DEVELOPMENT OF ENERGY BASED SIGNATURES FOR DECIPHERING PROKARYOTIC GENOME ORGANIZATION

by

GARIMA KAHNDELWAL

Department of Chemistry

Submitted

In fulfilment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

HAUZ KHAS, NEW DELHI, INDIA

NOVEMBER, 2012

Dedicated to my Parents

Certificate

This is to certify that the thesis entitled "Development of Energy Based Signatures for Prokaryotic Genome Organization" being submitted by Ms. Garima Khandelwal 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. Garima Khandelwal 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.

Dated:

Prof. B. JAYARAM
Department of Chemistry,
Indian Institute of Technology Delhi
New Delhi - 110016

Acknowledgements

The success of any project depends largely on the encouragement and support of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this phase

I owe my deepest gratitude to my supervisor, Prof. B. Jayaram, Department of Chemistry, IIT Delhi, for giving me an opportunity to do research under his able guidance. He helped me develop new ideas; think independently and develop a scientific perspective towards solving problems. His zeal towards his work, complete dedication towards research and mission for providing high quality science has left a deep impression on me. I thank Prof. Jayaram, for his invaluable support and for providing me with a research oriented environment equipped with state of the art facilities to complete this thesis work. I feel proud to be his student.

I'm extremely obliged to Prof. D. L. Beveridge, for providing me with a chance to work under his esteemed guidance. He opened new realms of knowledge and made science extremely easy to understand. His constant encouragement and appreciation has helped me fulfill the tasks assigned to me. I have a deep sense of respect towards him.

I thank all my teachers who have laid the foundation and helped me in growing both academically and personally during my life.

I would like to thank Department of Biotechnology, India for providing me with the fellowship during the tenure of my research. I would also like to acknowledge Department of Science and Technology, India for support towards my international travel.

I am grateful to all the past and present lab members of the Supercomputing Facility for Bioinformatics and Computational Biology, 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.

This acknowledgement would be incomplete without the special mention of few people, who have supported me through this journey through their own ways. I thank my friends Sana, Nisha, Dheerendra and Bibhas, for providing me with a home away from my home. They have been a pillar of emotional strength throughout this time. I also thank my seniors, KumKum Jain and Poonam Singhal, who gave immense support and direction during my nascent PhD years. I'm privileged to have Goutam, Srinivas and Cyril as colleagues, who

provided constant strength, making this journey an easier one. I'm privileged to have Priyanka, Tanya, Varsha and Ankita, who have been like sisters to me. The advices/ideas developed during the scientific discussions with Sahil, Avinash and Deepesh are greatly appreciated. I'm thankful to Bharat, Satya and Navneet for all the constant help they have provided me during my research.

I thank that special person in my life, Ashish, who has been a pillar of strength and comfort and has been more of a friend than a fiancée.

I'm indebted to my loving parents for their unconditional love, patience and encouragement towards my ordeal. Without their continuous support, it would have been impossible for me to arrive where I am today. I thank them for providing me with a sense of direction and the very best in life. I am also indebted to my brothers Vishnu and Anuj for their love, invaluable support and care that they have always shown for me. My naniji, mausiji, buaji, mamaji and cousins have also contributed much to this work in visible and invisible ways. This journey would have been very difficult without them.

Lastly, I thank Almighty God for the blessings that I have received throughout my life.

(Garima Khandelwal)

Abstract

DNA is the vital constituent of all organisms driving them through their course of sustenance. DNA molecule, consisting of just four bases and two strands, has been studied through various techniques and methods, to develop an understanding of its structural, dynamic and functional aspects. But, even after 60 years of the discovery of the double-helix, its language, viz., inference of function from an inspection of its sequence remains elusive.

The entire DNA content of an organism is termed as 'Genome'. The arrangement of various functional units (mRNA genes, tRNA genes, rRNA genes, promoters, exons, introns, splice sites etc.) on a genome is termed as genome organization. The problem of interpreting the nucleotide sequence data, to provide annotation on the position and functionality of these regions encompasses the area of genome analysis.

