

# **ADVANCED INFORMATION THEORY FOR THE ANALYSIS OF BIOLOGICAL DATA**

**NITHYA RAMAKRISHNAN**



**DEPARTMENT OF ELECTRICAL ENGINEERING  
INDIAN INSTITUTE OF TECHNOLOGY DELHI**

**DECEMBER 2017**

© Indian Institute of Technology Delhi (IITD), New Delhi, 2017

# **ADVANCED INFORMATION THEORY FOR THE ANALYSIS OF BIOLOGICAL DATA**

by

**NITHYA RAMAKRISHNAN**

**DEPARTMENT OF ELECTRICAL ENGINEERING**

Submitted

*in fulfillment of the requirements of the degree of Doctor of Philosophy*

to the



**INDIAN INSTITUTE OF TECHNOLOGY DELHI**

**DECEMBER 2017**

# Certificate

This is to certify that the thesis titled “**Advanced Information Theory for the Analysis of Biological Data**” being submitted by **Nithya Ramakrishnan** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, for the award of the degree of **Doctor of Philosophy** is the record of the bona-fide research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted either in part or in full to any other university or institute for the award of any degree or diploma.

Prof. Ranjan Bose

Department of Electrical Engineering

Indian Institute of Technology Delhi

# Acknowledgements

My journey of research at IIT Delhi has been a voyage of self-discovery, facilitated by the assistance and goodwill of several eminent and gracious people. At the very beginning, I would like to extend my heartfelt gratitude towards my research supervisor Prof. Ranjan Bose for the indefatigable support, patience and trust he had in me. Without his encouragement and scientific illumination in the right directions, this research work would not have been possible.

I would like to express my sincere thanks to my student research committee members (*viz.*) Prof. D. Sundar, Prof. Brejesh Lall and Prof. Manav Bhatnagar for many useful suggestions in this research. Their constant feedback helped this work scale new heights. I am also grateful to all the professors at IIT Delhi who have taught me the basics and advanced topics of the relevant study during the course-work. The enlightenment gained from their classes helped me lay down the foundation of this research work. In particular, I would like to thank Dr. Vivekanandan Perumal from the Kusuma School of Biological Sciences, IIT Delhi, who taught me the fundamentals of epigenetics, that forms a major part of this research. I would also like to thank all members and staff of the Electrical Engineering Department and Bharti School of Telecommunication

Technology and Management for their support during my presentations and other research-related activities.

My parents, parents-in-law and family have provided me unrelenting support throughout my PhD. I am grateful for their support in this work and in my life. My husband has been my constant companion in this wonderful expedition from its very inception. I cannot thank him enough for his corroboration in this work. My daughter Bhargavi inspires me every day with her smile as much as with her curiosity. She is both the beginning and means of this research.

I would also like to thank my friends, students and fellow researchers from whom I have learned a lot. They have all stimulated me to do my best always. I truly believe that this research work is the outcome of unflinching enthusiasm, perseverance, support, and the good wishes of people all around the world.

Nithya Ramakrishnan

# Abstract

Research in the areas of Computational Biology and Bioinformatics has gained a lot of momentum in recent years. The expansion in the mathematical and engineering techniques related to genetic data combined with the significant research progression in the related fields of biology have led to a surge of advancements in the interdisciplinary fields. Of late, we have seen the emergence of epigenetic data as equally or more important than the genetic data, due to its impact on an individual's metabolism, the associated diseases and personalized therapy. Our research work elucidated in this thesis encompasses both genetic data and epigenetic data analyses using information theoretic tools.

In the first part of our research work, we employ an important physio-chemical property of molecules (*viz.*) dipole moment, in the computation of the three-dimensional entropy of a given DNA sequence. The resultant measure which is a form of structural entropy, is then analyzed for differences in sequences belonging to coding regions and non-coding regions of the DNA. This technique of segmentation has been tried on several annotated and unannotated genes and the results are presented.

In addition to entropy, we explore additional measures of information theory such as superinformation to analyze the dipole-moment orientations. Superinformation can be understood as the “entropy of entropies” and has been proved to be a more suitable information theoretic measure than entropy. We explore dipole moment superinformation in both the double stranded and single stranded forms of DNA and employ it in the intron-exon segmentation problem.

