

**COMPUTATIONAL STUDIES ON
TERTIARY STRUCTURE PREDICTION
OF SMALL PROTEINS AND
ENERGETICS OF FOLDING**

DEBARATI DASGUPTA



**DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY DELHI
MAY 2018**

©Indian Institute of Technology Delhi (IITD), New Delhi, 2018

COMPUTATIONAL STUDIES ON TERTIARY STRUCTURE PREDICTION OF SMALL PROTEINS AND ENERGETICS OF FOLDING

by

Debarati DasGupta

Department of Chemistry

Submitted

**in fulfillment of the requirements of the degree of Doctor of Philosophy
to the**



**Indian Institute of Technology Delhi
May 2018**

Dedicated to my Parents

Certificate

This is to certify that the thesis entitled, “*Computational Studies on Tertiary Structure Prediction of Small Proteins and Energetics of Folding*”, being submitted by **Miss Debarati DasGupta** to the Indian Institute of Technology, Delhi for the award of the degree of **Doctor of Philosophy** in Chemistry, is a record of bonafide research work carried out by her. Debarati DasGupta has worked under my guidance and supervision and has fulfilled the requirements for the submission of this thesis, which to my knowledge has reached the requisite standard. The results contained in this dissertation have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

Dr. B. Jayaram
Professor
Department of Chemistry
Indian Institute of Technology Delhi

Acknowledgements

My Ph.D. journey is a life-changing experience and would not have been possible without the help of many people. Foremost, I am deeply indebted to my supervisor, Prof. B. JAYARAM, for introducing me to the most fascinating field of protein folding. I would like to express whole-heartedly my sincere gratitude and appreciations for the effort he has made to groom me as a researcher since the day I joined him as a graduate student.. I am also grateful to my SRC members, head and faculties of Department of Chemistry, Indian Institute of Technology Delhi for helping and supporting me during my research career and for providing the necessary facilities in the department. I am grateful to all supporting staff at the Chemistry department, IIT Delhi for their kind help and cooperation. I would like to thank my friends Mr. Pradeep Pant, Miss Ruchika Bhat, and Miss Amita Pathak for being there in times of need. They have made my stay in the lab memorable and I shall be indebted to them for their kind and supportive nature. I would also like to thank my hostel room mate and friend Mamta Yadav and Meghashree Padhan for being there with me in times of distress. I would also thank Varun who has always been a source of inspiration. Thanks MV! I would like to thank Dr Kalyan K. Bhattacharjee (Deputy Registrar IIT Delhi) for being a source of inspiration and for helping me out in several administrative matters. Without his intervention I would not have got the chance to travel abroad and stand on a foreign soil. I am indebted to you sir.

Parents are the pillars of one's strength and I would take this opportunity to thank them with all my heart. My dad has always been a source of constant moral support and has always fulfilled my requirements both in monetary and mental terms. His long telephonic consolations during paper rejections and other academic failure has transformed me to a much stronger person, ready to take up any challenge in life. My mother's constant boost up, and encouraging words has brought me so far, considering the fact that I am not a "very highly intelligent person" and tend to sweat it out to gain something in life. A very big thanks to my 'godly' parents, who according to me are the best that I could ever get in my life. I would also like to acknowledge Late Professor Ranjit Kr. Roy, my grandfather, who had long back imprinted the "chemical love" in me and it is because of his divine blessings that I could pursue the subject Chemistry in my higher studies. In childhood days, I have spent a lot of time with him till he passed away. As a very young kid, scanning curiously through his Organic chemistry text books, instantly germinated in me the love for this subject and made me passionate about Chemistry. My grandmother Late Mrs. Gouri Roy also deserves immense

recognition for tutoring my cooking skills. Being a brilliant cooking legend of her times, years spent with her in my childhood has made me love cooking and it has been and still is the sole passion and recreation in my life when I am not doing PhD work.

I would like to thank my husband and best friend and confidant Saikat, who has tolerated me these years in spite of my frequent bouts of harsh behavior in times of distress. He has apparently been a source of constant inspiration as to how dedicated a research scholar can be and I have learnt a lot from him in these years.

