DEVELOPMENT AND ASSESSMENT OF COMPUTER ALGORITHMS FOR THE PREDICTION OF TERTIARY STRUCTURES OF SMALL PROTEINS

by

KUMKUM BHUSHAN

Department of Chemistry

SUBMITTED

IN FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

to the

INDIAN INSTITUTE OF TECHNOLOGY, DELHI

INDIA

October, 2007

Certificate

This is to certify that the thesis entitled, "Development and assessment of computer algorithms for the prediction of tertiary structures of small proteins", being submitted by Ms. Kumkum Bhushan 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. Kumkum Bhushan 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

Acknowledgements

It is a pleasure and honor for me to thank Prof. B. Jayaram, my supervisor and the present Head of the Department of Chemistry, IIT Delhi, who helped me in climbing the rocky terrain of research during the past five and half years. I would like to take this opportunity to express my gratitude to him for allowing me to do research under his guidance and for keeping his confidence in my capabilities. His scientific judgment is impeccable and he helped me in developing a scientific attitude by his patience and motivation. He pushed me to think independently and develop new ideas towards protein structure prediction. I am deeply indebted to Prof. B. Jayaram for providing me with state of the art facilities and a research oriented environment along with his encouragement and support.

I would like to thank all my colleagues (present and past) from the Supercomputing Facility for Bioinformatics and Computational Biology and the Computer Modeling Lab of the Department of Chemistry, IIT Delhi for their help and ccoperation during my research work and associated activities.

I would also like to thank all the past and present Heads, DRC Chairmen and the faculty and staff members of the Department of Chemistry, IIT Delhi for providing their help, support and suggestions during this project.

I would like to express my sincere thanks to the Council for Scientific and Industrial Research for their financial assistance.

I wish to thank all my friends (from post graduation) especially Poonam, Arti, Sharani, Prosenjit and Pravin for always being there for me.

I greatly appreciate the love, patience and encouragement received from my grandparents, parents, in-laws and brother. I would also like to thank my colleague, friend and husband, Dr. Tarun Jain for being there in times of distress and for never losing his patience with me. This journey would have been very difficult or rather impossible without all of them.

Finally, I would like to thank Lord Almighty for giving me strength and courage to face the challenges laid out before me.

Abstract

The prediction of tertiary structure (three dimensional functional form) of a protein starting from its amino acid sequence information alone is one of the fundamental unsolved problems in computational biology / molecular biophysics. The folding of protein molecules with a large number of degrees of freedom spontaneously into a unique three-dimensional structure is of scientific interest intrinsically and also due to its application in structure based drug design endeavors, design of biocatalysts as well as designer proteins. The genome sequencing projects have lead to an explosion in the protein sequence databank as well. The cost and time factors involved in experimental methods for protein tertiary structure determination urge for a rapid solution. The computational methods can be divided into comparative modeling and *ab initio* methodologies. Both methodologies have certain disadvantages, the comparative modeling approaches are fast but are limited by the size and diversity of the structural database whereas the *ab initio* methodologies are accurate but time consuming. Thus to maintain the rigor in calculations and to overcome the limitations, *de novo* methodologies are adopted.

The aim of this thesis work has been towards development of an energy based protocol for limiting the search space of tertiary structures of small globular

proteins. This has been achieved through a combination of physics based potentials with biophysical filters and a Protein Regularity Index (*ProRegIn*) to arrive at 10 plausible candidate structures starting from sequence and secondary structure information. The methodology has been validated on 50 small globular proteins consisting of 2-3 secondary structural elements and at least one or more structures within 3-6 Å RMSD (root mean square deviation) from the native have been obtained in each case. To incorporate structural features in proteins such as beta sheet formation (via hydrogen bonding), disulphide bridge formation (covalent interactions) and metal ion coordination (as in zinc fingers) constraint minimization have been utilized. Explicit solvent molecular dynamics simulations are carried out on some candidate structures to propel the system towards better side chain packing. The protocol has been automated in the form of a web server (http://www.scfbio-iitd.res.in/bhageerath) and is made freely accessible for usage by the scientific community.

This thesis is divided into eight chapters. Chapter 1 gives an introduction of the protein tertiary structure prediction methodologies. An overview of the computational methodologies (comparative modeling, *ab initio* as well as *de novo*) is also presented. In addition, a brief account of the evaluation of the prediction methodologies carried out biannually (CASP) is presented. Chapter 2 describes the methodology developed as a part of this thesis for the prediction of

tertiary structures of proteins starting from its sequence and secondary structure information. The modular form of the methodology has been explained which allows for the automation of the methodology as a web server. In Chapter 3, an index based on the regularity observed in the loop regions of the 7351 unique proteins from the database is presented. The trends observed and the formulation and validation of a filter for selection of native-like structures is also described. Chapter 4 describes the results obtained from the methodology explained in Chapter 2 on 50 small globular proteins. A comparison of the results obtained with the comparative modeling approaches as well as two other *de novo* web servers is also presented. Chapter 5 explains the protocol devised for the further refinement of structures obtained from the methodology for the proteins with beta sheets, disulphide linkages and metal ions. Chapter 6 explains the protocol followed for carrying out the explicit solvent molecular dynamics simulations on a few systems. The methodology explained in Chapter 2 has been automated in the form of a web server which is described in Chapter 7. In addition, each of the modules has been made into an independent utility which are also described here. Chapter 8 summarizes the above thesis work and gives some interesting perspectives on protein tertiary structure prediction.

Contents