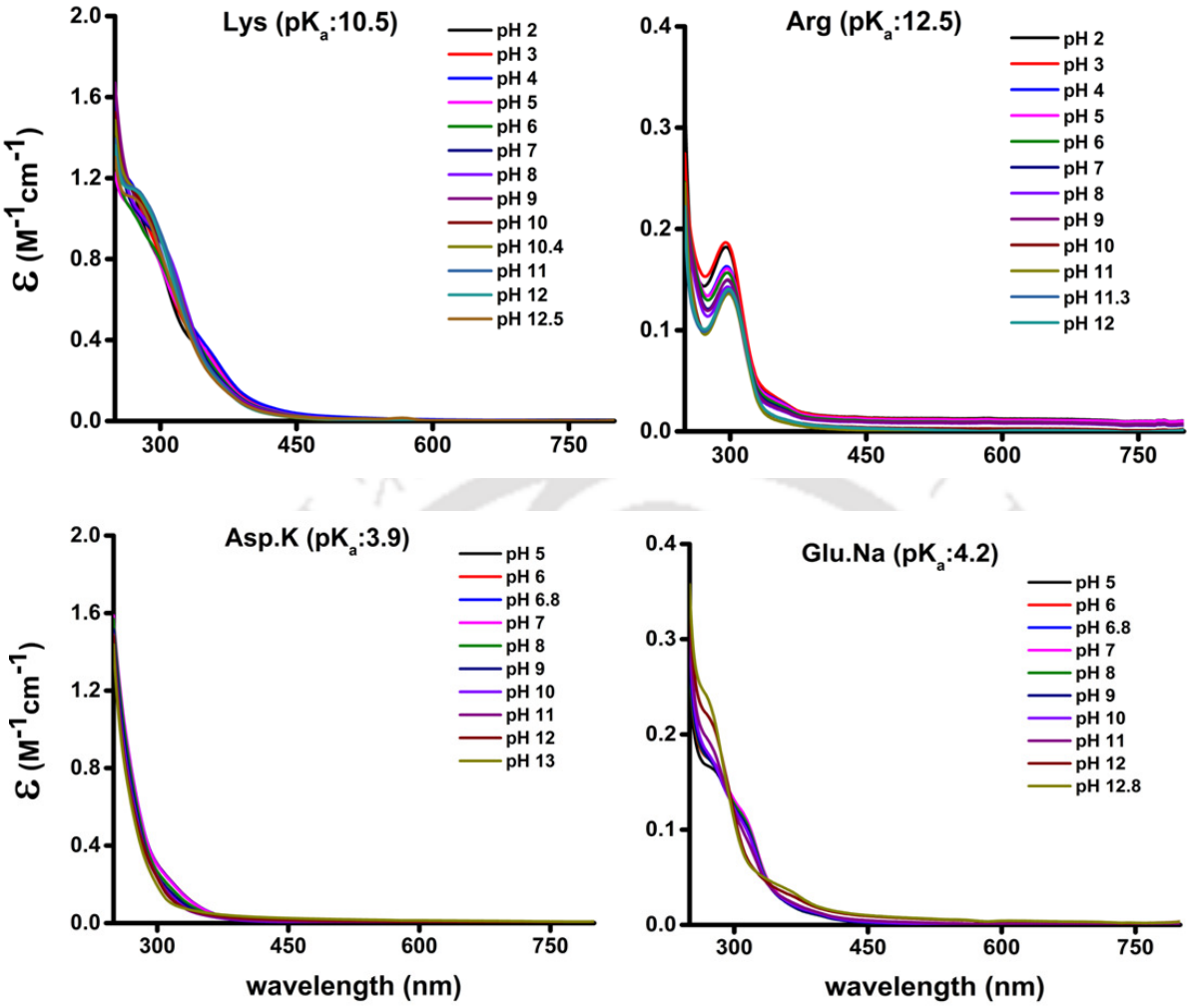


Figure 3.3.3B: Absorption spectra of different charged amino acids at different pH

pH dependent studies on each charged amino acid revealed that the pH had little influence on the absorption spectra of charged amino acids (Figure 3.3.3B). Lys and Arg have pK_a values which are very high (10.5 and 12.5 respectively). Each of them show miniscule changes in the absorption spectra with decrease in pH.

Asp.K and Glu.Na (pK_a 3.9 and 4.2 respectively) also do not show significant changes in absorption spectra with change in pH. Glu.Na does show some change in spectral profile but at very high pH values of 12 and 12.8.

However, His did show a siginificant increase in absorption intensity when the pH increased from 12 to 13 (Figure 3.3.3C). His with pK_a value of 6 is expected to be deprotonated at an extreme pH 12.

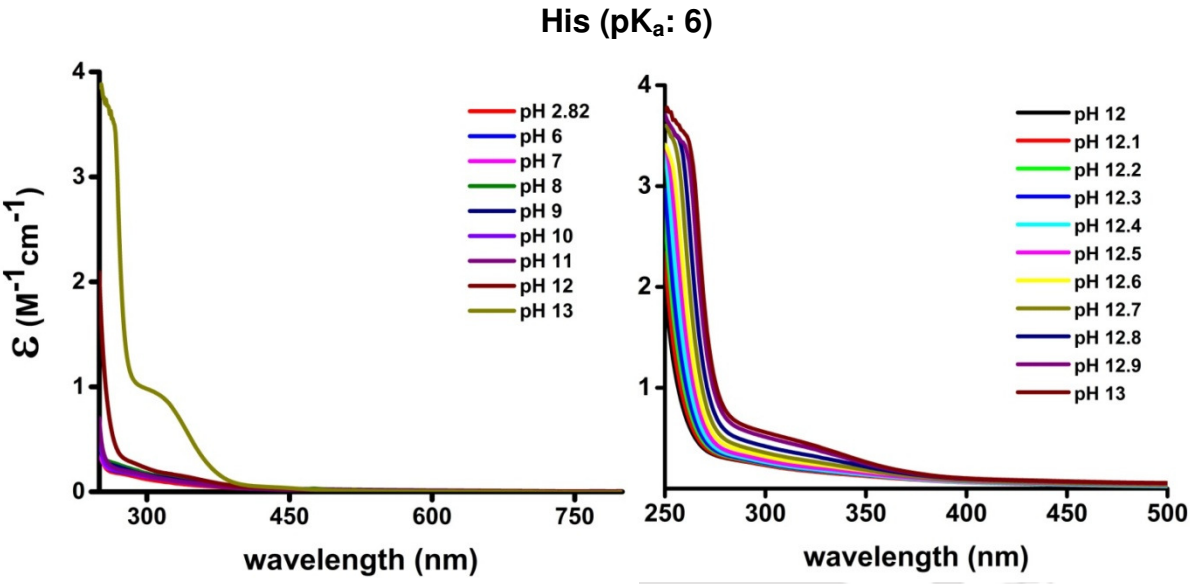


Figure 3.3.3C: Absorption spectra of Histidine at different pH

The insensitiveness of absorption spectra of charged amino acids to pH suggests that proton transfer does not play a major role in the initial photo excitation process.

3.3.4 pH dependent studies of poly-L-Lys.HBr

Poly-L-Lys has been reported to show changes in its absorption spectrum (180-250 nm) with change in pH due to change in its structure[128]. At 25 °C, it assumes a random coil structure at pH 6 and helical structure at pH 10.8. It forms an irreversible β-form at pH 10.8, 52 °C[91].

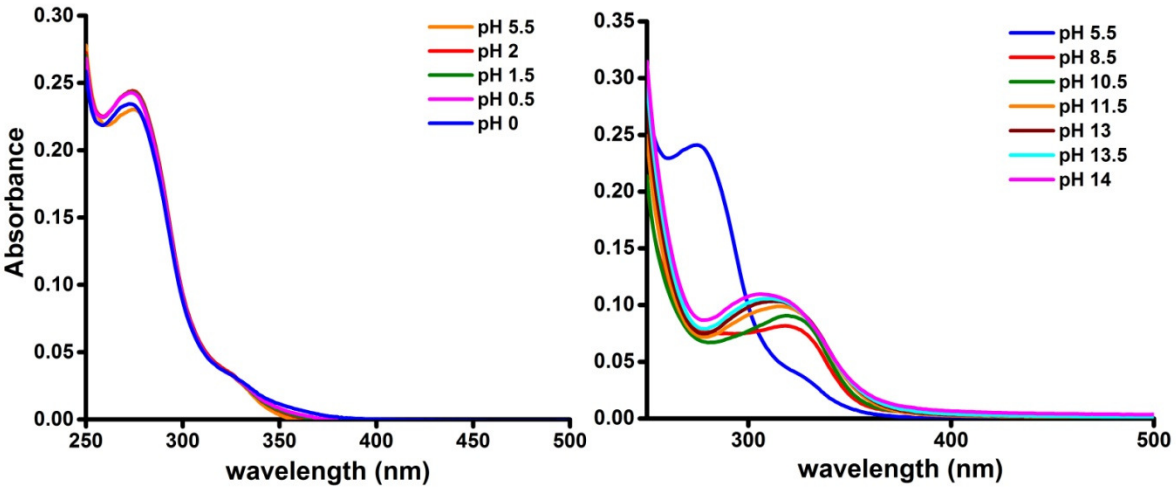


Figure 3.3.4: Absorption spectra of Poly-L-Lys.HBr at acidic pH (left) and at basic pH (right). The concentration of poly-L-Lys used was 10 mg/mL in deionised water (pH 5.5)

We studied the pH dependent change in absorption spectrum of poly-L-Lys.HBr in the UV-Vis region (250-500 nm). Poly-L-Lys.HBr does not show any change in the absorption spectra even at extreme acidic pH values (Figure 3.3.4). However, a clear shift in the absorption spectra is observed when the pH of the sample shifts to the basic side. The peak around 270 nm seems to disappear at basic pH and a new peak around 320 nm is observed. This occurs due to change in the structure of poly-Lys from random coil to helix.

Thus along with changes seen in the absorption spectrum of poly-Lys in the far UV region, it also shows changes in its absorption spectra in the UV-Vis region when the pH is shifted to the basic side.

3.3.5 Effect of KCl on Lys absorption

To understand the effect of salt on Lys spectra we recorded the absorption spectra of Lys in various concentrations of KCl. However, the Lys spectra did not show any variation with increasing salt concentration (Figure 3.3.5A). However, as shown in Figure 3.3.5B the intensity decreased slightly in MES buffer (pH 6.15), NaH_2PO_4 buffer (pH 7) and Tris buffer (pH 8.3).

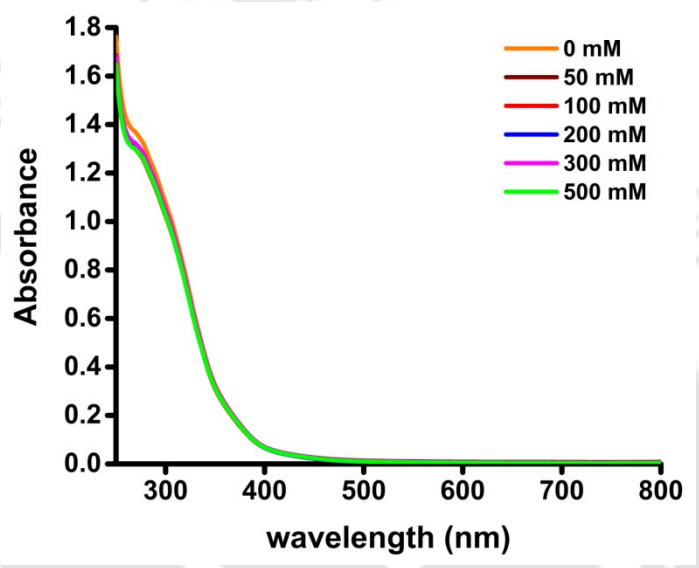


Figure 3.3.5A: Absorption spectra for Lys (1M) in different concentrations of KCl

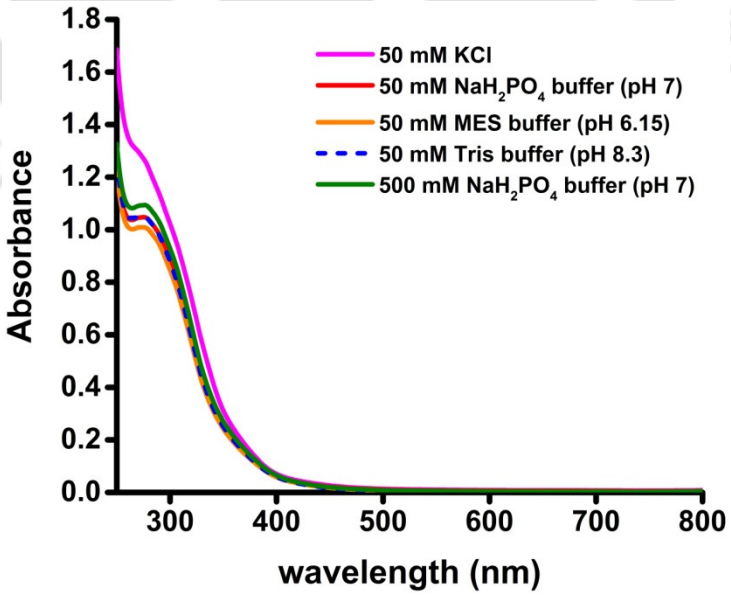


Figure 3.3.5B: Absorption spectra for Lys (1M) in different buffers

3.3.6 Effect of D_2O exchange on Lys absorption

Previous work by Homchaudhuri and Swaminathan hypothesized the role of water molecules in the unusual absorption features of Lys.HCl (chapter 1). To identify other factors which might affect the absorption spectra of Lys, we recorded its spectra in D_2O and water at different time points. The idea was to see the effect of proton exchange by deuterium on the absorption spectra of Lys. However, as shown in Figure 3.3.6, the spectra at zero hour are similar for both the samples. Similar observation holds true for the 24 hours samples as well.

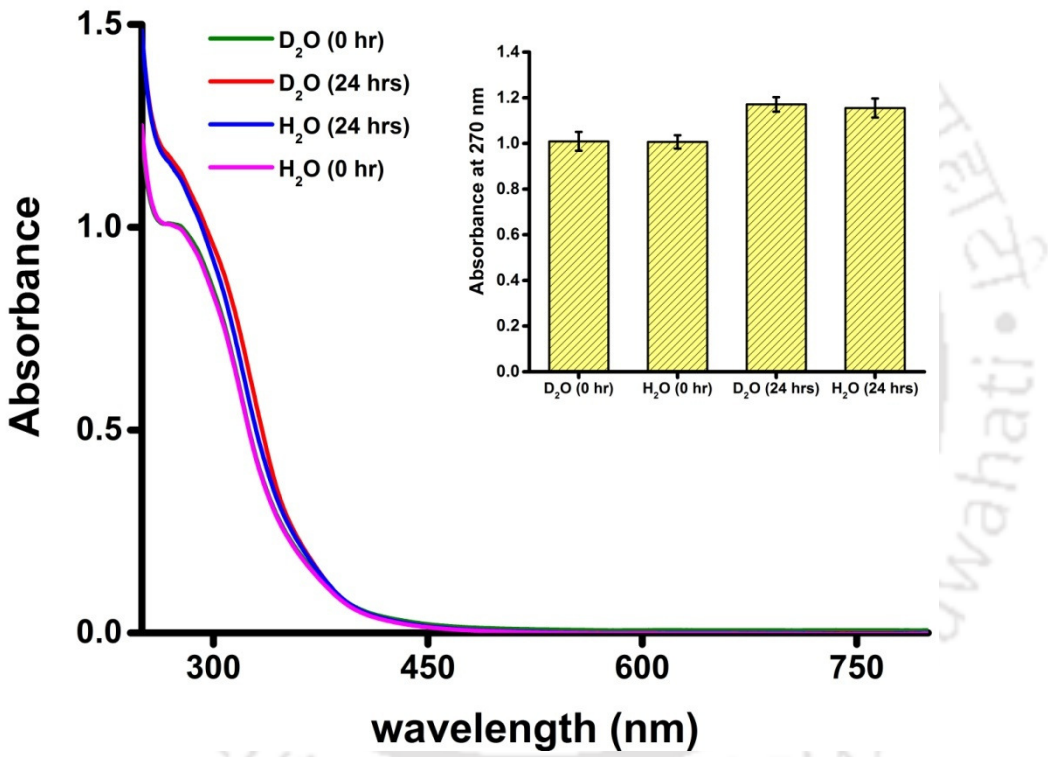


Figure 3.3.6: Absorption spectra for Lys (1M) in D_2O and H_2O at 0 and 24 hours incubated at room temperature. *Inset* shows comparison of absorbance at 270 nm

This suggests that deuterium exchange does not affect the absorption spectra of Lys. The increase in absorption intensity after 24 hours (in both H_2O and D_2O samples) might be attributed to increased interactions among the Lys residues over the period of time (may be due to formation of aggregates as shown by Homchaudhuri and Swaminathan)[107].

3.3.7 Effect of addition of Glu.Na on absorption of Lys.HCl

Lys.HCl and Glu.Na were mixed in different ratios in order to see the effect of mixture on the absorption spectra of Lys.HCl and Glu.Na. The idea was to see if oppositely charged amino acid affects the absorption spectra of Lys.HCl. The different samples prepared are shown in Table 3.3.7

Table 3.3.7: Samples with different ratios of Lys.HCl and Glu.Na

Samples	Concentration of Glu.Na (M)	Concentration of Lys.HCl (M)
1	0	0.5
2	0.1	0.5
3	0.3	0.5
4	0.5	0.5
5	0.5	0

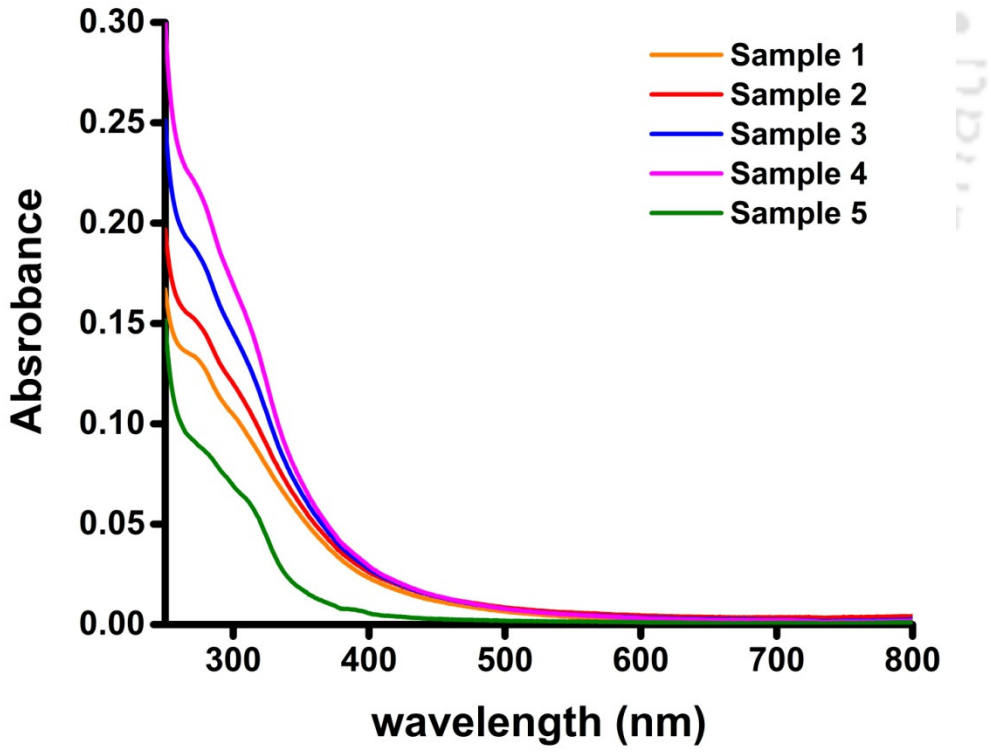


Figure 3.3.7A: Absorption spectra for different samples of Lys.HCl and Glu.Na in 0.1 N NaOH

As the amount of Glu.Na added to Lys.HCl increased, there was an increase in the absorption intensity (Figure 3.3.7A) with sample 4 (containing same concentration of Glu.Na and Lys.HCl) showing the maximum absorption intensity. This was an additive affect (Figure 3.3.7B) as the spectra for sample 4 (observed) is the addition of the individual spectrum of Lys.HCl (sample 1) and Glu.Na (sample 5) respectively (calculated spectra).

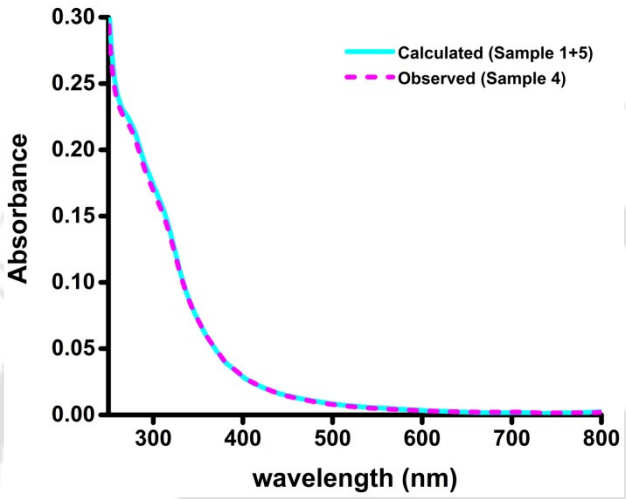


Figure 3.3.7B: Comparison of calculated spectra (sample 1+ 5) with the observed spectra (sample 4) in deionised water

The absorbance at 270 nm was not affected by change in pH, as all the samples showed similar intensities in water, 0.1 N NaOH and 0.1 N HCl (Figure 3.3.7C).

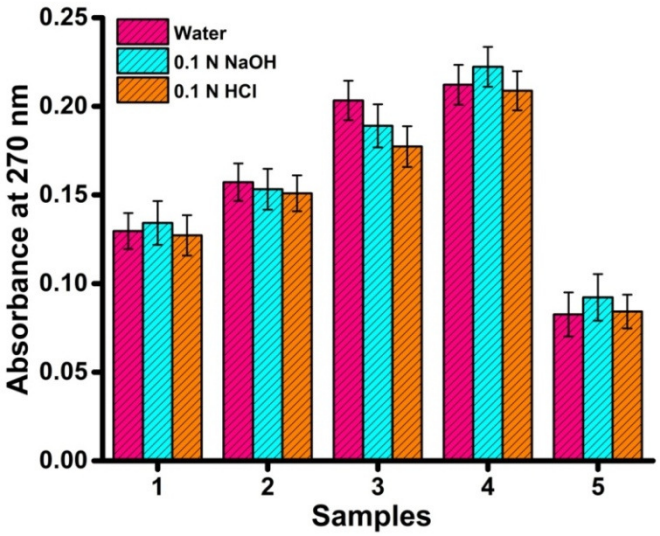


Figure 3.3.7C: Comparison of absorption at 270 nm for Lys.HCl and Glu.Na mixtures in various conditions

3.4 Conclusions:

1. Charged amino acids are special as they show absorption features in the 250-450 nm region when compared with their uncharged counterparts.

2. Lys.HCl shows absorption intensities ~6 times lower than pure Lys which supports the role of participation of charged head group behind the observed spectral signatures.

3. The absorption spectra of charged amino acids were insensitive to pH changes. Lys absorption is also insensitive to D_2O exchange thereby negating the role of proton transfer in the initial photo excitation process.

4. The absorption spectra of poly-L-Lys.HBr showed a significant shift from 270 nm to 320 nm in the UV-Vis region with increase in pH from 5.5 to 8.5.

3.5 Implications of the work:

Among the 20 naturally occurring amino acids, only aromatic amino acids have been shown to absorb in the UV-Visible region. We report for the first time absorption features of all non-aromatic amino acids up to 800 nm. Systematic control studies with high concentration non-aromatic amino acid solutions demonstrate significant absorption beyond 250 nm for charged amino acids. Together with the traditional absorption spectroscopy techniques, the ability of charged amino acids to absorb in UV-Visible region can open up a new spectral range in which the protein structure and dynamics can be monitored. The structure for most of the intrinsically disordered proteins is unknown. Since they are rich in charged amino acids, this absorption technique could hold a promise to investigate the properties of such proteins.

CHAPTER 4

QUANTIFICATION OF PROTEIN COMPOSITION AND VISUAL REPRESENTATION OF PROTEINS

4.1 Introduction:

The role of interaction between Lys residues in unusual absorption of proteins such as HSA, Calf thymus histone has been highlighted by Homchaudhuri and Swaminathan. (For details please refer to chapter 1). However these studies were focused on wavelength regions beyond 300 nm as wavelengths shorter than 300 nm would be dominated by strong absorption arising from aromatic amino acids in these proteins. Studies on small peptides (chapter 2) and non-aromatic amino acids (chapter 3), suggest the plausible role of charged amino acids behind the unique absorption features. To understand the underlying mechanisms involved behind these unusual absorption features it was important to carry out studies on whole protein.

A protein of known 3D structure which is devoid of any aromatic amino acids and rich in charged amino acids would be an ideal candidate to investigate the anomalous absorption. The idea was to carry out theoretical studies on such a protein so as to get insights on the type of transitions involved behind the unusual Lys spectral features. However, identifying such a protein in the vast sea of protein databases is a humungous task. The number of protein sequences listed in databases like Swissprot run in several millions and are ever increasing[129]. It is almost impossible to hunt for a protein of interest (with desired amino acid content) in such a vast repertoire based on the current available methods. Current approaches which store the amino acid sequences as one letter alphabetical code make the above task tedious and computationally intensive. Scanning and sorting millions of protein sequences represented as string of characters is a complex task. This chapter deals with formulation of numerical approach to capture the sequence composition of a protein.

It is observed that for a protein composed of 20 natural amino acids, the number of unique naturally occurring sequences for N residue protein is far less that $(20)^N$ possible protein variants. For example a 100 residue protein containing all 20 amino acids can have $(20)^{100}$ unique protein sequences. However in reality the actual unique population of naturally occurring 100 residue proteins is far less than this number. A numerical value coding the composition of a protein presents an opportunity to enumerate all the unique compositions. The numbers so representing the sequence composition of vast array of proteins can be sorted and searched for hunting a sequence of desired composition.

We have developed a novel method which can yield quantitative information on the amino acid composition of proteins. In this approach, each amino acid was first assigned with a unique prime integer number derived from hydrophobicity scale. Most polar residue, Arg was assigned a value 2, while most non-polar Ile was assigned prime integer value 1831. The prime numbers were used to generate a unique sequence ID by calculating the prime product (ProtID) for a given protein sequence. As factorization of a prime product yields unique factors, the ProtID at once contains all necessary information on the population of different amino acid residues in the sequence. Since ProtID can be an astronomical number for big proteins, it can be simplified by taking its base 2 logarithm, to yield protein sequence PS-Score. PS-Score can be used to quantitatively segregate protein sequences if sequence length is known. The PS-Score is directly linked to population of different amino acids in the protein. However, the protID or PS-Score dissolves all information on the exact sequence of amino acids from –N to –C terminal in the protein. To compensate for loss of protein sequence information in ProtID, we introduce protein sequence maps that display protein sequence information as a visual graphic in a variety of ways.

Various ways to represent protein composition has been tried by many workers. Recently Bao and co-workers[130] have proposed complex prime numerical representation (CPNR) of amino acids for protein function comparison, inspired by the similarity between a pattern among prime numbers and the number of codons of amino acids. They assigned unique prime numbers (2-67) and number 1 from Met to Lys based on number of codons required for each amino acid.

Resonant Recognition Model (RRM) is a widely used model which interprets the information in the linear sequence of amino acids by transforming a protein sequence into a numerical series and then into the frequency domain using Fourier Transform[131]. In the RRM the protein primary structure is represented as a numerical series by assigning to each amino acid a physical parameter value relevant to the protein's biological activity[132,133]. Electro-ion interaction pseudopotential (EIIP)[134,135] is also widely adopted method for the numerical representation of amino acids. It calculates the energy of the delocalized electrons of each amino acid residue which results in a numerical series representing the distribution of the free electrons energies along the protein molecule[136].

Numerous ways to graphically represent the proteins have also been tried for a long time. An exhaustive review on "Graphical representation of Proteins" by Plavsic and co-workers[137] lists various types of methods employed. Proteins have been represented in a variety of complex ways. A 2-D graphical representation of protein sequences based on nucleotide triplet codons has been derived for similarity analysis of protein sequences[138]. Proteins have been represented as magic circles[139] by placing 20 natural amino acids on the periphery of the unit circle. Proteins have also been represented as star like graphs which have a single central vertex and numerous branches[140]. Spectrum like representation of proteins, Amino Acid Adjacency-Count Matrix, Representation of Proteins as Walks in Space are some other methods by which protein sequences can be depicted[137].

4.2 Materials and Methods:

4.2.1 Materials:

The protein databank (http://www.rcsb.org) and Uniprot (http://www.uniprot.org/) served as repositories from which we sampled a total of 7,39,781 annotated protein sequences. Different proteins present across multiple organisms: *E.coli* (23,012 proteins), *S.cerevisiae* (6,721 proteins), *Mus musculus* (16,776 proteins), *Arabidopsis thaliana* (14,754 proteins) and *Homo sapiens* (23,012 proteins); the dark proteome of various classes of organisms: Bacteria (23,579 proteins) , Virus (7,308 proteins), Eukaryota (36,158 proteins), Archea (1,608 proteins) and Human (4,342 proteins); the non-dark proteome of various classes of organisms Bacteria (3,07,912 proteins), Virus (9,158 proteins), Eukaryota (1,42,412 proteins), Archea (17,618 proteins) and Human (15,731 proteins); Extremophiles: Acidophiles (9,130 proteins), Alkaliphiles (12,247 proteins), Halophiles (37,045 proteins) and Thermophiles (27,691 proteins); different functional classes of proteins: Membrane proteins (9,133 proteins), Nucleic acid binding proteins (3,949 proteins), Receptor proteins (1,641 proteins), Transport proteins (1,269 proteins), Signaling proteins (1,381 proteins) and Structural proteins (728 proteins); and different classes of human enzymes Oxidoreductases (546 proteins), Transferases (1,521 proteins), Hydrolases (1,601 proteins), Ligases (374 proteins), Isomerases (112 proteins) and Lyases (61 proteins) were examined.

4.2.2 Methods:

4.2.2.1 Determination of Hydrophobicity scales for amino acids

To segregate the proteins on the basis of their amino acid content we devised a new method which was based on the hydrophobicity scale of all the 20 amino acids. There exist several methods in literature which describe the hydrophobicity scales of amino acids. Each method arrives at a different scale to quantify the hydrophobicites of amino acids. No universal scale exists to quantify the hydrophobicites. Attempts to compare and to utilize different scales can be confusing, because different scales have different numerical values and ranges[141].

A vast majority of these scales involve the use of different organic solvents. The first major scale, developed by Nozaki and Tanford, used ethanol and dioxane solvents to model the protein interior, and proposed a hydrophobicity scale for nine amino acids. This model is based on free energy of transfer of amino acids from water to ethanol[142].

Fauchere and Pliska, using N-acetyl-amino acid amides and octanol–water partitioning were one of the first to use both a complete set of amino acids and derivatized amino acids[143]. Eisenberg gave a consensus scale wherein the scales were combined by averaging the normalized hydrophobicities[144] for each residue over the five scales of Tanford, Blomberg[145], Janin[146], Chothia[147] and Wolfenden[148]. Eisenberg and McLachlan also used the accessible surface area of the protein to calculate the Hydrophobicity scales of all the amino acids[149]. Biswas and co-workers provide an elaborate comparison of different Hydrophobicity scales used in the past[150]. The most used hydrophobocity scale is that of Kyte & Doolittle who combined accessible surface area measurements with water–vapor partitioning. This method takes into account the hydrophobic properties of an extensive library of proteins[151].

We combined the four most commonly used methods of Kyte & Doolittle, Eisenberg, Tanford and Fauchere & Pliska for generating the hydrophobicity scales for the 20 amino acids.

Table 4.2.2.1A: Hydrophobicity scales for different methods with maximum (blue) and minimum (red) values highlighted for each method

Amino acids	Kyte & Doolittle	Eisenberg	Tanford	Fauchere & Pliska
Ala	1.80	0.62	0.62	0.31
Arg	**-4.50**	**-2.53**	**-2.53**	**-1.01**
Asn	-3.50	-0.78	-0.78	-0.6
Asp	-3.50	-0.90	-0.09	-0.77
Cys	2.50	0.29	0.29	1.54
Gln	-3.50	-0.85	-0.85	-0.22
Glu	-3.50	-0.74	-0.74	-0.64
Gly	-0.40	0.48	0.48	0
His	-3.20	-0.40	-0.40	0.13
Ile	**4.50**	**1.38**	1.38	1.8
Leu	3.80	1.06	1.53	1.7
Lys	-3.90	-1.50	-1.50	-0.99
Met	1.90	0.64	0.64	1.23
Phe	2.80	1.19	1.19	1.79
Pro	-1.60	0.12	0.12	0.72
Ser	-0.80	-0.18	-0.18	-0.04
Thr	-0.70	-0.05	-0.05	0.26
Trp	-0.90	0.81	0.81	**2.25**
Tyr	-1.30	0.26	0.26	0.96
Val	4.20	1.08	**1.80**	1.22

As shown in the Table 4.2.2.1A, each of the four methods has different numerical scale for hydrophobicity of amino acids. Also some of the methods had redundant values for few amino acids (For e.g. same value of -3.50 for Asn, Asp, Glu and Gln in the Kyte and Doolittle method). It was therefore important to carefully assign the hydrophobicity values to each amino acid.