Experimental genome analysis techniques have not been able to keep pace with the exponential increase in the sequence data as these are extremely time and resource consuming. With the spate of genome sequence data, *in silico* genome annotation has become an integral part of the genome analysis and sequencing projects. A variety of genome annotation methods have been reported in the past, utilizing either similarity with the existing data or training of

statistical/mathematical models on the available data or a combination of both. Also, most of the genome analysis methods focus only on the protein-coding regions (mRNA genes). The probability of extension of these methods to other functional regions utilizing the same parameters is also very less as they lack the physico-chemical characterization of DNA.

The present thesis attempts to develop physico-chemical approaches to solve the problem of genome annotation and presents the necessary parameters, models and methods to discriminate various functional regions of DNA. The approach draws strength from energetic contributions of DNA, such as hydrogen-bonding, stacking and solvation energies as obtained from molecular simulations. The methods have been thoroughly validated on large datasets. This thesis is divided into eight chapters.

Chapter I gives an introduction to genome organization and an overview of different computational methodologies for genome analysis. It provides information about the current scenario of genomic data, and the tools to analyze them. This chapter also gives a general idea about the physico-chemical approach utilized in this thesis.

Chapter II describes the development of hydrogen-bonding, stacking and solvation energy parameters for dinucleotides. Methods for determining the thermodynamic stability of DNA in terms of their melting temperatures have also

been developed. The reliability of the parameters and the methods has been verified by their ability in predicting the melting temperatures of oligonucleotides. The correlation coefficients obtained between predicted and experimental melting temperatures in all the cases are $\geq$ 0.98 and the average error in prediction is lower than any of the existing melting temperature prediction methods for oligonucleotides.

Chapter III deals with the problem of computational melting temperature prediction of long DNA sequences extending to the level of genomes. This chapter explains the procedure to calculate thermodynamic stability of long DNA sequences of any length in terms of their melting temperatures along with the variation of stability (melting profiles) along the sequence. The correlation coefficient obtained between the predicted and experimental melting temperatures for genomic sequences is ~0.99. It also illustrates that the thermodynamic stability of experimentally verified promoter region is generally lower as compared to their genic counterparts as observed in the case of 496 *Escherichia coli* promoter-gene sequences.

Chapter IV discusses a physico-chemical approach to identify protein-coding regions in prokaryotic genomes. It describes the development of a protocol for predicting new genes utilizing the melting profiles as obtained by the methodology laid out in Chapter III. Prediction results on 16 prokaryotic systems of varied GC-

content are also presented. The reliability of prediction has been checked by an application to a synthetic genome (Synthia) and the error is found to be just about 2%. It also highlights the fact that there is a serious need for new approaches in genome analysis methods as substantial number of new genes/annotations have been predicted even in highly annotated systems such as *Escherichia coli* and *Bacillus subtilis.*

Chapter V discusses the role of physico-chemical properties of DNA sequences as a guide to developing insights into both coding and non-coding RNA genes. It describes the development of energy-based fingerprints for various RNA genes (mRNA, tRNA and rRNA). The separation of t-RNA and m-RNA genes utilizing solvation energies is very clear, with 99% accuracy in over 1500 prokaryotic systems. Separation of mRNA, tRNA and rRNA genes is also evident at the genomic level clearly indicating presence of physico-chemical signatures of different functional units.

Chapter VI provides a solution for the development of sequence based statistical-thermodynamic model of DNA, tractable for any sequence length. The model is further harnessed for determining the thermodynamic stability variation in terms of nucleotide stability of the DNA. The model is in excellent agreement with the experiment and gives a correlation coefficient of -0.97 between the computed average free energy per base pair and the experimental melting temperatures of

oligonucleotides. The nucleotide stability profiles for promoter and gene sequences of *Escherichia coli* follow the same trends as observed with the melting profiles.

Chapter VII addresses the issues encountered with eukaryotic genome analysis. As the methods developed in the previous chapters are based on the physico-chemical properties of DNA, these should be extendable to eukaryotic systems as well. This chapter inspects the capability of the thermodynamic stability methodology (both melting and statistical thermodynamic) in terms of identifying the exons and introns in eukaryotic gene sequences with encouraging results.

Finally, Chapter VIII provides the summary and perspectives emerging from this thesis work and some ideas for future explorations in this regard. The strong message that comes out of this thesis work is that the diverse functional units on genomic DNA carry unique physico-chemical signatures.

Contents