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

DEBARATI DASGUPTA

DEPARTMENT OF CHEMISTRY

INDIAN INSTITUTE OF TECHNOLOGY DELHI

MAY 2018

COMPUATATIONAL 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 डडेस ऑफ़ इंटेंसिव रिसर्च. तेह स्ट्रक्चरल डेटाबेसस हैवे बीन एनरिचेड विथ ठोसँदस ऑफ़ नई प्रोटीन स्ट्रूक्टइर्स स्टार्टिंग ऑफ विथ ा मेरे अकॉउंट ऑफ़ ५ स्रुक्टुरेस इन १९७०, थे लैटेस्ट अपडेटेड वर्शन ऑफ़ RCSB होस्ट्स ा मैमथ नंबर ऑफ़ १३६४७२ प्रोटीन स्रुक्टुरेस . तहसी है बीन सोलेली पॉसिबल बिकॉज़ ऑफ़ थे ेटेंसीवे कॉस्ट, टीम एंड लबोर इनवेस्टेड बी एक्सपेरिमेन्टलिस्ट्स. होवेवेर जीनोम सिक्वेंसिंग प्रोजेक्ट्स हैवे एडवांस्ड अग्रेस्सिवेली, लीडिंग तो ान इवन लार्जर रिपॉजिटरी ऑफ़ ुनिप्रोत-कब डेटाबेस ओवरफ्लूडिंग विथ ~९.८ मिलियन प्रोटीन सीक्वेंसेस. तहसी सी क्लेअर्ल्य १०० टाइम्स मोरे थान थे नंबर ऑफ़ स्रुक्टुरेस अवेलेबल. थुश, इन शार्ट, तेरे सी ान एवर इन्क्रेअसिंग नीड तो गेनेराते मोरे स्रुक्टुरेस सुच तहत तेह सीक्वेंस स्ट्रक्चर गैप कैन बे मिनिमिज़ेड. नॉलेज ऑफ़ टेरतीआर्य स्रुक्टुरेस ऑफ़ प्रोइटिसँ इस एसेंशियल फॉर फंक्शन नोटेशन, फॉर मैकेनिज्म ेलुसीदतिओं ऑफ़ एन्ज़इम्स , फॉर प्रोटीन डिज़ाइन एंड फॉर कंप्यूटर एडिड ड्रग डिज़ाइन. तसु तेरे इस ा प्रेसिंग नीड फॉर नेवेर स्ट्रक्चर प्रेडिक्शन मेथोडोलोजिज़ बोथ अब िनीतिओ एंड टेम्पलेट बेस्ड. A नेवेर क्लास ऑफ़ हाइब्रिड अप्प्रोचेस हैवे कम्प्लीमेंटेड अद्वंसमनेट्स इन स्ट्रक्चरल बायोलॉजी. थे कर्यो एम् फील्ड है आल्सो मेड सम नोटेवार्थी मेथोडोलॉजिकल अचीवमेंट्स विथ जनरेटिंग क्रूड रेसोलुशन स्रुक्टुरेस ऑफ़ २८ ऐंस्ट्रॉम (पड़ब ID १ड़३े) इन १९९९ तो अल्ट्रा हाई रेसोलुशन २.८ ऐंस्ट्रॉम स्ट्रक्चर (PDB ईद ६ब्दफ) इन २०१७. थुश है इममेंसेलय हेल्पेद थे स्ट्रक्चर प्रेडिक्शन फील्ड तो ुटीलीज़े कर्यो EM बेस्ड कंस्ट्रीन्ट्स एंड इम्प्रूव थे क्वालिटी ऑफ़ कंप्यूटर प्रेडिक्टेड मॉडल्स.

आईटी है बीन नोन सीन्स डडेस तहत प्रोटीन्स अरे चरक्टेरिज़ेड बी ा वेल डिफाइंड ३दिमेंशनल स्ट्रक्चर, कम्यांलय रेफ़रीद तो अस थे नेटिव स्टेट एंड आईटी इस थिस नेटिव स्ट्रक्चर एंड इतस डायनामिक्स व्हिच डिरेक्टस प्रोटीन फंक्शन. इंसपिते ऑफ़ तेह स्ट्रक्चरल डाटा, आईटी इस उंक्लेअर अस तो व्हाट इस थे ओरिजिन ऑफ़ थे यूनिक बिओलॉजिकली रिलेवेंट स्ट्रक्चर? चैप्टर २ ऑफ़ थे थीसिस अत्तेम्प्ट्स ात आना;ीज़िंग थे ओरिजिनस ऑफ़ कन्वर्जेन्स ऑफ़ प्रोटीन स्रुक्टुरेस ुसिंग Ramachandran मैप्स अस थे स्टॉर्टिंग पॉइंट. कस्निडरिंग तेह फाई साई स्पेस ऑफ़ प्रोटीन्स कैलकुलेशन ऑफ़ कन्फोर्मेशनल स्पेस लीडस् तो तेह रिजल्ट तहतस्ट्रेटिंग फ्रॉम त्रिपेप्टिड्स एंड बियॉन्ड तेरे इस ा रिडक्शन इन कन्फोर्मेशनल स्पेस.

चैप्टर ३ ुटीलीज़ेस थे कांसेप्ट ऑफ़ हायर order Ramachandran मैप्स तो प्रेडिक्ट ा करसे ग्रेन्ड स्रुक्टुरेस ऑफ़ स्माल प्रोटीन्स एंड इस ान एप्लीकेशन ोफ़्ते इन्फेरेंसेस फ्रॉम थे फॉर्मर चैप्टर. थे मेथडोलोग्य एंड वेबसर्वर डेवलप्ड क्रिस्टनेड "रम्२ट्स" कनाटिन्स बोथ थे मेथोडोलोग्य डेस्क्रिबेद. आईटी इस वलिदातेड़ ोें १५० स्माल प्रोटीन्स एंड इस अबले तो गेनेराते स्रुक्टुरेस वीथिन % अंगस्ट्रोम्स.

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

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

चैप्टर ६ प्रेजेंट्स ा समरी एंड पर्सपेक्टिव ऑफ़ थे वर्क कर्रिएद आउट इन थिस थीसिस.

Table of Contents

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

Chapter 6: Summary and Perspectives

152

List of Figures

List of Tables