We proceeded in the following way to assign the hydrophobicity values:

a) In all methods Arg was normalized to -100 since it had the most minimum value across all the methods and the remaining amino acids were scaled accordingly (Table 4.2.2.1B)

Table 4.2.2.1B: Original (O) and Normalized (N) hydrophobicites values for different methods

Amino acids	Kyte & Doolittle		Eisenberg		Tanford		Fauchere & Pliska	
	O1	N1	O2	N2	O3	N3	O4	N4
Ala	1.80	40.00	0.62	24.51	0.62	24.51	0.31	30.69
Arg	-4.50	**-100.00**	-2.53	**-100.00**	-2.53	**-100.00**	-1.01	**-100.00**
Asn	-3.50	-77.78	-0.78	-30.83	-0.78	-30.83	-0.6	-59.41
Asp	-3.50	-77.78	-0.90	-35.57	-0.09	-3.56	-0.77	-76.24
Cys	2.50	55.56	0.29	11.46	0.29	11.46	1.54	152.48
Gln	-3.50	-77.78	-0.85	-33.60	-0.85	-33.60	-0.22	-21.78
Glu	-3.50	-77.78	-0.74	-29.25	-0.74	-29.25	-0.64	-63.37
Gly	-0.40	-8.89	0.48	18.97	0.48	18.97	0	0.00
His	-3.20	-71.11	-0.40	-15.81	-0.40	-15.81	0.13	12.87
Ile	4.50	100.00	1.38	54.55	1.38	54.55	1.8	178.22
Leu	3.80	84.44	1.06	41.90	1.53	60.47	1.7	168.32
Lys	-3.90	-86.67	-1.50	-59.29	-1.50	-59.29	-0.99	-98.02
Met	1.90	42.22	0.64	25.30	0.64	25.30	1.23	121.78
Phe	2.80	62.22	1.19	47.04	1.19	47.04	1.79	177.23
Pro	-1.60	-35.56	0.12	4.74	0.12	4.74	0.72	71.29
Ser	-0.80	-17.78	-0.18	-7.11	-0.18	-7.11	-0.04	-3.96
Thr	-0.70	-15.56	-0.05	-1.98	-0.05	-1.98	0.26	25.74
Trp	-0.90	-20.00	0.81	32.02	0.81	32.02	2.25	222.77
Tyr	-1.30	-28.89	0.26	10.28	0.26	10.28	0.96	95.05
Val	4.20	93.33	1.08	42.69	1.80	71.15	1.22	120.79

b) Average normalized values were calculated by addition all the normalized values (column 2, Table 4.2.2.1C) for all the four methods and then taking average (column 3, Table 4.2.2.1C) over the four methods.

c) This scale was further normalized so that Arg was assigned a value of -5 (by dividing the average normalized values with 20) and the rest amino acids were scaled accordingly (column 4, Table 4.2.2.1C). This scale was then used for further analysis.

Table 4.2.2.1C: Average normalized values and the rescaled values for different amino acids

Amino acid	Total normalized values (N1+N2+N3+N4)	Average normalized values (N1+N2+N3+N4)/4	Rescaled values
Ala	119.70	29.93	1.50
Arg	-400.00	-100.00	-5.00
Asn	-198.84	-49.71	-2.49
Asp	-193.15	-48.29	-2.41
Cys	230.96	57.74	2.89
Gln	-166.75	-41.69	-2.08
Glu	-199.64	-49.91	-2.50
Gly	29.06	7.26	0.36
His	-89.86	-22.47	-1.12
Ile	387.31	96.83	4.84
Leu	355.13	88.78	4.44
Lys	-303.26	-75.82	-3.79
Met	214.60	53.65	2.68
Phe	333.52	83.38	4.17
Pro	45.22	11.30	0.57
Ser	-35.97	-8.99	-0.45
Thr	6.23	1.56	0.08
Trp	266.80	66.70	3.34
Tyr	86.71	21.68	1.08
Val	327.96	81.99	4.10

A graphical comparison of all the methods along with the rescaled values calculated by us is shown in Figure 4.2.2.1

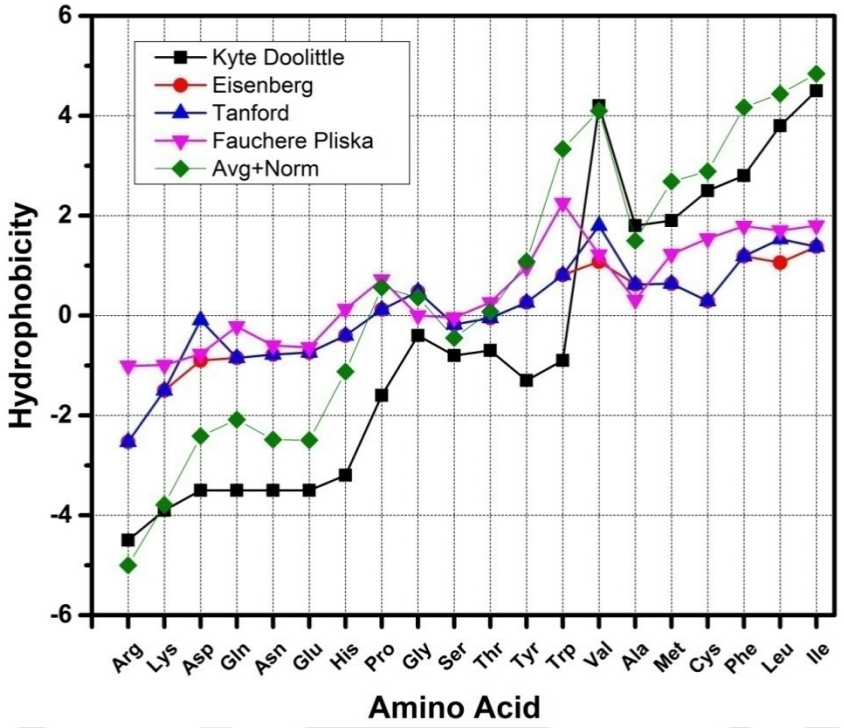


Figure 4.2.2.1: Different hydrophobicity scales for amino acids along with the averaged and normalized values calculated on the basis of four scales (green curve)

4.2.2.2 Assignment of unique prime numbers to each amino acid

After the average hydrophobicity scale (H-Index), was arrived at in the range of -5 (Arg) to +4.84 (Ile) for the 20 amino acids (column A, Table 4.2.2.2), the next target was to assign unique prime numbers to each amino acid. The aim was that each amino acid has a unique prime number and the prime number should also reflect the hydrophobicity content of the amino acid. Keeping these points in mind we proceeded to assign prime numbers.

Table 4.2.2.2: Prime numbers assignment for each amino acid

	A	B	C	D	E	F
Amino acids	**H-Index**	**H-Index + 6**	**$2^{(H\text{-}Index\ +6)}$**	**Assigned Prime Numbers**	**Log_2 (Prime number)**	**% deviation $\frac{(E-B)}{B} X\ 100$**
Arg (R)	-5.00	1.00	2	**2**	1.00	0.00
Lys (K)	-3.79	2.20	4.62	**5**	2.32	5.45
Glu (E)	-2.50	3.50	11.34	**11**	3.46	-1.14
Asn (N)	-2.49	3.51	11.42	**13**	3.70	5.41
Asp (D)	-2.41	3.58	12.00	**17**	4.09	14.24
Gln (Q)	-2.08	3.91	15.09	**19**	4.24	8.43
His (H)	-1.12	4.87	29.37	**29**	4.86	-0.21
Ser (S)	-0.45	5.55	46.86	**47**	5.55	0.00
Thr (T)	0.08	6.07	67.55	**67**	6.07	0.00
Gly (G)	0.36	6.36	82.32	**83**	6.38	0.31
Pro (P)	0.57	6.56	94.69	**97**	6.60	0.61
Tyr (Y)	1.08	7.08	135.66	**137**	7.10	0.28
Ala (A)	1.50	7.49	180.55	**181**	7.50	0.13
Met (M)	2.68	8.68	410.84	**409**	8.68	0.00
Cys (C)	2.89	8.88	473.40	**479**	8.90	0.23
Trp (W)	3.34	9.33	645.84	**647**	9.34	0.11
Val (V)	4.10	10.09	1097.11	**1097**	10.10	0.10
Phe (F)	4.17	10.16	1151.27	**1151**	10.17	0.10
Leu (L)	4.44	10.43	1388.35	**1381**	10.43	0.00
Ile (I)	4.84	10.84	1834.73	**1831**	10.84	0.00

Since the assigned H-Index scale (column A, Table 4.2.2.2) consists of negative values and therefore assigning prime numbers was not possible, the first step therefore was to convert these numbers to positive integers. This was done by addition of '6' since the smallest integer was '-5' (see column B, Table 4.2.2.2). However many of values were numerically close for different amino acids, it was therefore not possible to assign unique prime numbers to each amino acid while retaining the hierarchy of hydrophobicity. To expand the

range, we raised each number to the power of two. This enabled us in generation of a wide range of numbers from 2 to 1834.73 (column C, Table 4.2.2.2). We then assigned prime numbers closest to each of this number for respective amino acids (column D, Table 4.2.2.2). The aspect that each of these numbers reflects the hydrophobicity content of each amino acid was confirmed by calculating the $\log_2$ values (column E, Table 4.2.2.2) for each prime number. The $\log_2$ values were very close to the (H-Index + 6) values (column B, Table 4.2.2.2) with the sole exception of Glu. This is also reflected by very little deviation (column E, Table 4.2.2.2) of the assigned prime integers from the H-index values for all amino acid with Asp having the maximum deviation ~14%. This affirmed that the prime numbers were unique and also reflected the respective hydrophobicity content in them for each amino acid. These prime numbers were then later used to calculate ProtID and PS-Score for a wide range of protein sequences across different organisms.

4.2.2.3 Calculation of ProtID and PS-Score

The ProtID is defined as the product of assigned prime numbers of amino acids present in a protein sequence. Since ProtID can be a very big number for large protein sequences, the amino acid content information was stored in PS-Score which was calculated by taking the base 2 logarithm of ProtID. Averaging the PS-Score with respect to the sequence length yields the average PS-Score which can serve as an indicator for average hydrophobicity of amino acids in the sequence. All these parameters were calculated by codes written in Mathematica, Version 10.4.1(see Appendix).

4.2.2.4 Generation of different visual patterns for proteins

ProtID and PS-Score were calculated in order to yield information about the amino acid content for protein sequences. However, the sequence information of amino acids is lost during calculation of protID. To address this issue, different approaches to represent a protein as visual graphic was explored. Proteins sequences were represented as bands in an electrophoretic gel, circles, triangles, and so on. The bar pattern of protein sequence was generated using Mathematica, while other representations were made manually in Microsoft PowerPoint.

4.3 Results and discussion:

4.3.1 ProtID and PS-Scores for some proteins

ProtID, PS-Score and average PS-Score was calculated according to the method mentioned above. Some examples of ProtID and PS-Score for few proteins are shown in Table 4.3.1. The base 10 and base 36 values for ProtID are an alternate way of representing the amino acid content information for a protein. Base 10 representation is an approximation from which the original prime factors cannot be retrieved. Since ProtID is constructed using prime numbers, prime factorization of the ProtID would yield the combination of prime numbers of which the ProtID is made of which in turn tells us about the amino acid composition of a given protein sequence. The ProtID can serve as a *coarse grain* identity tag of the protein composition. ProtID will permit proteins with nearly similar amino acid composition to be grouped together when sorted in a hierarchy based on composition. Proteins that have exactly identical amino acid composition will have the same ProtID. As expected ProtID and PS-Score are functions of sequence length. As the sequence length increases both the PS-Score and ProtID increases. Average PS-Score (calculated by dividing the PS-Score with sequence length) normalizes the effect of sequence length. Since the most charged residue (Arg) was allotted the lowest prime number (2) and the least charged residue was allotted the highest prime number (1831), the lower the average PS-Score for a protein, higher is the population of charged residues in it and vice-versa. Therefore, just by looking at the average PS-Score value one can get a quick idea on the polarity of amino acid composition for a given protein sequence.

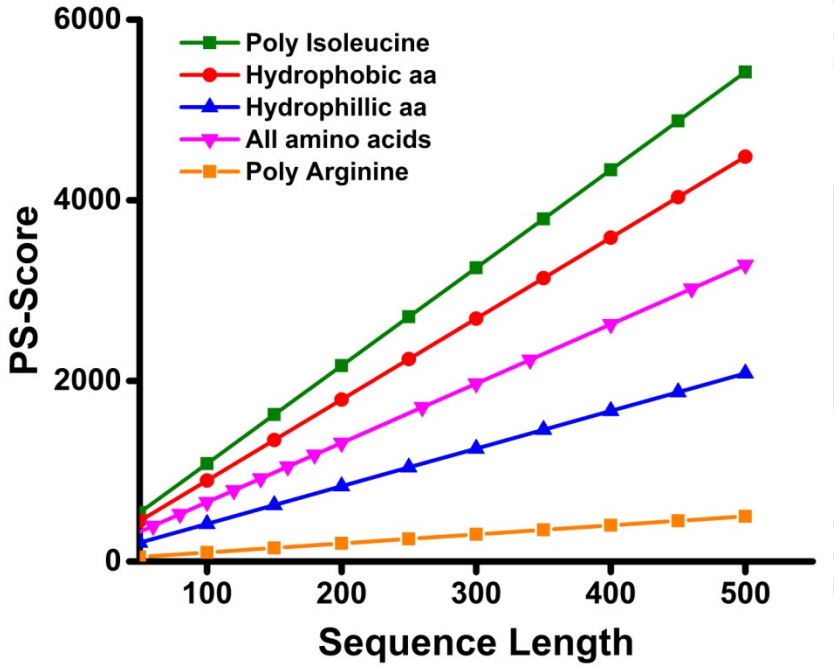
Table 4.3.1: ProtID and PS-Score for some proteins

Protein Name [Sequence length]	Sequence	ProtID	ProtID (Base 10)	ProtID (Base 36)	Log 2 (ProtID) or PS-Score
Cardiac phospholamban (PLB) [52]	MEKVQYLTRSAIRRASTIEMPQQARQK LQNLFINFCLILICLLLICIIVMLL	60472416733961068066704934776051866127954934483877486081736054355027550349007127367400480560151241132916396293342400	$6.047*10^{116}$	15850zqikdxrmhxseu41a5z7trnhcg6zyxb89yvdft32q6a0eonymjfgrv5wdvoddgs8vhwt074g	387.94
Apolipoprotein A-II [100]	MKLLAATVLLLTICSLEGALVRRQAKE PCVESLVSQYFQTVTDYGKDLMEKVK SPELQAEAKSYFEKSKEQLTPLIKKAGT ELVNFLSYFVELGTQPATQ	71639197879527242573982930537996879506639519559757714880424542126273302796874467696611077963339257934742589553040792400133688549675786519637204774213028976209660792733156257804641261528039752695312500	$7.163*10^{199}$	4g84yfhybkvyuzbulokg6ma3no126cym1pd4w6u0js4qwa2qtzfrtqohc5ywi9jitvlzv6jh8u9sqb5gkgakuaba9mdza85wlh4czqu8hh458y8ryngj7m7tnufmssigk	663.90
Ubiquitin-conjugating enzyme E2 K [200]	MANIAVQRIKREFKEVLKSEETSKNQI KVDLVDENFTELRGEIAGPPDTPYEGG RYQLEIKIPETYPFNPPKVRFITKIWHPN ISSVTGAICLDILKDQWAAAMTLRTVL LSLQALLAAAEPDDPQDAVVANQYKQ NPEMFKQTARLWAHVYAGAPVSSPEY TKKIENLCAMGFDRNAVIVALSSKSWD VETATELLLSN	28293825221915884347522325589975095039088131951179535634162083811944420937658236037852693931921885863821875254848748716595332970978126897319021253406051192490394808258184405968562727474079074535712362757089764642793077236480033778870525701998303041830197069196256389114550609902883094807635824238421687519312845748660023604829829385928812722547842992355563876985715653809790592977078125000000000	$2.829*10^{393}$	icc1hf55mnaterjn4ovk9bo3c7woccu78mleznq8mri3f6dkmiocs1e38y8amb6r3qxt41u1wruoqadooopu4j3tn6zsrurhc4ndgpe0163cng8lhoklzm98xzq8q0kjx8qx9zh59f7pi95u8e51a5ik7umdcr6qrblq3umx26wlgzsuwoiynzzsyw3xrk8nkmnwg5o3nq0j7vecuu99z84whjyucks60x40vax8ym608xtnqjsnrup0pgvls	1307.01

4.3.2 PS-Score for synthetic samples

We calculated PS-Scores for some synthetic samples in which we considered sequences of different lengths (50-500) made up of i) only Arg residues (Poly Arginine), ii) only Ile residues (Poly Isoleucine), iii) first 10 amino acids in the Hydrophobicity scale (hydrophillic aa), iv) last 10 amino acids in the Hydrophobicity scale (hydrophobic aa) and v) all the 20 amino acids (all amino acids).

As shown in Figure 4.3.2, PS-Score increases linearly with increase in sequence length. The slopes depict the average PS-Score for each sample.



Equation	y = a + b*x				
Protein type	Poly Isoleucine	Poly Arginine	Hydrophobic amino acids	Hydrophillic amino acids	All amino acids
Intercept	7.85E-05	0	7.89E-05	-1.77E-04	-1.62E-04
Slope	10.8384	1	8.96532	4.16709	6.5662

Figure 4.3.2: PS-Score vs. Sequence length for some synthetic samples. The table shows the slope and intercepts values for each plot

Poly Arg has the least slope/ average PS-Score of 1, while Poly Ile has the maximum slope of 10.83. The slopes for the remaining samples lie between these two extreme samples. The slopes for poly Arg and poly Ile serve as boundaries between which the slopes of any given protein sequence is expected to lie. All these synthetic samples serve as standards against which we can compare the PS-Score/Average PS-Score values and get an idea about the composition of a given protein sequence.

4.3.3 Identification of proteins rich in charged amino acids

The average PS-Score was used to identify few proteins from Uniprot (Table 4.3.3A) and PDB (Table 4.3.3B) which were rich in charged amino acids. We also calculated the number of Lys pairs which were within 10 Å distance for proteins in the PDB database since earlier studies had speculated the interaction among ε-NH_2 of two Lys residues to be important.

Table 4.3.3A: Examples of proteins rich in charged amino acids from Uniprot

Uniprot ID	Total residues	Avg. PS-Score	No. of charged and aromatic amino acids
Q8WVK2	105	4.48	W=0,Y=3,F=4, K=14,E=18,D=6,R=39, H=2
Q8N9E0	248	4.58	W=3,Y=5,F=1, K=61,E=30,D=11,R=20, H=4
Q9NWB6	273	4.63	W=0,Y=0,F=3, K=33,E=54,D=7,R=53, H=3
P20962	102	4.66	W=0,Y=0, F=0, K=13,E=39,D=8,R=4,H=0
Q5BKY9	247	4.71	W=2,Y=5,F=1, K=61,E=30,D=12,R=16, H=4

We were able to identify proteins which were highly rich in charged amino acids from the Uniprot database. The average PS-Scores for all these proteins were quite small. Since, none of these proteins had a PDB structure; we could not get information about the interaction among the charged amino acids in these proteins. Also to characterize the nature of the transitions involved using theoretical approaches, we needed a protein with a known 3D structure. We therefore carried out similar search in the PDB database.

Table 4.3.3B: Examples of proteins rich in charged amino acids from PDB

PDB ID	Total residues	Avg. PS-Score	No. of charged and aromatic amino acids	No. of Lys pairs within 10 Å
2LXY	67	5.45	W=0,Y=0, F=0, K=17,E=17,D=0,R=2,H=0	14
2DJV	78	5.73	W=0,Y=0,F=0, K=14,E=6,D=3,R=1, H=0	8
1C6W	33	5.12	W=0,Y=0,F=0, K=7,E=2,D=2,R=4, H=1	6
2LCM	28	5.92	W=0,Y=0,F=0, K=6,E=0,D=1,R=4, H=0	4
1DMD	31	5.79	W=0,Y=0,F=0, K=6,E=2,D=0,R=0, H=0	4
2KTC	34	5.93	W=0,Y=0,F=0, K=6,E=3,D=1,R=0, H=1	4
2KSG	48	6.39	W=0,Y=0,F=0, K=7,E=3,D=6,R=0, H=1	4
4MT2	62	6.36	W=0,Y=0,F=0, K=8,E=1,D=3,R=0, H=0	4
2L36	24	6.56	W=0,Y=0,F=0, K=6,E=0,D=0,R=0, H=0	3

All the proteins mentioned in Table 4.3.3B are devoid of any aromatic amino acids. The protein with PDB ID: 2LXY (Alpha 3C) with an average PS-Score of 5.45 seemed to be a potential candidate for our further studies.

It had the maximum number of Lys pairs within 10 Å. This protein is also rich in other charged amino acids as well. Based on this information we decided to carry further studies on α_3C, the details of which are mentioned in Chapter 5. For complete sequence and structure of this protein please refer to Figure 5.3.2A in Chapter 5.

Since we established a method using which we were able to identify proteins which were rich in charged amino acids, we decided to take forward this approach to proteins which belong to various classes of organisms to find potential applications of this method.

4.3.4 Analysis of PS-Score across proteomes of different organisms

We analyzed the proteins across different organisms in order to understand the amino acid distribution of in them. Since number of proteins studied was large, even a small difference in the PS-Scores or average PS-Scores can be considered significant.

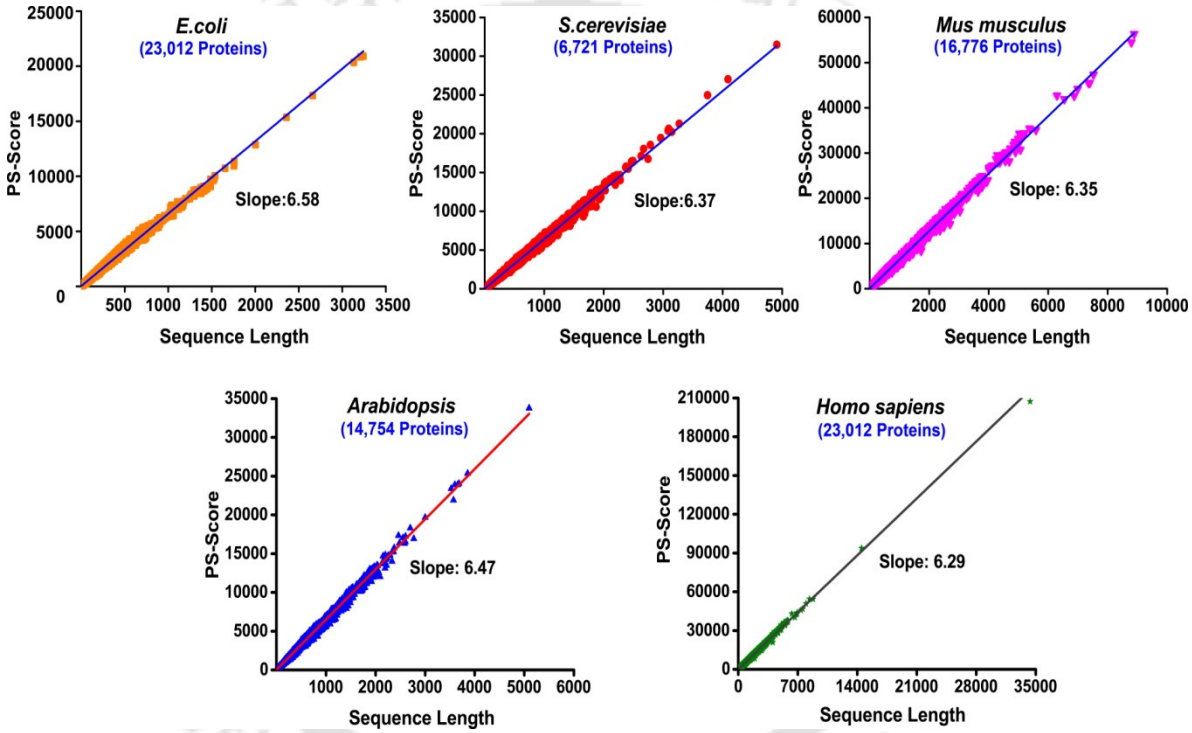


Figure 4.3.4A: PS-Score vs. Sequence length plots for proteomes of different organisms

The average PS-Score obtained from the slope of PS-Score vs. Sequence length plot for all proteins was largest for *E.coli* (6.58) among the organisms studied, while *Homo sapiens* had the least value of average PS-Score (6.29). The proteins in human show 0.95 fold decrease when compared with the average PS-Score of all amino acids (6.56). The proteins of *E.coli* had almost similar values to that of the average PS-Score of all amino acids (Figure 4.3.4B). This means that the *E.coli* has proteins which contain almost all amino acids in equal proportion in their sequence.

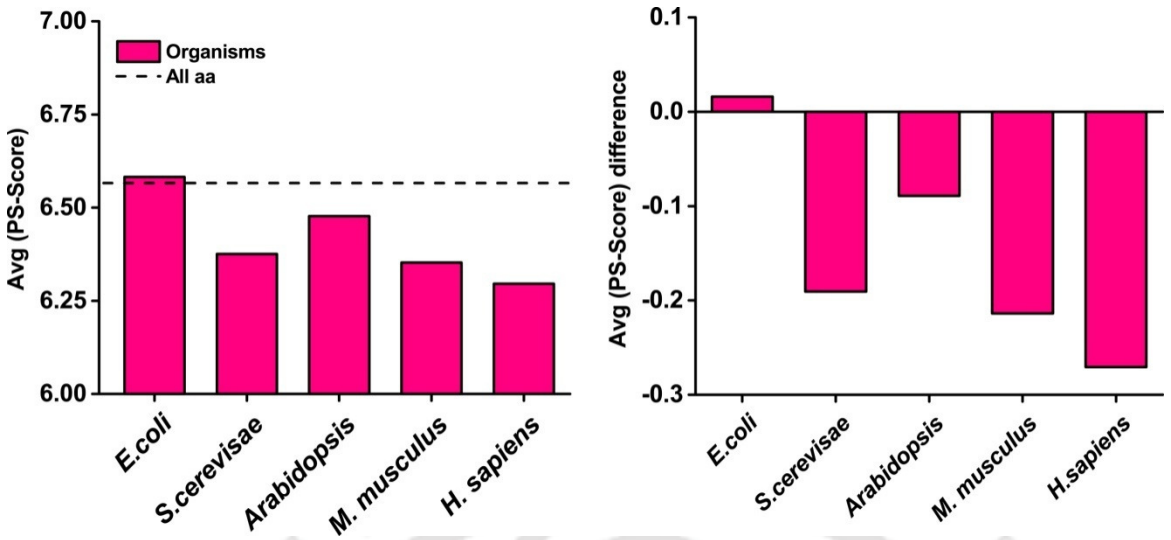


Figure 4.3.4B: Comparison of average PS-Scores of all organisms with the average PS-Score of all amino acids (left) and difference in average PS-Scores among organisms (right) compared to all amino acids

The difference plot (Figure 4.3.4B, right) clearly shows that average PS-Score values are lower than all amino acids for Human, Mouse, *Arabidopsis* and *S.cerevisae*. This means the proteins in these organisms have more population of polar amino acids.

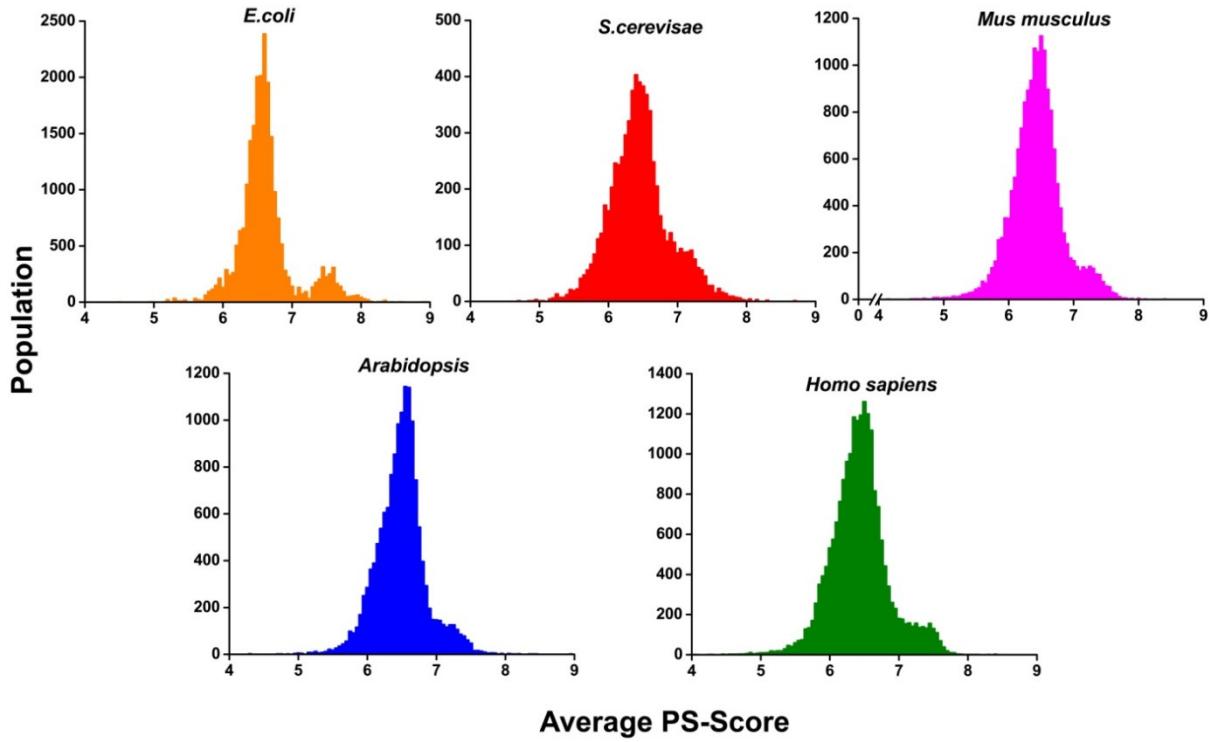


Figure 4.3.4C: Histogram distribution of average PS-Score for different organisms

To get an idea of how the average PS-Score is distributed across a population of proteins, we calculated the histogram distribution of the average PS-Score within a given organism. As shown in Figure 4.3.4C, *E.coli* shows two distributions. The larger one ranges from ~6-7 while a smaller population of proteins have average PS-Scores of ~7-8. This means that while majority of proteins in *E.coli* are rich in polar amino acids, subsequent to a gap in average PS-Score a small population comprising of proteins with more hydrophobic amino acids also exists. All other organisms show similar range of ~5-8 but with no noticeable gaps. The histogram distribution helps to easily identify the range of average PS-Scores seen in a protein population and one can then pick up the protein composition of interest based on this distribution.

4.3.5 Analysis of PS-Score across proteomes of extremophiles

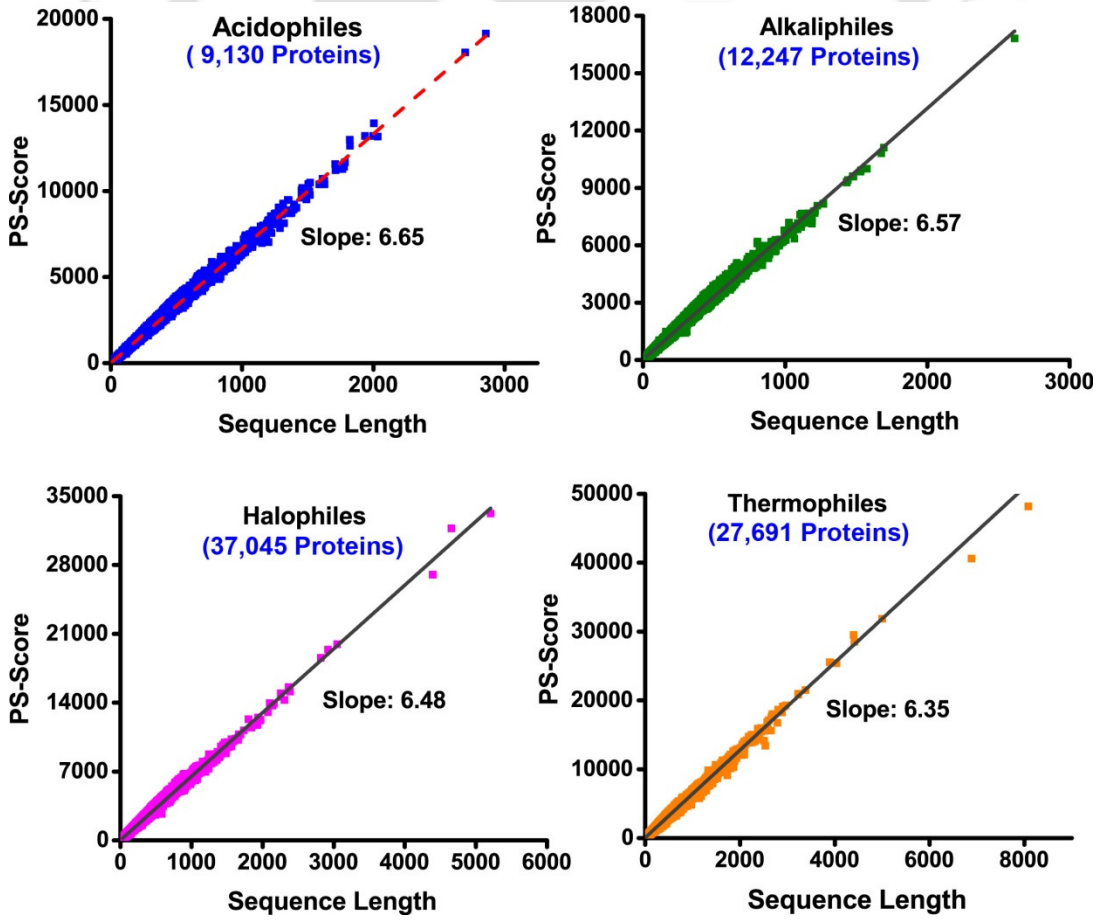


Figure 4.3.5A: PS-Score vs. Sequence length plot for proteomes of different extremophiles

Extremophiles are organisms which survive in extreme environments[152,153] such as extremes of temperature (thermophiles)[154], pH (acidophiles/alkaliphiles)[155,156] and salt (halophiles)[157]. These organisms have different mechanisms by means of which they are able to survive in such extreme conditions. We chose proteins from four different classes of extremophiles to analyse the amino acid composition in the proteins of each class of extremophile. Among all the extremophiles, acidophiles had the maximum slope of 6.65 while thermophiles had the least slope 6.35 (Figure 4.3.5A).

