

**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:

  
(B. JAYARAM) 15/3/02

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*

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 within 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

<i>Certificate</i>	i
<i>Acknowledgements</i>	ii
<i>Abstract</i>	iv
<i>List of Figures</i>	x
<i>List of Tables</i>	xii
<i>Abbreviations</i>	xiv

### ***Chapter I. Introduction***

1.1	Introduction	...	2
1.2	Protein folding problem	...	3
1.3	Why fold proteins?	...	4
1.4	Forces stabilizing folded proteins	...	8
1.5	Protein folding pathways/models	...	8
1.6	Experimental techniques for protein structure prediction	...	9
1.7	Computational strategies for protein structure prediction	...	11
1.8	<i>Ab initio</i> protein structure prediction	...	13
1.9	Potential energy functions	...	15
1.10	Previous work in the area of <i>ab initio</i> structure prediction	...	18
1.11	Challenges ahead	...	20

### ***Chapter II. A Computational Pathway***

2.1	Introduction	...	23
2.2	A computational pathway	...	24
2.3	From sequence to generation of 3-D representation of a linear polypeptide chain	...	27
2.4	From linear chain to structure with preformed secondary structures	...	28

2.4.1	Main-chain Ramachandran angle analysis	...	28
2.4.2	Side-chain rotamer angle analysis	...	29

### ***Chapter III. Protein trial structure generation***

3.1	Introduction	...	36
3.2	Methodology	...	38
3.2.1	Grid based sampling method	...	38
3.2.2	Dihedral based sampling method-1	...	39
3.2.3	Dihedral based sampling method-2	...	39
3.3	Results	...	41
3.4	Discussion	...	42
3.5	Conclusions	...	44

### ***Chapter IV. Biophysical filters***

4.1	Introduction	...	46
4.2	Theory and Methodology		
4.2.1	Biophysical filters	...	47
4.2.2	Clash removal	...	56
4.2.3	Energy minimization	...	56
4.3	Results and Discussion	...	57
4.5	Conclusions	...	59

### ***Chapter V. An empirical scoring function for ranking trial structures***

5.1	Introduction	...	61
5.2	Theory and Methodology		

5.2.1	Empirical scoring function	...	62
5.2.2	Assessment of the scoring function on publicly available decoys	...	65
5.2.3	Assessment of the scoring function on homology models	...	68
5.3	Results	...	70
5.5	Discussion	...	78
5.6	Conclusions	...	80

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

6.1	Introduction	...	82
6.2	Methodology	...	82
6.3	Results		
6.3.1	Trial structure generation	...	83
6.3.2	Filtering the trial structures	...	84
6.3.3	Clash removal and energy minimization	...	84
6.3.4	Energy scans using empirical scoring function	...	85
6.3.5	Monte Carlo optimization	...	85
6.3.6	Characterization of the selected structures	...	86
6.4	Discussion	...	91
6.5	Computational time required	...	92
6.6	Conclusions	...	93

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

7.1	Introduction	...	95
7.2	Methodology	...	95

7.2.1	Ramachandran angle analysis for sheet region and generation of initial structure	...	97
7.2.2	Generation of trial structures and application of Biophysical Filters	...	100
7.2.3	Removal of clashes and energy minimization	...	100
7.2.4	Ranking and selection of 100 lowest energy structures	...	100
7.2.5	Minimization with distance restraints	...	101
7.3	Results	...	101
7.4	Conclusions	...	104

### ***Chapter VIII. Summary and Perspectives***

8.1	Summary	...	106
8.2	Perspectives and suggestions for future work	...	107
	<i>References</i>	...	109
	<i>Appendices</i>	...	139

