Heuristic Based Search for Protein Structure Prediction

Mahmood Abdur Rashid
B.S.M.E., M.S.C.S.

Institute for Integrated and Intelligent Systems
Science, Environment, Engineering and Technology
Griffith University, Brisbane, Australia

Submitted in fulfilment of the requirements of the degree of
Doctor of Philosophy

December 19, 2013

Statement of Originality

This work has not previously been submitted for a degree or diploma to any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Mahmood A. Rashid

Brisbane, December 19, 2013

Dedicated to my parents and family for their unconditional love and support

Acknowledgements

This work has benefited greatly from the input and support of many people over the past four years. I would like to express my gratitude to everyone who contributed to it in one way or other.

First and foremost, I would like to thank my supervisors Professor Abdul Sattar, and Dr. Md Tamjidul Hoque. They have all been really helpful and supportive throughout my PhD candidature. Whenever I had a problem, I always got them beside me. I would also like to thank Dr. Bela Stantic, Dr. Guido Governatori and Dr. Duc Nghia Pham who have helped me in numerous occasions throughout my candidature.

A big thank must go to Dr. M.A. Hakim Newton. I appreciate for his valuable time, innovative ideas, and research experiences that he shared with great collaboration for last four years. Without his continuous guidance it was difficult for me to overcome the tough times in the PhD pursuit. His willingness to share his knowledge and readiness to help others are unequaled.

I am also thankful to other PhD Candidates: Mr. Swakkhar Shatabda, Mr. Iman Dehzangi, Ms. Thach-Thao Nguyen Duong, Mr. Kushal Ahmed, Ms. Zahra Shahabi Kargar, Mr. Yi Fan, Mr. Shaowei Cai, Mr. Roozbeh Derakhshan who shared the lab wonderfully to create a great working environment. I am also thankful to Mr. Lukas Folkman and Dr. Trent Higgs for their collaborations. However, I must want to give a special thank to Swakkhar for his ever helpful attitude.

Furthermore, I would like to acknowledge Griffith University and NICTA for their financial support during this study and for providing the financial supports to attend several conferences. I also acknowledge these two institutions for providing wonderful working environment with sufficient logistic supports. I am very much thankful to Ms. Natalie Dunstan (IIIS) and Ms. Sarah Simms (NICTA) for their extreme supports regarding logistics and official requirements during past four years.

I thank my friends in Brisbane for the fantastic time we have had, and my friends on the other side of the world for keeping in touch despite the distance.

Finally, I want to thank my parents and family. I am ever grateful to my wonderful boys Fardeen and Farhaan, and my forbearing wife Salma Naz for their unbounded sacrifice during my study.

Abstract

This thesis presents our heuristic based search approaches for protein structure prediction. Proteins are essentially sequences of amino acids. Every protein adopts a specific three-dimensional (3D) structure to perform a specific task. The function of a protein depends on its 3D native structure that has the minimum free energy level. Proteins are the fundamental component of all living cells. However, misfolded proteins cause fatal diseases. Therefore, protein structures are very important in biotechnology and drug design. Hence, protein structure prediction (PSP) has emerged as a very important multi-disciplinary research problem.

Given a protein's amino acid sequence, the protein structure prediction problem is to find a three dimensional structure of the protein such that the total free energy amongst the amino acids in the sequence is minimised. *In-vitro* laboratory methods used in PSP are very time-consuming, cost-intensive, and failure-prone. On the other hand, computational predictive methods are NP-hard even when conformation models are simplified by using low-resolution energy function and discretised lattice-based structures.

There are three computational approaches for protein structure prediction: *homology modeling*, *protein threading* and *ab initio* approach. The prediction quality of *homology modeling* depends on the sequential similarity with proteins that have previously known structures. On the other hand, protein threading also known as fold recognition, is based on the structural similarity with the previously known fold families. Our work is based on the *ab initio* approach that only depends on the amino acid sequence of the target protein. The computational methods found in the literature are largely based on optimisation algorithms. The widely used algorithms in PSP are genetic algorithms, memetic algorithms, ant colony optimisation, particle swarm optimisation, simulated annealing, stochastic local search, and Monte Carlo simulation.