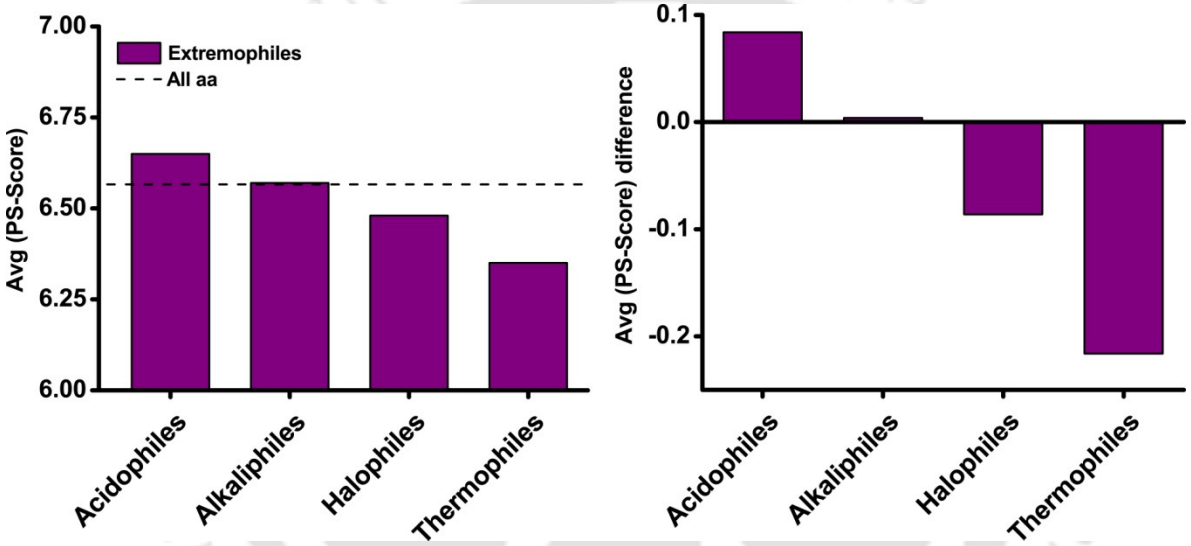


Figure 4.3.5B: Comparison of average PS-Scores of all extremophiles with the average PS-Score of all amino acids (left) and difference in average PS-Scores among extremophiles (right) compared to all amino acids

The difference plots with respect to the average PS-Score value of all amino acids showed that thermophiles differed the most (Figure 4.3.5B). Alkaliphiles have average PS-Score similar to all amino acids which suggests that they mostly contain proteins which have distribution of all the 20 amino acids.

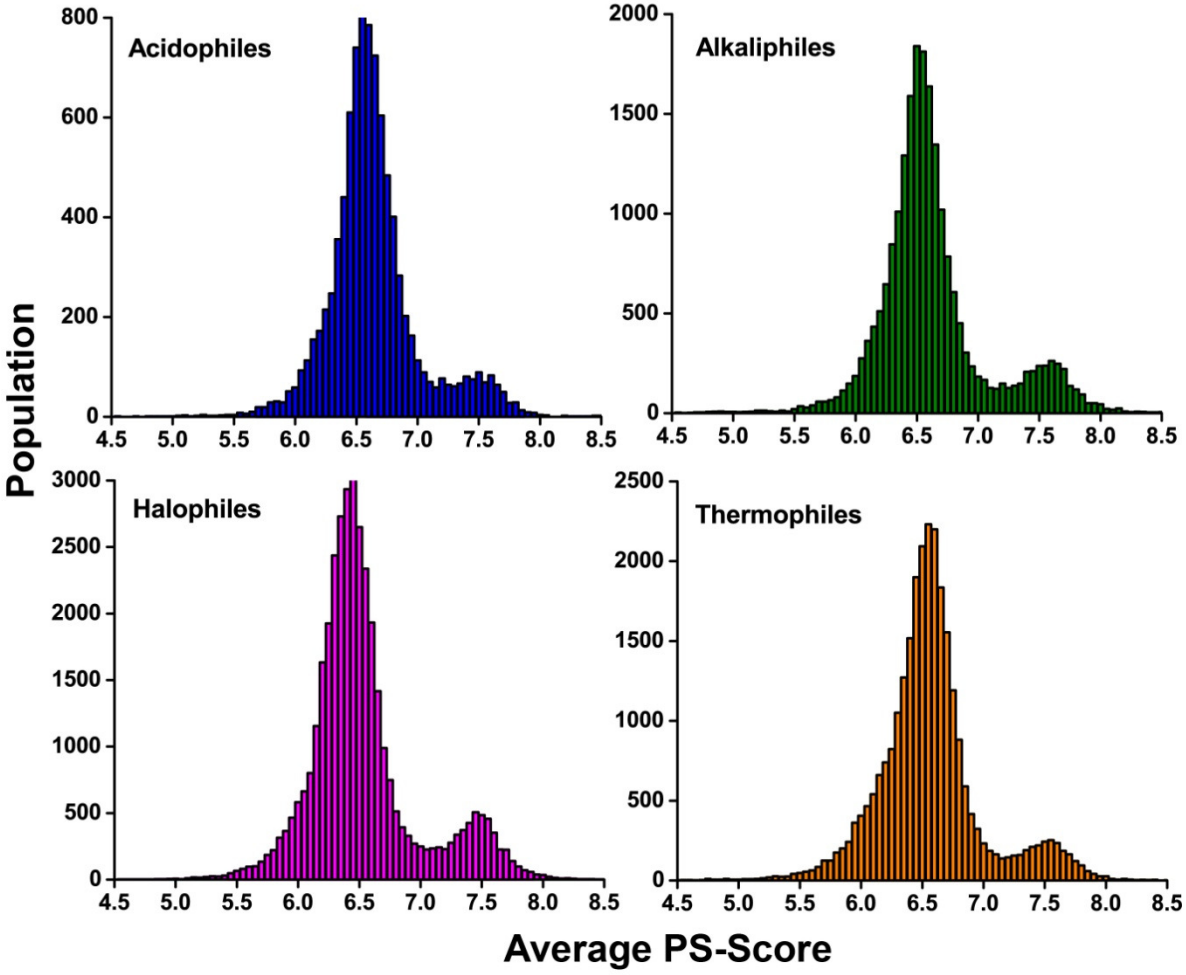


Figure 4.3.5C: Histogram distribution of average PS-Score for different classes of extremophiles

If we look at the population of average PS-Scores in extremophiles, halophiles contain the maximum population of proteins in the 5-7 range. The population depicts a bimodal distribution of average PS-Scores (5-7 and 7-8) for every class of extremophile (Figure 4.3.5C). This means that while most of the proteins contain all the amino acids in their protein sequences, small populations of proteins also have hydrophobic amino acids as well.

4.3.6 Analysis of PS-Score across different functional classes of proteins

Human beings are known to contain different types of proteins which perform a wide variety of functions. These functional classes contain a varied distribution of amino acids.

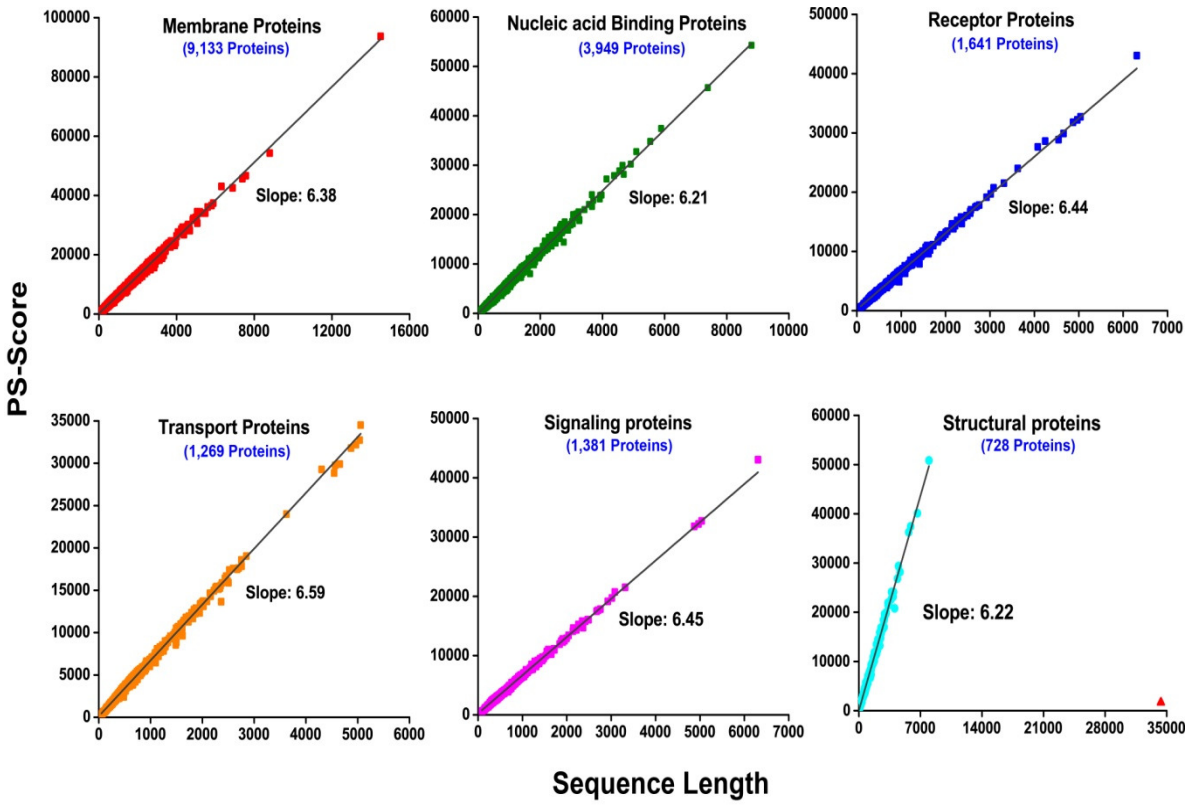


Figure 4.3.6A: PS-Score vs. Sequence length plot for different functional classes of proteins

We analysed six different functional classes of proteins in humans. Among these classes, transport proteins had the maximum slope of 6.59 while the nucleic acid binding proteins showed the lowest slope of 6.21 (Figure 4.3.6A). Since the nucleic acid (NA) binding proteins are known to contain charged amino acids, they show the lowest value. This property is also depicted by the difference plots (Figure 4.3.6B) where nucleic acid binding proteins differ the most from the average PS-Score of all amino acids. Structural proteins (with an outlier of PS-Score 1651.7) were also seen to have almost similar PS-Score as the nucleic acid binding proteins. Transport proteins had almost same average PS-Score as that for all amino acids. Similar slopes in signaling and receptor proteins (6.45 and 6.44 respectively) indicate that these classes of protein have similar composition of amino acids.

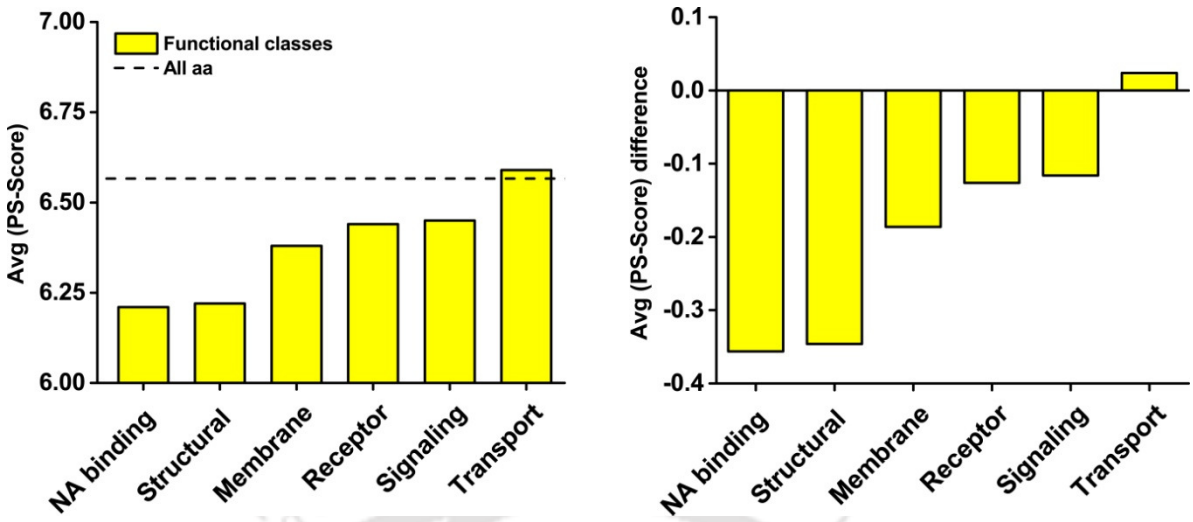


Figure 4.3.6B: Comparison of average PS-Scores of all functional classes with the average PS-Score of all amino acids (left) and difference in average PS-Scores among the functional classes (right) compared to all amino acids

The histogram distributions for all the functional classes of proteins (Figure 4.3.6C) are quite different. Membrane and nucleic acid binding proteins show similar profiles with nearly bimodal distributions of the average PS-Score (5-7 and 7-8). Transport proteins display unimodal distribution with average score ranging from ~5.5-7.5. Along with a large population of proteins with average PS-Score ~6-7, transport proteins also have a high population of proteins with average PS-Score 7-7.5. Receptor proteins also show a bimodal distribution. While majority of receptor proteins have average PS-Score ~6.5, there is also a large section of proteins with average PS-Score of about ~7-7.5. This indicates that there are a large number of receptor proteins which are rich in hydrophobic amino acids. Signaling proteins show a multimodal histogram distribution. The larger populations of this class of proteins have average PS-Scores between 7-7.5, while proteins with the score between 6-7 occupy a lower population. Structural proteins nearly have a unimodal distribution with majority of the proteins having average PS-Score lying in between 6-6.5.

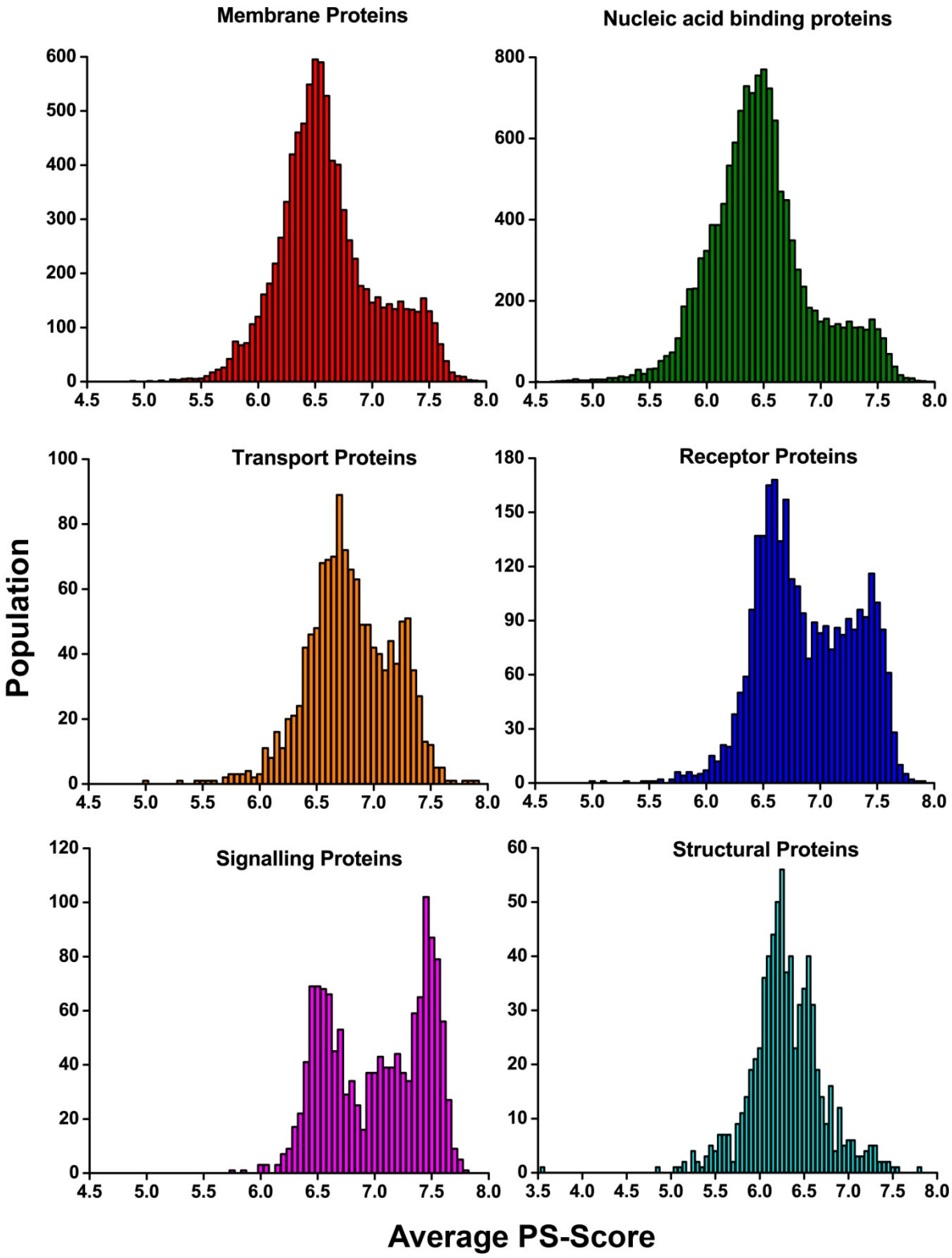


Figure 4.3.6C: Histogram distribution of average PS-Score for different functional classes of proteins

4.3.7 Analysis of PS-Score across different enzyme classes

Human beings possess six different classes of enzymes, with each class of enzyme performing a specific function. The amino acid composition in each of these classes was analyzed using our method.

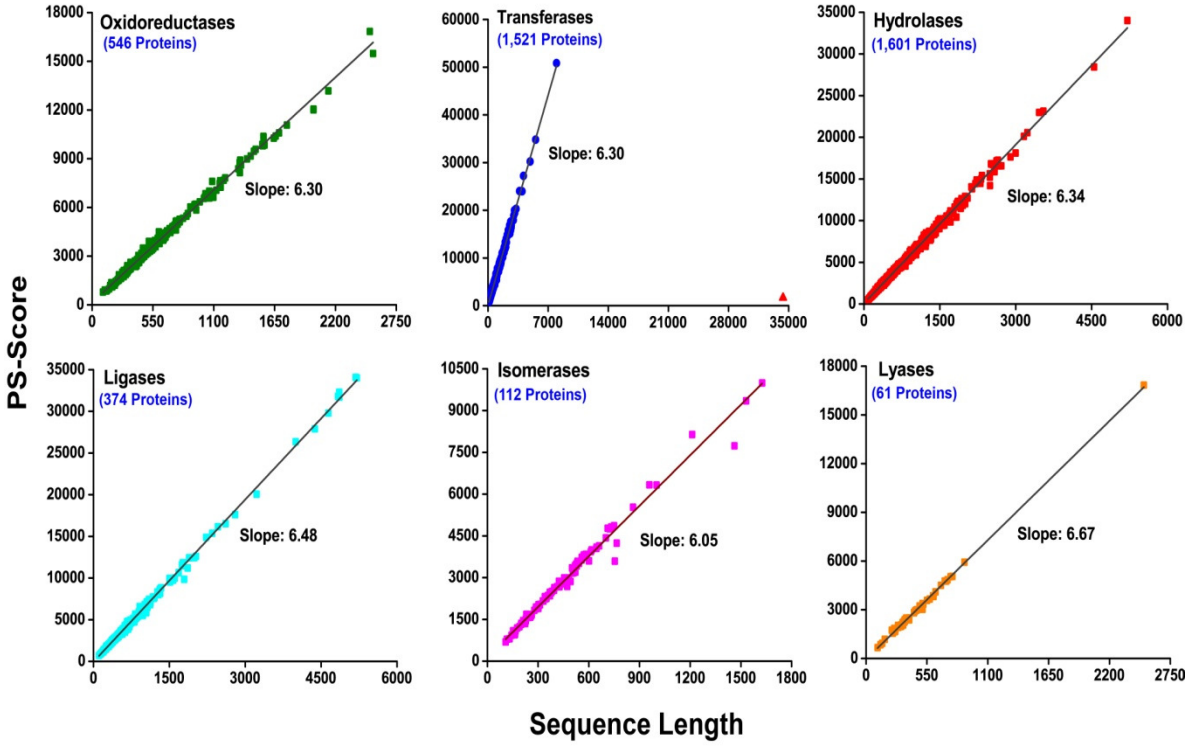


Figure 4.3.7A: PS-Score vs. Sequence length plot for different classes of Enzymes

Among the six classes of enzymes, Lysases which catalyze non-hydrolytic addition/removal of atoms from substrates had the maximum slope (6.67) while Isomerases which perform isomerization reactions had the lowest slope of 6.05. Both Oxidoreuctases (catalyzes oxidation\reduction reaction) and Transferases (carries out transfer of one functional group from one substance to another) had exactly the same slope of 6.30. They had an outlier with a PS-Score of 1651.71 and sequence length of 34,350 amino acids. Ligases which catalyze joining of two molecules had a slope of 6.48. Hydrolases which carry out hydrolysis reactions had a slope of 6.34 (Figure 4.3.7A).

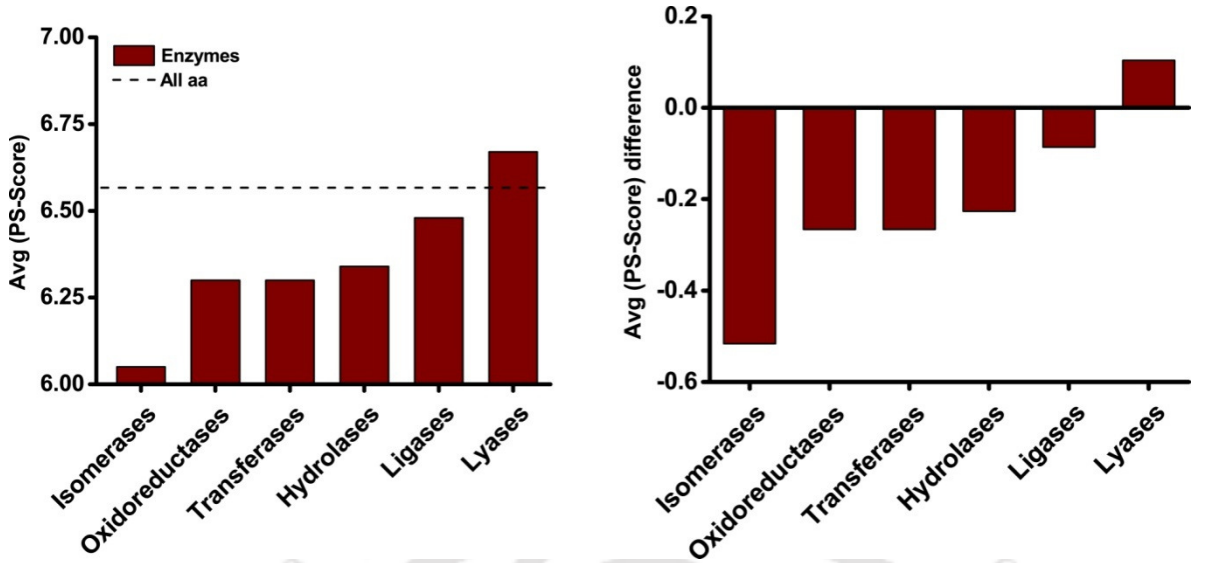


Figure 4.3.7B: Comparison of average PS-Scores of all classes of enzymes with the average PS-Score of all amino acids (left) and difference in average PS-Scores among different enzyme classes (right) compared to all amino acids

Comparison of average PS-Score with that of all amino acids (Figure 4.3.7B), showed the maximum negative difference for Isomerases while Lyases had a larger average PS-Score than all amino acids. All other classes of enzymes had negative difference, which meant that their average PS-Scores were less than that of all amino acids.

The histogram distribution for Oxidoreductases showed a unimodal distribution, with maximum population of proteins having the average PS-Scores ranging from 6-7. Ligases, Transferases and Hydrolases showed similar distribution profiles, with all the three classes of enzymes having peak at the average PS-Score of 6.5. For Isomerases, maximum number of proteins had an average PS-Score of 6.5 along with some scattered population on both sides of the distribution. The Lyases distribution also peaked at the score of 6.5 and there were few protein populations which had higher average PS-Scores of 7-7.5.This suggests that while majority of proteins in all the six classes of enzymes have an even distribution of all the amino acids, some fraction of the protein population in Isomerases are also rich in charged amino acids while some proteins belonging to the Lyase class are rich in hydrophobic amino acids (Figure 4.3.7C).

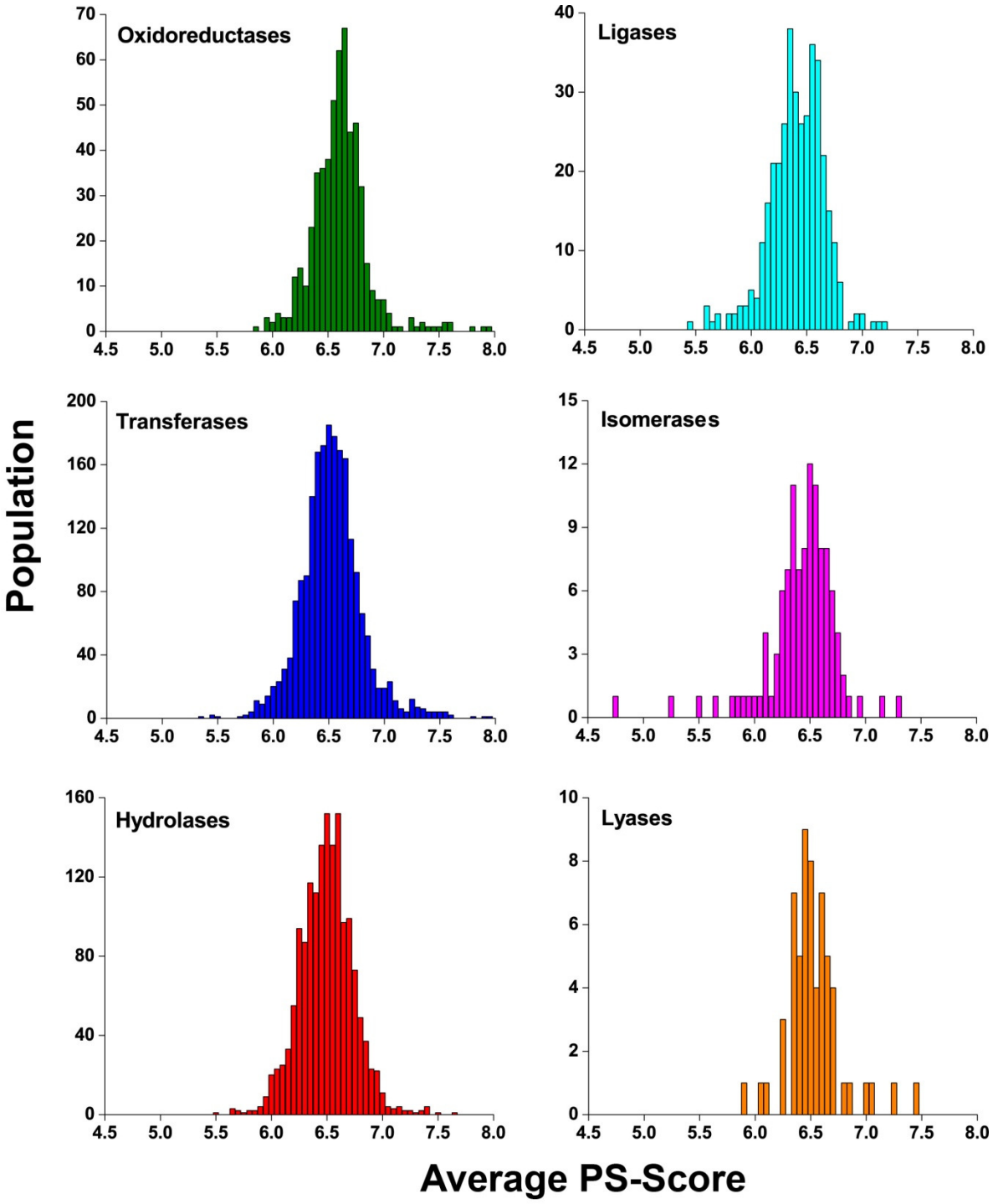


Figure 4.3.7C: Histogram distribution of average PS-Score for different classes of enzymes

4.3.8 Analysis of PS-Score across dark and non-dark proteomes

Proteome refers to the entire set of proteins coded by an organism. However, there are several proteins whose structural features and thus functions are not well understood as they are not observed by experimental structure determination and are inaccessible to homology modelling as well. These set of proteins are referred to as the "Dark Proteome" and these seem to play important role in an organism. Scientists are trying best to understand this part of the proteome. Understanding the dark proteome will provide insights into the function of the protein sequence that are not known currently[158].

Perdigão and co-workers provide a detailed study on the "dark" proteome of various classes of organisms. They found that for 546,000 Swiss-Prot proteins, 44–54% of the proteome in eukaryotes and viruses were "dark", compared to only ~14% in archaea and bacteria. They also showed that, most of the dark proteome could not be accounted for by conventional explanations, such as intrinsic disorder or transmembrane regions. They also reported that dark proteins fulfil a wide variety of functions such as disulphide bonding, proteolytic cleavage etc[159].

In an attempt to gain an insight into the amino acid composition in the proteins in dark proteome, we analyzed the dark proteome in various classes of organisms and compared them with non-dark proteome. Any additional information on the dark proteome would definitely aid in better understanding of such a diverse class of proteins.

The PS-Score analysis for the dark proteome showed that bacteria had the maximum slope of 6.67. Archea and virus also had a higher slope of 6.65 and 6.51 respectively (Figure 4.3.8A). This suggests that these dark proteome of these organisms are rich in hydrophobic amino acids. Eukaryota and human had a lower slope (6.20 and 6.29 respectively) when compared to others. This indicates that these organisms contain proteins which are rich in charged amino acids. Among the non-dark proteome, virus had the maximum slope of 6.70 while human had the least slope 6.29 (Figure 4.3.8B). Another feature to be noted is that the human dark and non-dark proteome have the same slope which hints that probably they have the same average amino acid composition.

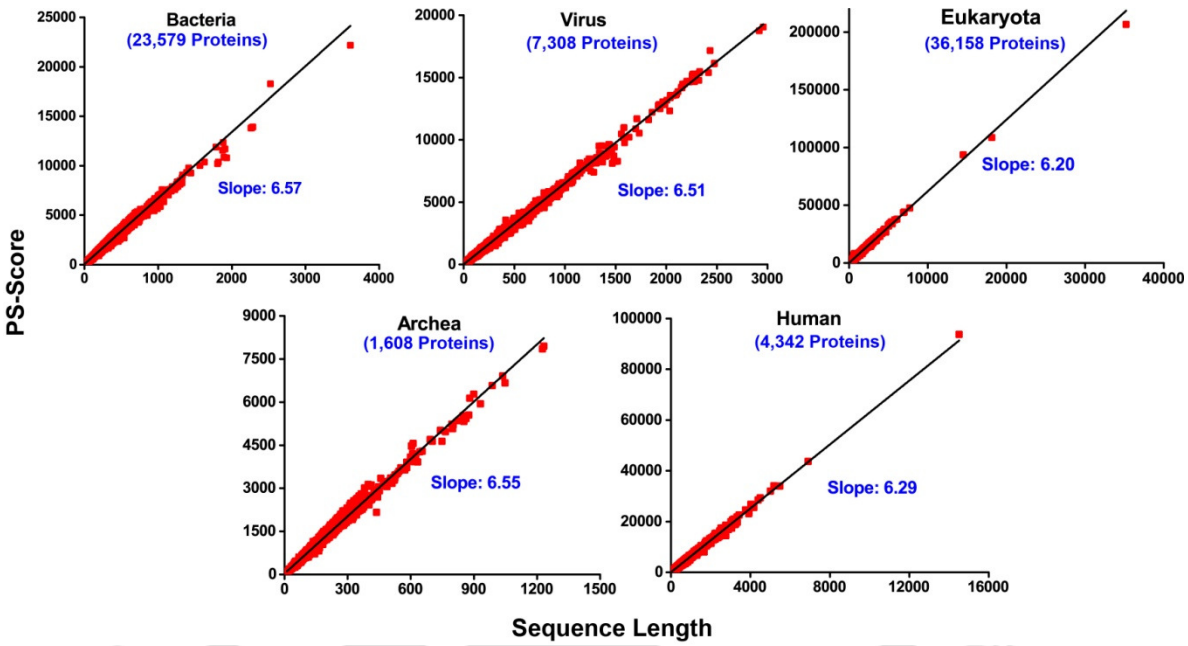


Figure 4.3.8A: PS-Score vs. Sequence length plot of dark proteome for different classes of organisms

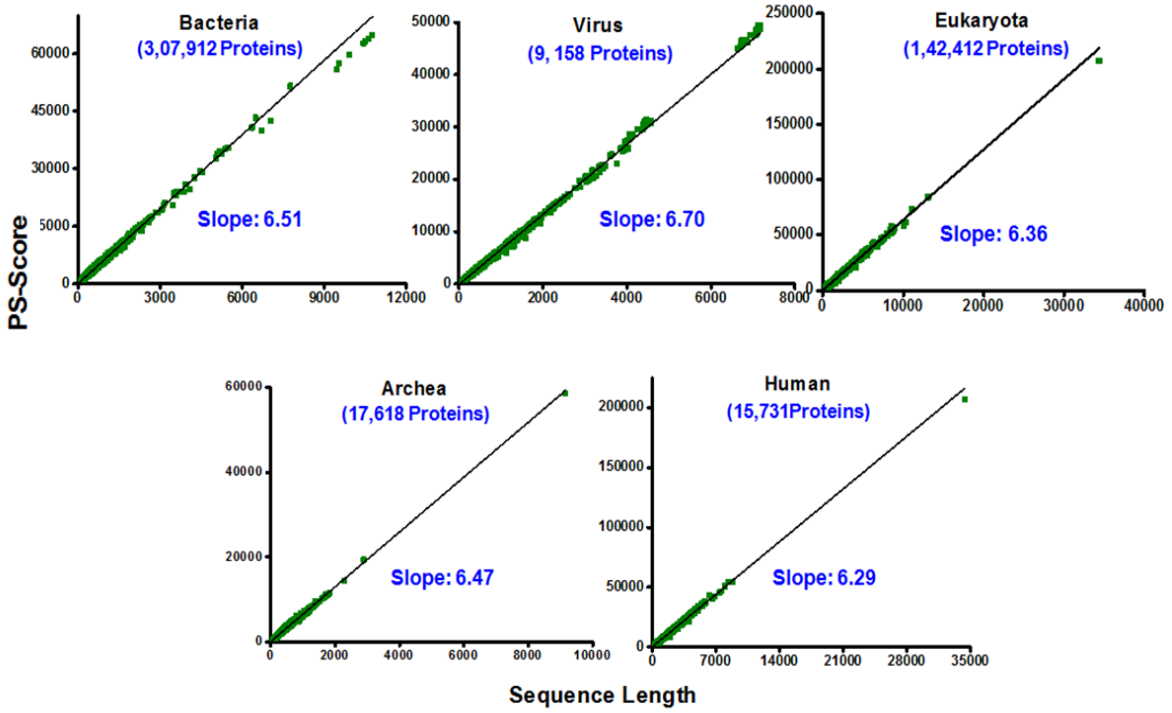


Figure 4.3.8B: PS-Score vs. Sequence length plot of non-dark proteome for different classes of organisms

Comparison of the average PS-Score with that of all amino acids score (6.56), shows that both the dark and non-dark proteome of eukaryota and human have average PS-Scores which are less than that of all amino acids (Figure 4.3.8C). Also to notice is that, the dark proteome of eukaryota has a much lower slope when compared to its non-dark proteome. This means that the dark proteome of eukaryota are rich in charged amino acids.

