DEVELOPMENT OF A HOMOLOGY/*AB INITIO* HYBRID METHODOLOGY FOR SAMPLING NEAR-NATIVE PROTEIN CONFORMATIONS

PRIYANKA DHINGRA

DEPARTMENT OF CHEMISTRY

INDIAN INSTITUTE OF TECHNOLOGY DELHI

NOVEMBER 2014

DEVELOPMENT OF A HOMOLOGY/*AB INITIO* HYBRID METHODOLOGY FOR SAMPLING NEAR-NATIVE PROTEIN CONFORMATIONS

by

PRIYANKA DHINGRA

DEPARTMENT OF CHEMISTRY

Submitted

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

INDIAN INSTITUTE OF TECHNOLOGY DELHI

November 2014

Dedicated to my Maa and Dad

Thanks for your immense love, support and faith

Certificate

This is to certify that the thesis entitled, "**Development of a Homology/ *Abinitio* Hybrid Methodology for Sampling Near-native Protein Conformations**", being submitted by **Ms. Priyanka Dhingra** 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. Ms. Priyanka Dhingra 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.

Prof. B. Jayaram
Department of Chemistry
Indian Institute of Technology Delhi

Dated:

New Delhi

Acknowledgements

Ph.D. is an enticing and challenging journey with frequent moments of failure, stress along with fruits of success and excitement. This journey would not have been possible without the help, support and guidance of many people. I would like to take this opportunity to thank them for helping me climb this rocky terrain of research.

Foremost, I would like to express my deepest gratitude and sincere thanks to my Ph.D. supervisor Prof. B. Jayaram, Department of Chemistry, IIT Delhi for his fundamental role in my doctoral work. Prof. Jayaram believed in my capabilities and provided me with an opportunity to do research under his able guidance. I got associated with him in 2007 as a M.Sc. project student and later joined as a Ph.D. in the Department of Chemistry. He introduced me to the fascinating world of proteins and imbibed within me a desire to fold protein. He allowed me to think independently and develop new skills and ideas. His scientific judgment, passion, motivation and patience helped me in developing a scientific attitude. I am deeply indebted to him for providing me with state-of-the-art facilities and research oriented environment. Prof. Jayaram provided me ample opportunities to share and present my research work with the wider scientific community. His encouragement and support has been driving force throughout this journey.

I would like to thank Dr. Vivekanandan Perumal, School of Biological Sciences, IIT Delhi for his help, support and guidance during our collaboration on a project. I would like to thank Prof. Aditya Mittal, School of Biological Sciences, IIT Delhi for his teachings, scientific inputs, and words of encouragement during the Ph.D. work.

I would like to thank all the past and present Dean of Students, Dean of PG section, Dean IRD, IIT Delhi for their cooperation and support. I would like to thank the past and present Heads, DRC Chairmen, and faculty and staff members for their suggestions, help and support.

I am thankful to the Department of Biotechnology, Government of India for the financial support and International travel grants for attending INCOB, Malaysia. I am thankful to Department of Science and Technology, Government of India for their financial support in attending CASP10 conference in Gaeta, Italy.

A healthy and interactive research environment is instrumental for successful research. My past and present colleagues at the Computer Modeling Lab, Department of Chemistry and Supercomputing Facility for Bioinformatics and Computational Biology, IIT Delhi have contributed significantly in my journey. I would like to thank my senior Dr. Kumkum Bhushan for her guidance, scientific inputs and support. Special thanks to Bharat and Satya for their efforts and valuable inputs in my research, working with them was a constant learning process. I am thankful to Shashank for providing necessary computational resources. I am thankful to Protein Folding team members, Avinash, Rahul, Goutam, Ankita, Abhilash and Ashutosh. I would like to thank Preeti Mam and Sanjeev for the administrative support during this period and Srinivas for his cheerful and motivating attitude.

I am very grateful to my best friend, roommate and colleague Tanya for her constant support, patience, love, affection, encouragement, guidance and valuable scientific inputs in the past 9 years of our friendship. Without her this journey would not have been possible. I enjoyed the time we spent together in IIT, the late night ramp walks, coffee breaks, discussions and outings.

Very special thanks to my parents and grandma for their immense support, unconditional love and encouragement. I am at loss of words to thank my Dad for his tremendous faith in me and my capabilities. He was and will always be my first teacher, who mentored me in pursuing my chosen goals. Thanks to my loving Maa and grandma for their blessings, love and care. I wish to thank my loving sister Yatika and brother-in-law Himanshu for having faith in me and always being there for me. I am blessed to have two angels in my life my niece Gunnu and nephew Vishu. My angels made my each home visit so rejuvenating and fun filled. Thanks to my loving husband Vivek for helping me cover the last mile of this journey. The confidence, love and support, which he has placed in me will always be a driving force.

Lastly I would like to thank Almighty God (Maa Bhagwati) for giving me strength, and courage to face the challenges of life. Without her blessings, this work would not have been possible.