Last but not the least, it is Lord Shiva who has given me constant courage and strength to face the harsh challenges in life and it is only because of His kindness and love that I have got such beautiful devoted parents and grandparents in my life. Without His blessings, nothing in life would have been possible. Thank you God.

Debarati DasGupta

Abstract

Protein structure prediction field has made some very impressive advancements in the last seven decades of intensive research. The structural databases have also been enriched with thousands of new protein structures. Starting off with a mere count of 5 structures in 1970, the latest updated version of RCSB hosts a mammoth number of 136472 protein structures. This has been solely possible because of the extensive cost, time and labour invested by experimentalists. However, genome sequencing projects have advanced aggressively, leading to an even larger repository of UniprotkB database over flooding with ~9.8 million protein sequences. This is clearly 100 times more than the number of structures available. Thus, in short, there is an ever-increasing need to generate more structures such that the sequence-structure gap can be minimized. Knowledge of tertiary structures of proteins is essential for function annotation, for mechanism elucidation of enzymes, for protein design and for computer aided drug design. Thus there is a pressing need for newer structure prediction methodologies, both *ab initio* and template based. A newer class of hybrid approaches have complemented advancements in structural biology. The cryo-EM field has also made some noteworthy methodological achievements with generating crude resolution structures of 28 Å (PDB ID 1D3E) in 1999 to ultra-high resolution 2.8 Å structure (PDB ID 6BDF) in 2017; this has immensely helped the structure prediction field to utilize cryo-EM based constraints and improve the quality of computer predicted models. In this thesis work, the primary goal was to develop a new methodology to analyse the molecular origins of convergence and uniqueness in protein structures, simultaneously use the same for creation of an automated computational protocol for predicting tertiary structures of small water soluble monomeric proteins. Structure refinement was the next step taken.

It has been known since decades that proteins are characterized by a well-defined three dimensional structure, commonly referred to as the native state and it is this native structure and its dynamics which directs protein function. In spite of the surge in structural data, it is unclear as to what is the origin of the unique biologically relevant structure? Chapter 2 of the thesis attempts at analysing the origins of convergence of protein structures using the Ramachandran Map as the starting point. Considering the ϕ, ψ space of proteins, calculation of conformational space leads to the result that starting from tripeptides and beyond, there is a reduction in conformational space,

which eventually leads to a unique three dimensional structure. Thus protein folding can be thought of as a convergent problem.

Chapter 3 entitled “From Ramachandran maps to tertiary structures of proteins” utilizes the concept of higher order Ramachandran maps to predict a coarse grained structures of small proteins and is an application of the inferences from the former chapter. The chapter accomplishes two challenging tasks, representation of a complex 3D structure and prediction of a tertiary structure starting from an input amino acid sequence. Protein structures are complex and visualization using several softwares are the only resort to appreciate their inherent complexity. However, it would be a fantastic challenge if one could use a pen and paper to illustrate a protein 3D structure in a simplified fashion without losing out on critical structural information. This has been achieved in this chapter, wherein a complicated structure has been successfully represented as an alphanumeric string (termed as 3D→1D representation). The next challenge was that, given an input amino acid sequence, is it possible to enumerate these alphanumeric strings for the sequence, without having any prior structural data. The specification of the alphanumeric string for an input amino acid sequence has been achieved using higher order tripeptide libraries (termed as 1D→3D structure prediction). The methodology and webserver developed, christened “RM2TS” contains both the methodologies described. The methodology is validated on 150 small proteins and is able to generate structures within 5Å from the native. For new sequences having no known sequence homologs, it is expedient to use the RM2TS methodology for tertiary structure prediction. It takes approximately only 30 minutes to predict the structure of a small protein.

The structure prediction community has been able to bring together in most cases, a correct topology structure varying in structural similarity from the native within 1-6 Å RMSD range. However, high resolution protein structures ~1-2 Å are an integral starting point for a wide range of applications as protein function annotation and structure based drug design. Thus, beyond structure prediction, protein structure refinement is an inevitable step towards creation of a high resolution computer predicted structure. Hence the next chapter (chapter 4) of the thesis focuses on adoption of an efficient implicit solvent based MD based protocol for structure refinement. The protocol has been successfully validated on 136 predicted structures, with an average RMSD gain of 1.3 Å. The methodology reported shows a decent success rate of 79%. Furthermore, it does not deteriorate the quality of the input modeled structure. Further improvements to the methodology

are envisioned such that the accuracies can be pushed further and structures can be driven to the 1-2 Å bin routinely.