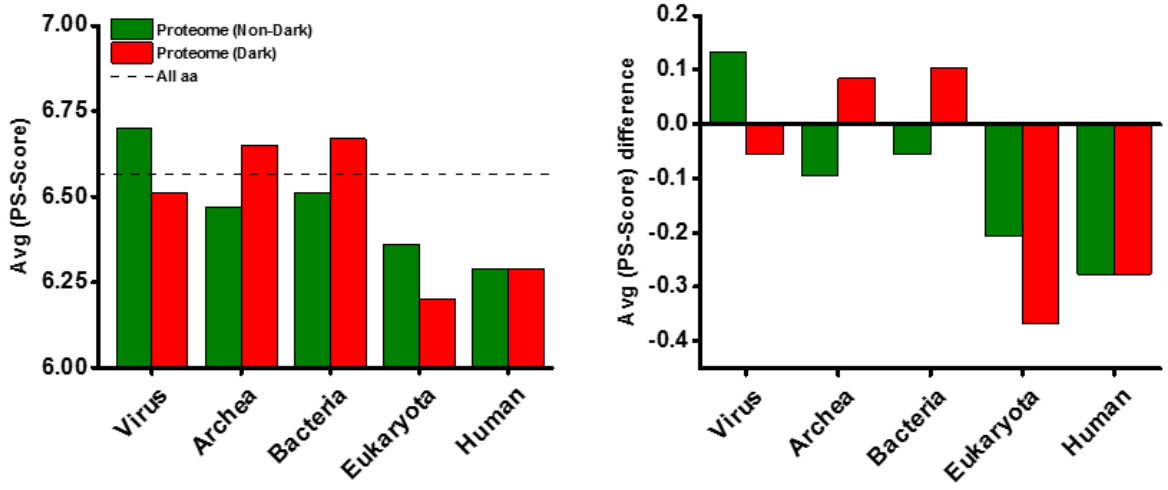


Figure 4.3.8C: Comparison of average PS-Scores for dark and non-dark proteome with the average PS-Score of all amino acids (left) and difference in average PS-Scores for dark and non-dark proteomes (right) compared to all amino acids

In case of bacteria the average PS-Score for the non-dark proteome is lower than all amino acids score, while the dark proteome has a higher value. This means that the dark proteome of bacteria are rich in proteins which contain hydrophobic amino acids. The same is true for the proteins of archea as well. The virus proteome however differs with the non-dark proteome having a much higher value (6.70) than the value of all amino acids (6.56). This suggests that the non-dark proteome of virus are rich in proteins containing hydrophobic amino acids. Further it is evident that both dark and non-dark proteomes of Eukaryota and Human have distinctly different amino acid composition compared to lower organisms like Virus, Bacteria and Archea (Figure 4.3.8C right).

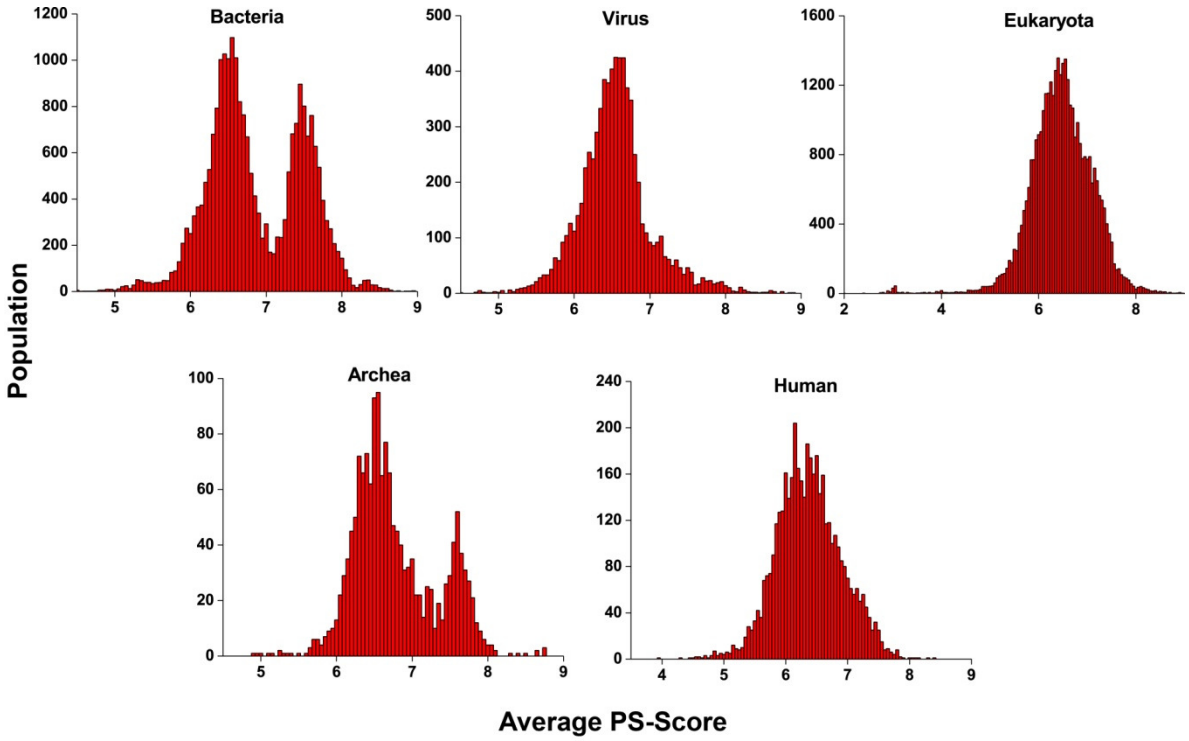


Figure 4.3.8D: Histogram distribution of average PS-Score of dark proteome for different classes of organisms

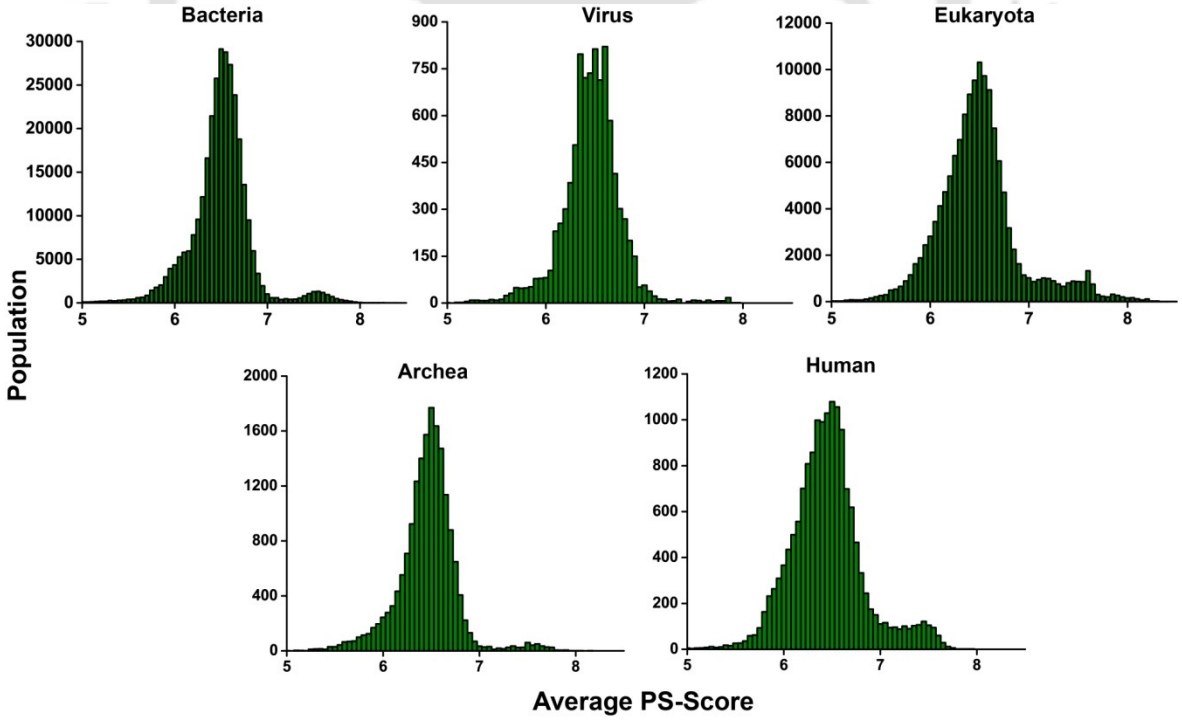


Figure 4.3.8E: Histogram distribution of average PS-Score of non-dark proteome for different classes of organisms

The histogram distributions show varied features for all the organisms. The dark proteome of bacteria has a broad distribution with scores ranging from 5-8.5. It has a bimodal distribution with scores of 6-7 and 7-8. Both of these the populations have almost equal distribution while the histogram distribution of the non-dark proteome of bacteria has a bimodal distribution with a peak average PS-Score of around 6.5. It also contains a very small but significant population with scores between 7 to 8 (Figure 4.3.8D). The histogram distributions for both dark and non-dark proteome of virus are not exactly similar with peak score being around 6.5. The virus dark proteome however shows a broader distribution with scores ranging from 5-8. The dark proteome of eukaryota have maximum population with average PS-Scores between 6-8, whereas the non-dark proteome have maximum population between 6-7. There is a small population with scores 7-8 as well. The dark proteome of archea shows a bimodal distribution with majority of population having the average PS-Scores between 6-7, while another set of population has scores which lie between 7-8. There are very few proteins which also have scores of ~8.5. This feature is not seen in the non-dark archea proteome which shows a unimodal distribution with scores ranging from 6-7 with miniscule population having scores between 7-8. The human dark proteome has a unimodal distribution with peak score ~6.5. The non-dark proteome shows a bimodal profile with addition of a small population which have scores between 7-7.75. Thus although the average PS-Score for dark and non-dark regions of human proteome were identical, their histogram distributions reveal clear region of dissimilarity in the population (Figure 4.3.8E).

This simple methodology using average PS-Score therefore allows one to quickly compare the dark and non-dark proteomes of organisms and gain information about the amino acid composition in them.

4.3.9 Visual representation of proteins

The use of PS-Score and average PS-Score methodology gives us an idea about the amino acid composition within a protein. However the information about the sequence of amino acids in a given protein cannot be deciphered from it. To circumvent this we designed some novel ways of representing the proteins sequences which would give information about the amino acid composition along within sequence for a given protein. Few examples of such representations are given below. In all cases amino acids are arranged in order of decreasing polarity.

4.3.9.1 Protein sequence represented as electrophoretic band profile

Figure 4.3.9.1A: Protein sequence map for Human Myoglobin (154 residues)

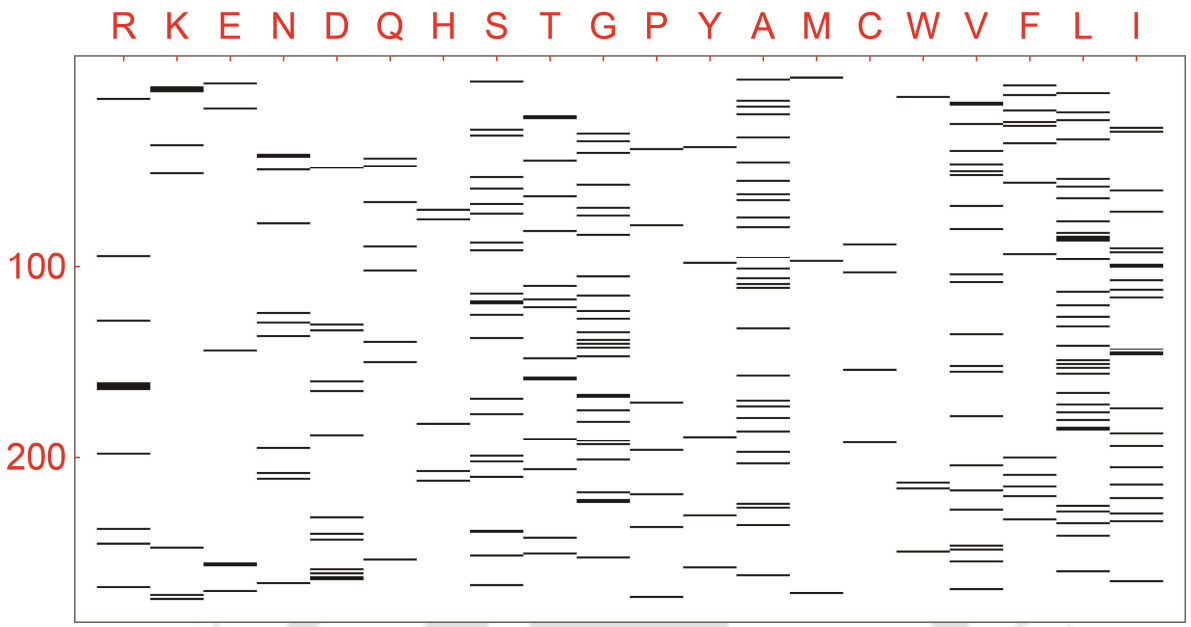


Figure 4.3.9.1B: Protein sequence map for Human Aquaporin (269 residues)

This pattern displays each amino acid in a protein as bands similar to that seen in electrophoresis (Figure 4.3.9.1A and 4.3.9.1B). The top row depicts all the 20 amino acids while the vertical axis on left depicts the residue number of amino acid. Consecutive similar amino acids appear as thicker bands in the pattern. One can therefore build on the amino acid sequence of a protein by looking at the order in which these bands appear against a given amino acid.

4.3.9.2 Protein sequences as two dimensional grid

Different protein sequences were depicted as two dimensional grids, which depict the position and population of each amino acid in the sequence. The numbers below amino acids depict the position of the respective amino acid in the protein sequence. Examples for Human Myoglobin and Alpha Synuclein are depicted in Figure 4.3.9.2.

Human Myoglobin (154 residues)

R	K	E	N	D	Q	H	S	T	G	P	Y	A	M	C	W	V	F	L	I
32	17	07	13	05	09	25	04	40	02	23	104	20	01	111	08	11	34	03	22
140	35	19	133	21	27	37	52	68	06	38	147	58	56		15	14	44	10	31
	43	28	146	45	92	49	59	71	16	89		67	132			18	47	12	76
	46	39		54	114	65	93	96	24	101		72	143			29	107	30	87
	48	42		61	117	82	109		26	121		85				69	124	33	100
	51	53		123	129	83	118		36			91				102	139	41	108
	57	55		127	153	94	145		66			95				115	152	50	112
	63	60		142		98			74			126						62	113
	64	84				120			75			128						70	
	78	86							81			131						73	
	79	106							122			135						77	
	80	110							125			144						90	
	88	137							130									105	
	97	149							151									116	
	99								154									136	
	103																	138	
	119																	150	
	134																		
	141																		
	148																		

Human Alpha Synuclein (140 residues)

R	K	E	N	D	Q	H	S	T	G	P	Y	A	M	C	W	V	F	L	I
	06	13	65	2	24	50	09	22	07	108	39	11	01			03	04	08	88
	10	20	103	98	62		42	33	14	117	125	17	05			15	94	38	112
	12	28	122	115	79		87	44	25	120	133	18	116			16		100	
	21	35		119	99		129	54	31	128	136	19	127			26		113	
	23	46		121	109			59	36	138		27				37			
	32	57		135	134			64	41			29				40			
	34	61						72	47			30				48			
	43	83						75	51			53				49			
	45	104						81	67			56				52			
	58	105						92	68			69				55			
	60	110							73			76				63			
	80	114							84			78				66			
	96	123							86			85				70			
	97	126							93			89				71			
	102	130							101			90				74			
		131							106			91				77			
		137							111			107				82			
		139							132			124				95			
												140				118			

Figure 4.3.9.2: Protein sequence map for Human Myoglobin (top) and Alpha Synuclein (bottom)

4.3.9.3 Proteins as triangles, bars on opposite strands and circles

The 2D grid approach was further extended to depict protein sequences in interesting visual graphics such as triangles, lateral strings/bars and circles. A few examples of such representations are shown here.

One way of representing proteins is as an equilateral triangle where one side of the triangle contains the 10 most polar amino acids, while the other side has the most non-polar amino acids and the bars above them represent the position and density of each amino acid in a given protein (Figure 4.3.9.3A). Protein can also be represented as two lateral strings which are anti-parallel to each other (Figure 4.3.9.3B). The upper string contains the 10 most polar amino acids (Arg to Gly) while the lower string contains the non-polar amino acids (Pro to Ile). Circular representations depict protein sequences as clocks wherein the polarity of amino acids decreases as one goes in the clock wise direction (Figure 4.3.9.3C) A quick visual glance at all these representations instantly conveys the density of charged and uncharged amino acids in a given protein sequence.

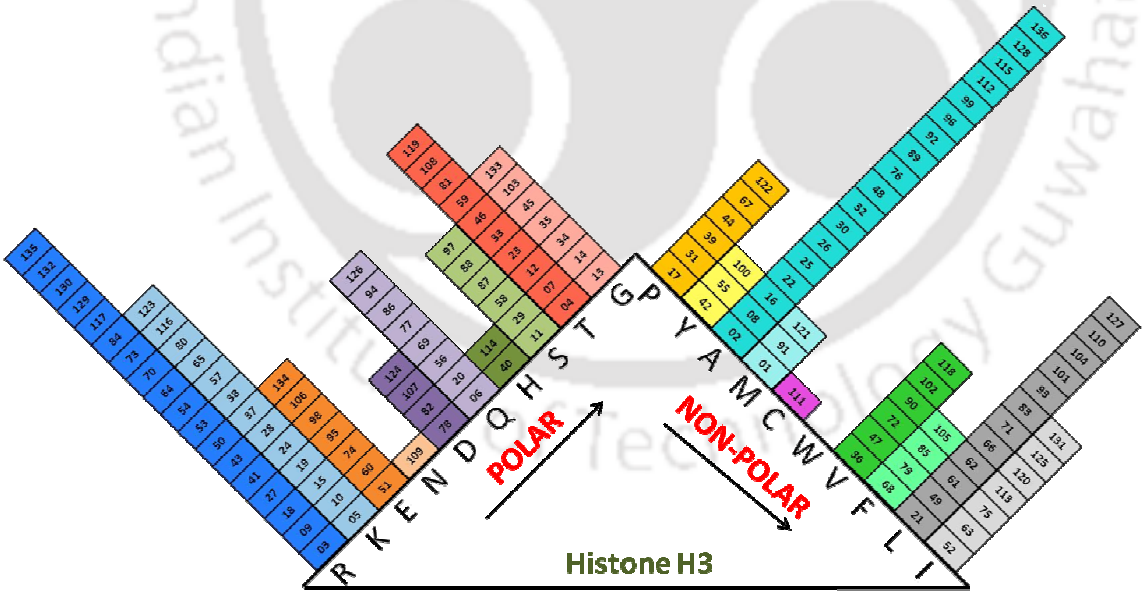


Figure 4.3.9.3A: Representation of Histone H3 as a triangle

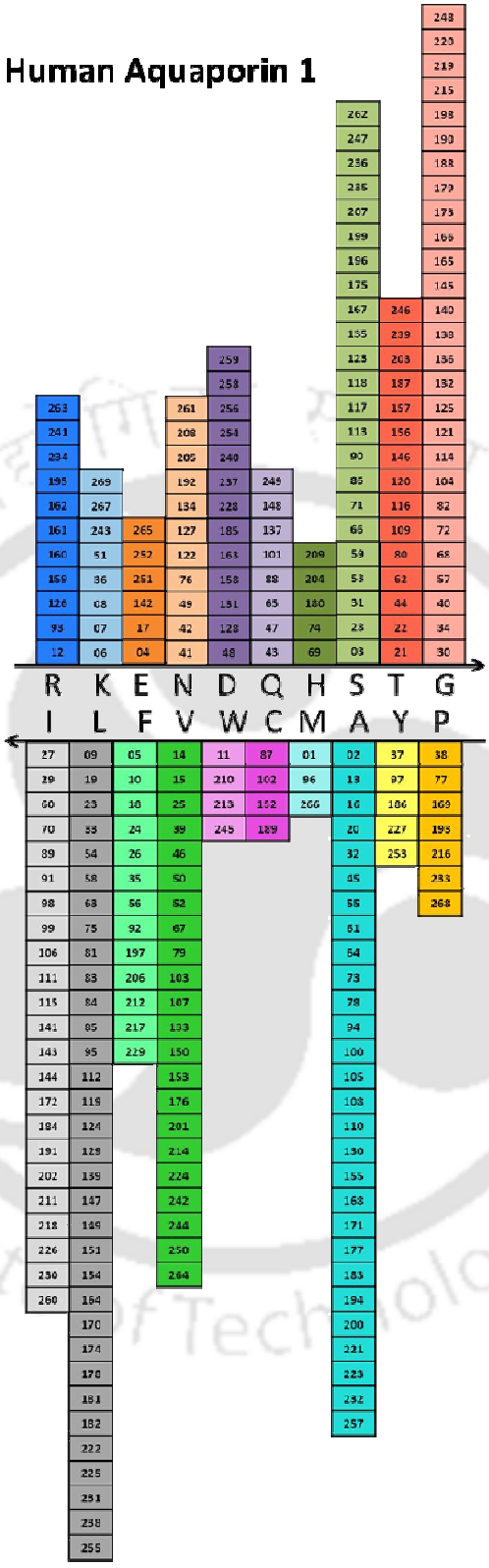


Figure 4.3.9.3B: Representation of Human Aquaporin 1 (269 residues) as bars on opposite strands

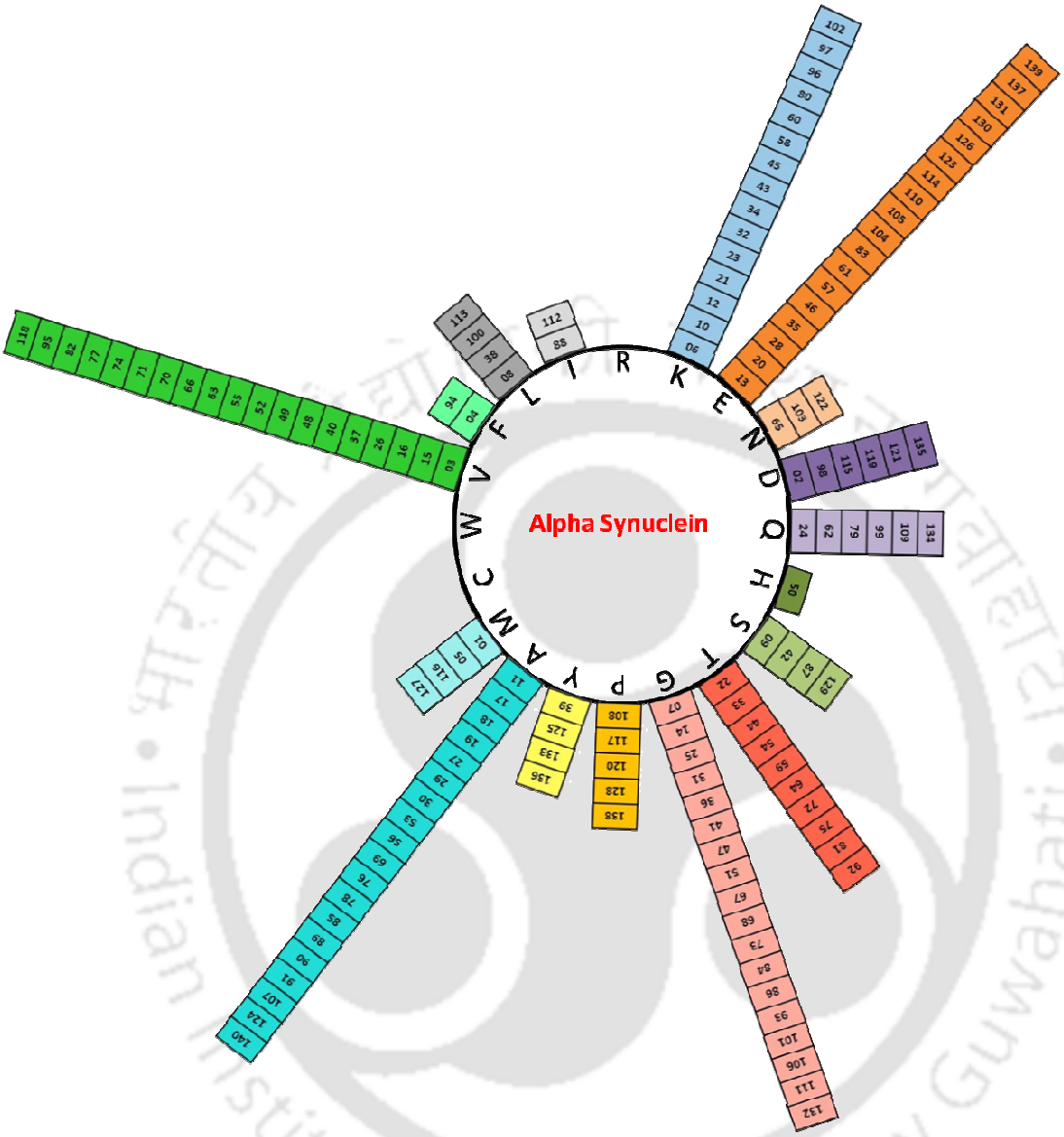


Figure 4.3.9.3C: Representation of Human Alpha Synuclein (140 residues) as a circular graphic

4.4 Conclusions:

1. A new method using prime numbers was devised which gives us a numerical parameters like protID, PS-Score and average PS-Score. This was used to identify proteins rich in charged amino acids.

2. A synthetic protein (Alpha 3C) rich in charged amino acids and devoid of aromatic amino acids suited for our studies was identified using the PS-Score methodology.

3. Analysis of amino acid content and histogram distribution was carried out for wide variety proteins. These provided interesting insights into the composition of proteins across organisms.

4. Novel tool to represent proteins as visual graphic was developed which equips us for a quick visual analysis of protein content and helps in comparison of various protein sequences.

4.5 Implications of the work:

In the hunt for proteins rich in charged amino acids, we devised a new scoring methodology which can segregate proteins on the basis of their polarity. Each amino acid was numerically coded with a unique prime integer which was assigned using the hydrophobicity index of all amino acids. Using the prime number assignments, we can compute the product (ProtID) of all amino acids in sequence and its base 2 logarithm value (PS-Score). The ProtID quantitatively stores the protein sequence composition as the prime product factorization yield unique factors. The PS-Score is a cumulative sum of the hydrophobicity index of all constituent amino acids in the protein. This score presents a convenient handle to sort all polypeptide sequences in a proteome in a hierarchy, reveal their average hydrophobicity and display the latter as a histogram distribution against population of entire proteome. This technique can be used as a big data analysis tool for proteomes which contain millions of protein sequences.

CHAPTER 5

INVESTIGATIONS ON α_3C PROTEIN: EXPERIMENTAL AND THEORETICAL ASPECTS

5.1 Introduction:

Previous studies have reported unusual absorption beyond 320 nm in Lys-rich proteins (chapter 1) but the exact origin/nature of such transitions have remained elusive. Also our studies have showed unique absorption features of charged amino acids in comparison to their uncharged counterparts (chapter 3). These observations motivate a systematic investigation of protein absorption spectra beyond 320 nm and its dependence on charged amino acid content. Given the strong prominent spectral features of Trp, Tyr and Phe, it is desirable to investigate the UV-Vis absorption spectra in a model protein devoid of aromatic amino acids.

Alpha 3C (α_3C) discovered by our search (chapter 4) is an excellent protein to investigate these unusual spectral signatures. It is a small (67 residue), monomeric synthetic protein devoid of aromatic amino acids and rich in charge amino acids. It is a three helix bundle rich in charged amino acids (54% of the sequence) which comprise of 17 Lys, 17 Glu, and 2 Arg residues[160]. Further, several Lys residues in α_3C are in close proximity making it likely that Lys associated spectra may be observable here.

We carried out systematic experimental and theoretical studies (MD simulation and electronic structure calculations) on α_3C in order to gain better insights into the mechanisms involved. **Theoretical studies were carried out in collaboration with Dr. Ravindra Venkatramani at Tata Institute of Fundamental Research, Mumbai. I carried out MD simulations and related analysis on α_3C while Imon Mandal (Graduate student in Dr. Venkatramani's lab) carried out electronic structure calculations using Time Dependent Density Functional Theory (TDDFT).** Elaborate calculations on the extracted amino acid fragments from α_3C reveal that the unique spectral features arise due to charge transfer transitions involving the charged amino acids (Lys and Glu) of α_3C and the peptide backbone.

5.2 Materials and Methods:

5.2.1 Materials:

Full length plasmid for α_3C cloned in a modified pET-32b vector was a generous gift from Dr. Cecilia Tommos, University of Pennsylvania. GenElute^{TM} Plasmid miniprep kit (PLN70), SDS-PAGE Gel loading buffer (G2526), BSA (A3059), HEWL (L6876), HSA (A1887), poly-L-Lys hydrochloride (P2658, Mol wt: 15,000-30,000 Da), Phenylmethylsulfonyl fluoride (PMSF) (P7626), Calcium chloride (C8106), Ampicillin sodium salt (A8351), Isopropyl β-D-1-thiogalactopyranoside (IPTG) (I5502), Acrylamide (A3553), Ammonium persulphate (APS) (A3678) and N,N,N′,N′ Tetramethylethylenediamine (TEMED) (T7024) were purchased from Sigma Aldrich. Sinapic acid (85429) was purchased from Fluka. Human Alpha Thrombin (C00129) was obtained from Haematologic Technologies Inc. Nickel-Nitrilotriacetic acid (Ni-NTA) agarose (30210) was procured from Qiagen. Luria Bertani (LB) (M1245), Luria Agar (M557), Terrific Growth Medium (G004), Imidazole (GRL1864), Copper (II) sulphate pentahydrate (MB238), Potassium sodium tartrate tetrahydrate (GRM598), Folin's Reagent (ML059) and Magnesium chloride (TC186) were procured from HiMedia. Tris buffer (17714), Sodium chloride (93206), Sodium hydroxide pellets (93102), Sodium carbonate anhydrous (17844), Sodium dodecyl sulfate (SDS) (18419) were purchased from Merck.

5.2.2 Experimental Methods: Expression, purification and characterization of α_3C protein

5.2.2.1 Competent cell preparation

Two types of competent cells were prepared. *E.coli* DH5-α which was used for plasmid isolation and *E.coli* BL-21 (DE3) which was used for protein expression. Either a single bacterial colony was picked up from a Luria Agar plate or a very small volume of inoculum from previously prepared competent cell preparation was taken and inoculated into 5 mL of LB medium. The primary culture was allowed to grow overnight at 37 °C at 180 rpm. 500 µL (1% of the final culture volume) of this starter culture was used to inoculate 50 mL of LB medium. This culture was allowed to grow at 37 °C at 180 rpm till the OD of the culture

reached approximately to 0.3 or 0.4 at 600 nm. This culture was then incubated in ice for 10 minutes. Cells were harvested by centrifugation at 3000 rpm for 10 minutes at 4 °C. The supernatant was discarded and cells were thoroughly resuspended gently by pipetting/inverting in 15 mL of ice cold 80 mM $MgCl_2$+ 20 mM $CaCl_2$ solution. Cells were pelleted again at 3000 rpm for 10 minutes at 4 °C. The supernatant was discarded and 900 µL of 0.1 M $CaCl_2$ and 100 µL of glycerol were added to the pellet. The cells were re-suspended and kept as aliquots of 100 µL in microcentrifuge tubes. The competent cells were stored at -80 °C until further use. All the materials were autoclaved and all the steps were carried out in laminar air flow to maintain proper sterile conditions.