We then stride into the area of epigenetics. Focusing on DNA methylation analysis, we study the impact of aberrant methylations in cancer. We formulate “methylation entropy” based on the intensities of DNA methylation data in healthy and tumor samples. Applying various mathematical transforms to the methylation probabilities, we enhance regions of methylation to obtain modified methylation entropies. We then plot the PDFs (Probability Density Functions) of the methylation entropies and modified methylation entropies to observe if the healthy and tumor samples differ in their methylation distributions. A classifier trained based on the methylation entropies and modified methylation entropies is found to yield superior values of specificity and sensitivity. This analysis is based not just on the intensity of methylation, but also on the distribution of methylation in CpG islands. The focus in our work is KIRC (Kidney Renal Clear Cell Carcinoma), which is a near fatal cancer of the urogenital tract.

To delve deeper into the epigenetic data analysis, we use other different measures of information theory in the likes of Kullback-Leibler distance and Jensen-Shannon distance. These distance measures are based on an “average healthy methylation distribution” which is computed using several healthy samples’ methylation

distributions. The distance of a specific sample's methylation distribution from the average healthy methylation distribution is computed, both symmetrically (using Jensen-Shannon distance) and asymmetrically (using Kullback-Leibler distance). The distributions of these distance measures are analyzed for various bin-sizes of the histograms. They are then used in the segregation of healthy and tumor samples using a trained classifier. These results are studied for specific sets of significant genes like tumor-suppressor-genes or oncogenes or the global set of genes in the epigenome.

These techniques and investigations can be enhanced to study various other forms of cancer or other epigenetic diseases. Towards the end of the thesis, we provide suggestions for further research in the domains of introns and exons as well as computational epigenetics.

## सार

कम्प्यूटेशनल बयोलॉजी और बयो-इनफॉर्मैटिक्स के क्षेत्रों में अनुसंधान ने हाल के वर्षों में बहुत तरक्की की है। जीव विज्ञान के संबंधित क्षेत्रों में महत्वपूर्ण शोध प्रगति के साथ आनुवांशिक डेटा से संबंधित गणितीय और इंजीनियरिंग (अभियांत्रिक) तकनीकों के विस्तार ने अंतःविषय क्षेत्रों में प्रगति को जन्म दिया है। हाल ही में, आनुवंशिक डेटा की तुलना में एपिजेनेटिक डेटा के उद्भव को समान या अधिक महत्वपूर्ण माना गया है, और उसकी वजह है इन्सान का चयापचय, संबंधित बीमारियां और व्यक्तिगत चिकित्सा पर इसके प्रभाव के कारण। इस थीसिस में वर्णित हमारे शोध के काम में आनुवांशिक डेटा और एपिजेनेटिक डेटा दोनों को शामिल किया गया है जो सूचना सैद्धांतिक उपकरणों का उपयोग कर रहा है।

हमारे अनुसंधान कार्य के पहले भाग में, हम एक डीएनए अनुक्रम के त्रि-आयामी एंट्रोपी की गणना में, अणुओं की एक महत्वपूर्ण भौतिक-रासायनिक गुण (जैसे) द्विध्रुविय आघूर्ण का उपयोग करते हैं। परिणामस्वरूप माप जो संरचनात्मक एन्ट्रोपी का एक रूप है, उसके बाद डीएनए के कोडिंग क्षेत्रों और गैर-कोडिंग क्षेत्रों से संबंधित दृश्यों

में अंतर के लिए विश्लेषण किया गया है। विभाजन की इस तकनीक को कई व्याख्यायित जीन और अस्पष्ट जीन पर लगाया गया है और परिणाम प्रस्तुत किए गए हैं।

एंद्रोपी के अलावा, हम सूचना सिद्धांत के अतिरिक्त उपायों का पता लगाया है, जैसे कि विशिष्ट सूचना जिस से हमें द्विध्रुविय आघूर्ण उन्मुखीकरण का विश्लेषण कर सकते हैं। विशिष्ट सूचना को "एंद्रॉपियों का एंड्रोपी" के रूप में समझा जा सकता है और एंड्रोपी की तुलना में अधिक उपयुक्त जानकारी सैद्धांतिक माप साबित हुई है। डीएनए के दोहरे और एकल गुंथे हुए रूपों में हम द्विध्रुवीय आघूर्ण की विशिष्ट सूचना की खोज करते हैं और इंद्रॉन-एक्सॉन विभाजन समस्या में इस का प्रयोग करते हैं।

