DEVELOPMENT OF AN *AB INITIO* PHYSICO-CHEMICAL MODEL FOI ANALYZING PROKARYOTIC GENOMES

by

POONAM SINGHAL

Department of Chemistry

THESIS SUBMITTED IN FULFILMENT

OF THE REQUIREMENTS FOR

THE DEGREE OF

DOCTOR OF PHILOSOPHY

to the

INDIAN INSTITUTE OF TECHNOLOGY, DELHI

HAUZ KHAS, NEW DELHI, INDIA

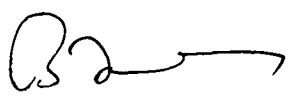
May, 2008

Dedicated to my parents

Certificate

This is to certify that the thesis entitled "Development of an *ab initio* Physico-Chemical Model for Analyzing Prokaryotic Genomes" being submitted by Ms. Poonam Singhal 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. Poonam Singhal 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.



Prof. B. JAYARAM
Head, Department of Chemistry,
Indian Institute of Technology, Delhi

Acknowledgements

Guru Bhramaha, Guru Vishnu, Guru Devo Maheshwara,

Guru Sakshat Param Bhramaha, Tasmai Siri Gurve Namah.

The old saying is as true as it was in the ancient time. The credit of this project goes to my ***Guru****, who made me to learn new things and taught me to take up new challenges in studies as well as in life. I am deeply obliged to my research supervisor and Head, Prof. B. Jayaram, Department of Chemistry, Indian Institute of Technology New Delhi, for giving me an opportunity to pursue my research interests in his group, his valuable suggestions, constant encouragement, keen interest and affectionate attitude during the course of this work. He has been a pillar of support to me during my research period and has always been available to give advice and to guide me. He is a great human being, who cares for his students.*

I would like to acknowledge the financial support from Department of Biotechnology, Government of India and IIT for attending the international conference at Albany, USA. I am highly grateful to Prof. D. L. Beveridge and Dr. Surjit B. Dixit for their valuable scientific help and making me feel like home during my stay in USA.

I am grateful to CSIR, New Delhi for providing me with the fellowship during the tenure of my research.

Thanks are due to all the teachers who taught me at any point of time and have laid the foundation and helped me in growing both at academic as well as personal level.

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 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 owe a lot to my friends. I feel privileged to have friends like Kumkum, Arti, Teena, and Garima.

I wish to express my deepest sense of gratitude and reverence to my loving and respectable parents for their unconditional love, encouragement and patience to pursue my interests. Without their initial momentum and continuous support, it would have been impossible for me to arrive where I am today. I thank them for providing me the best of education and making me realize the virtues of life. I am also indebted to my sisters Anchal, Jyoti and brother Sachin for their love, invaluable support and care that they have always shown for me. My mausi, mausaji and cousins Amit, Ankit and Nitin have also contributed much to this work in visible and invisible ways. This journey would have been very difficult or rather impossible without them.

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.

Abstract

The genome sequence represents the book of life. Buried in this large volume are genes, which are scattered as small DNA fragments throughout the genome and comprise a small percentage of the total text. Finding these indistinct 'needles' in a vast genomic 'haystack' can be extremely challenging.

The computational gene identification problem is an endeavor to interpret nucleotide sequences, in order to provide an annotation on the location and hopefully the functional class of protein-coding genes (mRNA) as well as genes for other RNAs (rRNA, tRNA etc.). This problem is of self-evident importance and remains far from full resolution. With the advent of whole-genome sequencing projects, there is considerable use for algorithms which scan genomic DNA sequences to find genes, particularly those that encode for proteins. Even though there is no substitute to experimentation in determining the exact locations of genes in a genome sequence, a prior knowledge of the approximate location of genes will hasten the process to a great extent apart from saving huge amount of laboratory time and resources. Several algorithms have been reported in the past for analysis of prokaryotic genomes based on sophisticated, statistical and mathematical methods. A physico-chemical understanding of the characteristics of the genes is lacking. The present thesis attempts to address this issue.

This thesis work presents an *ab initio* model for gene prediction in prokaryotic genomes based on physico-chemical characteristics of double helical trinucleotide sequences corresponding to the 64 codons calculated from molecular dynamics (MD) simulations. The methodology combines filters based on stereochemical properties of protein sequences and frequencies of occurrences of codons via their corresponding amino acids from 175000 Swissprot proteins to reduce false positives. The methodology has been validated on 372 prokaryotic genomes with sensitivity, specificity and correlation coefficients averaged over 356208 genes and an equal number of frame-shifted genes (non-genes) reaching 97.50%, 97.20% and 94.25% respectively. The protocol has been automated in the form of a web server (http://www.scfbio-iitd.res.in/chemgenome2) 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 gene prediction methodologies. An overview of the computational methodologies (extrinsic, intrinsic, hybrid and comparative genomics approaches) is also presented. Chapter 2 describes the methodology developed as a part of this thesis for the evaluation of DNA sequences and prediction of protein-coding genes from the whole genome sequence of the organisms. The modular form of the methodology has been explained which allows for the automation of the methodology as a web server. Chapter 3 describes the results obtained from the gene evaluation

methodology explained in Ch. 2 on 372 prokaryotic genomes. An evaluation of the methodology with the virus genomes is also presented. Chapter 4 describes the results obtained from the gene prediction methodology explained in Ch. 2 on 372 prokaryotic genomes. Chapter 5 describes the performance appraisal of the gene evaluation methodology for eukaryotic genomes. Chapter 6 explains the protocol devised for extracting the physico-chemical information hidden in DNA sequences. The methodology explained in Ch. 2 has been automated in the form of web servers which are described in Chapter 7. Finally in Chapter 8, summary and some perspectives emerging from this thesis work on genome analysis are provided.

Contents