5.2.2.2 Transformation

Competent cells taken out from -80 °C were thawed in ice for 10 minutes. 1-2 µL of the desired DNA was added to an aliquot of the competent cells and was incubated in ice for 30 minutes. One aliquot of the competent cells served as negative control in which no DNA was added. Meanwhile the water bath was switched on and temperature was set at 42 °C. Heat shock was given to the cells for 60 seconds at 42 °C after which they were immediately transferred to ice and incubated for another 2-3 minutes. 800 µL of sterile LB media (without any antibiotic) was then added to the cells and was incubated at 180 rpm for 45 minutes at 37 °C. The cells were then centrifuged at 13,000 g for 10 minutes. Supernatant was discarded and the pellet was resuspended in 100 µL of fresh LB broth. The transformants were then plated on a LB agar plate containing Ampicillin. The plates was inverted and kept in an incubator for 12 hours at 37 °C. All the materials were autoclaved and all the steps were carried out in laminar air flow to maintain proper sterile conditions.

5.2.2.3 Plasmid Isolation

GenElute™ Plasmid miniprep kit was used to isolate DNA from the transformed *E.coli* DH5-α cell. A transformed colony was picked up to inoculate 5 mL of LB media containing 5 µL of Ampicillin (0.1% of the final culture volume). This starter culture was allowed to grow overnight at 37 °C at 180 rpm in an incubator. 200 µL (1% of the final culture volume) of this starter culture was used to inoculate 20 mL of LB medium containing 20

µL of Ampicillin. This culture was allowed to grow at 37 ℃ at 180 rpm till the OD of the culture reached approximately to 0.4 or 0.5 at 600 nm. It was then divided in 5 microcentrifuge tubes each containing 2 mL of the culture. All the steps from here on were performed at room temperature. The cells were pelleted down by centrifugation at 13,000 g for 2 minutes. Supernatant was discarded and the remaining culture was added to the tubes and cells were pelleted down. The cells were now resuspended by vortexing /pipetting in 200 µL of Resuspension Solution (with RNase solution) in each tube till the solution became homogenous. The resuspended cells were then lysed by adding 200 µL of lysis solution and the contents were mixed immediately by gentle inversion (6-8 times) until the mixture became clear and viscous. It was made sure that the lysis reaction did not exceed 5 minutes as prolonged lysis may permanently denature the plasmid DNA. The cell debris was precipitated by adding 350 µL of Neutralization/Binding solution. The tubes were inverted gently for 4-6 times and the cell debris was pelleted by centrifugation at 12,000 g for 10 minutes. The supernatant from each tube was then carefully transferred to a fresh tube and pooled together. In the meanwhile, the GenElute Miniprep Binding Coulmns were inserted into a microcentrifuge tubes provided with the kit. 500 µL of the Column Preparation Solution was added to each miniprep column and centrifuged at 12,000 g for 1 minute and the flow through was discarded. The clear lysate (supernatant collected earlier) was now transferred to these columns and centrifuged at 12,000 g for 1 minute. The flow through liquid was discarded. This was followed by addition of 500 µL of the Optional Wash Solution (with ethanol) and centrifugation at 12,000 g for 1 minute. The flow through liquid was discarded. 750 µL of diluted Wash solution was then added to the column and centrifuged at 12,000 g for 1 minute. The flow through liquid was discarded and the contents were again centrifuged at 12,000 g for 2 minutes without any additional Wash Solution to remove excess ethanol. The columns were now transferred to fresh collection tubes. 100 µL of Elution Solution was then added to the columns and it was allowed to incubate for 2 minutes. The desired plasmid DNA was eluted by centrifugation at 12,000 g for 1 minute. The plasmid DNA which was present in the Flow through liquid (eluate) was collected and stored at -20 ℃ until further use. The sequence of the plasmid was later confirmed by sequencing by SciGenom Labs Private Ltd., Kerala, India.

5.2.2.4 Protein Expression

The protein was expressed as thioredoxin fusion using a modified pET32b vector transformed into BL-21(DE3) cells. One isolated colony of transformed BL-21 (DE3) cells was picked up from the LB agar plate and inoculated in 5 mL of LB media containing 5 µL of Ampicillin. The starter culture was allowed to grow overnight at 37 °C, 180 rpm in an incubator. 500 µL of the starter culture was then used to inoculate another 5 mL of LB media containing 5 µL of Ampicillin. The culture was allowed to grow for 3-4 hours and then it was used to inoculate 400 mL of Terrific Broth containing 400 µL of Ampicillin. The culture was grown at 37 °C at 180 rpm in an incubator till the OD of the culture reached 0.6 at 600 nm. The culture was then induced with 1 mM (final concentration for the induction of the desired gene) Isopropyl β-D-1-thiogalactopyranoside (IPTG). The culture was induced for 4 hours at 37 °C at 180 rpm in an incubator. The induced cells were harvested by centrifugation at 4000 rpm for 20 minutes at 4 °C. The pellet was stored at -20 °C till purification.

5.2.2.5 Protein Purification

The pelleted cells were resuspended in lysis buffer (see Appendix) and phenylmethylsulfonyl fluoride (1% of the final volume) was added to avoid non-specific protease cleavage. The resuspended cells were then treated with Hen egg-white lysozyme (300 µg/ mL) for 30 min at 4 °C. Lysis of cells was done in ice by sonication at 35% output, with cycles of 4 seconds on and 10 seconds off pulse for about 20 minutes. Cell lysate was clarified by centrifugation at 10000 rpm for 20 minutes at 4 °C. The supernatant was then filtered with 0.45 µm filter. The filtrate containing the protein of interest was further purified using Ni-NTA affinity chromatography.

5.2.2.5.1 Ni-NTA Affinity Chromatography

About 2 mL of Ni-NTA agarose beads were taken in a column (Column Volume: 10 mL) for cells obtained from 1 L culture. The beads were equilibrated with 2-3 column volumes of lysis buffer to remove ethanol. The cell lysate was then added and kept for binding with the beads for 4-5 hours at 4 °C on an end to end rocker. The unbound lysate was allowed to

pass through the column collected as flow through. The beads bound to the lysate were then washed first with lysis buffer and then with wash buffer (see Appendix). All the different wash fractions were collected. The thioredoxin fusion protein was eluted by running the gradient of imidazole from 50 mM to 400 mM. All these fractions were collected as elute fractions and run on SDS-PAGE. The fractions containing comparatively pure protein were pooled together and dialyzed against thrombin cleavage buffer (see Appendix) overnight at room temperature.

5.2.2.5.2 Thrombin cleavage

The concentration of the fusion protein was determined spectrophotometrically using the molar extinction coefficient of 13,980 $M^{-1}cm^{-1}$ for the fusion protein. Thrombin was added to the fusion protein with thrombin/protein ratio 1:2000 (w/w). The resulting mixture was then dialyzed against thrombin cleavage buffer for 16 hours at room temperature. The digested mixture was passed over a column containing Ni-NTA agarose beads pre-equilibrated with thrombin cleavage buffer. The contents were kept for binding for 4 hours at room temperature to remove the His-tagged thioredoxin and any remaining undigested fusion products. Cleaved product was collected, dialyzed against deionised water and stored as lyophilized powder for further use. The protein purification steps were monitored by SDS-PAGE and the purity of the samples was evaluated by Mass spectrometry and reverse phase HPLC.

5.2.2.5.3 SDS-PAGE

The method SDS-Polyacrylamide Gel Electrophoresis was carried out according to the protocol reported in Sambrook and Russel, Molecular Cloning-A laboratory manual[161]. 15% resolving gels (see Appendix) were prepared for running all the samples. Protein samples were prepared in Gel loading buffer followed by boiling at 95 ℃ for 5 minutes and were separated on Mini-PROTEAN^{R} Tetra Electrophoresis System (Make: Bio-Rad). All the gels were run for 2 hours at 80 Volts. The gels were stained with colloidal Coomassie stain. After proper staining the gels were then de-stained overnight in the de-staining solution (see Appendix) and the approximate protein size was determined by comparing the migration of the protein band with those of standard molecular markers.

5.2.2.6 Protein Estimation

The concentration of protein was determined using Lowry's method[33]. BSA was used for plotting the standard curve. Fresh stock solution of BSA (1 mg/mL) was prepared in deionised water. Various dilutions ranging from 0.05 mg/mL to 1 mg/mL were prepared from this stock solution (Final volume: 1 mL). 200 µL of sample from each dilution of BSA was added to different tests tubes wrapped with Aluminum foil. α_3C was taken in a separate tube. All the samples were taken in duplicates. Deionised water was taken as blank. 2 mL of Reagent I (see Appendix) was added to each of the tubes and mixed thoroughly using a Vortexer. All the tubes were incubated at room temperature for 10 minutes. 200 µL of Reagent II (see Appendix) was then added and the contents were stirred. The samples were then incubated for 30 minutes at room temperature in dark. The amount of protein in each sample was determined spectrophotometrically by recording absorbance at 650 nm. The approximate concentration of α_3C was determined by the slope obtained from the standard plot of BSA.

Protein concentration for α_3C was also cross verified by far-UV absorbance[162] using equation:

$$Protein\ concentration\ (\mu g/mL) = 144\ (A_{215} - A_{225}) \quad (5.1)$$

where, A_{215} and A_{225} are absorbance at 215 nm and 225 nm respectively.

5.2.2.7 Mass Spectrometry

A saturated solution of Sinapic acid (matrix) was made in TA-30 solvent. TA-30 solvent was prepared by mixing 0.1% TFA with Acetonitrile in 7:3 ratios. The contents were then sonicated for 15 minutes in a water bath. The contents were then centrifuged at 13,000 rpm for 15 minutes at room temperature. The resultant supernatant was then stored at 4 °C for further use. The protein of interest was then dissolved in the matrix in ratio of 1:2. Around 2 µL of this sample was then spotted on the MALDI plate and was allowed to dry. The mass of the protein was then measured on a MALDI-TOF machine (Make: Daltonics Bruker).

5.2.2.8 Reverse Phase HPLC

The purity of the samples was further evaluated by reverse phase HPLC (Make: Waters 600E RP-HPLC). Enable C18G column with particle size of 5 µm and column size of 250 X 4.6 mm was used. Binary solvent system were used, solvent A (0.1 % TFA in Water) and solvent B (0.1 % TFA in Acetonitrile). The samples were run with linear Water/Acetonitrile gradient of 20-70% Acetonitrile over 30 minutes with a flow rate of 1 mL per minute.

5.2.2.9 Absorption Spectroscopy

Absorption spectra for α_3C were recorded at different concentrations according to the procedure mentioned in Chapter 2. The α_3C protein was dissolved in deionised water and the absorption spectra were recorded for different concentrations (5-105 µM) of the protein. Pure deionised water was kept as blank control for the measurements. Spectra were acquired with multiple scans (3-5) and averaged subsequently. For recording spectra at 85°C, Varian Cary-100 double beam spectrophotometer equipped with a Peltier-based sample temperature controller was used. The sample was thermally equilibrated at high temperature for at least 30 minutes prior to recording absorption spectra.

5.2.2.10 Circular Dichroism

Circular dichroism measurements were carried out at different temperatures on a spectropolarimeter (Make: Jasco, Model: J-1500, Jasco Inc., Maryland, USA). The scan was recorded from 300- 190 nm with data pitch of 0.1 nm, bandwidth of 2 nm; thinning scale was kept at 9. Three scans were recorded for each sample and deionised water served as blank in all the cases. Quartz Cuvette with 1 mm path length with transmission range up to 190 nm was used for recording all the measurements.

5.2.3 Computational Methods:

5.2.3.1 Molecular dynamics (MD) simulations of α_3C

Molecular dynamics simulation was performed on α_3C using the NAMD program[163] (version 2.9) and the CHARMM27 force field[164]. The initial structure used in the simulations was an NMR derived structure (PDB code: 2LXY) captured with Mercaptophenol ligated at the C32 site. The ligand was removed during processing to carry out simulations of Mercaptophenol free α_3C. The PDB structure had 31 frames of α_3C and we chose frame 15 (this frame had the maximum number of Lys residues within 10 Å of each other) as the reference structure for simulations. First hydrogen atoms were added to the structure using the psfgen utility in VMD program[165] and the protein was solvated (TIP3P water model) inside a rectangular water box of dimensions ~67x56x60 $Å^3$. The system was neutralized by adding 2 Cl^- ions. Our simulations employed periodic boundary conditions with the Particle Mesh Ewald method for describing electrostatic interactions. The van der Waals forces were calculated with the use of a switching function that has 10 Å as switching distance and 12 Å cutoff. The equilibration protocol comprised of an initial 10000 step energy minimization step, followed by gradual heating from 0 K to 300 K over 50 ps (steps of 6 K/1 ps) and thermal equilibration at 300 K for another 50 ps. These steps were initially performed with the protein heavy atoms fixed (unconstrained hydrogen atoms, ions and waters) and then with the harmonic constraint of 25 kcal/mol/$Å^2$ on the protein heavy atoms. This was followed by constant pressure and constant temperature (NPT) equilibration run for 150 ps to stabilize the density of the system at 1 atm pressure and temperature 300 K. The pressure of the system was maintained by the Nose–Hoover method in combination with Langevin Dynamics to control the temperature of the system. The NPT protocol was repeated 3 more times by lowering the harmonic constraints set to 12, 6 and 3 kcal/mol/$Å^2$. Then an unconstrained 200 ps NPT (NPT-free) run was performed at 300 K to equilibrate the system. Finally a 110 ns MD NPT production run was carried out generating snapshots at interval of 2 ps. The α_3C protein structure was found to be stable in the 3-helix bundle form along the trajectory.

5.2.3.2 Electronic Structure Calculations

We computed the absorption spectra of amino acids within α_3C using 100 protein snapshot structures sampled from the last 100 ns of the MD production run. For each snapshot the atomic coordinates of specific amino acid fragments (monomers and dimers) were extracted and the dangling bonds were capped using the psfgen module with modified C terminus (CHO group) and N terminus (NH_2 group) in VMD. Absorption spectra were calculated then for each geometry using Time Dependent Density Functional Theory (TDDFT)[166,167] with a CAM-B3LYP[168] functional and the 6-31++G (d) basis set on all atoms in the Gaussian 09 program[169]. All the calculations were carried out with a vacuum dielectric. Difference electron density plots were calculated using the Multiwfn 3.3.8 software[170] and visualized in GaussView 5.0[171]

5.2.3.3 Characterization of transitions

Two measures were used to characterize the transitions as charge transfer (CT) transitions or non-CT transitions. The first measure is the average hole-electron separation distance, Δr:[172],

$$\Delta r = \frac{\sum_{i,j}(K_i^j)^2 \, |\langle \Phi_j | \boldsymbol{r} | \Phi_j \rangle - \langle \Phi_i | \boldsymbol{r} | \Phi_i \rangle|}{\sum_{i,j}(K_i^j)^2} \tag{5.2}$$

where ϕ is the orbital wavefunction and the index i and j go over all occupied and vacant molecular orbitals respectively. Here $K_i^j = P_i^j + Q_i^j$ where P_i^j and Q_i^j are excitation (i→ j) and de-excitation (j← i) configuration coefficients. The second measure is the distance between the centroid of the hole and electron distribution D_{CT}[173,170] and defined as:

$$D_{CT} = \sqrt{(D_{CT,X})^2 + (D_{CT,Y})^2 + (D_{CT,Z})^2} \tag{5.3}$$

With

$$D_{CT,\alpha} = | \int \alpha \rho^{electron}(r) dr \; - \int \alpha \rho^{hole}(r) dr \; | \tag{5.4}$$

where, the index α represents cartetsian components (X,Y,Z) and $\rho^{electron/hole}$ is the electron/hole density distribution.

The two measures were calculated within Multiwfn 3.3.8 and then the following conditions are used to categorize the transitions: i) for $\Delta r > 2$ Å and $D_{CT} > 1$ Å, the excitation was classified as a CT transition wherein the hole and electron distributions are spatially separated, ii) for $\Delta r = < 2$ Å or if $\Delta r > 2$ Å and $D_{CT} < 1$ Å, the transition was classified as non-CT transitions.

5.3 Results and Discussion:

5.3.1 Purification and characterization of α_3C

α_3C is cloned between the BamH1 and EcoR1 sites of the modified expression vector pET32b. The sequence of the cloned plasmid was confirmed by sequencing using the T7 forward and reverse primers.

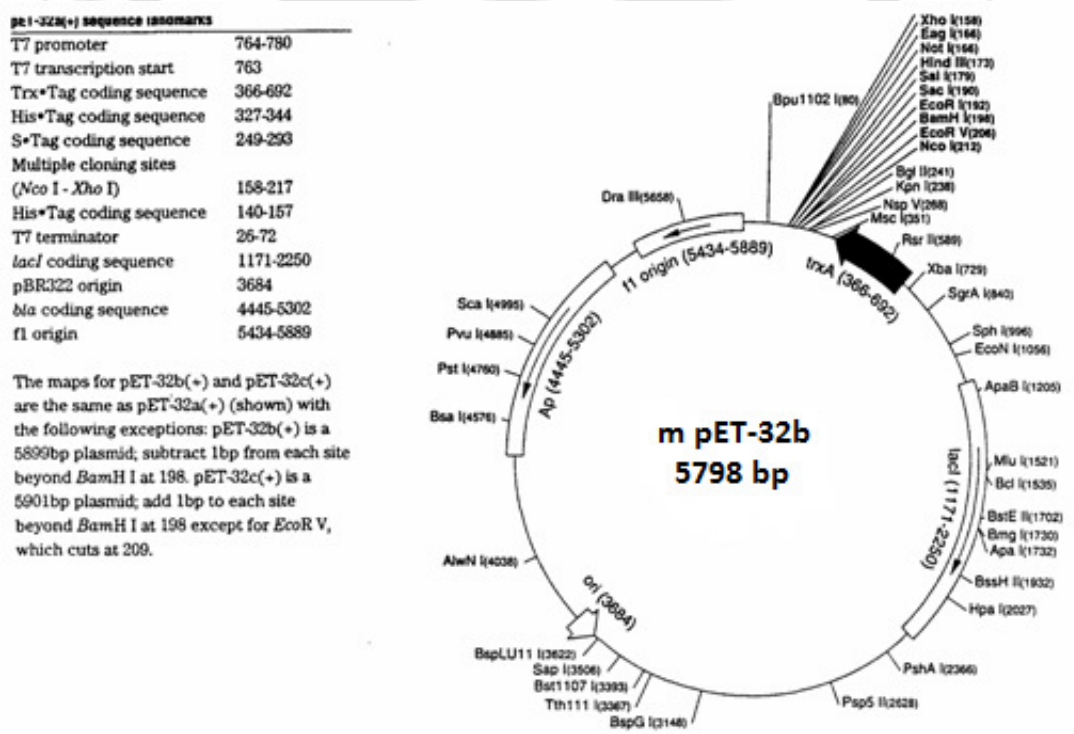


Figure 5.3.1A: Vector map for modified pET-32b used for cloning of α_3C (*Courtesy: Dr. Cecilia Tommos, University of Pennsylvania, USA*)

The protein was expressed as discussed earlier. Subsequently, the purification for protein was carried out. The purification of protein was monitored at each step by SDS-PAGE (Figure 5.3.1B). Purity was further confirmed by reverse phase HPLC (Figure 5.3.1C) and the mass ascertained by Electrospray Mass Spectrometry to be 7462.883 Da (Figure 5.3.1D). Approximate yield of the protein after purification was about 5 mg per L of culture.

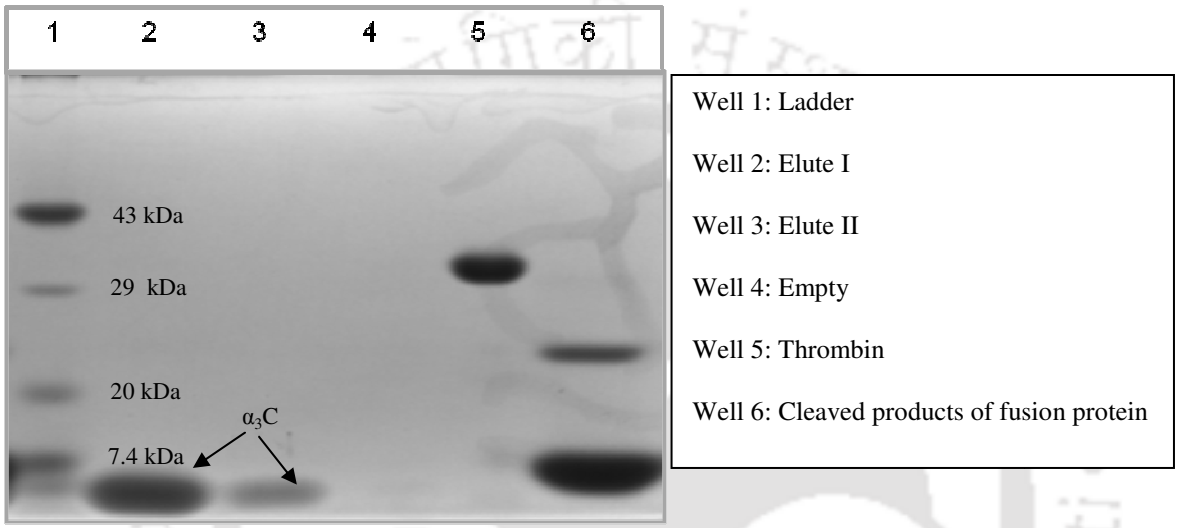


Figure 5.3.1B: 15 % SDS-PAGE to show purified band of α_3C (well 1 and 2)

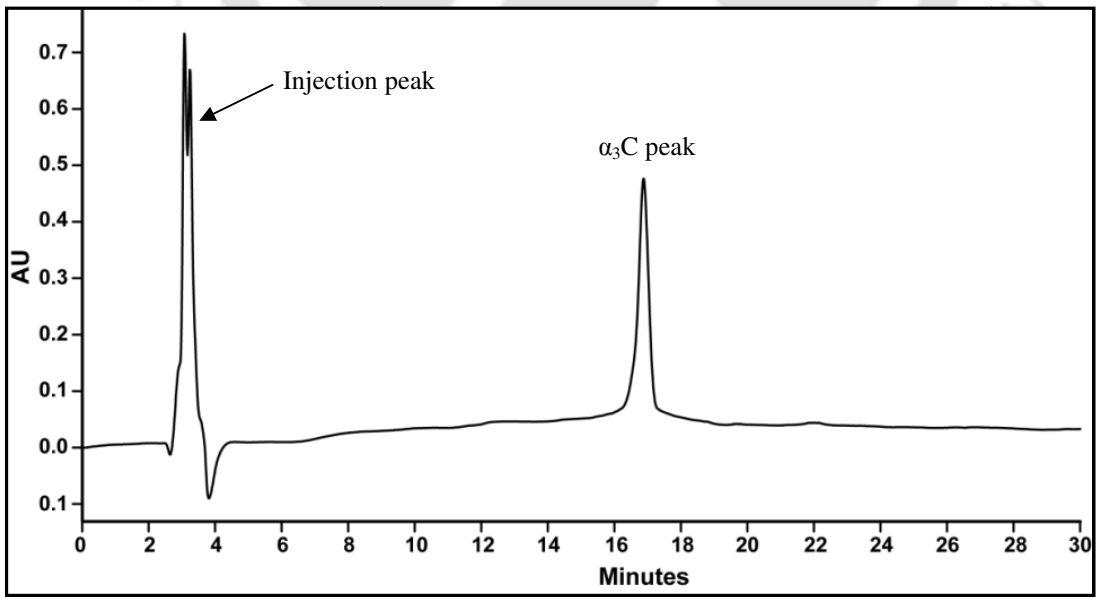


Figure 5.3.1C: Reverse phase HPLC profile for α_3C

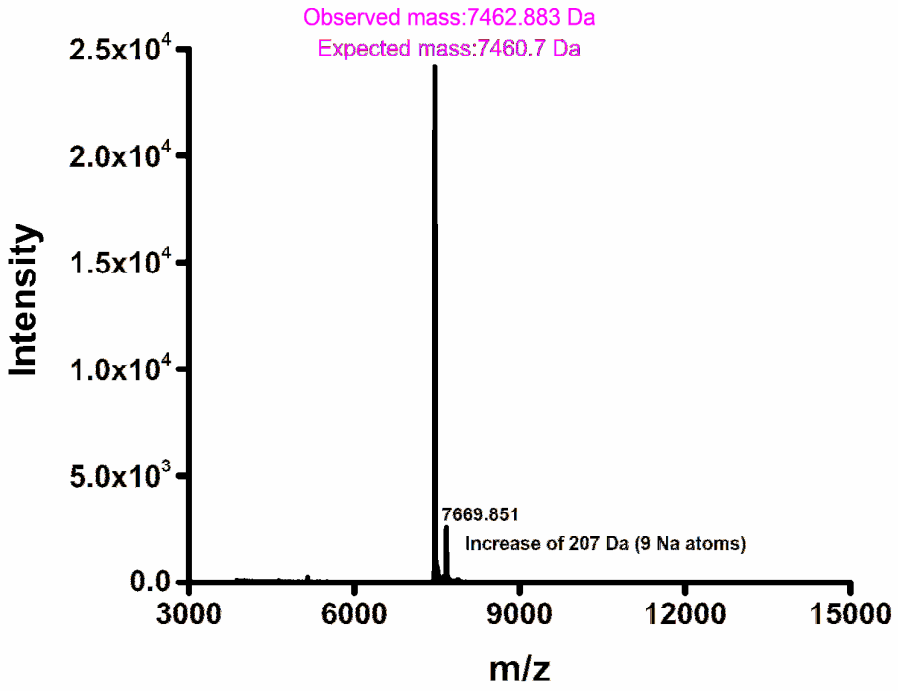


Figure 5.3.1D: Mass spectrum of α_3C

5.3.2 Absorption spectrum of α_3C

GSRVKALEEK$_{10}$VKALEEKVKA$_{20}$LGGGGRIEEL$_{30}$KKKCEELKKK$_{40}$IEELG
GGGEV$_{50}$KKVEEEVKKL$_{60}$EEEIKKL

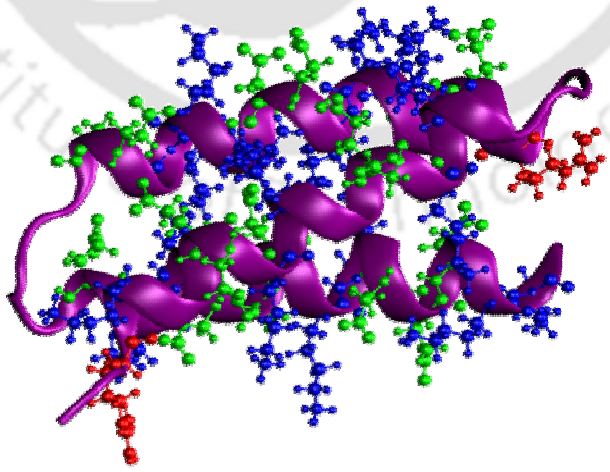


Figure 5.3.2A: Amino acid sequence and structure (PDB Code: 2LXY) of α_3C

We investigated the UV-Vis spectra between 250-800 nm for different solution concentrations of α_3C ranging from 5-105 μM. The molar extinction coefficient (Figure 5.3.2B) reveals moderate absorption (ε = 7338 ± 191 $M^{-1}cm^{-1}$ at 250 nm) features in the 250-300 nm region which diminish gradually with a distinctive long tail that extends into the visible region (ε = 964 ± 129 and 501 ± 66 $M^{-1}cm^{-1}$ at 450 and 800 nm, respectively).

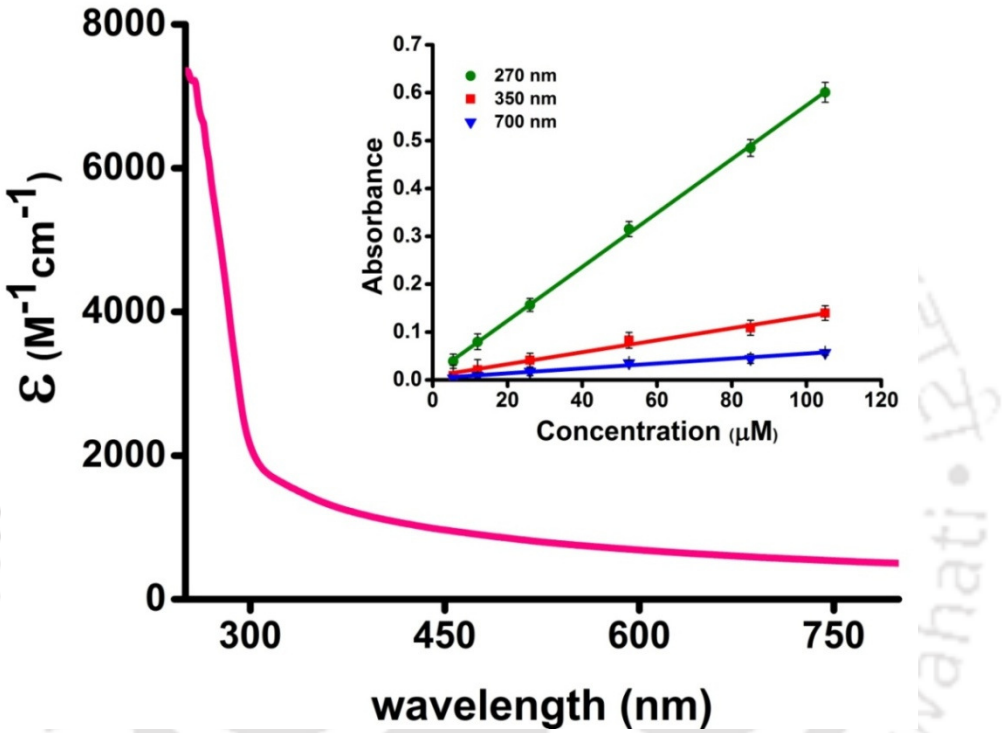


Figure 5.3.2B: Absorption spectrum of α_3C in deionised water. *Inset* shows the concentration dependence of absorbance at different wavelengths

The absorbance at different wavelengths increases linearly with concentration (Figure 5.3.2B-*Inset*) which argues against any contribution arising from protein intermolecular interactions to the spectra. Indeed, the monomer proteins are likely to be greater than 20 nm apart (calculated on the basis of volume occupied per molecule in 1 mL solution) from each other, on average, at 105 μM concentration. While absorption above 320 nm was seen previously in proteins rich in charged amino acids such as HSA (ε = 1546 $M^{-1}cm^{-1}$ at 325 nm)[111], the spectra was masked by strong contributions from Trp and Tyr residues. In this regard, α_3C clearly stands out as it is completely devoid of aromatic amino acids and rich in charged amino acids. Thus, the spectral features of α_3C, even in the 250-300 nm range, are

novel as they do not arise from aromatic chromophores. The α_3C absorption profile up to 350 nm matches well with features seen previously for HSA and calf thymus histone and the demonstration that it can extend up to 800 nm for a protein lacking aromatic or active site chromophores is unprecedented.

Table 5.3.2: Comparison of molar extinction coefficients for various proteins (having varying percentage of charged amino acids) at different wavelengths. The numbers in square brackets indicate the standard deviation for n=3-5

Protein (Total residues)	**Number and (fraction) of charged amino acids**	**Molar extinction coefficient ($M^{-1}cm^{-1}$)***				
		315 nm	**350 nm**	**450 nm**	**600 nm**	**750 nm**
HSA (585)	197 (33.6%)	2481 [276]	1460 [191]	512 [136]	333 [140]	199 [24]
α_3C (67)	36 (54%)	1727 [164]	1396 [158]	964 [129]	686 [98]	537 [75]
HEWL (129)	27 (21%)	191 [8]	<50#	<50#	<50#	<50#

*measured in deionised water, # too low to be measured accurately

α_3C and HSA are both rich in charged amino acids and both display absorption features beyond 320 nm. In α_3C, 50% of the protein is comprised of Lys and Glu residues alone. In contrast, HEWL which lacks significant absorption features beyond 320 nm has only 20% of charged residues in its sequence (Table 5.3.2). Taken together, these results suggest that charged amino acids may be key players in the observed spectral properties in α_3C.

5.3.3 Negation of Rayleigh Scattering

To evaluate if the observed spectral features above 300 nm are arise due to Rayleigh scattering, we compared the absorption spectrum with the simulated scatter. The simulated scatter was obtained by plotting the absorption spectrum of α_3C as λ^{-4} dependence. As shown in equation 5.5[174], Rayleigh scattering follows λ^{-4} dependence[175,176]. Therefore, any sample having the scattering component should overlap the simulated scattering curve.

$$\sigma_s = \frac{2\pi^5}{3} \frac{d^6}{\lambda^4} \left(\frac{\eta^2-1}{\eta^2+2}\right)^2 \tag{5.5}$$

where, σ_s is the Rayleigh scattering cross-section, d is the diameter of the particles, λ is the wavelength of light and $\eta = \frac{\eta_{particle}}{\eta_{medium}}$ is the relative refractive index.