हम फिर एपिजेनेटिक्स के क्षेत्र में आगे बढ़ते हैं। डीएनए मेथिलेशन विश्लेषण पर ध्यान केंद्रित करते हुए, हम कैंसर में असामान्य मेथिलेशन के प्रभाव का अध्ययन करते हैं। हम स्वस्थ और ट्यूमर के नमूनों में डीएनए मेथिलेशन डेटा की तीव्रता के आधार पर "मेथिलेशन एंड्रोपी" तैयार करते हैं। मेथिलेशन संभावनाओं के लिए विभिन्न गणितीय परिवर्तनों को लागू करने के लिए, हम मेथिलेशन के क्षेत्रों को संशोधित मेथिलेशन एंड्रोपी प्राप्त करने के लिए बढ़ाते हैं। हम मेथिलेशन एंड्रोपी के पी.डी.एफ (प्रोबेबिलिटी डेंसिटी फ़ंक्शन) की रूप-रेखा करते हैं और मेथिलेशन एंड्रोपी को संशोधित करने के लिए निरीक्षण करते हैं, जानने के लिए कि क्या स्वस्थ और अर्बुद के नमूने उनके मेथिलेशन वितरण में विभिन्न होते हैं। मेथिलेशन एंड्रोपी और संशोधित मेथिलेशन एंड्रोपी के आधार पर प्रशिक्षित वर्गीकारक ठोसता और संवेदनशीलता की श्रेष्ठतर मूल्यों को प्रदर्शित करता है। यह विश्लेषण केवल मेथिलेशन की तीव्रता पर

आधारित नहीं है, सी.पी.जी द्वीपों में मेथिलेशन वितरण पर भी आधारित है। हमारा काम के.आई.आर.सी (किडनी रीनल क्लियर सेल कार्सिनोमा) पर केंद्रित है, जो मूत्रजनित पथ का घातक कैंसर है।

एपीजेनेटिक डेटा विश्लेषण को गहराई से जानने के लिए, हम क्लेबैक-लीबेल दूरी और जेन्सेन-शैनन दूरी जैसे सूचना सिद्धांत के अन्य विभिन्न उपायों का उपयोग करते हैं। ये दूरीमाप एक "औसत स्वस्थ मेथिलेशन वितरण" पर आधारित हैं जिसकी गणना स्वस्थ नमूनों के मेथिलेशन वितरण का उपयोग करके की जाती है। औसत स्वस्थ मेथिलेशन वितरण से विशिष्ट नमूने के मेथिलेशन वितरण की दूरी को संतुलित रूप से (जेन्सेन-शैनन दूरी का उपयोग करके) और असममित रूप से (कुल्बैक-लीब्लेर दूरी का उपयोग करके) गणना की जाती है। हिस्टोग्राम के विभिन्न बिन-आकार के लिए इन दूरी उपायों के वितरण का विश्लेषण किया जाता है। तब उन्हें एक प्रशिक्षित वर्गीकारक का उपयोग करके स्वस्थ और ट्यूमर के नमूने के अलगाव में उपयोग किया जाता है। इन परिणामों का उल्लेख महत्वपूर्ण जीन के विशिष्ट संग्रहों जैसे कि एपिजेनोम में ट्यूमर-सप्रेस-जीन या ओंकोजिन या जीन के सार्वत्रिक समूह के लिए किया जाता है।

इन तकनीकों और जांचों को कैंसर के अन्य अन्य रूपों या अन्य एपिजेनेटिक बीमारियों के अध्ययन के लिए बढ़ाया जा सकता है। थीसिस के अंत में, हम इंटरॉन-एक्सॉन के ज्ञानक्षेत्र के साथ-साथ कम्प्यूटेशनल एपिजेनेटिक्स में आगे के शोध के लिए सुझाव प्रदान करते हैं।