While the protein structure prediction field is advancing drastically in the last few decades, parsing the free energies of folding, is not yet amenable to experiment. The fifth chapter of the thesis sheds light on the net free energy stabilizing the folded state vis-à-vis the role of each component e.g., electrostatics, van der Waals, hydrophobicity solvent effects to name a few. The work done, attempted to evaluate the free energies of folding of 35 proteins and rationalize a component analysis of the free energy contributors. It has thus been able to synthesise a consensus from the various divergent views on the forces dominating protein folding. It also hints at a subtle fold specific nature of free energy components. (Chapter 5)

Chapter 6 presents a summary and perspective of the work carried out in this thesis.

सार

प्रोटीन स्ट्रक्चर प्रेडिक्शन फील्ड है मेड सम वैरी इम्प्रेससिवे अडवांसमेंट्स इन थे लास्ट सेवन डेडस ऑफ इंटेंसिव रिसर्च. तेह स्ट्रक्चरल डेटाबेसस हैवे बीन एनरिचेड विथ ठोसँदस ऑफ नई प्रोटीन स्ट्रक्चर्स स्टार्टिंग ऑफ विथ ा मेरे अकॉउंट ऑफ ५ खुकुरेस इन १९७०, थे लैटेस्ट अपडेटेड वर्शन ऑफ RCSB होस्ट्स ा मैमथ नंबर ऑफ १३६४७२ प्रोटीन खुकुरेस . तहसी है बीन सोलेली पॉसिबल बिकॉज ऑफ थे ेरेसीवे कॉस्ट, टीम एंड लबोर इनवेस्टेड बी एक्सपेरिमेंटलिस्ट्स. होवेवेर जीनोम सिक्वेंसिंग प्रोजेक्ट्स हैवे एडवांसड अग्रेसिवेलेली, लीडिंग तो ान इवन लार्जर रिपॉजिटरी ऑफ ुनिप्रोट-कब डेटाबेस ओवरफ्लूडिंग विथ ~९.८ मिलियन प्रोटीन सीक्वेंसेस. तहसी सी क्लेअरल्य १०० टाइम्स मोरे थान थे नंबर ऑफ खुकुरेस अवेलेबल. थुश, इन शार्ट, तेरे सी ान एवर इन्क्रेअसिंग नीड तो गेनेराते मोरे खुकुरेस सुच तहत तेह सीक्वेंस स्ट्रक्चर गैप कैन बे मिनिमिजेड. नॉलेज ऑफ टेरतीआर्य खुकुरेस ऑफ प्रोइटिसँ इस एसेशियल फॉर फंक्शन नोटेशन, फॉर मैकेनिज्म ेलुसीदतिओं ऑफ एन्जइम्स , फॉर प्रोटीन डिजाइन एंड फॉर कंप्यूटर एडिड ड्रग डिजाइन. तसु तेरे इस ा प्रेसिंग नीड फॉर नेवेर स्ट्रक्चर प्रेडिक्शन मेथोडोलोजिज बोथ अब िनीतिओ एंड टेम्पलेट बेस्ड. A नेवेर क्लास ऑफ हाइब्रिड अप्रोचेस हैवे कम्प्लीमेंटेड अडवंसमेंट्स इन स्ट्रक्चरल बायोलॉजी. थे कर्यो एम् फील्ड है आल्सो मेड सम नोटेवार्थी मेथोडोलॉजिकल अचीवमेंट्स विथ जनरेटिंग क्रूड रेसोलुशन खुकुरेस ऑफ २८ ऐंस्ट्रॉम (पडब ID १३३०) इन १९९९ तो अल्ट्रा हाई रेसोलुशन २.८ ऐंस्ट्रॉम स्ट्रक्चर (PDB ईद ६ब्दफ) इन २०१७. थुश है इममेंसेलय हेल्पेद थे स्ट्रक्चर प्रेडिक्शन फील्ड तो ुटीलीजे कर्यो EM बेस्ड कंस्ट्रिन्ट्स एंड इम्प्रूव थे क्वालिटी ऑफ कंप्यूटर प्रेडिक्टेड मॉडल्स.