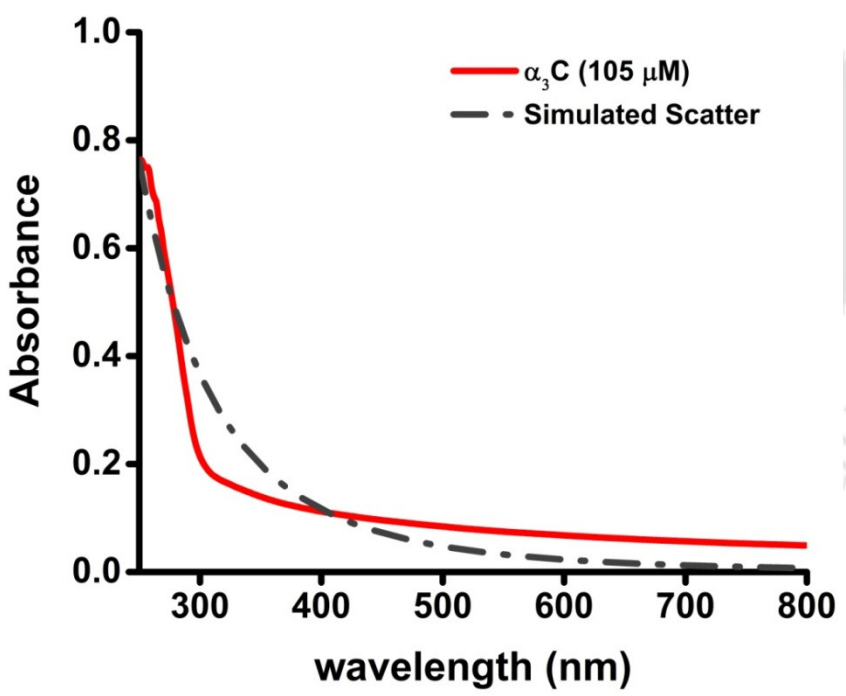


Figure 5.3.3: Comparison of absorption spectrum of the α_3C protein with the simulated scatter

We ruled out the presence of scattering artefacts behind the observed long tail (300-800 nm) in the absorption spectra of the protein based on the poor overlap of the observed spectra with a simulated Rayleigh scattering profile (Figure 5.3.3).

5.3.4 Role of protein fold

To understand the role of interaction among amino acids behind the observed spectral features in α_3C, we studied the absorption spectrum of the amino acids which constitute α_3C and compared it with the absorption spectrum of the intact protein. Unfolding reagents such as Urea and Guanidine hydrochloride could not be employed as they have significant absorption in the UV region[177].

Table 5.3.4: Population of constituent amino acids in α_3C sequence and their individual concentrations in 105 µM of protein

Sl. No.	Amino Acid	Count in α_3C	Concentration (µM)
1	Alanine (A)	3	315
2	Arginine (R)	2	210
3	Asparagine (N)	0	0
4	Aspartic Acid (D)	0	0
5	Cysteine (C)	1	105
6	Glutamine (Q)	0	0
7	Glutamic acid (E)	17	1785
8	Glycine (G)	9	945
9	Histidine (H)	0	0
10	Isoleucine (I)	3	315
11	Leucine (L)	8	840
12	Lysine (K)	17	1785
13	Methionine (M)	0	0
14	Phenylalanine (F)	0	0
15	Proline (P)	0	0
16	Serine (S)	1	105
17	Threonine (T)	0	0
18	Tryptophan (W)	0	0
19	Tyrosine (Y)	0	0
20	Valine (V)	6	630
Total number of amino acids		**67**	

To examine the role of the protein fold in producing the observed spectral features we studied the absorption spectra of solution mixtures of α_3C amino acids in proportions present in the protein. Table 5.3.4 shows the concentrations of various amino acids for amino acid concentrations corresponding to 105 μM protein solution.

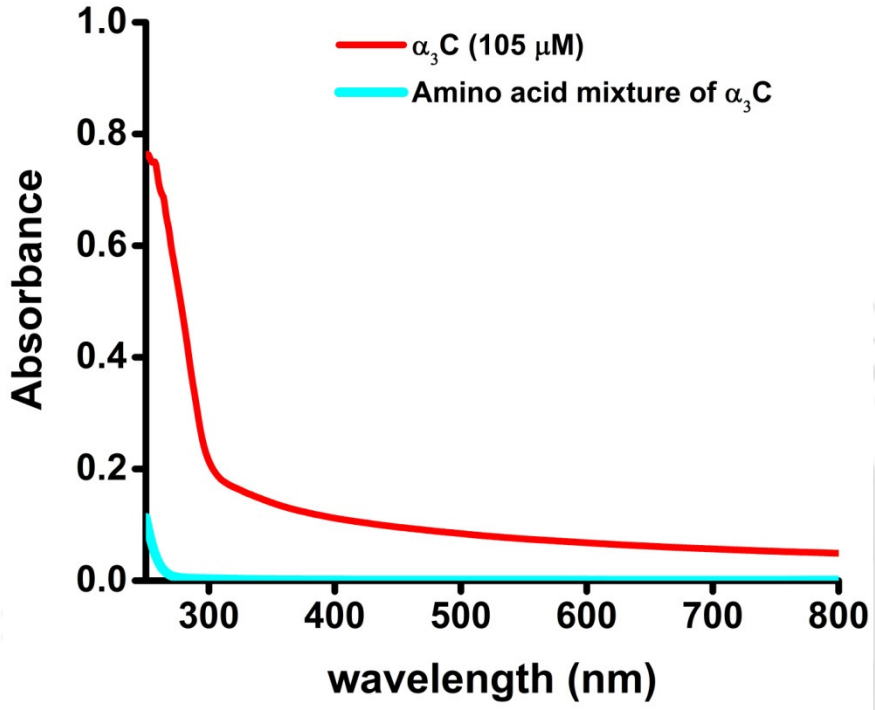


Figure 5.3.4A: Comparison of absorption spectrum of the intact α_3C protein with constituent amino acids in α_3C (105 μM)

The mixture of amino acids with composition described in Table 5.3.4 however does not show any significant absorption in the 250-800 nm region (Figure 5.3.4A). This indicates that the three dimensional fold of the protein has an important role in bringing the amino acid residues in close proximities which then translates into the observed spectral features.

A comparison of molar extinctions coefficients (ε) of pure Lys amino acid solutions (1.42 $M^{-1}cm^{-1}$ at 270 nm) with that for α_3C (5808 $M^{-1}cm^{-1}$ at 270 nm) shows that a ~4000 fold enhancement in absorptivity is achieved by the α_3C protein structure at 270 nm. We note, however, that α_3C also contains other charged amino acids (Glu and Arg) besides Lys which may also contribute to the enhanced absorption intensities in the α_3C protein.

Also there are notable differences in the spectral features of α_3C versus the absorption profiles of the charged amino acids. The long tail (beyond 320 nm) observed in the protein spectrum is curtailed to ~400 nm in solutions of charged amino acids (chapter 3). In contrast, poly-L-Lys hydrochloride shows absorption features which extend till 500 nm (Figure 5.3.4B). Thus, a possible role of the peptide backbone and the protein fold in the origin of tail spectra between 400-800 nm (Figure 5.3.2B) is anticipated.

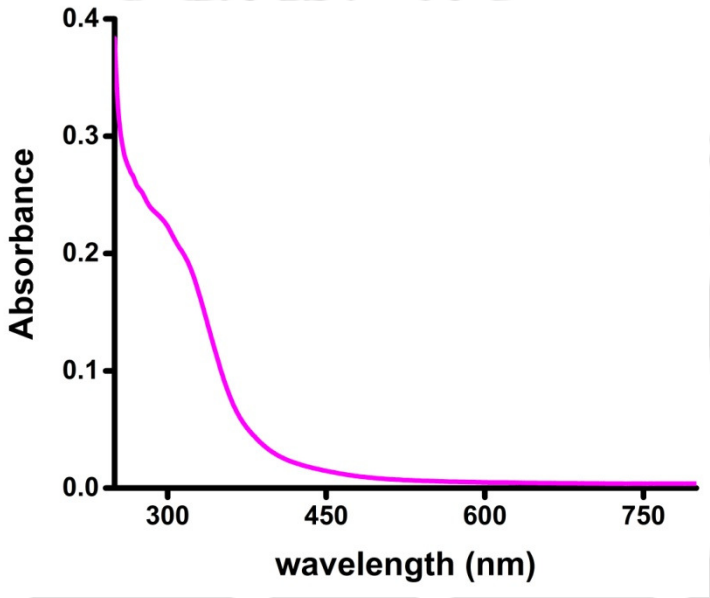


Figure 5.3.4B: Absorption spectrum of poly-L-Lys hydrochloride (12.5 mg/mL) in deionised water.

5.3.5 MD simulations of α_3C reveal interactions between Lys and Glu

The NMR structure for the α_3C protein (Figure 5.3.2A) show several Lys and Glu residue pairs placed in close proximity. We thus investigated the interactions of Lys/Glu side chains within the α_3C protein fold using classical atomistic MD simulations of the solvated protein. As discussed previously even at the maximum concentration (105 μM) of α_3C employed in our experiments we expect α_3C to remain in monomer form. Accordingly, our simulations comprised of a single α_3C molecule immersed in water box of volume ~22500 $Å^3$ with periodic boundary conditions.

An enlarged view of one of the helices of α_3C with one Lys and Glu residue is shown in Figure 5.3.5A. As the protein is rich in Lys and Glu we wanted to investigate how these amino acids interact among themselves and with each other.

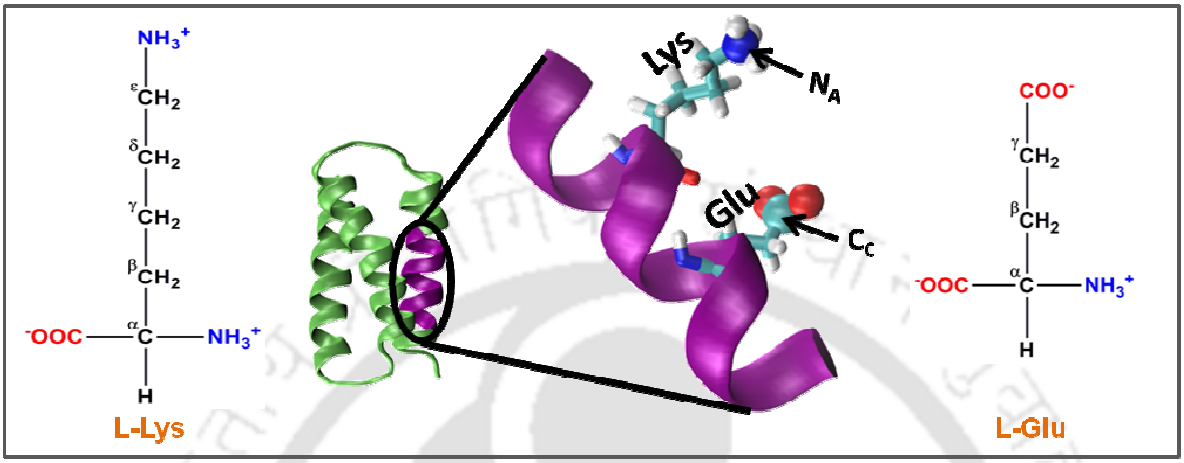


Figure 5.3.5A: The α_3C protein (green) and an enlarged view of one of its helical segments (purple) containing a Lys and Glu residue. The Lys amino nitrogen (N_A) and Glu carboxylate carbon (C_C) atoms are marked. The structures of L-Lys and L-Glu are also shown

We generated radial distribution function (RDF) plots from the MD simulation trajectory which capture the distribution of separations for: 1) Lys amino nitrogen (N_A-N_A) atom pairs, 2) Glu carboxylate carbon and Lys amino nitrogen (C_C-N_A) atom pairs, and 3) Glu carboxylate carbon (C_C-C_C) atom pairs.

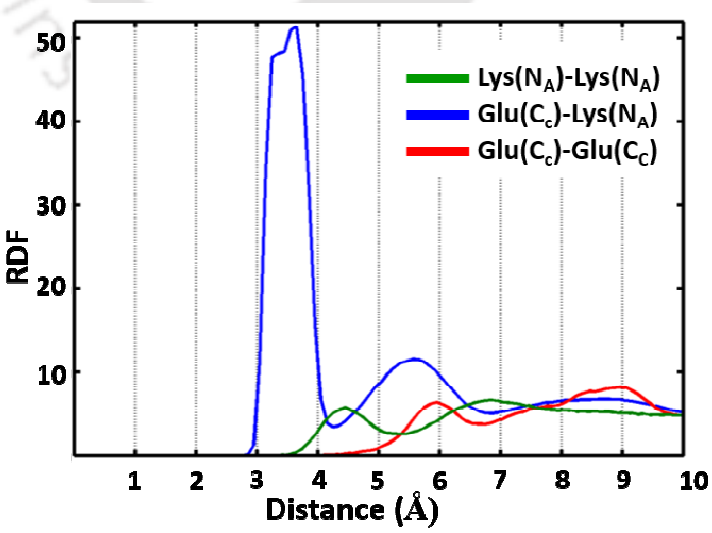


Figure 5.3.5B: Radial distribution Function (RDF) plots for N_A-N_A, C_C-N_A, and C_C-C_C atom pairs of Lys and Glu

The Lys N_A-N_A RDF plots show peaks around 4.5 Å and 7 Å (Figure 5.3.5B) which is surprising as two positively charged side chains ideally should repel each other. The Glu-Lys C_C-N_A RDF shows a peak around 3.5 Å. This peak corresponds to strong salt bridge interactions between the Lys amino group and the Glu carboxylate group. The Glu C_C-C_C RDF plot shows peaks at ~6 Å and ~9 Å, indicating weaker interactions of the Glu side chains relative to that between Lys side chains.

Further, we created 2-D contact maps (Figure 5.3.5C) displaying the average separations of N_A-N_A, C_C-N_A, and N_A-C_C atom pairs over the MD trajectory representing Lys-Lys, Glu-Lys, and Glu-Glu side chain interactions respectively.

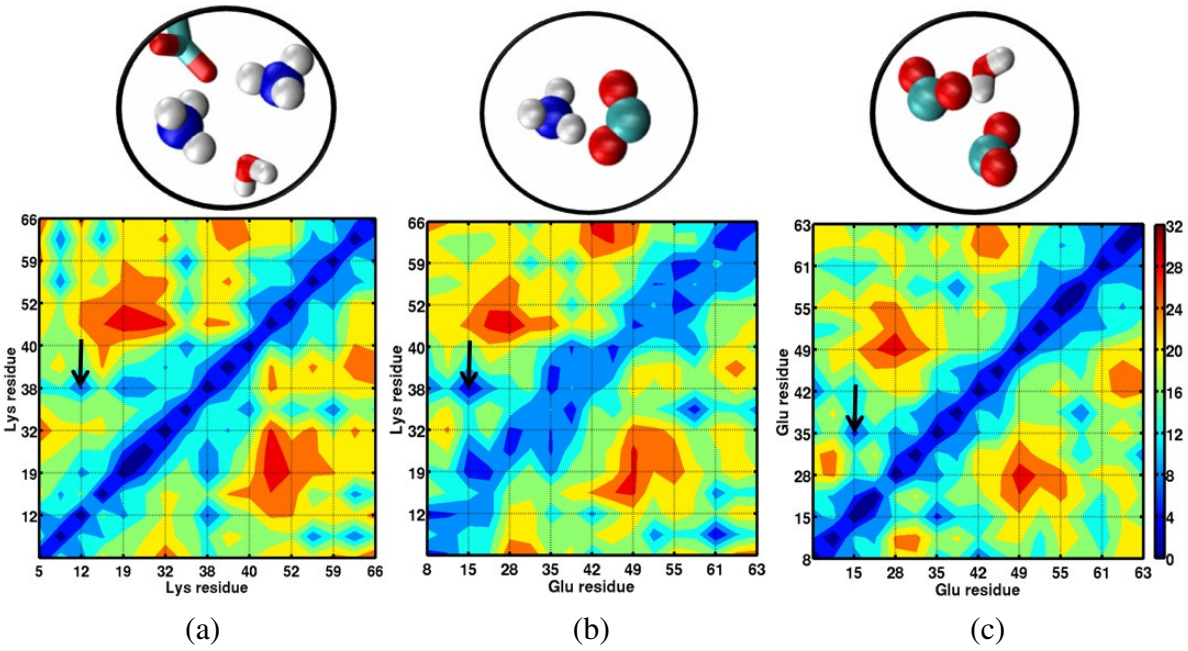


Figure 5.3.5C: Contact maps (pair wise distances averaged over the MD trajectory) for (a) N_A-N_A, (b) C_C-N_A, and (c) C_C-C_C atom pairs. Representative interactions for the region marked in the contact maps are explicitly showed in the circled images extracted from MD snapshots.

Contact maps also reveal multiple sets of Lys-Lys amino group interactions wherein the average N_A-N_A separation is lower than 7 Å over the ~100 ns MD trajectory. Interactions among the Lys-Glu side chains and Glu-Glu side chains are also seen in the contact maps. Visualization of the dynamics of Lys residues during the MD trajectory reveals that the interactions of Lys amino groups are mediated either by water molecules or by Glu residues or by both (Figure 5.3.5C circled images). Glu residues can indirectly mediate Lys-Lys side

chain interactions by screening the Lys charge. Water molecules mediate Lys-Lys side chain interactions through hydrogen bonding and by screening the Lys amino group charges as hypothesized previously[111]. As many as three water molecules may be involved in mediating the Lys side chain interactions (Figure 5.3.5D).

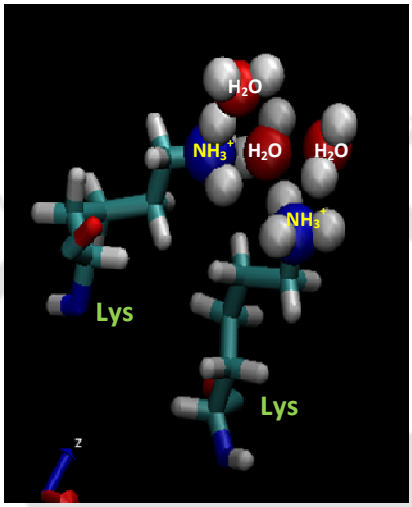


Figure 5.3.5D: A Lys pair (residue 12 and 38) bridged by three water molecules in α_3C

5.3.6 Computed UV-Vis absorption spectra for Lys and Glu monomers

The α_3C protein is rich in both Glu and Lys and we carried out TDDFT electronic structure calculations on these charged amino acids sampled from MD simulations of α_3C to simulate their absorption spectra between 200-800 nm. The spectra for Gly were calculated as control. The strategy of applying electronic structure calculations to MD sampled structures has proven to be effective for calculating UV-Vis spectra and electronic couplings for charge transfer in organic molecules[178,179,180].

The simulated spectrum of Lys monomer shows transitions in the same spectral range (200-250 nm) as the Gly control. In contrast, the Glu monomer displays transitions up to 450 nm (Figure 5.3.6A).

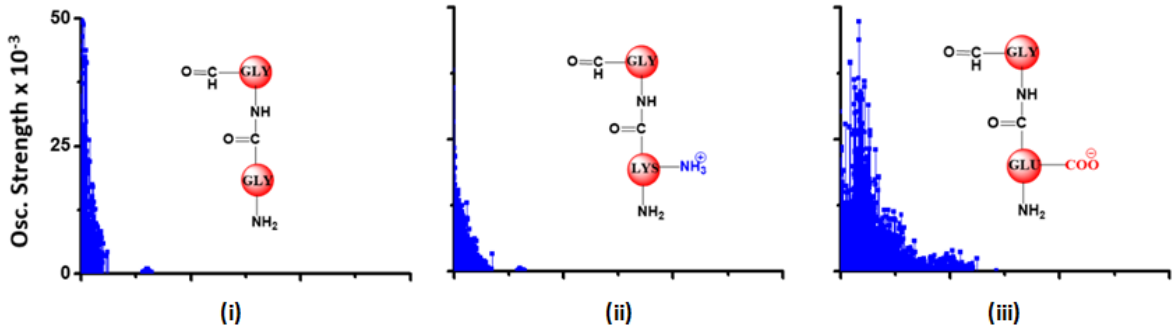


Figure 5.3.6A: Simulated absorption spectra of (i) Gly, (ii) Lys and (iii) Glu monomer. The chemical structures drawn represent monomer fragments used in the calculations

We further visualized the selected transitions of Lys, Glu and Gly through difference density plots which show the location of hole (red) and electron (green) density on each amino acid fragment (Figure 5.3.6B). Although the monomer spectra for both Lys and Gly are similar, there are fundamental differences between the transitions in Lys/Glu and the Gly control.

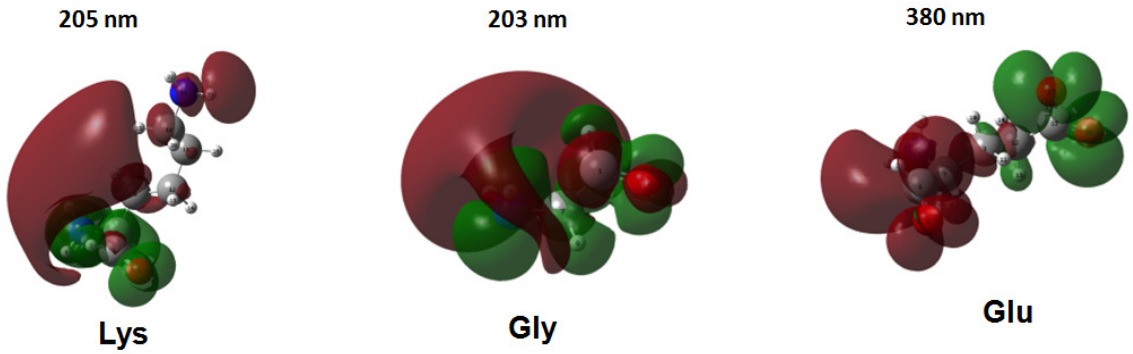


Figure 5.3.6B: Representative difference density plots with donor (green) and acceptor (red) states for various monomers

Difference density plots for Lys transitions around 205 nm show the hole density (red) delocalized on the peptide backbone atoms and the electron density (red) delocalized over the charged amino group and side chain of Lys. Calculations revealed that this was a CT transition between the Lys backbone atoms to the charged amino group. In comparison, the Gly control possess no CT transitions with both hole and electron densities spatially localized on the same set of backbone atoms. For Glu monomer the states are well separated with donor state located on the carboxylate group of Glu while the acceptor states

located on the Glu backbone (Figure 5.3.6B). The presence of a positive charge on the Lys side chain amino group shifts the unfilled orbitals of Lys to the amino group, thereby populating it with charge acceptor states. Further, the aliphatic portion of the Lys side chain creates a sigma bridge separating the charge donor states on the backbone from the charge acceptor states. Thus, well-separated photo-induced CT transitions should be characteristic and unique to amino acids with charged side chains and their derivatives.

Given, the sensitivity of CT transitions to the nature of the charge donor/acceptor states and the nature of the side chain bridge separating them, each charged amino acid is expected to show distinct absorption features. Indeed, the Glu spectra (Figure 5.3.6A) is rich in CT transitions, and shows transitions over a much greater spectral range which extends into the visible wavelength range (up to 450 nm) relative to Lys. Further, since the Glu side chain has a negatively charged carboxylate group, the direction of charge transfer during photo excitation is opposite that for Lys i.e., from the side chain carboxylate group to the polypeptide backbone. Differences between the charged side chain groups (carboxylate COO^- for Glu vs. amino NH_3^+ for Lys), different extents of hyper conjugation involving the charged groups, the presence of lone pair electrons for Glu, and the shorter side chain for Glu (two CH_2 links vs. four CH_2 links in Lys), all contribute towards the difference in spectral features for Lys and Glu monomers.

Thus charged amino acids, Lys and Glu, in monomeric form produce characteristic CT transitions. However, we note that the intensities and spectral range of transitions for monomeric Lys and Gly are very similar and these amino acids are not distinguishable on the basis of their absorption spectra. The electronic properties of the monomeric Lys/Glu chromophores was neither able to explain the full spectral range of the transitions seen in high concentration Lys solutions (250-400 nm) nor that seen for the α_3C protein (extending up to 800 nm). We therefore explored higher order intramolecular interactions between the charged amino acids within α_3C which shed light on the role of the protein fold in dramatically extending the spectral range of Lys/Glu CT transitions.

5.3.7 Computed UV-Vis absorption spectra for various dimers

TDDFT based spectra were generated from 200-800 nm for various Lys-Lys, Glu-Glu and Glu-Lys residue dimers sampled from the α_3C MD trajectory. The head groups of the amino acids can interact from dimers which are nearest neighbors (NN) or from dimers which are distally separated (DS) as shown in Figure 5.3.7.

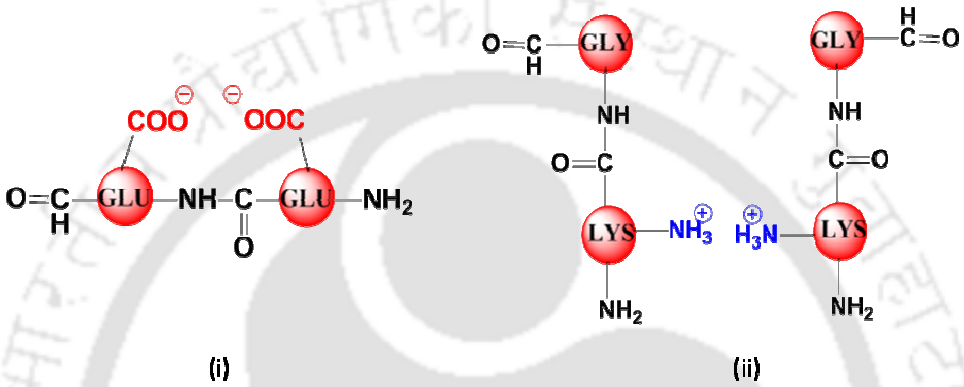


Figure 5.3.7: Dimer model representing (i) nearest neighbor in sequence (NN) Glu residue pair interactions and (ii) dimer pair model representing distally separated in sequence (DS) Lys residue pair interactions

In these calculations an extended dimer for the backbone was retained for DS fragments wherein the backbone of each of the two residues of a DS pair was extended to include one of the adjacent peptide units. To better represent the polypeptide backbone in the protein environment, Lys/Glu pairs were extracted along with its adjacent residue and then the adjacent residue in the fragment was mutated to Gly.

All the dimers showed remarkable difference when compared to their respective monomers. The Lys-Lys dimers showed additional transitions above 250 nm which extended up to 450 nm which did not exist in the Lys monomer spectra. Also Lys-Lys spectra could be clearly distinguished from the Gly-Gly spectrum which was not seen in the monomers. Other dimers also showed spectra which extended in to the visible region when compared to their respective monomer spectra.

5.3.7.1 Lys-Lys Dimers

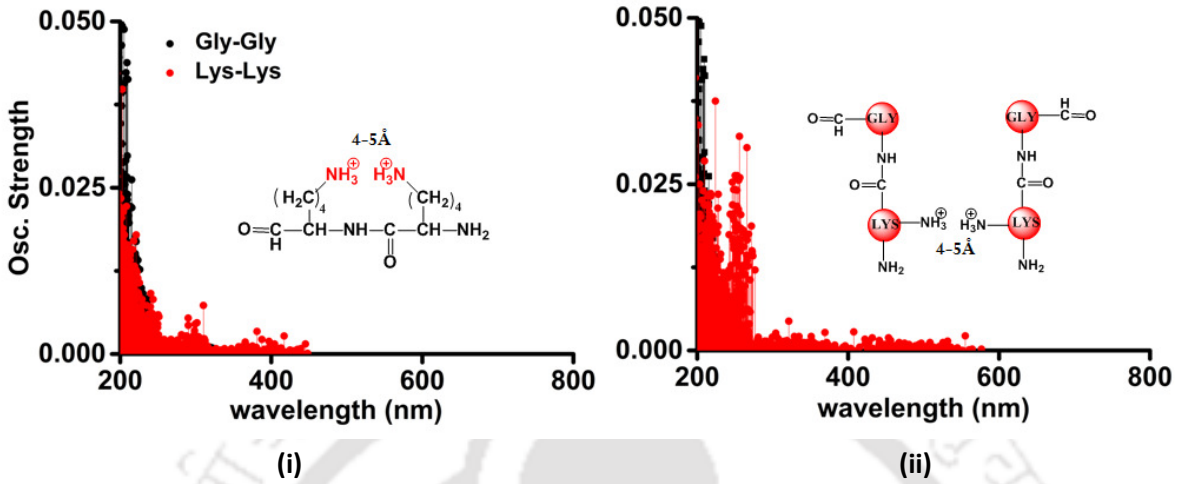


Figure 5.3.7.1A: Simulated absorption spectra for Lys-Lys (i) NN pairs and (ii) DS pairs for strong interactions

The spectral range for DS Lys-Lys pairs (Figure 5.3.7.1A) extended up to 550 nm when there are strong interactions (4-5 Å) between the ε-NH_3^+ moieties of two distally separated Lys residues. For both NN and DS pairs these transitions were curtailed to ~350 nm when the interactions among the Lys residues are weak i.e., when the distance between the head groups of two Lys residues is between 6-7 Å (Figure 5.3.7.1B).

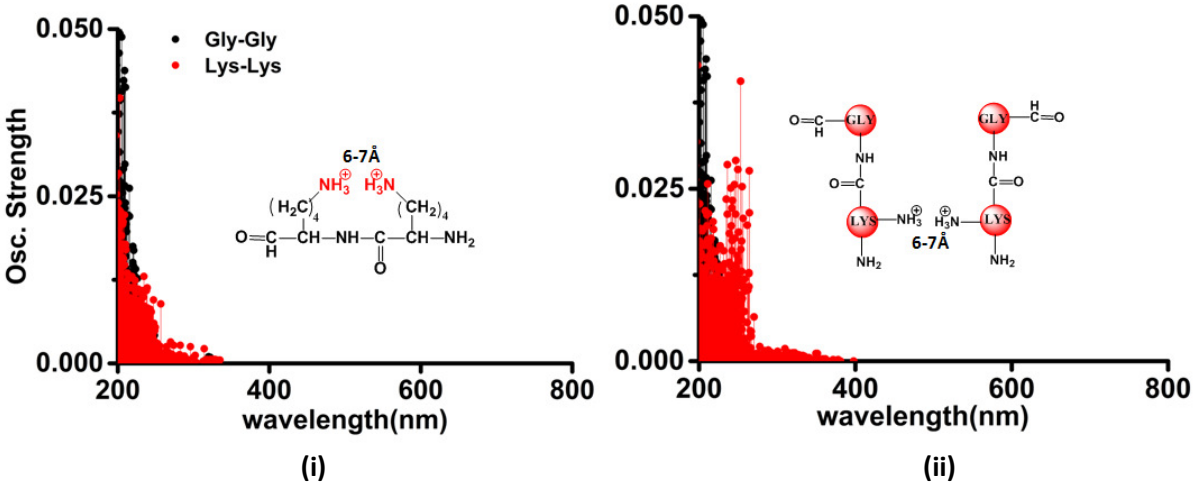


Figure 5.3.7.1B: Simulated absorption spectra for Lys-Lys (i) NN pairs and (ii) DS pairs for weak interactions

5.3.7.2 Glu-Glu Dimers

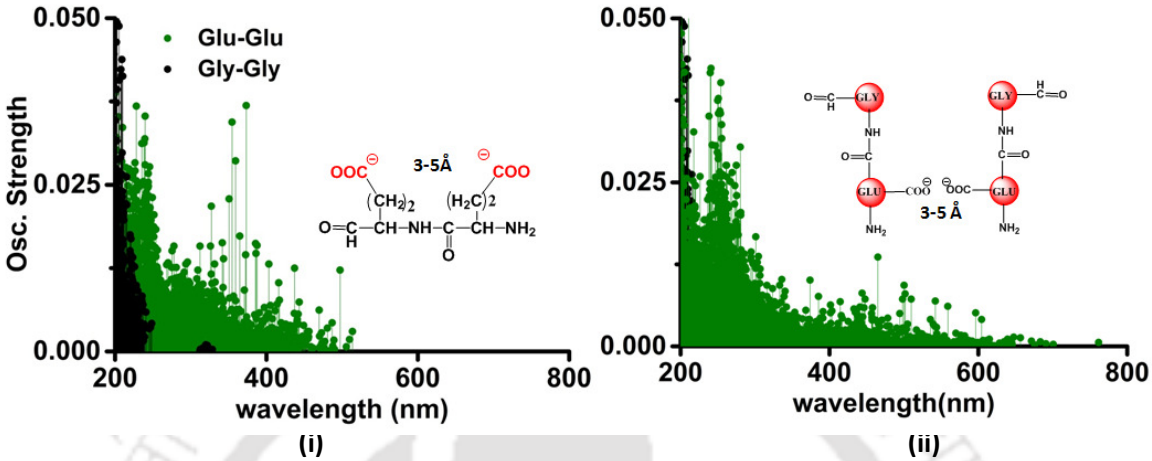


Figure 5.3.7.2A: Simulated absorption spectra for Glu-Glu (i) NN pairs and (ii) DS pairs for strong interactions

Side chain interactions in Glu-Glu dimer created new long wavelength transitions which extend the absorption range up to 600 nm (Figure 5.3.7.2A). The spectrum was curtailed up to 450 nm for strong interacting Glu-Glu NN pairs. Weakly interacting Glu-Glu pairs also showed similar absorption profiles (Figure 5.3.7.2B).

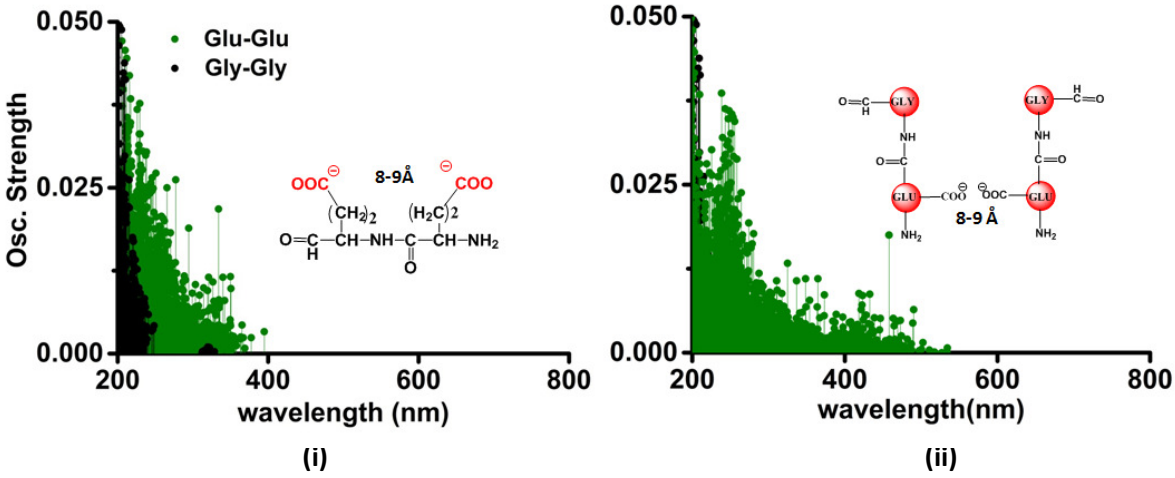


Figure 5.3.7.2B: Simulated absorption spectra for Glu-Glu (i) NN pairs and (ii) DS pairs for weak interactions

5.3.7.3 Glu-Lys Dimers

The interactions between Lys-Glu residue pairs led to starkly different spectral profiles from that produced by Lys-Lys and Glu-Glu interactions. The transitions were curtailed to 350 nm in strongly interacting Glu-Lys pairs (Figure 5.3.7.3A) whereas weakly interacting Glu-Lys residues produced prominent transitions up to 800 nm (Figure 5.3.7.3B).