# Table of Contents

|                                                  |              |
|--------------------------------------------------|--------------|
| <b>Certificate</b>                               | <b>i</b>     |
| <b>Acknowledgements</b>                          | <b>ii</b>    |
| <b>Abstract</b>                                  | <b>iv</b>    |
| <b>List of Figures</b>                           | <b>xiii</b>  |
| <b>List of Tables</b>                            | <b>xvi</b>   |
| <b>List of Abbreviations</b>                     | <b>xviii</b> |
| <b>1 Introduction to Biological Data</b>         | <b>1</b>     |
| 1.1 Introduction .....                           | 1            |
| 1.2 Overview of Biological Components.....       | 2            |
| 1.2.1 Molecular Structure of DNA .....           | 2            |
| 1.2.2 Double-helical Structure of the DNA .....  | 3            |
| 1.2.3 Introns and Exons .....                    | 4            |
| 1.3 Epigenetics .....                            | 5            |
| 1.4 DNA Methylation .....                        | 8            |
| 1.5 Abnormal DNA Methylation and Cancer .....    | 9            |
| 1.6 Information Theory for Biological Data ..... | 10           |
| 1.7 Research Objectives .....                    | 12           |
| 1.8 Thesis Layout and Contributions .....        | 13           |

|          |                                                                                        |           |
|----------|----------------------------------------------------------------------------------------|-----------|
| 1.9      | Summary .....                                                                          | 15        |
| <b>2</b> | <b>Literature Survey</b>                                                               | <b>16</b> |
| 2.1      | Introduction.....                                                                      | 16        |
| 2.2      | A Classical Problem.....                                                               | 17        |
| 2.3      | Bio-physical Methods for Intron-Exon Segmentation .....                                | 18        |
| 2.4      | Previous DNA Structural Analysis Techniques.....                                       | 19        |
| 2.5      | Our Research Work in Intron-Exon Segmentation .....                                    | 20        |
| 2.6      | Cancer Epigenetics and DNA Methylation.....                                            | 21        |
| 2.7      | Entropy and Computational Techniques in Cancer .....                                   | 22        |
| 2.8      | Computational Approaches for DNA Methylation Analysis using Entropy ...                | 23        |
| 2.9      | Analyses in Specific Types of Tumor (KIRC – Kidney Renal Clear Cell<br>Carcinoma)..... | 24        |
| 2.10     | Distance Measures in Information Theory.....                                           | 25        |
| 2.11     | Our research work in DNA Methylation and KIRC .....                                    | 26        |
| 2.12     | Biological Databases .....                                                             | 27        |
| 2.13     | Summary .....                                                                          | 28        |
| <b>3</b> | <b>Dipole Moment Based Structural Entropy for DNA Analysis</b>                         | <b>29</b> |
| 3.1      | Introduction.....                                                                      | 29        |
| 3.2      | Base Angles in DNA.....                                                                | 30        |

|       |                                                                         |           |
|-------|-------------------------------------------------------------------------|-----------|
| 3.3   | Dipole Angular Entropy.....                                             | 33        |
| 3.3.1 | Definitions .....                                                       | 33        |
| 3.3.2 | Sample Computation of Dipole Moment Angles .....                        | 34        |
| 3.3.3 | Computation of Dipole Angular Probabilities .....                       | 36        |
| 3.3.4 | Steps for computation of dipole angular entropy.....                    | 37        |
| 3.4   | Dipole Angular Spectral Entropy.....                                    | 38        |
| 3.4.1 | Formulation of Dipole Angular Spectral Entropy .....                    | 38        |
| 3.4.2 | Steps to calculate Dipole Angular Spectral Entropy.....                 | 40        |
| 3.5   | Experiments and Results .....                                           | 41        |
| 3.5.1 | Data and Experimental Procedure Details .....                           | 41        |
| 3.5.2 | Results Analysis (Twist Angle $36^\circ$ ).....                         | 42        |
| 3.5.3 | Results Analysis (Twist Angle – ( $20^\circ$ to $50^\circ$ )).....      | 43        |
| 3.5.4 | Probability Density Function (PDF) Curves.....                          | 45        |
| 3.6   | Accuracy and Efficiency .....                                           | 46        |
| 3.6.1 | Accuracy of Separation - Definition.....                                | 46        |
| 3.6.2 | Accuracy Results .....                                                  | 48        |
| 3.6.3 | Computational Efficiency .....                                          | 49        |
| 3.7   | Experiments on Random Annotated and Unannotated Gene Sequences.....     | 50        |
| 3.7.1 | Annotated Gene Sequences .....                                          | 51        |
| 3.7.2 | Experiments on Unannotated Genes.....                                   | 51        |
| 3.7.3 | Threshold Selection .....                                               | 52        |
| 3.8   | Summary .....                                                           | 54        |
| 4     | <b>Dipole Moment Based Structural Superinformation for DNA Analysis</b> | <b>55</b> |