आईटी है बीन नोन सीन्स डेडस तहत प्रोटीन्स अरे चरक्टेरिजेड बी ा वेल डिफाइंड ३दिमेंशनल स्ट्रक्चर, कम्प्यालय रेफरीद तो अस थे नेटिव स्टेट एंड आईटी इस थिस नेटिव स्ट्रक्चर एंड इतस डायनामिक्स विच डिरेक्टस प्रोटीन फंक्शन. इंसपिते ऑफ तेह स्ट्रक्चरल डाटा, आईटी इस उंकलेअर अस तो व्हाट इस थे ओरिजिन ऑफ थे यूनिक् बिओलॉजिकली रिलेवेंट स्ट्रक्चर? चैप्टर २ ऑफ थे थीसिस अत्तेम्प्ट्स ात आना;ीजिंग थे ओरिजिनस ऑफ कन्वर्जेन्स ऑफ प्रोटीन खुकुरेस ुसिंग Ramachandran मैप्स अस थे स्टॉर्टिंग पॉइंट. कस्निडरिंग तेह फाई साई स्पेस ऑफ प्रोटीन्स कैलकुलेशन ऑफ कन्फोर्मेंशनल स्पेस लीडस् तो तेह रिजल्ट तहतस्ट्रेटिंग फ्रॉम त्रिपेटिड्स एंड बियाँन्ड तेरे इस ा रिडक्शन इन कन्फोर्मेंशनल स्पेस.

चैप्टर ३ ुटीलीजेस थे कांसेप्ट ऑफ हायर order Ramachandran मैप्स तो प्रेडिक्ट ा करसे ग्रेन्ड खुकुरेस ऑफ स्माल प्रोटीन्स एंड इस ान एप्लीकेशन ोफते इन्फेरेसेस फ्रॉम थे फॉर्मर चैप्टर. थे मेथडोलोग्य एंड वेबसर्वर डेवलपड क्रिस्टनेड "रम्-ट्स" कनाटिन्स बोथ थे मेथोडोलोग्य डेस्क्रिबेद. आईटी इस वलिदातेड ोँ १५० स्माल प्रोटीन्स एंड इस अबले तो गेनेराते खुकुरेस वीथिन % अंगस्ट्रोम्स.

बियाँन्ड स्ट्रक्चर प्रेडिक्शन, प्रोटीन स्ट्रक्चर रेफिनेमेंट इस ान इनएविटेबल स्टेप टुवर्ड्स क्रिएशन ऑफ ा हाई रेसोलुशन कंप्यूटर प्रेडिक्टेड स्ट्रक्चर. हंस थे नेक्स्ट चैप्टर (चैप्टर ४) फोकसेस ोँ एडॉप्शन ऑफ ान एपिफिसिएंट इम्प्लिसित साल्वेंट बेस्ड मद बेस्ड प्रोटोकॉल फॉर स्ट्रक्चर रेफिनेमेंट. थिस मेथड गिवेस ान अवेअरगे रंसद गेन ऑफ १.३ अंगस्ट्रोम्स.

व्हिले थे प्रोटीन स्ट्रक्चर प्रेडिक्शन फील्ड इस एडवांसिंग दृस्टिकल्ल्य, इन तेह लास्ट फ्यू डेडस पॉसिंग तेह फ्री ेनेर्गीएस ओफ फोल्डिंग इस नॉट येत ामनाबले तो एक्सपेरिमेंट. चैप्टर ५ शेड्स लाइट ोँ तेह नेट फ्री एनर्जी स्टैबिलिजिंग थे नेटिव स्टेट विस्-ा-विस् थे रोले ोफीच कॉम्पोनेन्ट. थे वर्क दोने अटेम्पटेड तो ेवलुएट थे फ्री ेनेर्गीएस ऑफ फोडलिंग ऑफ ३५ स्माल प्रोटीन्स एंड रेशनलाइज थे कपोनेन्ट्स एनालिसिस ऑफ थे फ्री एनर्जी कंट्रीब्युटर्स.