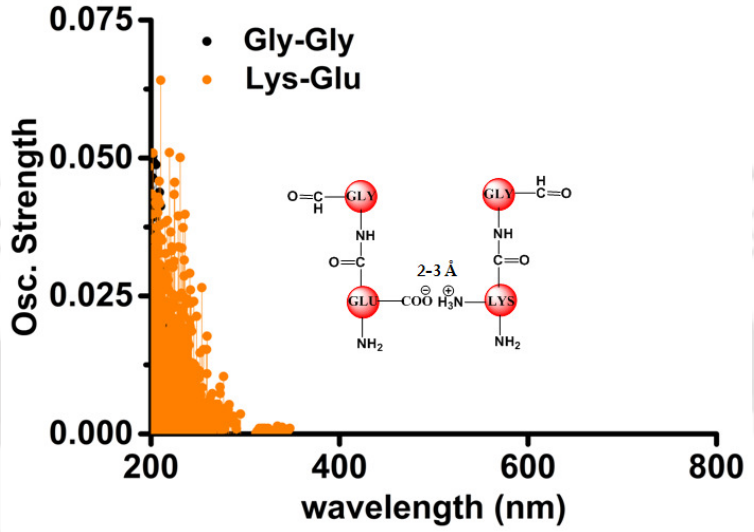


Figure 5.3.7.3A: Simulated absorption spectra for Glu-Lys DS pairs for strong interactions. No pairs were found for NN mode of interaction

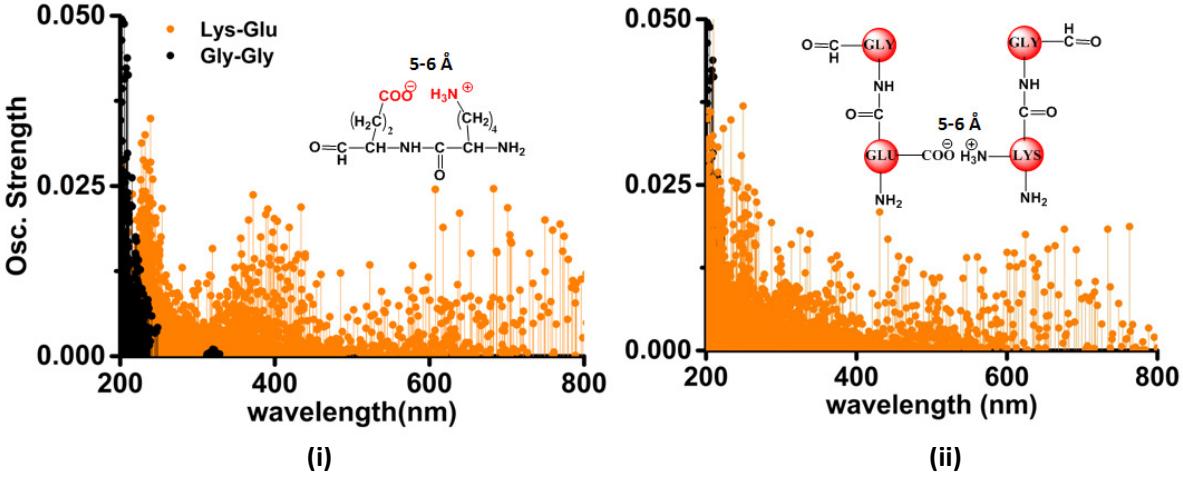


Figure 5.3.7.3B: Simulated absorption spectra for Glu-Lys (i) NN pairs and (ii) DS pairs for weak interactions

Strong interactions between Lys amino and Glu carboxylate groups could create a neutral moiety through the formation of a salt bridge which could suppress the CT transitions above 300 nm. Thus, in contrast to the Lys-Lys/Glu-Glu dimers, extension of CT spectral range towards long wavelengths is inversely proportional to the strength of DS Glu-Lys interactions.

5.3.7.4 Other Dimers

We extended our calculations of absorption profiles for NN Lys-AA (AA=Ala,Val,Ile,Cys and Leu) pairs which together with the Lys-Lys, Glu-Glu, and Glu-Lys pairs represent all Lys containing NN dimer species present in α_3C.

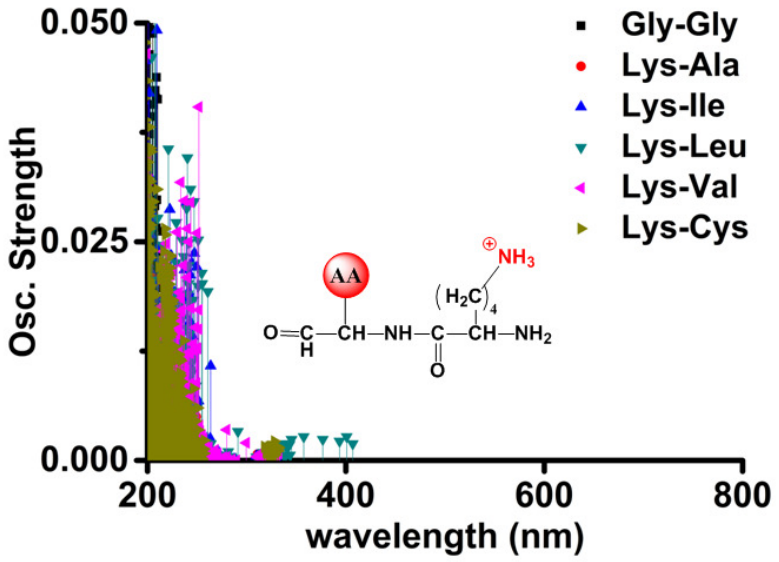


Figure 5.3.7.4: Simulated absorption spectra for all NN Lys-AA dimers

Other than the charged amino acid dimers (Lys-Lys, Glu-Glu, and Glu-Lys) studied, no other dimer species showed significant absorption beyond 300 nm (Figure 5.3.7.4). Thus, the electronic structure calculations unambiguously show that the long tail absorption of α_3C between 300-800 nm arises from the associations between Lys and Glu amino acid side chains.

5.3.8 Nature of transition in dimers

Some representative donor and acceptor states are shown for the Lys-Lys, Glu-Glu and Glu-Lys dimers. An analysis of the spectra for Lys-Lys and Glu-Glu interacting pairs (both DS and NN) revealed mostly CT transitions beyond 250 nm. For both DS and NN pairs, we found CT transitions between the extended backbone and the amino/carboxylate group of Lys/Glu residues.

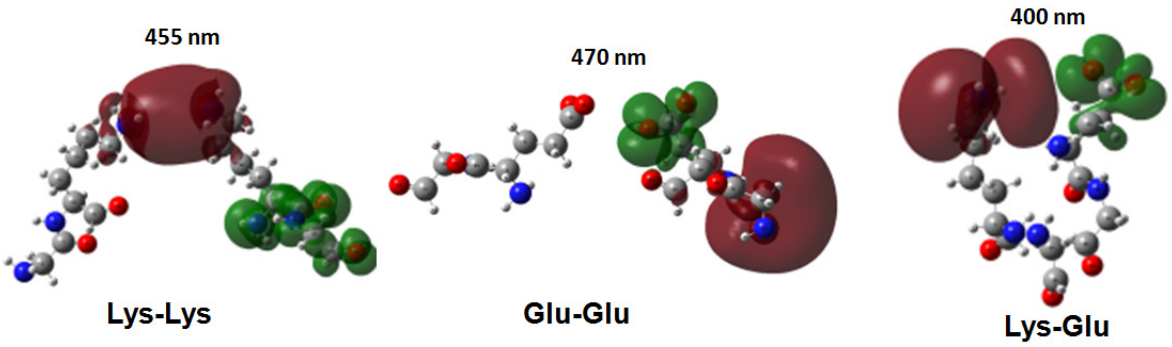


Figure 5.3.8: Representative difference density plots with donor (green) and acceptor (red) states for various dimers

For CT transitions in Lys-Lys pairs, the charge acceptor states (red) tend to delocalize between the interacting amino groups (Figure 5.3.8). For Glu-Glu pairs, the direction of CT was opposite to that seen for Lys-Lys pairs with the charge donor states (green) located on side chain carboxylate groups and acceptor states on the backbone.

For Glu-Lys pairs the positively charged amino group of Lys stabilizes electron acceptor states, whereas the negatively charged carboxylate group of Glu is electron rich and acts as a charge donor state. For Glu-Lys interactions, distinct and relatively intense CT transitions from the Glu carboxylate group to the Lys amino group are produced.

Thus taken together, charged side chain amino/carboxylate groups in Lys/Glu residue pairs can greatly extend the spectral range of CT transitions observed for Lys/Glu monomers up to 800 nm. These transitions are therefore able to explain the long tail observed for α_3C in experiments.

5.3.9 Sensitivity of UV-Vis absorption of α_3C to temperature and pH

The ranges of the absorption profile as well as intensities of CT bands in the spectral profile are sensitive to the strength of interactions between Lys/Glu side chains. Thus, perturbations of the protein tertiary fold which disrupt the Lys/Glu side chain interactions should expectedly modify the UV-Vis absorption spectral profile of α_3C. To verify this, we recorded CD and absorption spectra for α_3C over a temperature range of 25-85 °C. The CD spectra (Figure 5.3.9A) revealed that the protein retains significant fraction of α-helical structure and does not get completely denatured even at temperatures as high as 85 °C.

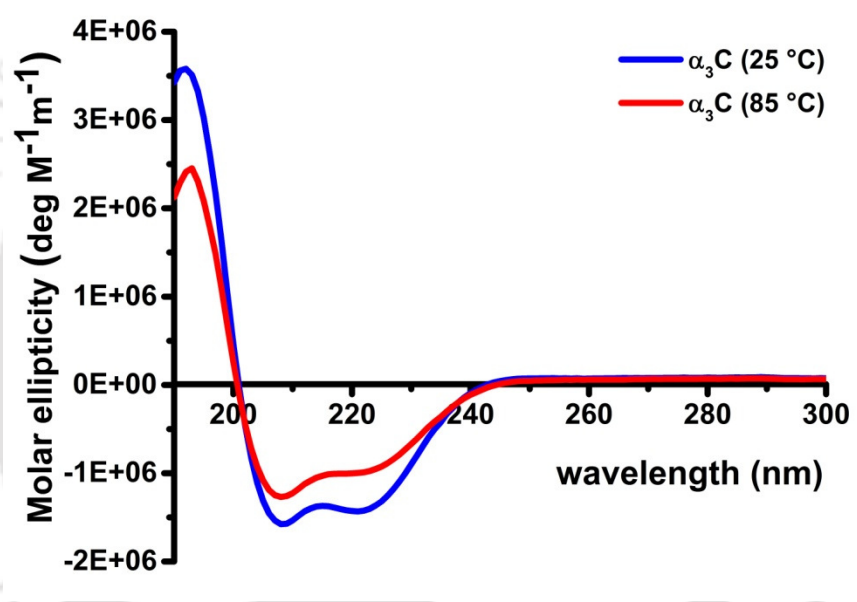


Figure 5.3.9A: CD spectra α_3C at 25 °C and 85 °C in deionised water

In contrast, the absorption profile of α_3C showed much more sensitivity to temperature (Figure 5.3.9B) increasing by 1.2-2 folds between 250-400 nm. The largest absorbance increases occurred beyond 300 nm where the CT transitions are dominant.

To ascertain the effect of pH we altered the pH of the medium to extreme limits (pH 1 and 13), so that the protein contained only one charged species. Figure 5.3.9B shows that absorption in the range 310–800 nm has nearly diminished (a dramatic >70% dip), both at pH 1.5 and 13, in comparison to the spectrum at pH 5.5.

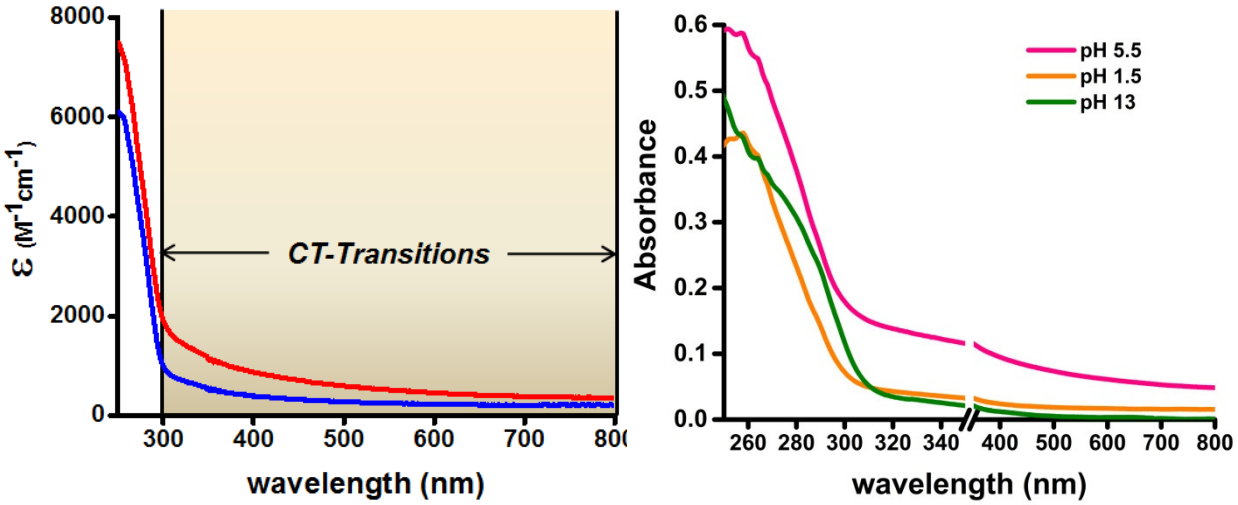


Figure 5.3.9B: Comparison of molar extinction coefficient of α_3C at 25 ℃ and 85 ℃ in deionised water (left);) Absorption spectra of α_3C (85 µM) at different pH. An expanded scale is specifically shown to highlight changes at short wavelengths. Scale has a break between 350-351 nm (right)

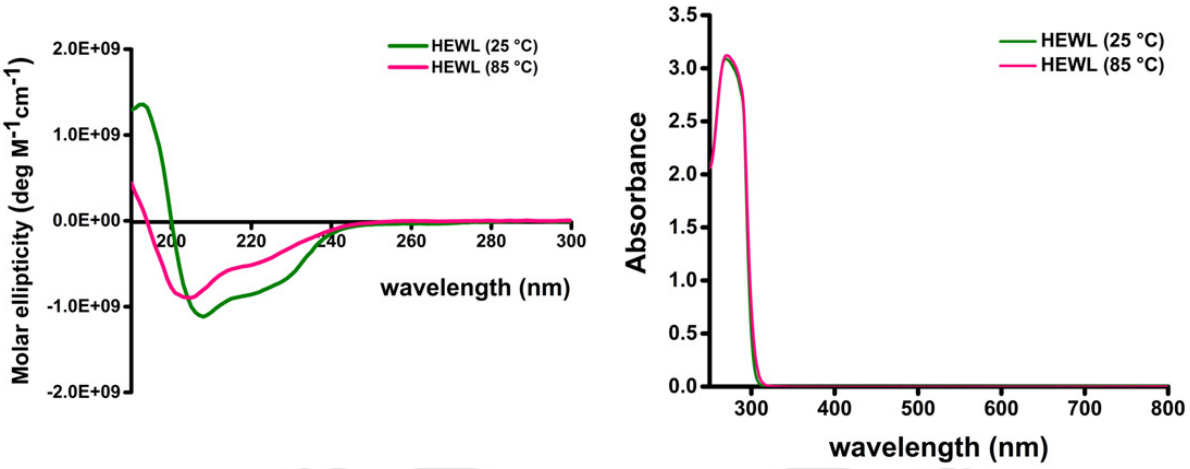


Figure 5.3.9C: CD spectra for HEWL at 25 ℃ and 85 ℃ (left) and absorption spectra of HEWL (150 µM) at 25 ℃ and 85 ℃ (right) in deionised water

In contrast for, HEWL which comprises of 20% charged amino acids does not show any absorption features beyond 300 nm. Thermal perturbation at 85 ℃ has no effect on its absorption spectral profile although its CD spectrum is affected. In fact the absorption spectra for both at 25 ℃ and at 85 ℃ overlap each other (Figure 5.3.9C).

Further, we saw that the changes in spectral profile were non-uniform across the wavelength range probed with temperature (Figure 5.3.9D), whereas changes in protein

concentration induced uniform and linear changes in the absorption intensities across the wavelength range. Thus, the observed changes in absorption spectra hint at perturbation of Glu-Glu and Lys-Glu contacts in α_3C at higher temperatures.

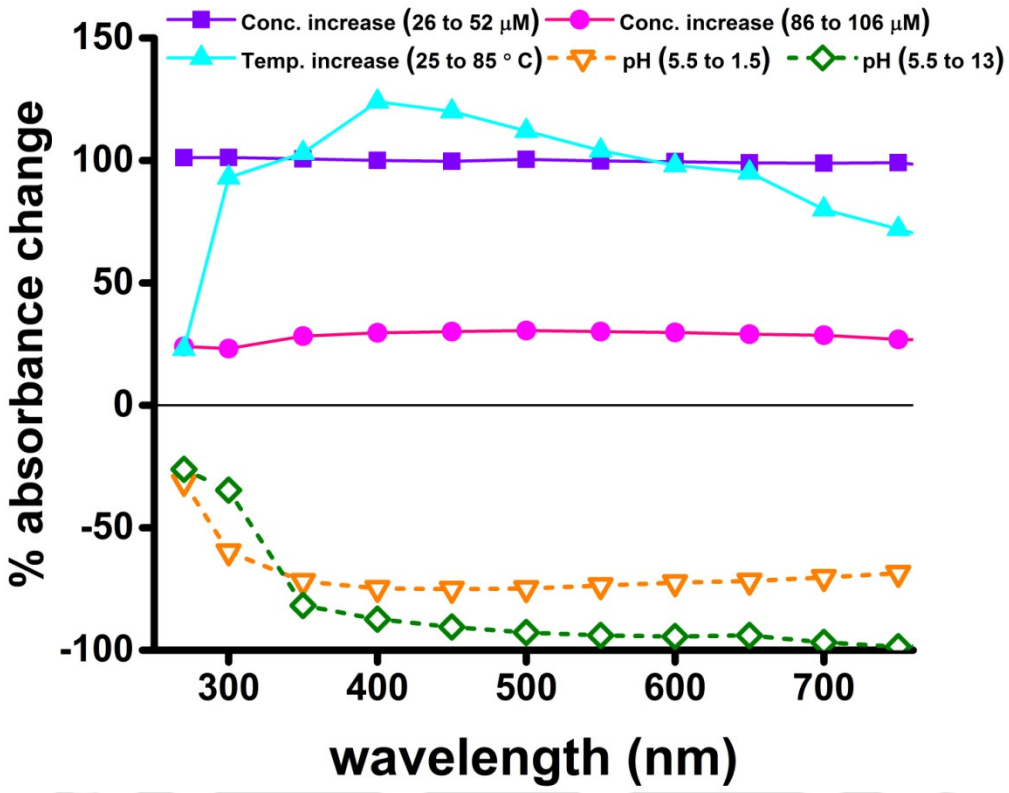


Figure 5.3.9D: Percent change in absorbance of α_3C at different wavelengths with increase in temperature (25 °C to 85 °C) compared with increase in concentration (26 μM to 52 μM and 85 to 105 μM) and change in pH (5.5 to 1.5 and 5.5 to 13)

In summary, the pH variations clearly validate the critical role played by charged Lys–Lys, Glu–Glu, and Lys–Glu interactions contributed by the protein fold to the a_3C absorption in the near UV-visible range. The temperature variation, on the other hand, emphasizes the sensitivity of the a_3C absorption intensity to perturbation of tertiary Lys–Lys, Glu–Glu and Lys–Glu sidechain contacts.

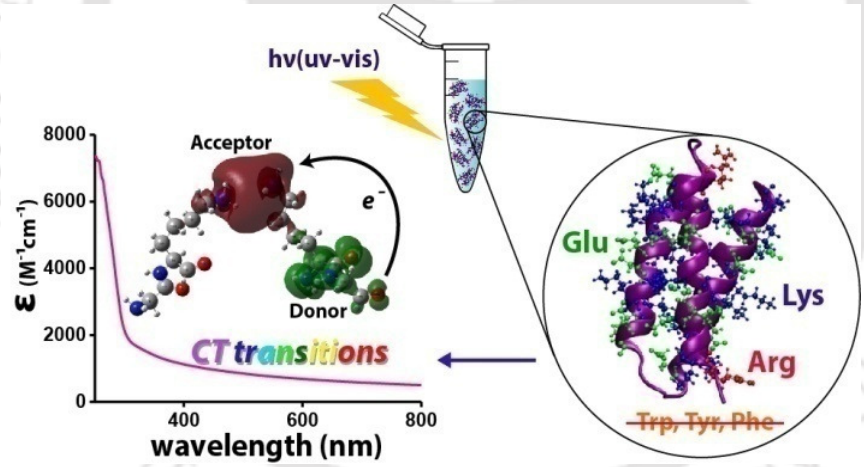
5.4 Conclusions:

1. The α_3C protein displays unusual absorption features beyond 250 nm extending all the way to 800 nm despite being devoid of any aromatic amino acids.

2. Our calculations corroborate the previous speculation on the role of water molecules bridging two Lys residues in the origin of the novel spectra.

3. Theoretical calculations show that the novel spectra arise due to intramolecular interactions among the Lys and Glu residues of α_3C.

4. Charge transfer transitions involving Lys and Glu residues of α_3C along with polypeptide backbone are the origin of these unique spectral features beyond 300 nm. This work can be summarized in a graphical representation as:



5. Glu is a more potent chromophore in comparison to Lys. This also explains higher absorption values for peptides 7, 8 and 9 (chapter 2) which were devoid of Lys residues.

5.5 Implications of the work:

Using α_3C as a model protein, distinctive UV-Visible absorption between 250-800 nm has been shown. Classical MD simulations of α_3C reveled dynamic interactions between multiple charged side chains of Lys and Glu residues present in the protein. TDDFT calculations on charged amino acid residues sampled from MD trajectories of α_3C reveal that the absorption features of α_3C may arise from charge transfer transitions involving spatially proximal Lys and Glu residues. The report of absorption features beyond 350 nm in a monomeric protein lacking aromatic chromophores is presented for the first time. This investigation significantly expands the pool of spectroscopically active bimolecular chromophores which can be employed to probe dynamics of biologically relevant proteins rich in charged amino acids. It also opens up doors for using a very simple technique like UV-Visible absorption spectroscopy to study the structure of proteins.

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

CHAPTER 6

CONCLUDING REMARKS

Among the 20 naturally occurring amino acids, only the aromatic amino acids i.e., Trp, Tyr and Phe (chromophores present in the side chains of aromatic amino acids) have been attributed as chromophores behind the UV-Visible absorption in proteins. The characteristic UV-Visible electronic absorption arising from aromatic amino acids is typically seen around 225-280 nm of the electromagnetic spectrum. Apart from aromatic amino acids, peptide bond shows characteristic absorption features at around 190-220 nm in proteins and peptides. A protein or a peptide, devoid of any aromatic amino acid is therefore expected to remain optically silent beyond 250 nm. Previous studies on L-Lysine monohydrochloride, poly-L-Lys and proteins rich in Lys residues such as HSA, report unusual absorption features in the 250-350 nm region. The work reported in this thesis work takes an account of these interesting observations in order to find out the underlying mechanisms behind the novel spectral signatures arising from non-aromatic amino acids and peptides. Establishing a quantitative link between non-aromatic amino acid content in the protein and the UV-Vis absorption features would open up a new spectral window to probe prominent proteins of biomedical relevance such as nucleic acid binding proteins, intrinsically disordered proteins etc.

With our studies on small amine containing aliphatic compounds we have shown that the $-NH_2$ moiety indeed plays an important role in the origin of spectral signatures near UV region. Studies on Lys containing short peptides have shown that intra-molecular association between the Lys residues within a peptide enhances the spectrum. Through our studies on high concentration solutions all non-aromatic amino acids we have shown that charged amino acids are special as they show significant absorption above 350 nm when compared to their uncharged counterparts. The work reported in this thesis is able to demonstrate the ability of charged residues either in amino acid form or within extended peptide chains to absorb in the near UV region. Given the strong prominent spectra features of Trp, Tyr and Phe, it was desirable to investigate the UV-Vis absorption spectra within a model protein rich in charged amino acids and devoid of these aromatic amino acids. To achieve this goal we devised a new scoring scheme for amino acids to allow a better differentiation of the general amino acid composition based on their polarity. The magnitude of the assigned prime number was based on the hierarchy of the amino acid in the hydrophobicity scale. Based on this scheme various numerical parameters like protID,

PS-Score and average PS-Score were introduced which helped us to identify various proteins rich in charged amino acids. The ProtID quantitatively stores the protein sequence composition as the prime product factorization yield unique factors. The PS-Score is a cumulative sum of the hydrophobicity index of all constituent amino acids in the protein. Using this approach we were able to identify a synthetic monomeric protein (α_3C) rich in charged amino acids and devoid of any aromatic amino acids. In addition, this strategy was employed to analyze the amino acid content of proteomes for several organisms including the much unknown dark proteome. At present there are no such tools available which equip us with a numerical tool to study proteins across different classes of organisms. This score presents a convenient handle to sort all polypeptide sequences in a proteome in a hierarchy, reveal their average hydrophobicity and display the latter as a histogram distribution against population of entire proteome. Further, we have also developed novel tools to represent proteins as visual graphics which help in quick analysis of amino acid content along with the sequence in a given protein. We believe our method to numerically encode amino acids, opens up new avenues to analyze sequence information across multiple proteomes including the dark proteome and aids the big data analysis of several proteomes.

In a joint experimental and theoretical study on α_3C as a model protein, we have demonstrated that monomeric proteins lacking aromatic amino acids can display significant UV-Vis absorption features between 250-300 nm with a long tail that can extend into the visible region of the electromagnetic spectrum. Our computational analysis on Lys and Glu amino acids extracted from MD generated structures of α_3C revealed CT transitions in these amino acids. MD simulations revealed dimer and higher order interactions between Lys and Glu residues which were found to strongly modulate the CT absorption spectral profile. Taken together, the novel assignment of CT transitions to the 250-800 nm region in the absorption profile of proteins opens up a new spectral window to develop intrinsic spectral markers to monitor structure and dynamics of proteins rich in charged amino acids which can complement traditional techniques based on aromatic chromophores. Further, the spectral profile broadly overlaps with the emission profile of fluorescent chromophores (such as Trp) or dyes. Thus, in addition to monitoring the absorption profile changes directly, the decay kinetics of fluorescent probes may also be used as a spectral marker to follow the dynamics and interactions of charged amino acids within protein folds.

APPENDIX

α_3C purification buffers

Lysis buffer

tris(hydroxymethyl)aminomethane, pH 8	20 mM
NaCl	500 mM

Wash buffer

tris(hydroxymethyl)aminomethane, pH 8	20 mM
NaCl	500 mM
Imidazole	20 mM

Elution buffer

tris(hydroxymethyl)aminomethane, pH 8	20 mM
NaCl	500 mM
Imidazole	50-400 mM

Thrombin cleavage buffer

tris(hydroxymethyl)aminomethane, pH 8	20 mM
NaCl	500 mM
Calcium chloride	2.5 mM

Components of SDS-PAGE

Components for 15% Resolving Gel

Solutions	Resolving Gel (10 mL)	Stacking Gel (5 mL)
Deionised water	2.3 mL	3.4 mL
30% Acrylamide	5 mL	830 µL
1.5 M tris(hydroxymethyl)aminomethane	2.5 mL (pH 8.8)	630 µL (pH 6.8)
10% SDS	100 µL	50 µL
10% Ammonium persulphate (APS)	100 µL	50 µL
N,N,N′,N′ Tetramethylethylenediamine (TEMED)	12 µL	8 µL

De-Staining solution, 100 mL

Water	60 mL
Methanol	30 mL
Acetic acid	10 mL

Solutions for Lowry method

Solution A (2 % Sodium carbonate in 0.1 % NaOH), 100 mL

NaOH	0.47 g
Na_2CO_3	2 g

Solution B (2.37 % Potassium sodium tartrate in water), 50 mL

$KNaC_4H_4O_6 \cdot 4H_2O$	1.185 g

Solution C (1.56 % Copper sulphate in water), 50 mL

$CuSO_4.5H_2O$	0.78 g

REAGENT I: 48 mL of solution A + 1 mL of solution B + 1 mL of solution C

REAGENT II: 1 part of Folin's reagent + 1 part of water (Stored in dark)

Reagents I and II were prepared freshly before use.

Mathematica codes

a) Code for calculating protID

```
i=1831;
l=1381;
f=1151;
v=1097;
w=647;
c=479;
m=409;
a=181;
y=137;
p=97;
g=83;
t=67;
s=47;
h=29;
q=19;
d=17;
n=13;
e=11;
k=5;
r=2;

StringRiffle[Characters[ToLowerCase["MQPAALLGLLGATVVAAVS
SMPVDIRNHNEEVVTHCIIEVLSNALLKSSAPPITPECRQVLKKNGKELKNEE"]]]

m q p a a l l g l l g a t v v a a v s s m p v d i r n h
n e e v v t h c i i e v l s n a l l k s s a p p i t p e
c r q v l k k n g k e l k n e e
```

protID:

```
16632269288098678823581571296656608435961834809358332772
59632937491498639839017819702057548997424850283570234681
52548507155433528216429412936125 00
```

```
BaseForm[%,36]
```

protID (Base 36)

```
31xizlnc0lglk0qvhuk3way6gmy3fejhy5sbqvzm82gt9zhzc8vbqz3t
d0t4cou9lxvxyygzrnvzfat9kzphjah9bmewgk
```

b) Code for calculating PS-Score and Average PS-Score

```
data=
Import["\Users\RSW\Dropbox\Saumya_Thesis\Dark_Proteome\Human\Human_Dark.xlsx",{"Data",1}]
```

	A	B	C	D	E	F	G	H	I	J
1	Entry	Organism	Protein na	Length	Sequence					
2	P01858	Homo sap	Phagocytc	4	TKPR					
3	P02729	Homo sap	Urine glyc	8	CEHSHDGA					
4	P02728	Homo sap	Erythrocyt	10	CEGHSHDHGA					
5	P01358	Homo sap	Gastric jui	10	LAAGKVEDSD					
6	P22103	Homo sap	Pneumadi	10	AGEPKLDAGV					
7	P69208	Homo sap	Morphoge	11	QPPGGSKVILF					
8	Q9NRI7	Homo sap	Putative p	21	MAAACRCLSLLLLSTCVALLL					
9	Q6NVV0	Homo sap	Putative n	33	MLLAAVGDDELTDSEDESDLFHEELEDFYDLDL					
10	P0C5K6	Homo sap	Putative t	33	MSPPSSMCSPVPLLAAASGQNRMTQGQHFLQKV					
11	Q9NRI6	Homo sap	Putative p	33	MATVLLALLVYLGALVDAYPIKPEAPGEDAFLG					
12	Q8WZA8	Homo sap	Putative g	35	MIPGNPSPGADLAVSKHFFSLSWFCGLLLLESKQK					
13	Q86Y28	Homo sap	B melanor	39	MAAGAVFLALSAQLLQARLMKEESPVVSWWLEPEDGTAL					

```
Dimensions[data]
{4342,5}

i=1831.;
l=1381.;
f=1151.;
v=1097.;
w=647.;
c=479.;
m=409.;
a=181.;
y=137.;
p=97.;
g=83.;
t=67.;
s=47.;
h=29.;
q=19.;
d=17.;
n=13.;
e=11.;
k=5.;
r=2;
```

```
psscoreList={};

Do[AppendTo[psscoreList,Log2[ToExpression[StringRiffle[Characters[ToLowerCase[data[[x,5]]]]]]]],{x,2,4343}];

Export["\Users\RSW\Dropbox\Saumya_Thesis\Dark_Proteome\Human\Human PS-Score.xlsx",psscoreList]
```

Human PS-Score.xlsx

	A
1	15.98793
2	45.59621
3	56.82923
4	61.41645
5	64.74935
6	79.61228
7	177.1136
8	216.0954
9	220.954
10	254.6034
11	249.9667
12	280.436

```
avgpsscoreList ={};

Do[AppendTo[avgpsscoreList,Log2[ToExpression[StringRiffle[Characters[ToLowerCase[data[[x,5]]]]]]/data[[x,4]]],{x,2,4343}];

Export["\Users\RSW\Dropbox\Saumya_Thesis\Dark_Proteome\Human\Human_Avg_PS-Score.xlsx",avgpsscoreList]
```

Human_Avg_PS-Score.xlsx

	A
1	3.996983
2	5.699527
3	5.682923
4	6.141645
5	6.474935
6	7.23748
7	8.43398
8	6.548347
9	6.695575
10	7.715254
11	7.141907
12	7.190667

c) Code for calculating histogram distributions

```
data=Import["/Users/RSW/Dropbox/Thesis/Saumya_Thesis/FUNCTIONAL_CLASSES/Avg_PS_Score_Receptor_proteins.xlsx",{"Data",1}];
```

1	7.127917
2	6.384491
3	7.073855
4	6.82536
5	6.473025
6	6.604948
7	7.007692
8	7.127811
9	6.023008
10	6.132113
11	6.646177
12	7.629772

```
myList ={};

Do[AppendTo[myList,data[[x, 1]]],{x, 1, 1641}];

data1= HistogramList[myList,{0.05}];

Export["/Users/RSW/Dropbox/Thesis/Saumya_Thesis/FUNCTIONAL CLASSES/Histogramlist_Receptor_proteins.xlsx"]
```

Histogramlist_Receptor_proteins.xlsx

5	5.05	5.1	5.15	5.2	5.25	5.3	5.35	5.4	5.45
1	0	1	0	0	0	1	0	0	1

LIST OF PUBLICATIONS AND CONFERENCES

LIST OF PUBLICATIONS

1. Prasad, S. Mandal, I. Singh, S. Paul, A. Mandal, B. Venkatramani, R. and Swaminathan, R., Near UV-Visible electronic absorption originating from charged amino acids in a monomeric protein. *Chemical Science*., 2017 (DOI: 10.1039/C7SC00880E)

2. Prasad, S. and Swaminathan, R., Measuring the diffusion of fluorescent dye or protein inside living cells. *Current Science*., 2013, **105**, 1549-1561

3. Prasad, S and Swaminathan, R., Protein Sequence Score and Protein Sequence Maps: Numerical and Graphical tools to analyze Protein Sequences in Biological Databases. (*Manuscript under preparation*)

Conferences/Workshop attended

International

1. Prasad, S., Mandal, I., Paul, A., Mandal, B.,Venkatramani, R. and Swaminathan, R., Investigation of Novel Spectroscopic Features in the Near Ultraviolet Region Arising from Non-Aromatic Amino Acids in Peptides and Proteins. Poster presented at the 60th Biophysical Society meeting, Los Angeles, February 2016. Abstract published in Biophysical Journal., 110, Issue 3, Supplement 1, p489a, 2016

National

2. Workshop on "Intellectual Property Rights (IPR)" at IIT Guwahati, November 2016

3. Prasad, S., Mandal, I., Paul, A., Mandal, B.,Venkatramani, R. and Swaminathan, R., Investigation of Novel Spectroscopic Features in the Near Ultraviolet Region Arising from Non-Aromatic Amino Acids in Peptides and Proteins. Poster presented at the Optics within Life Sciences (OWLS) meeting, Tata Institute of Fundamental Research, Mumbai, March 2016

4. Prasad, S., Paul, A., Mandal, B.,Venkatramani, R. and Swaminathan, R., Role of Lysine residues on the Origin of Novel Absorption Spectra in the Near Ultraviolet region in

Proteins devoid of aromatic residues. First prize for poster presented at the South Asian Workshop on Optics & Photonics (SAWOP), IIT Guwahati, November 2015

5. Prasad, S., Paul, A., Mandal, B.,Venkatramani, R. and Swaminathan, R., Role of Lysine residues on the Origin of Novel Absorption Spectra in the Near Ultraviolet region in Proteins devoid of aromatic residues. Poster presented at the National Symposium on Biophysics, Jamia Milia Islamia, New Delhi, February 2015

6. Prasad, S., Paul, A., Mandal, B.,Venkatramani, R. and Swaminathan, R., Investigation of Novel Spectroscopic Features in the Near Ultraviolet Region Arising from Non-Aromatic compounds and peptides devoid of aromatic amino acids. Poster presented at the National workshop on Fluorescence and Raman Techniques, Indian Institute of Science Education and Research, Pune, December 2014.