This thesis focuses on genetic algorithms and stochastic local search algorithms because they are found promising in conformational search. A genetic algorithm (GA) is a population based optimisation algorithm inspired from nature. Genetic algorithms mimic genetic mutation, recombination and survival for the fittest principles. On the other hand, local search (LS) algorithms, starting from a single initial solution, move from one solution to another to find a better solution in the vicinity of the current solution. However, for large sequences, the performance of both algorithms deteriorate and the search often get stuck in plateaus and/or local minima. In general, the success of GA and LS methods crucially depends on the balance of diversification and intensification of the search. Any further progress to these algorithms require addressing the above issues appropriately.

In this thesis, we introduce a new population based algorithm (GA^+) under the GA framework for simplified PSP. We use hydrophobic-polar (HP) energy model on 3D FCC lattice to simplify the problem. In GA^+, we use *i*) an exhaustive generation approach to diversify the search; *ii*) a novel hydrophobic core-directed macro-mutation operator to intensify the search; and *iii*) a random-walk based stagnation recovery technique.

In HP model, protein structures have hydrophobic cores (H-core) that hide the hydrophobic amino acids from water and expose the polar amino acids to the surface to be in contact with the surrounding water molecules. Taking H-core formation as an objective, in this thesis, we present a new spiral search algorithm for PSP. The spiral search algorithm, denoted by SS-Tabu, works in a spiral fashion within a hydrophobic-core directed local search that is guided by tabu meta-heuristic.

In a separate work, we present a parallel processing technique to expedite exploration by starting spiral search from different points. In this approach, a number of random initial solutions are generated and distributed to different threads. We allow each thread to run for a pre-defined period of time. The improved solutions from the threads are merged together and a number of solutions are selected for next round processing.

As we find no single algorithm suits the best for the protein structure prediction problem, we have tried a hybrid search technique to mix the power of stochastic local search and population based genetic algorithms to gain improvements. In this approach, we embed our spiral search algorithm within the GA^+ framework. The spiral search improves the individuals within the GA^+ population and quickly builds a rich population.

A detailed 20×20 energy model could better capture the behaviour of the actual energy function than a very simple 2×2 HP energy model. In contrast, the HP energy model could effectively bias the search towards certain promising directions. We have used the HP model with a detailed 20×20 model in a mixed fashion in GA+ and parallel spiral search frameworks and found that the mixed setting works more effectively than using the energy models separately.

We evaluated our approaches on a set of benchmark proteins using FCC lattice, and the 2×2 HP and the detailed 20×20 energy models. We experimentally show that our approaches significantly outperform the current state-of-the-art approaches for the same lattice and energy models.

In conclusion, this thesis presents a new local search algorithm, a parallel local search framework, a hybrid search framework and several enhancements made to the GA framework. These algorithms and enhancements improve protein structure prediction in lattice based structures and low resolution energy models.

List of publications

The scientific contributions of this research have been fully or partly published the following peer-reviewed journals and proceedings:

Journal Publications

[J1] **M. A. Rashid**, M. A. H. Newton, M. T. Hoque, and A. Sattar; Mixing Energy Models in Genetic Algorithms for On-Lattice Protein Structure Prediction; *BioMed Research International*, 2013.

[J2] **M. A. Rashid**, S. Shatabda, M. A. H. Newton, D. N. Pham, M. T. Hoque, and A. Sattar; Spiral Search: A Hydrophobic-Core Directed Local Search for Simplified PSP on 3D FCC Lattice; *BMC Bioinformatics*, 2013.

[J3] S. Shatabda, M. A. H. Newton, **M. A. Rashid**, D. N. Pham, and A. Sattar; The Road Not Taken: Retreat and Diverge in Local Search for Simplified Protein Structure Prediction; *BMC Bioinformatics*, 2013.

[J4] **M. A. Rashid**, M. A. H. Newton, M. T. Hoque, and A. Sattar; A parallel framework for multi-point spiral search in *ab initio* protein structure prediction; *Advances in Bioinformatics*, 2014.