चैप्टर ६ प्रेजेन्ट्स ा समरी एंड पर्सपेक्टिव ऑफ थे वर्क करिंएद आउट इन थिस थीसिस.

Table of Contents

| Title | Page No |
|---|----------------|
| Certificate | I |
| Acknowledgements | III |
| Abstract | V |
| Table of Contents | VIII |
| List of Figures | XI |
| List of tables | XV |
| Chapter 1: Introduction | |
| 1.1 Introduction | 1 |
| 1.2 Protein folding problem | 6 |
| 1.3 Computational methods for protein structure prediction | 6 |
| 1.3.1 Comparative modelling | 7 |
| 1.3.2 Ab initio Modeling | 9 |
| 1.3.2.1 Molecular dynamics simulation of protein structures | 10 |
| 1.4 Scope of the thesis | 12 |
| 1.5 References | 13 |
| Chapter 2: Protein Folding is a convergent problem | |
| 2.1 Introduction | 22 |
| 2.2 Methodology | 23 |

| | |
|----------------------------|----|
| 2.3 Results and discussion | 24 |
| 2.4 Conclusions | 31 |
| 2.5 References | 33 |

Chapter 3: From Ramachandran Maps to Tertiary Structures of Proteins

| | |
|---|----|
| 3.1 Introduction | 38 |
| 3.2 Theory and Methodology | 41 |
| 3.2.1 Development of a Method to Convert a Tertiary Structure to an Alphanumeric String (3D → 1D Mapping) | 41 |
| 3.2.2 Development of a Method to Specify an Alphanumeric String for an Input (Query) Amino Acid Sequence | 44 |
| 3.3 Results and discussion | 47 |
| 3.4 Comparison with homology models | 54 |
| 3.5 Web utility | 56 |
| 3.6 Conclusions | 57 |
| 3.7 References | 59 |

Chapter 4: Protein Structure Refinement using Molecular Dynamics Simulations

| | |
|----------------------------|-----|
| 4.1 Introduction | 68 |
| 4.2 Methodology | 75 |
| 4.3 Results and discussion | 77 |
| 4.4 Conclusions | 85 |
| 4.5 References | 115 |

**Chapter 5: A component analysis of the free energies of folding of 35 proteins: A
consensus view on the thermodynamics of folding at the molecular level**

| | |
|---|-----|
| 5.1 Introduction | 124 |
| 5.2 Methodology | 130 |
| 5.3 Results and discussion | 137 |
| 5.3.1 A consensus view | 137 |
| 5.3.2 Are the energy components fold specific | 139 |
| 5.3.3 A perspective from crystal structures on fold specific signatures | 141 |
| 5.3.4 Analysis on internal waters | 144 |
| 5.3.5 Intramolecular entropy calculations | 144 |
| 5.3.6 Extension of the methodology on a larger system | 146 |
| 5.3.7 Single point mutations | 146 |
| 5.4 Conclusions | 149 |
| 5.5 References | 151 |

Chapter 6: Summary and Perspectives

| | |
|-------------------------------|-----|
| Conclusion and Future aspects | 177 |
| List of publications | 180 |
| Biodata | 181 |
| | 152 |