7. Workshop on "Statistical Data Analysis" at Indian Statistical Institute (ISI), Kolkata, September 2014

8. Prasad, S., Garg, J., Seshadri, V., and Swaminathan, R., Expression, purification and Aggregation studies on Islet Amyloid Polypeptide. Poster presented at the National Symposium on "Frontiers of Biophysics, Biotechnology & Bioinformatics, University of Mumbai, December 2013

9. Prasad, S., Garg, J., Seshadri, V., and Swaminathan, R., Expression, purification and Aggregation studies on Islet Amyloid Polypeptide. Poster presented at the National Fluorescence Workshop, Saha Institute of Nuclear Physics, December 2012

REFERENCES

1. Cooper, A. *Biophysical Chemistry*. *Royal Society of Chemistry* (2011).

2. Lesk, A. M. *Introduction to Protein Architecture : The Structural Biology of Proteins*. (Oxford University Press, 2001).

3. Rodger, A. & Sanders, K. Biomacromolecular Applications of UV-Visible Absorption Spectroscopy. *Encyclopedia of Spectroscopy and Spectrometry (Edition 1)* 130 (1999).

4. Campbell, I. D. & Dwek, R. A. *Biological spectroscopy*. (Benjamin/Cummings Pub. Co., 1984).

5. Pavia, D. L., Lampman, G. M., Kriz, G. S. & Vyvyan, J. R. *Introduction to Spectroscopy*. (Cengage learning, 2013).

6. Andrews, D. L. Electromagnetic Radiation. *Encyclopedia of Spectroscopy and Spectrometry (Third Edition)* (2017).

7. Thiele, S. & Salzer, R. in *Handbook of Spectroscopy* (Wiley-VCH Verlag GmbH & Co. KGaA, 2005).

8. Noyes, W. The Correlation of Spectroscopy and Photochemistry. *Rev. Mod. Phys.* **5,** (1933).

9. Ingle, J. D. J. & Crouch, S. R. *Spectrochemical analysis*. (Prentice Hall, Inc., 1988).

10. Faust, B. in *Modern Chemical Techniques* (The Royal Society of Chemistry, 1997).

11. Pearse, R. W. B. Spectroscopy: 3. The application of spectroscopy to photochemistry. *Reports Prog. Phys.* **3,** 382 (1936).

12. Flammer, J., Mozaffarieh, M. & Bebie, H. *Basic Sciences in Opthalmology: Physics and Chemistry*. (Springer, 2013).

13. Hu, X. & Schulten, K. How nature harvests sunlight. *Physics Today* 28 (1997).

14. Metzler, D. E. *Biochemistry: The Chemical Reactions of Living Cells*. (Academic Press Inc. Ltd., 1977).

15. Atkins, P. W. & Paula, J. de. *Physical Chemistry for the Life Sciences*. (W.H. Freeman and Company, 2006).

16. Rodger, A. & Wormell, P. Absorption Spectroscopy: Practical Aspects. *Encyclopedia of Biophysics* 33 (2013).

17. Lakowicz, J. R. & Masters, B. R. *Principles of Fluorescence Spectroscopy, Third Edition*. (Springer Science+Business Media, 2008).

18. Hollas, J. M. *Modern Spectroscopy*. (John Wiley & Sons, Ltd, 2004).

19. Cantor, C. R. & Schimmel, P. R. *Biophysical Chemistry Part II: Techniques for the study of biological structure and function*. (W. H. Freeman and Company, 1980).

20. Smith, G. S. *An Introduction to Classical Electromagnetic Radiation*. (Cambridge University Press, 1997).

21. Engel, T., Reid, P. & Hehre, W. *Physical Chemistry*. (Pearson Education, Inc., 2013).

22. Owen, T. *Fundamentals of UV-visible Spectroscopy: A primer*. (Hewlett Packard, 1996).

23. Atkins, P. & Friedman, R. *Molecular Quantum Mechanics*. (Oxford University Press, 2005).

24. Parson, W. W. *Modern Optical Spectroscopy: With Exercises and Examples from Biophyscis and Biochemistry*. (Springer-Verlag, 2009).

25. Lindon, J. C., Tranter, G. E. & Koppenaal, D. W. *Encyclopedia of Spectroscopy and Spectrometry (Third edition)*. (Elsevier Ltd., 2017).

26. Glasel, J. A. & Deutscher, M. P. *Introduction to Biophysical Methods for Protein and Nucleic Acid Research*. (Academic Press, Inc., 1995).

27. Hills, A. E. Spectroscopy in Biotechnology Research and Development. *Encyclopedia of Spectroscopy and Spectrometry (Third Edition)* (2017).

28. Sudha, P. D. C. *Pharmaceutical Analysis*. (Pearson Education India, 2013).

29. Wüthrich, K. NMR studies of structure and function of biological macromolecules (Nobel Lecture). *J. Biomol. NMR* **27,** 13 (2003).

30. Tsai, C. S. *Biomacromolecules: Introduction to Structure, Function and Informatics*. (John Wiley & Sons Inc., 2007).

31. Perkampus, H.-H. *UV-VIS Spectroscopy and Its Applications*. (Springer-Verlag, 1992).

32. Yanari, S. & Bovey, F. A. Interpretation of the Ultraviolet Spectral Changes of Proteins. *J. Biol. Chem.* **235,** 2818 (1960).

33. Aitken, A. & Learmonth, M. P. *The Protein Protocols Handbook*. (Humana Press, 2002).

34. Rouessac, F. & Rouessac, A. in *Chemical Analysis: Modern Instrumentation Methods and Techniques* (John Wiley & Sons, Ltd., 2007).

35. Chang, R. *Physical Chemistry for the Biosciences*. (University Science Books, 2005).

36. Nilapwar, S. M., Nardelli, M., Westerhoff, H. V. & Verma, M. in *Methods in Enzymology* (Elsevier, 2011).

37. Rodger, A. UV Absorbance Spectroscopy of Biological Macromolecules. *Encyclopedia of Biophysics* 2714 (2013).

38. Christian, G. D. *Analytical Chemistry*. (John Wiley and Sons Inc., 2004).

39. Kuhn, H. A Quantum-Mechanical Theory of Light Absorption of Organic Dyes and Similar Compounds. *J. Chem. Phys.* **17,** 1198 (1949).

40. Shoemaker, D. P., Garland, C. W. & Nibler, J. W. *Experiments in Physical Chemistry*. (McGraw-Hill, 1989).

41. Brooker, L. G. S., Keyes, G. H. & Williams, W. W. Colour and Constitution. V. The

Absorption of Unsymmetrical Cyanines. Resonance as a Basis for a Classification of Dyes. *J. Am. Chem. Soc.* **64,** 199 (1942).

42. Moog, R. S. Determination of Carbon-Carbon Bond Length from the Absorption Spectra of Cyanine Dyes. *J. Chem. Educ.* **68,** 506 (1991).

43. Beer. Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten (Determination of the absorption of red light in colored liquids). *Ann. der Phys. und Chemie* **86,** 78 (1852).

44. Lambert, J. H. & DiLaura, D. L. *Photometry, or, On the measure and gradations of light, colors, and shade : Translation from the Latin of Photometria, sive, De mensura et gradibus luminis, colorum et umbrae*. (Illuminating Engineering Society of North America, 2001).

45. Andrews, D. L. *Perspectives in Modem Chemical Spectroscopy*. (Springer-Verlag, 1990).

46. Leermakers, P. A. & Vesley, G. F. Organic Photochemistry and the Excited State. *J. Chem. Educ.* **41,** 535 (1964).

47. Homocianu, M., Airinei, A. & Dorohoi, D. O. Solvent Effects on the Electronic Absorption and Fluorescence Spectra. *J. Adv. Res. Phys.* **2,** 1 (2011).

48. Nicol, M. F. Solvent Effects on Electronic Spectra. *Appl. Spectrosc. Rev.* **8,** 183 (1974).

49. McConnell, H. Effect of Polar Solvents on the Absorption Frequency of n→π Electronic Transitions. *J. Chem. Phys.* **20,** 700 (1952).

50. Bayliss, N. S. The Effect of the Electrostatic Polarization of the Solvent on Electronic Absorption Spectra in Solution. *J. Chem. Phys.* **18,** 292 (1950).

51. Baldini, F., Chester, A. N., Homola, J. & Martellucci, S. *Optical Chemical Sensors*. (Springer, 2006).

52. Cotton, S. *Lanthanide and Actinide Chemistry*. (John Wiley & Sons, Ltd, 2006).

53. Daniel C Harris, M. D. B. *Symmetry and Spectroscopy: An Introduction to Vibrational and Electronic Spectroscopy*. (Oxford Univeristy Press, 1978).

54. Crosby, G. A. Spectroscopic Investigations of Excited States of Transition-Metal Complexes. *Acc. Chem. Res.* **8,** 231 (1975).

55. Laporte, O. & Meggers, W. F. Some Rules of Spectral Structure. *J. Opt. Soc. Am. Rev. Sci. Instruments* **11,** 459 (1925).

56. Kleinschmidt, K. M., Dobson, J. F. & Doddrell, D. M. Vibronic Coupling and Spin Relaxation in Discrete Spin-1/2 Transition-Metal Complexes. *Chem. Phys. Lett.* **68,** 115 (1979).

57. Sridharan, K. *Spectral Methods in Transition Metal Complexes*. (Elsevier Inc., 2016).

58. Reddy, S. L., Endo, T. & Reddy, G. S. in *Advanced Aspects of Spectroscopy* 3 (InTech, 2012).

59. Aspinall, H. C. *Chemistry of the f-Block Elements*. (Gordon and Breach Science, 2001).

60. Gschneidner, K. A. *Handbook on the physics and chemistry of rare earths: Lanthanides/Actinides: Chemistry*. (Elsevier, 1994).

61. Balzani, V. & Campagna, S. *Photochemistry and Photophysics of Coordination Compounds I*. (Springer, 2007).

62. Campagna, S., Puntoriero, F., Nastasi, F., Bergamini, G. & Balzani, V. in *Photochemistry and Photophysics of Coordination Compounds I* (Springer, 2007).

63. Reddy, K. H. *Bioinorganic Chemistry*. (New Age International (P) Limited, 2005).

64. Mason, R. Charge Transfer Processes in Biological Systems. *Discuss. Faraday Soc.* **27,** 129 (1959).

65. Schuster, G. B. Long-Range Charge Transfer in DNA: Transient Structural

Distortions Control the Distance Dependence. *Acc. Chem. Res.* **33,** 253 (2000).

66. Boxer, S. G. Mechanisms of Long-Distance Electron Transfer in Proteins: Lessons from Photosynthetic Reaction Centers. *Annu. Rev. Biophys. Biophys. Chem.* **19,** 267 (1990).

67. Jortner, J., Bixon, M., Langenbacher, T. & Michel-Beyerle, M. E. Charge transfer and transport in DNA. *Proc. Natl. Acad. Sci.* **95,** 12759 (1998).

68. Bertini, I., Drago, R. S. & Luchinat, C. *The Coordination Chemistry of Metalloenzymes: The Role of Metals in Reactions Involving Water, Dioxygen, and Related Species*. (Reidel Publishing Company, 1983).

69. Pierloot, K., Kerpel, J. O. a De, Ryde, U. & Olsson, M. H. M. Relation between the Structure and Spectroscopic Properties of Blue Copper Proteins. *J.Am.Chem.Soc.* **120,** 13156 (1998).

70. Tennent, D. L. & McMillin, D. R. A Detailed Analysis of the Charge-Transfer bands of a Blue Copper Protein. Studies of the Nickel (II), Manganese (II), and Cobalt (II) Derivatives of Azurin. *J. Am. Chem. Soc.* **101,** 2307 (1979).

71. McMillin, D. R. & Morris, M. C. Further perspectives on the charge transfer transitions of blue copper proteins and the ligand moieties in stellacyanin. *Proc. Natl. Acad. Sci.* **78,** 6567 (1981).

72. McMillin, D. R., Holwerda, R. A. & Gray, H. B. Preparation and Spectroscopic Studies of Cobalt (II) -Stellacyanin. *Proc. Natl. Acad. Sci.* **71,** 1339 (1974).

73. Janes, R. & Moore, E. *Metal-Ligand Bonding*. (The Royal Society of Chemistry, 2004).

74. Gregory, R. P. F. *Photosynthesis*. (Chapman and Hall, Inc., 1989).

75. King, B. A., Stanley, R. J. & Boxer, S. G. Excited State Energy Transfer Pathways in Photosynthetic Reaction Centers. 2. Heterodimer Special Pair. *J. Phys. Chem. B* **101,** 3644 (1997).

76. Meech, S. R., Hoff, A. J. & Wiersma, D. A. Role of charge-transfer states in bacterial photosynthesis. *Proc.Natl.Acad.Sci.* **83,** 9464 (1986).

77. Amesz, J., Neuberger, A. & Deenen, L. L. M. *New Comprehensive Biochemistry: Photosynthesis*. (Elsevier, 1987).

78. Green, B. R. & Parson, W. W. *Advances in Photosynthesis and Respiration: Light-harversitng Antennas in Photosynthesis*. (Springer Science + Business Media, 2003).

79. Warshel, A. & Parsone, W. W. Spectroscopic Properties of Photosynthetic Reaction Centers.1.Theory. *J. Am. Chem. Soc.* **109,** 6143 (1987).

80. Kosower, E. M. Additions to Pyridinium Rings. III. Chemical and Biochemical Implications of Charge-Transfer Complex Intermediates. *J. Am. Chem. Soc.* **78,** 3497 (1956).

81. Pullman, B. Molecular Associations in Biology. in *International Symposium held at the Institut de Biologie Physico-chimique, Fondation Edmond de Rothschild,Paris* (Academic Press Inc. Ltd., 1968).

82. Demchenko, A. P. *Ultraviolet Spectroscopy of Proteins*. (Springer-Verlag, 1986).

83. Serdyuk, I. N., Zaccai, N. R. & Zaccai, J. *Methods in Molecular Biophysics: Structure, Dynamics, Function*. (Cambridge University Press, 2007).

84. Quickenden, T. I. & Irvin, J. A. The ultraviolet absorption spectrum of liquid water. *J. Chem. Phys.* **72,** 4416 (1980).

85. Pattabhi, V. & Gautham, N. *Biophysics*. (Kluwer Academic Publishers, 2002).

86. Hunt, H. D. & Simpson, W. T. Spectra of Simple Amides in the Vacuum Ultraviolet. *J. Am. Chem. Soc.* **75,** 4540 (1953).

87. Ham, J. S. & Platt, J. R. Far U.V. Spectra of Peptides. *J. Chem. Phys.* **20,** 335 (1952).

88. Curtis, W. J. in *Methods of Biochemical Analysis* 61 (John Wiley & Sons Inc., 1985).

89. Imahori, K. & Tanaka, J. Ultraviolet absorption spectra of poly-L-glutamic acid. *J.*

Mol. Biol. **1,** 359 (1959).

90. I Tinoco, A. H. and W. T. S. Polyamino acids, polypeptides, and proteins: Proceedings of an international symposium. in *Polyamino acids, polypeptides, and proteins* (University of Wisconsin Press, 1962).

91. Rosenheck, K. & Doty, P. The Far Ultraviolet Absorption Spectra of Polypeptide and Protein Solutions and Their Dependence on Conformation. *Proc. Natl. Acad. Sci.* **47,** 1775 (1961).

92. Bent, D. V. & Hayon, E. Excited State Chemistry of Aromatic Amino Acids and Related Peptides. III. Tryptophan. *J. Am. Chem. Soc.* **97,** 2612 (1975).

93. Creed, D. The Photophysics and Photochemistry of the Near-UV Absorbing Amino Acids–I. Tryptophan and Its Simple Derivatives. *Photochem. Photobiol.* **39,** 537 (1984).

94. Wetlaufer, D. B. in *Advances in Protein Chemistry* (Academic Press Inc., 1963).

95. Bent, D. V. & Hayon, E. Excited State Chemistry of Aromatic Amino Acids and Related Peptides. I. Tyrosine. *J. Am. Chem. Soc.* **97,** 2599 (1975).

96. Steiner, R. F. & Weinryb, I. *Excited states of Proteins and Nucleic Acids*. (Plenum Press, 1971).

97. Grinspan, H., Birnbaum, J. & Feitelson, J. Environmental Effects of the Ultraviolet Absorption Spectrum of Tyrosine. *Biochim. Biophys. Acta* **126,** 13 (1966).

98. Creed, D. The Photophysics and Photochemistry of the Near-UV Absorbing Amino Acids-II. Tyrosine and its Simple Derivatives. *Photochem. Phorobiology* **39,** 563 (1983).

99. Antosiewicz, J. M. & Shugar, D. UV–Vis spectroscopy of tyrosine side-groups in studies of protein structure. Part 1: Basic principles and properties of tyrosine chromophore. *Biophys. Rev.* **8,** 151 (2016).

100. Antosiewicz, J. M. & Shugar, D. UV–Vis spectroscopy of tyrosine side-groups in

studies of protein structure. Part 2: Selected applications. *Biophys. Rev.* **8,** 163 (2016).

101. Bent, D. V. & Hayon, E. Excited State Chemistry of Aromatic Amino acids and Related Peptides. II. Phenylalanine. *J. Am. Chem. Soc.* **97,** 2606 (1975).

102. Otey, M. C. & Greenstein, J. P. Studies on polycysteine peptides and proteins. II. Apparent dissociation constants, and ultra-violet and infrared absorption spectra of isomeric cystinylcystine peptides. *Arch. Biochem. Biophys.* **53,** 501 (1954).

103. Schmid, F.-X. Biological Macromolecules : UV-Visible Spectrophotometry. *Encycl. Life Sci.* 1 (2001).

104. Aliverti, A., Curti, B. & Vanoni, M. A. in *Methods in Molecular Biology* 9 (Springer, 1999).

105. Dawson, R. M. C. *Data for Biochemical Research*. (Clarendon Press, 1989).

106. Karnaukhova, E. *et al.* Characterization of heme binding to recombinant α1-microglobulin. *Front. Physiol.* **5,** 1 (2014).

107. Homchaudhuri, L. & Swaminathan, R. Novel Absorption and Fluorescence Characteristics of L-Lysine. *Chem. Lett.* 844 (2001).

108. Chai, B., Zheng, J., Zhao, Q. & Pollack, G. H. Spectroscopic Studies of Solutes in Aqueous Solution. *J. Phys. Chem. A* **112,** 2242 (2008).

109. Segarra-Martí, J., Coto, P. B., Rubio, M., Roca-Sanjuán, D. & Merchán, M. Towards the understanding at the molecular level of the structured-water absorption and fluorescence spectra: a fingerprint of π-stacked water. *Mol. Phys.* **111,** 1308 (2013).

110. Degtyareva, O. V., Afanasiev, V. N., Khechinashvili, N. N. & Terpugov, E. L. Structure and Properties of Liquid L-Lysine monohydrochloride and L-Glycine under exposed to a low-intense Optical Radiation. *Fundam. Res.* **4,** 1 (2013).

111. Homchaudhuri, L. & Swaminathan, R. Near Ultraviolet Absorption Arising from Lysine Residues in Close Proximity: A Probe to Monitor Protein Unfolding and

Aggregation in Lysine-Rich Proteins. *Bull. Chem. Soc. Jpn.* **77,** 765 (2004).

112. Fasman, G. D. *Practical Handbook of Biochemistry and Molecular Biology*. (CRC Press, 1992).

113. Saha, A. & Yakovlev, V. V. Detection of picomolar concentrations of lead in water using albumin-based fluorescence sensor. *Appl. Phys. Lett.* **95,** 93 (2009).

114. Anand, U. & Mukherjee, M. Exploring the Self-Assembly of a Short Aromatic Aβ(16-24) Peptide. *Langmuir* **29,** 2713 (2013).

115. Mercato, L. L. del *et al.* Charge transport and intrinsic fluorescence in amyloid-like fibrils. *Proc. Natl. Acad. Sci.* **104,** 18019 (2007).

116. Amdursky, N. *et al.* Blue Luminescence Based on Quantum Confinement at Peptide Nanotubes. *Nano Lett.* **9,** 3111 (2009).

117. Guptasarma, P. Solution-state characteristics of the ultraviolet A-induced visible fluorescence from proteins. *Arch. Biochem. Biophys.* **478,** 127 (2008).

118. Shukla, A. *et al.* A novel UV laser-induced visible blue radiation from protein crystals and aggregates: Scattering artifacts or fluorescence transitions of peptide electrons delocalized through hydrogen bonding? *Arch. Biochem. Biophys.* **428,** 144 (2004).

119. Smith, A. M. *et al.* Fmoc-Diphenylalanine Self Assembles to a Hydrogel via a Novel Architecture Based on π-π Interlocked β-Sheets. *Adv. Mater.* **20,** 37 (2008).

120. Chan, F. T. S. *et al.* Protein amyloids develop an intrinsic fluorescence signature during aggregation. *Analyst* **138,** 2156 (2013).

121. Sharpe, S., Simonetti, K., Yau, J. & Walsh, P. Solid-State NMR Characterization of Autofluorescent Fibrils Formed by the Elastin-Derived Peptide GVGVAGVG. *Biomacromolecules* **12,** 1546 (2011).

122. Pinotsi, D. *et al.* Proton Transfer and Structure-Specific Fluorescence in Hydrogen Bond-Rich Protein Structures. *J. Am. Chem. Soc.* **138,** 3046 (2016).

123. Tannenbaum, E., Coffin, E. M. & Harrison, A. J. The Far Ultraviolet Absorption Spectra of Simple Alkyl Amines. *J. Chem. Phys.* **21,** 311 (1953).

124. Hesse, M., Meier, H. & Zeeh, B. *Spectroscopic Methods in Organic Chemistry*. (Georg Thieme Verlag, 1997).

125. Kaiser, E., Colescott, R., Bossinger, C. & Cook, P. I. Color test for Detection of Free Terminal Amino Groups in the Solid-Phase Synthesis of Peptides. *Anal. Biochem.* **34,** 595 (1970).

126. Sarin, V. K., Kent, S. B. H., Tam, J. P. & Merrifield, R. B. Quantitative Monitoring of Solid-Phase Peptide Synthesis by the Ninhydrin Reaction. **117,** 147 (1981).

127. Saidel, L. J., Goldfarb, A. R. & Waldman, S. The Absorption Spectra of Amino Acids in the Region Two Hundred to Two Hundred and Thirty Millimicrons. *J. Biol. Chem.* **197,** 285 (1952).

128. Myer, Y. P. The pH-Induced Helix-Coil Transition of Poly-L-Lysine and Poly-L-glutamic acid and the 238-mμ Dichroic Band. *Macromolecules* **2,** 624 (1969).

129. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28,** 45 (2000).

130. Chen, D., Wang, J., Yan, M. & Bao, F. S. A Complex Prime Numerical Representation of Amino Acids for Protein Function Comparison. *J. Comput. Biol.* **23,** 669 (2016).

131. Cosic, I. *The Resonant Recognition Model of Macromolecular Bioactivity : Theory and Applications*. (Springer-Verlag, 1997).

132. Cosic, I. Macromolecular Bioactivity: Is It Resonant Interaction Between Macromolecules?—Theory and Applications. *IEEE Trans. Biomed. Eng.* **41,** 1101 (1994).

133. Cosic, I., Hodder, A. N., Aguilar, M.-I. . & Hearn, M. T. W. Resonant recognition model and protein topography: Model studies with myoglobin, hemoglobin and

lysozyme. *Eur. J. Biochem.* **198,** 113 (1991).

134. Nair, A. S. & Sreenadhan, S. P. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation* **1,** 197 (2006).

135. Veljkovic, V. & Slavic, I. Simple General-Model Pseudopotential. *Phys. Rev. Lett.* **29,** 105 (1972).

136. Costic, I., Pirogova, E., Vojisavljevic, V. & Fang, Q. Electromagnetic Properties of Biomolecules. *FME Trans.* **34,** 71 (2006).

137. Randic, M., Zupan, J., Balaban, A. T., Vikic-Topic, D. & Plavsic, D. Graphical Representation of Proteins. *Chem. Rev.* **111,** 790 (2011).

138. Bai, F. & Wang, T. A 2-D graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.* **413,** 458 (2005).

139. Randić, M., Butina, D. & Zupan, J. Novel 2-D graphical representation of proteins. *Chem. Phys. Lett.* **419,** 528 (2006).

140. Randic, M., Zupan, J. & Vikic-Topic, D. On representation of proteins by star-like graphs. *J. Mol. Graph. Model.* **26,** 290 (2007).

141. Eisenberg, D. Three-Dimensional Structure of Membrane and Surface Proteins. *Annu. Rev. Biochem.* **53,** 595 (1984).

142. Nozaki, Y. & Tanford, C. The Solubility of Amino Acids and Two Glycine Peptides in Aqueous Ethanol and Dioxane Solutions : Establishment of a Hydrophobicity Scale. *J. Biol. Chem.* **246,** 2211 (1971).

143. Fauchere, J. L. & Pliska, V. Hydrophobic parameters of pi amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18,** 369 (1983).

144. Eisenberg, D., Weiss, R. M., Terwilliger, T. C. & Wilcox, W. Hydrophobic Moments and Protein Structure. *Faraday Symp. Chem. Soc* **17,** 109 (1982).

145. Von Heijne, G. & Blomberg, C. Trans-membrane Translocation of Proteins: The Direct Transfer Model. *Eur. J. Biochem.* **97,** 175 (1979).

146. Janin, J. Surface and inside volumes in globular proteins. *Nature* **277,** 491 (1979).

147. Chothia, C. The Nature of the Accessible and Buried Surfaces in Proteins. *J. Mol. Biol.* **105,** 1 (1976).

148. Wolfenden, R., Andersson, L., Cullis, P. M. & Southgate, C. C. B. Affinities of Amino Acid Side Chains for Solvent Water. *Biochemistry* **20,** 849 (1981).

149. Eisenberg, D. & McLachlan, A. D. Solvation energy in protein folding and binding. *Nature* **319,** 199 (1986).

150. Biswas, K. M., DeVido, D. R. & Dorsey, J. G. Evaluation of methods for measuring amino acid hydrophobicities and interactions. *J. Chromatogr. A* **1000,** 637 (2003).

151. Kyte, J. & Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **157,** 105 (1982).

152. Horikoshi, K. & Grant, W. D. (William D. . *Extremophiles : Microbial Life in Extreme Environments*. (Wiley, 1998).

153. Rothschild, L. J. & Mancinelli, R. L. Life in Extreme Environments. *Nature* **409,** 1092 (2001).

154. Grogan, D. W. Extreme Thermophiles. *Encyclopedia of Life Sciences* (2001).

155. Quatrini, R. . & Johnson, D. B. *Acidophiles: Life in Extremely Acidic Environments*. (Caister Academic Press, 2016).

156. Horikoshi, K. Alkaliphiles. *Encyclopedia of Life Sciences* (2008).

157. Ma, Y., Galinski, E. A., Grant, W. D., Oren, A. & Ventosa, A. Halophiles 2010: Life in Saline Environments. *Appl. Environ. Microbiol.* **76,** 6971 (2010).

158. Bhowmick, A. *et al.* Finding Our Way in the Dark Proteome. *J. Am. Chem. Soc.* **138,** 9730 (2016).

159. Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci.* **112,** 15898 (2015).

160. Tommos, C., Valentine, K. G., Martínez-Rivera, M. C., Liang, L. & Moorman, V. R. Reversible Phenol Oxidation and Reduction in the Structurally Well-Defined 2-Mercaptophenol-α3C Protein. *Biochemistry* **52,** 1409 (2013).

161. Sambrook, J. & Russell, D. W. *Molecular Cloning-A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, 2001).

162. Waddel, W. J. A simple ultraviolet spectrophotometric method for the determination of protein. *J. Lab. Clin. Med.* **48,** 311 (1956).

163. Phillips, J. C. *et al.* Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **26,** 1781 (2005).

164. MacKerell, A. D. *et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **102,** 3586 (1998).

165. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics* **14,** 33 (1996).

166. Runge, E. & Gross, E. K. U. Density-Functional Theory for Time-Dependent Systems. *Phys. Rev. Lett.* **52,** 997 (1984).

167. Marques, M. A. L. & Gross, E. K. U. Time-Dependent Density Functional Theory. *Annu. Rev. Phys. Chem.* **55,** 427 (2004).

168. Yanai, T., Tew, D. P. & Handy, N. C. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **393,** 51 (2004).

169. Frisch, M. J. *et al.* Gaussian 09. (2009).

170. Lu, T. & Chen, F. Multiwfn: A Multifunctional Wavefunction Analyzer. *J. Comput. Chem.* **33,** 580 (2012).

171. Dennington, R., Keith, T. & Millam, John.:Semichem Inc., Shawnee Mission, K. GaussView Version 5. (2009).

172. Guido, C. A., Cortona, P., Mennucci, B. & Adamo, C. On the Metric of Charge Transfer Molecular Excitations: A Simple Chemical Descriptor. *J. Chem. Theory Comput.* **9,** 3118 (2013).

173. Le Bahers, T., Adamo, C. & Ciofini, I. A Qualitative Index of Spatial Extent in Charge-Transfer Excitations. *J. Chem. Theory Comput.* **7,** 2498 (2011).

174. Howell, J. R., Siegel, R. & Menguc, M. P. *Thermal Radiation Heat Transfer*. (CRC Press, 2010).

175. Lord Rayleigh, F. R. S. On the Transmission of Light through an Atmosphere containing Small Particles in Suspension, and on the Origin of the Blue of the Sky. *Philos. Mag. Ser.* **47,** 375 (1899).

176. Hulst, V. de. H. C. *Light Scattering by Small Particles*. (John Wiley & Sons, 1957).

177. Klotz, I. M. & Askounis, T. Absorption Spectra and Tautomerism of Cyanuric Acid, Melamine and Some Related Compounds. *J. Am. Chem. Soc.* **69,** 801 (1947).

178. Rogers, D. M., Besley, N. A., O'Shea, P. & Hirst, J. D. Modeling the Absorption Spectrum of Tryptophan in Proteins. *J. Phys. Chem. B* **109,** 23061 (2005).

179. Stepanek, P. & Bour, P. Multi-scale modeling of electronic spectra of three aromatic amino acids: importance of conformational averaging and explicit solute-solvent interactions. *Phys Chem Chem Phys* **16,** 20639 (2014).

180. Kessler, J., Dračínský, M. & Bouř, P. Parallel Variable Selection of Molecular Dynamics Clusters as a Tool for calculation of Spectroscopic Properties. *J. Comput. Chem.* **34,** 366 (2013).