(Priyanka Dhingra)

Abstract

The tertiary structure prediction of a protein using amino acid sequence information alone is one of the fundamental unsolved problems in computational biology/ molecular biophysics known as "*The Protein Folding Problem*". Protein folding, considered to be a grand challenge problem in modern science and a holy grail of molecular biology, remains intractable even after six decades since the report of the first crystal structure. Proteins are the biomolecules that are involved in almost every biological process. Their functions range from maintaining cellular shape, organization, membrane potential, reaction catalysis, transport, signaling and immunity. Any consideration of protein function requires understanding of its three dimensional structure. Knowledge of protein tertiary structure has wide range of scientific applications in different areas of research such as structure based drug design, protein functional annotation, understanding mechanism of molecular recognition, protein design and engineering. The advent of human genome sequencing project has led to an explosion in the number of protein sequences in databanks. Despite the availability of over half a million sequences in UniProtKB/SwissProt database and around 37 million protein sequences in non-redundant (NR) database, there are only 99,775 X-ray and NMR structures in the protein data bank (PDB). This diverging gap between available sequences and structures calls for an immediate *in silico* solution. Computational methods such as homology modeling which rely on extracting information from the known structures in PDB have proved to be successful in predicting tertiary structures of sequences which share high sequence similarities. Algorithms such as fold recognition/threading have further supplemented in searching PDB for distant homologs, while *ab initio/ devo no* methods attempt to fill the gaps/missing links in the protein structure. Despite the progress made, the challenges faced by tertiary structure prediction strategies involve prediction of structure of

proteins with no significant sequence similarity, exploration of protein fold space to generate near-native conformations, refinement of low resolution protein models to atomistic details and finally, a single automated computational protocol to predict near-native protein candidate structures for a given amino acid sequence. In an attempt to meet these challenges, newer homology/*ab initio* hybrid approaches are being explored to solve the tertiary structure prediction problem.

The aim of this thesis was to develop an automated homology/ *ab initio* hybrid computational protocol for predicting tertiary structure of soluble monomeric proteins. Complementing the limitations of individual methodologies i.e. homology modeling and *ab initio*, the goal was to evolve a hybrid method for predicting tertiary structure of proteins. This has been achieved through a combination of individual modules, which includes (i) a robust *Bhageerath ab initio* protein structure modeling protocol for predicting tertiary structures of small globular proteins ($\leq$ 100 amino acids), (ii) a homology/*ab initio* hybrid method for exhaustive sampling of protein conformation space and (iii) a quantum mechanics (PM6) based protein loop bond angle optimization method for refining low resolution proteins models. All the individual modules are pipelined in the freely accessible *Bhageerath*-H web server (http://www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp). For a given amino acid sequence, *Bhageerath*-H web server predicts tertiary structures of soluble proteins without any size limitation. The web server was validated on 75 CASP10 targets and showed the ability to predict a structure in top 5 with TM-score $\geq$ 0.5 in 91% of the cases, while in 59% of the cases *Bhageerath*-H was able to predict a model in top 5 having a Cα RMSD (root mean square deviation) $\leq$ 5.0 Å from the native. Currently, *Bhageerath*-H is being fielded in the ongoing CASP11 experiment (1-May-14 to 18-July-14).

This thesis has been divided into six chapters. Chapter 1 gives an introduction and application of the field of protein tertiary structure prediction. An overview of the

computational methodologies (comparative/homology modeling, threading/fold recognition and *ab initio/de novo* prediction) is presented. A brief note on critical assessment of methods for protein tertiary structure (CASP) experiment and scope of the thesis is presented. Chapter 2 describes *Bhageerath ab initio/ de novo* method for protein tertiary structure prediction of small globular proteins. This chapter explains in detail the *Bhageerath* pathway for predicting structure of proteins with $\leq$ 100 amino acids and $\leq$ 5 secondary structural elements. The chapter discusses each of the eight computational modules and the patching algorithm for proteins with more than 3 secondary structural elements. Validation of the *Bhageerath* web server on 80 small globular proteins is presented. In Chapter 3, a homology/ *ab initio* hybrid methodology is proposed for effective sampling of protein conformation space especially for proteins >100 amino acids. Development of *Bhageerath*-H Strgen method for protein decoy generation, its validation on CASP9 dataset, performance comparison with different decoy generation methods and availability as a web tool is presented. Chapters 4 discusses the role of bond angles parameters as an important contributor to force field noise and indicate the systematic role played by bond angle in protein structure modeling. Overlooking their role can lead to structural inaccuracies. The chapter presents a quantum mechanical (PM6) method for protein loop bond angle optimization for refining low resolution protein models. Performance of the methodology on CASP10 dataset is shown. The methodologies explained in Chapters 2, 3 and 4 are streamlined along with few additional modules into *Bhageerath*-H, a fully automated web server for protein tertiary structure prediction. Testing of the web server on 75 CASP10 targets and evaluation of its performance with other software is presented in Chapter 5. Lastly, Chapter 6 summarizes the above thesis work and presents some perspectives on the evolving field of protein tertiary structure prediction.

Contents