List of Figures

| Figure No. | Figure Title | Page |
|------------|--|-----------|
| 1.1 | Levels of protein structure depicted (a) primary sequence, (b) secondary structure, (c) motifs, (d) tertiary structure and (e) quaternary structure of proteins | 2 |
| 1.2 | Size distribution of human proteome (data from Uniprot latest release) | 4 |
| 1.3 | Yearly growth in the number of protein structures (data from latest RCSB release) | 5 |
| 1.4 | Number of protein sequences in the UniprotKB/Swiss-Prot (shown in red) along with the growth in number of protein structures in RCSB | 5 |
| 1.5 | Basic flowchart of comparative modeling and <i>ab initio</i> modeling of protein structures | 7 |
| 1.6 | Basic flowchart of a molecular dynamics (MD) protocol | 12 |
| 2.1 | Pictorial definition of monomers, dimers, trimers, tetramers and pentamers | 23 |
| 2.2 | Allowed (green) versus disallowed (red) regions on higher order Ramachandran (HOR) maps. (a) For monomer (---Ala---) on a standard Ramachandran Map ²⁶ , (b) for a dimer (---Ala-Ser---) and (c) for a trimer (---Ala-Ser-Ala---). The structures are not curated based on strict filters and hence there is some noise anticipated in the allowed regions on the Map | 26 |
| 2.3 | The fraction of allowed regions versus the number of structures in RCSB starting from 1990 to 2017 showing that the fractional occupancy is not database growth dependent. The numbers below the abscissa (x-axis) indicate the number of structures in RCSB deposited till that specific year | 27 |
| 2.4 | Plot showing log conformational volume versus sequence length; from tripeptide onwards (ln C) there is a convergence observed as the sequence length grows | 29 |

| | | |
|-------------|---|-----------|
| | | |
| 2.5 | Coarse grained tertiary structure of a protein modeled using (a) zero order Ramachandran map values, (b) first order map values and (c) second order map values | 31 |
| 3.1 | phi (Φ) and psi (Ψ) angles defined on an extended polypeptide chain | 39 |
| 3.2 | Observed Φ and Ψ values (black dots) in a protein (PDB ID: 5TQL) comprising 316 amino acid residues, an “experimental” ⁶³ Ramachandran map | 40 |
| 3.3 | 3D to 1D mapping flowchart | 43 |
| 3.4 | An illustration of 3D (tertiary structure) \rightarrow 1D (alphanumeric string) mapping for chicken villin headpiece (PDB ID: 1VII) | 43 |
| 3.5 | “First order Ramachandran map” of dipeptides – a superposition of 400 individual dipeptide maps. Rows C1 to C27 refer to the conformations of the i^{th} residue, columns C1 to C27 refer to conformations of the $(i+1)^{\text{th}}$ residue. Each of the 729 (27x27) cells is color coded as green or red depending on their population in RCSB. The figure presents observed conformational frequencies of 14.7 million (Table 3.1) residues. The green area covers 95% of the observed population and red remaining 5% | 45 |
| 3.6 | Cumulative population (%) covered by the number of conformational classes for tripeptide | 46 |
| 3.7 | 1D to 3D mapping flowchart | 47 |
| 3.8 | Histogram showing the variation in sequence length (number of amino acid residues) versus number of sequences considered | 48 |
| 3.9 | A representation of the predicted structures (red) superposed on their natives (blue) along with their PDBID | 52 |
| 3.10 | Additional 50 protein sequences modeled using RM2TS methodology, showing modeled structure (red) and superimposed on native (blue). The average sequence length taken is 82, and having at least 2 secondary structural elements. RMSDs are in the range of 2-6 Å | 52 |
| 3.11 | van der Waals energy ($-E_{\text{vdw}}/kT$) of each tripeptide in each of the conformational classes plotted against logarithm of the observed population in RCSB | 53 |