Conference Publications

[C1] **M. A. Rashid**, M. A. H. Newton, M. T. Hoque, and A. Sattar; A Local Search Embedded Genetic Algorithm for Simplified Protein Structure Prediction ; *IEEE Congress on Evolutionary Computation (CEC)*, 2013.

[C2] S. Shatabda, M. A. H. Newton, **M. A. Rashid**, and A. Sattar; An Efficient Encoding for Simplified Protein Structure Prediction Using Genetic Algorithms; *IEEE Congress on Evolutionary Computation (CEC)*, 2013.

[C3] **M. A. Rashid**, M. A. H. Newton, M. T. Hoque, and A. Sattar; Collaborative Parallel Local Search for Simplified Protein Structure Prediction; *IEEE International Symposium on Parallel and Distributed Processing with Application*, 2013.

[C4] **M. A. Rashid**, S. Shatabda, M. A. H. Newton, D. N. Pham, M. T. Hoque, and A. Sattar; Random-Walk: A Stagnation Recovery Technique for Simplified Protein Structure Prediction; *ACM BCB*, 2012.

[C5] **M. A. Rashid**, M. A. H. Newton, D. N. Pham, M. T. Hoque, and A. Sattar; A New Genetic Algorithm for Simplified Protein Structure Prediction; *Australasian Joint Conference on Artificial Intelligence (AI)*, 2012.

List of acronyms

2D	–	Two Dimensional
3D	–	Three Dimensional
AA	–	Amino Acid
ACO	–	Ant Colony Optimisation
AIS	–	Artificial Immune System
Avg	–	Average
BM	–	20×20 energy model by Berrera *et al.* [1]
BH	–	Mixed energy model combining BM and HP energy models
BnB	–	Branch and Bound
C_α	–	Alpha Carbon
C_β	–	Beta Carbon
CASP	–	Critical Assessment of Techniques for Protein Structure Prediction
CBLS	–	Constraint Based Local Search
CG	–	Chain Growth Algorithm
CHCC	–	Compact Hydrophobic Core Construction
CLP	–	Constraint Logic Programming
CP	–	Constraint Programming
CPSP	–	Constraint-based Protein Structure Prediction
DFS	–	Depth First Search
DNA	–	Deoxyribonucleic Acid
EGA	–	Exhaustive Genetic Algorithm
EMC	–	Evolutionary Monte Carlo
FCC	–	Face-Centred-Cubic (lattice)
FD	–	Finite Domain
FM	–	Free Modeling
GA	–	Genetic Algorithm
GA^+	–	A new genetic algorithm
H	–	Hydrophobic
HC	–	Hill Climbing Algorithm
HCC	–	H-Core centre
HCP	–	Hexagonal Close Pack (lattice)
HGA	–	Hybrid Genetic Algorithm

HP	–	Hydrophobic-Polar (Hydrophilic)
IA	–	Immune Algorithm
LP	–	Linear Programming
LS	–	Local Search
LSEGA	–	Local Search Embedded Genetic Algorithm
LS-Tabu	–	Tabu guided Local Search
MC	–	Monte Carlo Algorithm
MD	–	Molecular Dynamics
MGA	–	Macro-mutation based Genetic Algorithm
MJ	–	20×20 energy model by Miyazawa and Jernigan [2]
MH	–	Mixed energy model combining MJ and HP energy models
MH-Tabu	–	Mixed heuristic local search guided by tabu
NMR	–	Nuclear Magnetic Resonance
NP	–	Non Polynomial
P	–	Polar (Hydrophilic amino acids)
PDB	–	Protein Data Bank
PPC	–	Prediction of Protein Complex
PSO	–	Particle Swarm Optimisation
PSP	–	Protein Structure Prediction
PSS	–	Parallel Spiral Search
RGA	–	Randomised Genetic Algorithm
RMSD	–	Root-Mean-Square Deviation
RNA	–	Ribonucleic Acid
SA	–	Simulated Annealing
SAW	–	Self-Avoiding-Walk
SVM	–	Support Vector Machine
SS	–	Spiral Search
SS-Tabu	–	Tabu guided Spiral Search
TM-score	–	Tempate Modeling score
WGA	–	Random-Walk based Genetic Algorithm

Table of Contents

List of Figures

List of Tables