AN ALL ATOM ENERGY BASED COMPUTATIONAL PROTOCOL FOR PREDICTING BINDING AFFINITIES OF PROTEIN-LIGAND COMPLEXES

by

TARUN JAIN
Department of Chemistry

THESIS SUBMITTED IN FULFILLMENT
OF THE REQUIREMENTS OF
THE DEGREE OF

DOCTOR OF PHILOSOPHY

to the

Indian Institute of Technology, Delhi
Hauz Khas, New Delhi
India

March, 2007

Certificate

This is to certify that the thesis entitled, "An All Atom Energy Based Computational Protocol for Predicting Binding Affinities of Protein-Ligand Complexes", being submitted by Mr. Tarun Jain 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 him. Mr. Tarun Jain 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.

Dated: 19/7/07

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

Acknowledgements

A journey is easier when you travel together. This thesis is a result of four and a half years of work whereby I have been accompanied and supported by many people. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

I wish to express a great debt of gratitude and respect to my supervisor, Prof B. Jayaram, Head, Department of Chemistry, IIT Delhi, for giving me an opportunity to do research under his able guidance. He helped me to think independently, develop ideas and inculcate a scientific attitude towards solving a problem. His enthusiasm and integral view on research, complete dedication towards his job and mission for providing only high quality work has made a deep impression on me. I thank Prof Jayaram, for providing a research oriented environment equipped with state of the art facilities to complete this thesis work. I feel proud to be his student.

I am grateful to all my lab members of the Supercomputing Facility for Bioinformatics and Computational Biology and the Chemistry Department Lab, for their help and cooperation received in completing my research work and other associated activities.

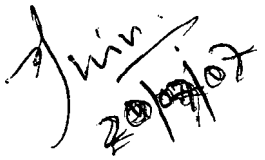
I thank all the faculty and staff of the Chemistry Department, IIT Delhi for their help and support received during this project.

I am extremely thankful to Prof R.S. Agarwal, Prof M.N. Gupta and Prof A. Ramannan for their guidance and encouragement during my Ph.D. days.

I express sincere thanks to the Department of Biotechnology, Govt. of India, Department of Science & Technology, Govt. of India and Dabur Research Foundation for their financial assistance.

I greatly appreciate the love, affection, patience and encouragement received from my best friend and wife, Kumkum. This journey would have been very difficult without her.

Finally, I dedicate this thesis to my parents, for whom words are not just enough to describe.



Abstract

Predicting the binding affinities of candidate molecules to proteins is an essential step in structure-based drug design for discovering new drug leads. After a validated target is chosen and its structure determined, new leads can be designed based on physico-chemical principles or chosen from a subset of small molecules that score well when docked *in silico* against the target. Available computational approaches are able to dock small molecules satisfactorily in the drug target but predicting accurate binding affinities still remains a major challenge for virtual screening in drug design. A multitude of methods at various levels of rigor and speed are available today for estimating binding affinities. Simple and fast methods employing severe approximations although useful may neglect important components to the binding free energy. More sophisticated methods are time consuming because they rely on sampling of the entire conformational and configurational space of the complex with the solvent, which in turn requires accurate force fields and simulation protocols. A good agreement / correlation with the experiment on a diverse set of systems can only be achieved if the methodology integrates all the vital components involved in the thermodynamics of binding.

The aim of this thesis work has been to develop atomic level computational protocols for predicting accurately affinities of ligands binding to non-metallo and zinc containing metalloproteins. This is achieved through development of an all atom energy

based empirical scoring function. The empirical free energy function developed comprises contributions from electrostatics with a sigmoidal dielectric function, van der Waals, hydrophobic and loss in conformational entropy of protein side chains. The protocols have been validated on a diverse set of 161 non-metallo protein-ligand complexes and 90 zinc containing metalloprotein-ligand complexes comprising 60 unique protein targets like; HIV-1 protease, alpha thrombin, carbonic anhydrase, matrix metalloproteinase, alcohol dehydrogenase etc. A high correlation of (R^2 = 0.85) r = 0.92 and (R^2 = 0.77) r = 0.88 for the predicted against the experimental binding affinities for non-metallo and zinc containing metalloprotein-ligand complexes respectively shows the robustness of the methodologies. Model validation, various statistical tests (R^2, q^2, S_{press}) and parameter analysis studies have been performed to test the predictive ability of the scoring function. All the results obtained are within acceptable limits, suggesting the utility of the methodology in structure-based drug design to develop and predict affinities of ligands binding to proteins. Heterogenity of the dataset on which the protocols have been validated and parameters obtained promises transferability to protein-ligand systems from different families of proteins, with different active sites and a variety of ligand architectures. The scoring functions have been web enabled as BAPPL and BAPPL-Z servers for free access to the scientific community to aid in the design of novel new molecular entities/leads for various therapeutic targets.

The thesis is divided into six chapters. Chapter 1 gives an introduction to structure-based drug design and the many roles of computation in drug discovery. A brief overview of virtual screening (docking and scoring), *de novo* design and

computational ADME/T prediction is presented. The chapter then deals with the various computational approaches for predicting binding affinities of protein-ligand complexes. Chapter 2 explains the theory of the binding process from a thermodynamic perspective. The empirical energy based scoring function developed in this study is then described. A general methodology for preparing a protein-ligand complex in a force field compatible manner is presented. In Chapter 3, the results of the scoring function in predicting the binding affinities of 161 non-metallo protein-ligand complexes are presented and discussed. This chapter describes in detail, the protein-ligand complex dataset, computational protocol, various model validation and parameter analysis studies used in this work. Chapter 4 describes the results of the computational methodology and the scoring function developed for predicting the binding affinities of 90 zinc containing metalloprotein-ligand complexes. The different computational approaches adopted to examine the sensitivity of results to the choice of force field parameters and dielectric treatments applied for system preparation and scoring are discussed. The web enabling of the scoring function for predicting the binding affinities of non-metallo and zinc containing metalloprotein-ligand complexes is presented in Chapter 5. Finally in Chapter 6, a summary and some perspectives emerging from this thesis work on drug design *in silico* are provided.

Contents