A COMPUTATIONAL PATHWAY FOR BRACKETING NATIVE-LIKE TERTIARY STRUCTURES FROM SEQUENCE AND SECONDARY STRUCTURAL INFORMATION OF SMALL ALPHA HELICAL GLOBULAR PROTEINS

by

POOJA NARANG

DEPARTMENT OF CHEMISTRY

THESIS

SUBMITTED

IN FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

to

INDIAN INSTITUTE OF TECHNOLOGY, DELHI

HAUZ KHAS, NEW DELHI

INDIA

DECEMBER 2005

CERTIFICATE

This is to certify that the thesis entitled "*A Computational Pathway for Bracketing Native-Like Tertiary Structures from Sequence and Secondary Structural Information of Small Alpha Helical Globular Proteins*" being submitted by Ms. Pooja Narang 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. Pooja Narang 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 other University or Institute for the award of any degree or diploma.

Date: 15/3/07

(B. JAYARAM)
Professor of Chemistry,
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi-110016
India

Acknowledgements

The thesis is a culmination of efforts carried out by me over the last five years in the direction of research in the challenging field of protein folding. Several people were part of the process and this is an opportune moment to thank every one of them. First off, I would like to express my sincere gratitude to my research supervisor Prof. B Jayaram, for introducing me to the exciting field as a beginner. His scientific judgement, motivation and patience helped me manage and persevere the often bumpy ride during my journey, whilst at the same time enjoying the little fruits of success that came along the way.

I am thankful to the Head of the Department of Chemistry, Prof. U K. Nadir, for his cooperation, and providing the necessary facilities to work with in the department.

I would like to acknowledge the financial support from Department of Biotechnology, Government of India; INSA and IIT for attending the international conference at Albany, USA.

My colleagues at the Supercomputing Facility for Bioinformatics and Computational Biology, IIT Delhi, have made my journey through the often wild and dark alleyways of science zestful and full of light. I am proud to be one of those privileged people who have seen the Supercomputing Facility growing as a National facility within a short course of time since its inception three years ago. I sincerely thank all of the members for their love and affection I received during my stay. The support and affection from my senior, Dr N. Latha is gratefully acknowledged. Special thanks are due to Surojit for intellectual discussions, enthusiasm, and his never-say-die attitude that helping me in holding a positive outlook towards research in particular and life in

general. I am also thankful to Saher for the scientific discussions and her encouragement at every step. I had a nice time in the company of Praveen with a cheerful attitude, Vidhu with his patience and calmness, Pankaj with enthusiasm and jovial nature. I also thank Gandimathi and Gurvisha for the love and affection I received from them, Samrat for instilling his down-to-earth attitude in me, Anuj for taking the jokes played on him in a right spirit and Shashank for always taking my side during heated discussions. I also thank Dr Shenoy, Kumkum, Poonam & Tarun for their company.

My boisterous hostelmates made life lighter and happier whenever I felt drudgeries coming along the way. My special thanks goes to Vidyottama for always being there, Arunima for her guidance from time to time and Anshu and Garima for always making themselves available for discussions over coffee.

I would like to thank my father and mother for their constant support and encouragement. My father always mentored me in pursuing the chosen goals relentlessly in life and my mother was, and always will be, my first teacher. They always missed me because of my few visits to home. I wish to thank my elder sister and brother-in-law, for always being there for me at times when I needed them the most. I would also like to thank my younger sister and brother for always having faith in me.

My husband Vinay has always been with me through the thick and thin of life and there are no words to describe his love, affection, care and being extremely patient in sharing each and every thought of mine and always keeping faith in me and my abilities.

Lastly I would like to thank Almighty God for giving me a direction and a sense of purpose in my life. Without His blessings, the work described herein would not have been possible.

Pooja Narang

Abstract

The protein folding problem has come to center stage with the growing genome databases on one hand and the pressing necessity for the discovery of structures of new drug targets for life threatening diseases on the other, and an early solution either via theory or experiment or both is expected. Despite advances in the field of bioinformatics for structure prediction and growing genomic and structural databases, the insurmountable time scale problem in simulating protein folding *in silico* continues to raise doubts whether a computational solution to the protein folding problem -categorized as an NP-hard problem- is within reach in the near future. With the advent of protein engineering experiments, structure prediction techniques, clusters and supercomputers, the field of protein structure prediction has progressed immensely. Combining some specially designed biophysical filters and vector algebra tools with *ab initio* methods, the thesis presents a promising computational pathway for bracketing native-like structures of small alpha helical globular proteins departing from secondary structural information. The automated pipeline is able to bracket a few structures to with in 3-5 Å of the native. Thus the formidable "needle in a hay-stack" problem is narrowed down to finding an optimal solution amongst a computationally tractable number of alternatives at least for small proteins (with less than 100 amino acids).

The thesis is divided into eight chapters. Chapter 1 provides introduction to the protein folding problem and different approaches for protein structure prediction. A computational pathway for predicting tertiary structures for proteins starting from the amino acid sequence and secondary structure information is presented in Chapter 2. This

chapter also presents the database analyses carried out on known representative proteins from the PDB for the dihedral angles, employed for the three dimensional model building of proteins. Chapter 3 concentrates on various methods for an efficient sampling of the conformational space of the polypeptide chain of given amino acid sequence. A comparative analysis of each method is provided. The conformations generated are assessed for their native-like characteristics based on a few Biophysical Filters. Development of these Biophysical Filters and their applicability in protein structure prediction attempts is described in Chapter 4. Chapter 5 focuses on the necessity for a reliable empirical scoring function for *ab initio* protein structure prediction. The performance of the proposed empirical scoring function for protein structure prediction is investigated via its applicability on several publicly available as well as some new decoy sets. The success of the proposed protocol in bracketing native-like structures for twelve small alpha helical globular proteins is demonstrated as a proof of concept in Chapter 6. Extension of the methodology to mixed alpha beta proteins is provided in Chapter 7. Finally a critical assessment of the proposed computational pathway together with further improvements envisioned is summarized in Chapter 8.

CONTENTS

Chapter I. Introduction

Chapter II. A Computational Pathway

Chapter III. Protein trial structure generation

Chapter IV. Biophysical filters

Chapter V. An empirical scoring function for ranking trial structures

Chapter VI. A case study of twelve small alpha helical globular proteins

Chapter VII. Extension of the methodology to mixed alpha/beta globular proteins

Chapter VIII. Summary and Perspectives

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

CASP	Critical Assessment of Structure Prediction
RMSD	Root Mean Square Deviation
PDB	Protein Data Bank
NMR	Nuclear Magnetic Resonance Spectroscopy
AMBER	Assisted Model Building with Energy Refinement
CHARMM	Chemistry at HARvard Molecular Mechanics
OPLS	Optimized Liquid State
GROMOS	GROningen MOlecular Simulation
MD	Molecular Dynamics
MM	Molecular Mechanics
SD	Steepest Descent
CG	Conjugate Gradient
GB	Generalized Born
PES	Potential Energy Surface
MC	Monte Carlo
MD	Molecular Dynamics
BPTI	Bovine Pancreatic Trypsin Inhibitor