| | | |
|-------------|--|------------|
| 3.12 | (a). Predicted structures (red) with (a) RM2TS (present) method and with (b) homology modeling, superposed on the native (blue) for a protein (PDB ID: 4UZX) | 54 |
| 3.13 | A snapshot of the output generated after 3D→1D representation | 56 |
| 3.14 | Screenshot of 1D→3D structure prediction module of RM2TS web-server | 57 |
| 4.1 | Pre-processor protocol for cleaning modeled structures before MD refinement | 77 |
| 4.2 | A schematic of the MD refinement protocol | 78 |
| 4.3 | Schematic showing the performance of the refinement methodology proposed as tested on 136 modeled structures. Details of improvement of each target is present in Table 4.1 | 79 |
| 4.4 | Schematic showing the performance of top ranked refinement servers, GalaxyRefine (22% improvement), i3DRefine (39% improvement), PREFMD (21% improvement), ModRefiner (56% improvement) and Tigress (40% improvement) | 86 |
| 4.5 | Left panel shows superimposition of native (blue) with modeled structure (red), and right panel shows superimposition of native (blue) with refined model (using the above proposed methodology) (green) for 136 small proteins. | 87 |
| 5.1 | A flow chart of the methodology adopted for estimating free energies of folding. | 133 |
| 5.2 | Correlation between computed free energies of folding and experiment for 35 proteins using a) MMGBSA and b) MMPBSA methods | 138 |
| 5.3 | A consensus view of the components contributing to the free energy of folding using (a) MMGBSA methodology (b) MMPBSA methodology. Bars on the graphs represent standard error of the mean. The terminology in the bar graph is explained as follows: INT (ΔH°_{int}) (deep pink), VDW (ΔH°_{vdw}) (yellow ochre), EEL (ΔH°_{EL}) (Light pink), HYP ($T \Delta S^{\circ}_{HYP}$) (blue), ENTROPY ($T \Delta S^{\circ}_{int}$) (deep green) and ΔG° (magenta) | 139 |
| 5.4 | Components contributing to the free energy of folding using MMGBSA methodology (a) all alpha, (b) all beta, (c) mixed and through MMPBSA methodology, (d) all alpha, (e) all beta (f) mixed. Bars on the graphs represent standard error of the mean. | 141 |

| | | |
|------------|---|------------|
| | The terminology in the bar graph is explained as follows: INT ($\Delta H^{\circ}_{\text{int}}$) (deep pink), VDW ($\Delta H^{\circ}_{\text{vdw}}$) (yellow ochre), EEL ($\Delta H^{\circ}_{\text{EL}}$) (Light pink), HYP ($T \Delta S^{\circ}_{\text{HYP}}$) (blue), ENTROPY ($T \Delta S^{\circ}_{\text{int}}$) (deep green) and ΔG° (magenta) | |
| 5.5 | Structural analysis on ~2500 proteins showing (a) Number of H-bonds per residue for different folds and (b) loss of accessible surface area per residue upon folding analysis for different folds. Bars on the graphs represent standard error of the mean. Number of proteins considered is 522, 839 and 1175 for α , β and $\alpha+\beta$ folds respectively | 144 |
| 5.6 | Correlation between experimental $\Delta\Delta G^{\circ}$ (kcal/mol) and $\Delta\Delta G^{\circ}_{\text{calculated}}$ kcal/mol for 33 mutants (data in Table 5.3)] | 149 |

List of Tables

| Table No. | Title | Page No. |
|------------|--|------------|
| 1.1 | List of different softwares in public domain for comparative modeling | 7 |
| 1.2 | List of softwares that perform ab initio based protein tertiary structure prediction | 9 |
| 2.1 | Observed occupancy of the ϕ, ψ space for a few oligopeptides. Data comprises 43612 protein structures and ~26.5 million amino acids residues taken from RCSB. The values reported are weighted according to their stoichiometries ¹¹ in proteins. Since some strict filters have not been applied, there is a greater span of allowed regions on the Ramachandran Map | 24 |
| 2.2 | Configurational volume of 'n'-mers | 28 |
| 2.3 | Logarithm of configurational volume of n-mers (A, B, C, D and E have been explained in Table 2.2) | 29 |
| 3.1 | Observed distribution of amino acid residues in RCSB in the defined 27 conformational classes | 41 |
| 3.2 | Details of the 100 small proteins investigated for structure generation with RM2TS (1D \rightarrow 3D) methodology | 49 |
| 4.1 | Performance of the refinement methodology as tested on 136 systems and evaluation score assessed by ProTSAV methodology | 80 |
| 5.1 | Conditions used while collecting experimental data from ProTherm Database | 134 |
| 5.2 | Proteins studied (35 systems) and their experimental (data collected from ProTherm Database ⁹⁸) and calculated free energies of folding | 135 |

| | | |
|-------------|--|------------|
| 5.3 | Calculated and experimental free energies of folding given as $\Delta\Delta G^\circ$ with respect to the native for 33 mutants belonging to three proteins (PDB ID 1BVC, 1MJC, 1STN) | 147 |
| A5.1 | The unscaled values of the calculated free energy components (using MMGBSA). First row has PDB ID followed by the component name. Last row has the net values of energy components. | 160 |