## LIST OF FIGURES

		<i>Page No.</i>
Fig 1.1	A pictorial representation of the <i>protein folding problem</i>	4
Fig 1.2	Biochemical classes of drug targets	5
Fig 2.1	A computational pathway for bracketing native-like structures of small alpha helical globular proteins	26
Fig 2.2	Generation of three-dimensional structure of the polypeptide chain in extended conformation from sequence information	27
Fig 2.3	Generation of representative structure with secondary structural elements	30
Fig 3.1	Dihedral sampling method, depicting tree representation of the conformational search problem	40
Fig 4.1(a)	Persistence length versus number of amino acids for ~1000 globular proteins	48
Fig 4.1(b)	Persistence length analysis on the dataset of ~1000 globular proteins	49
Fig 4.2(a)	Radius of gyration analysis on the dataset of ~1000 globular proteins	50
Fig 4.2(b)	Radius of gyration versus number of amino acids for ~1000 globular proteins	51
Fig 4.3	Hydrophobicity ratio analysis on the dataset of ~1000 globular proteins	53

Fig 4.4(a)	Packing fraction analysis on the dataset of ~1000 globular proteins	55
Fig 4.4(b)	Packing fraction versus number of amino acids for ~1000 globular proteins	55
Fig 5.1	Flowchart of the protocol followed for determining the energy rank of decoys vis-à-vis the native structure	68
Fig 5.2	Relative energy versus RMSD (Å) plots for decoy sets: (a) EMBL, (b) CASP1, (c) four-state reduced, (d) Lattice_ssfit, (e) Lmds, (f) Fisa, (g) Fisa_CASP3, (h) Hg_structal, (i) Semfold, (j) Rosetta, (k) CASP5, (l) Homology model built set	74-77
Fig 6.1	The lowest RMSD structure emerging from the proposed pathway superimposed on the corresponding native structure for proteins: (a) 1VII; (b) 1DV0; (c) 1GVD; (d) 1MBH; (e) 1GAB; (f) 1IDY; (g) 1PRV; (h) 1HDD; (i) 1BDC; (j) 1HP8; (k) 1BW6; (l) 2EZH	89-90
Fig 7.1	A pathway for bracketing native-like structures of mixed alpha-beta globular proteins	96
Fig 7.2	The lowest RMSD structure emerging from the proposed pathway superimposed on the corresponding native structure for proteins: (a) 1E0Q; (b) 1B03; (c) 1FME; (d) 1PMC; (e) 1BHI; (f) 1I6C	103

## LIST OF TABLES

Table 1.1	Some diseases arising from folding defects	7
Table 1.2	Statistics from PDB and Swiss-Prot databases	10
Table 1.3	A list of comparative modeling softwares for protein structure prediction	12
Table 2.1(a)	Average values for Ramachandran angles ( $\phi$ , $\psi$ ) for the helix, sheet and loop regions obtained from a database analysis of ~1000 globular proteins	31
Table 2.1(b)	The most probable values for Ramachandran angles for the loop regions obtained from a database analysis of ~1000 globular proteins	32
Table 2.2(a)	The most probable values for the side chain dihedrals in the helix region obtained from a database analysis of ~1000 globular proteins	33
Table 2.2(b)	The most probable values for the side chain dihedrals in the loop regions obtained from a database analysis of ~1000 globular proteins	34
Table 3.1	A comparison of trial structure generation methodologies and their performance on some test proteins	42
Table 3.2	Trial structure generation for proteins with short as well as long loops	43
Table 4.1	Number of structures selected with persistence length and radius of gyration filters and the lowest RMSD in each case	58
Table 5.1	The decoy sets studied and the total number of proteins in each decoy set prepared by different research groups	63
Table 5.2	Number of protein sequences and decoy structures investigated	

	in the present study	67
Table 5.3	Number of structures generated with different homology modeling softwares for each protein sequence	69
Table 6.1	Results emerging from the proposed computational pathway for bracketing native-like structures for small alpha helical globular proteins	87
Table 6.2	Characterization of the predicted native-like structures from the pathway	88
Table 6.3	A performance appraisal of different modeling softwares for protein structure prediction	92
Table 6.4	CPU time required for each step of the pathway for two representative proteins	93
Table 7.1	Average values for Ramachandran angles ( $\Phi$ , $\Psi$ ) for the sheet region obtained from the PDB database analysis	98
Table 7.2	The most probable values for the main chain dihedrals in the sheet region considering parallel and antiparallel strands separately, obtained from the PDB database analysis	99
Table 7.3	Results from the computational pathway for bracketing native-like decoys for six small proteins	102

## 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