|          |                                                                           |           |
|----------|---------------------------------------------------------------------------|-----------|
| 4.1      | Introduction.....                                                         | 55        |
| 4.2      | Superinformation .....                                                    | 56        |
| 4.3      | Dipole Moment Vectors in Double Stranded DNA .....                        | 57        |
| 4.4      | Dipole Base Angular Superinformation.....                                 | 58        |
| 4.4.1    | Calculation of Dipole Vector Angles for Base-Pairs in DNA .....           | 58        |
| 4.4.2    | Computation of Probabilities and Superinformation .....                   | 58        |
| 4.4.3    | Steps to compute Dipole Base Angular Superinformation .....               | 60        |
| 4.5      | Dipole Nucleotide Angular Superinformation in Single-Stranded DNA .....   | 61        |
| 4.5.1    | Computation of Dipole Nucleotide Angular Superinformation.....            | 61        |
| 4.5.2    | Steps for computation of Dipole Angular Nucleotide Superinformation ..... | 63        |
| 4.6      | Experiments and Results .....                                             | 63        |
| 4.7      | Results – Dipole Base Angular Superinformation .....                      | 64        |
| 4.8      | Results - Dipole Nucleotide Angular Superinformation.....                 | 65        |
| 4.9      | Results - Accuracy and Efficiency.....                                    | 65        |
| 4.10     | Sensitivity Results.....                                                  | 68        |
| 4.11     | Results of Experiments on Annotated Genes.....                            | 70        |
| 4.12     | Summary .....                                                             | 71        |
| <b>5</b> | <b>Analysis of DNA Methylation using Entropy in Cancer</b>                | <b>73</b> |
| 5.1      | Introduction.....                                                         | 73        |
| 5.2      | Epigenetics and DNA Methylation in Cancer.....                            | 74        |

|          |                                                                                  |           |
|----------|----------------------------------------------------------------------------------|-----------|
| 5.2.1    | Biological Background .....                                                      | 74        |
| 5.2.2    | Tumor Suppressor Genes and Oncogenes .....                                       | 76        |
| 5.2.3    | Kidney Renal Clear Cell Carcinoma (KIRC) .....                                   | 76        |
| 5.3      | Methylation Entropy .....                                                        | 77        |
| 5.3.1    | Methylation Alphabet .....                                                       | 78        |
| 5.3.2    | Computation of Methylation Probabilities .....                                   | 78        |
| 5.3.3    | Methylation Entropy Computation .....                                            | 80        |
| 5.4      | Modified Methylation Entropy .....                                               | 81        |
| 5.4.1    | Mathematical Transformations.....                                                | 81        |
| 5.4.2    | Training and Segregation.....                                                    | 82        |
| 5.5      | Experiments and Results .....                                                    | 84        |
| 5.5.1    | Extraction and Processing of Data.....                                           | 84        |
| 5.5.2    | Selection of Specific Genes .....                                                | 85        |
| 5.5.3    | Results and Analysis in Tumor Suppressor Genes .....                             | 86        |
| 5.5.4    | Results and Analysis in Oncogenes .....                                          | 89        |
| 5.5.5    | Discussion.....                                                                  | 92        |
| 5.6      | Summary .....                                                                    | 94        |
| <b>6</b> | <b>Analysis of DNA Methylation using Information Theoretic Distance Measures</b> | <b>96</b> |
| 6.1      | Introduction.....                                                                | 96        |
| 6.2      | Information Theoretic Distance Measures .....                                    | 97        |
| 6.2.1    | Kullback-Leibler Distance.....                                                   | 97        |

|          |                                                     |            |
|----------|-----------------------------------------------------|------------|
| 6.2.2    | Jensen-Shannon Distance .....                       | 98         |
| 6.3      | Experimental Techniques.....                        | 99         |
| 6.3.1    | Average Healthy Methylation Distribution .....      | 99         |
| 6.3.2    | Methylation Distribution of a Sample .....          | 101        |
| 6.3.3    | Training and Classification.....                    | 102        |
| 6.4      | Results and Discussion.....                         | 104        |
| 6.4.1    | Tumor Suppressor Genes.....                         | 104        |
| 6.4.2    | Oncogenes.....                                      | 107        |
| 6.4.3    | Global Methylation Distributions .....              | 108        |
| 6.4.4    | Other Inferences.....                               | 112        |
| 6.5      | Summary .....                                       | 114        |
| <b>7</b> | <b>Conclusion and Future Work</b>                   | <b>116</b> |
| 7.1      | Conclusion .....                                    | 116        |
| 7.1.1    | Introns and Exons .....                             | 116        |
| 7.1.2    | Future Directions - Intron-Exon Research.....       | 117        |
| 7.1.3    | Our Research Work in DNA Methylation .....          | 118        |
| 7.1.4    | Further Research in DNA Methylation and Cancer..... | 120        |
| 7.1.5    | Biological Data in Cryptography .....               | 121        |
| 7.2      | Summary .....                                       | 121        |
|          | <b>Bibliography</b>                                 | <b>122</b> |
|          | <b>Publications</b>                                 | <b>130</b> |
|          | <b>Technical Biography of Author</b>                | <b>131</b> |

# List of Figures

|                                                                                                                                                                                                                                                                |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1: Figure illustrating the molecular structure of DNA in its double-stranded form. ....                                                                                                                                                                      | 6  |
| 1.2: Double Helical Structure of the DNA. ....                                                                                                                                                                                                                 | 7  |
| 1.3: Figure illustrating the splicing process. ....                                                                                                                                                                                                            | 7  |
| 1.4: Figure illustrating cytosine methylation in the DNA molecule. ....                                                                                                                                                                                        | 11 |
| 1.5: Figure illustrating DNA methylation in the CpG islands. ....                                                                                                                                                                                              | 12 |
| 3.1: Schematic representation of the angles and the displacement between adjacent DNA base pairs. ....                                                                                                                                                         | 32 |
| 3.2: Pictorial representation of the dipole moment angles in an imaginary DNA sequence. ....                                                                                                                                                                   | 35 |
| 3.3: Histogram plots of the dipole angular entropy ( $H_{ang}$ ) with twist angle taken as $36^\circ$ for all the (a) exons and (b) introns in the data from the annotated portion of the human genome. ....                                                   | 43 |
| 3.4: Histogram plots of the dipole angular spectral entropy ( $H_{as}$ ) with twist angle taken as $36^\circ$ ....                                                                                                                                             | 44 |
| 3.5: Plots of the probability density function curves of the dipole angular entropies (given by (a)) and dipole angular spectral entropies (given by (b)) with twist = $36^\circ$ for the introns and exons in the annotated portion of the human genome. .... | 47 |

|                                                                                                                                                                                                                                                                                                               |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.6: Plot of the probability density function curves of the (a) dipole angular entropies and (b) dipole angular spectral entropies with twist angle as a random variable uniformly distributed between (20° -50°). The data includes the introns and exons in the annotated portion of the human genome. .... | 48  |
| 3.7: Figure showing the tapering down of the exonic regions in an annotated gene using dipole angular entropies. ....                                                                                                                                                                                         | 52  |
| 3.8: Plot showing the tapering down of the exons in an unannotated gene using the dipole angular entropies (Hang). ....                                                                                                                                                                                       | 53  |
| 4.2: Histograms and PDFs of the dipole nucleotide angular superinformation ( $H_{nas}$ ) of the introns and exons ....                                                                                                                                                                                        | 69  |
| 4.3: Plot illustrating how the dipole Base Angular superinformation technique can be used to identify the exon regions of an annotated gene. ....                                                                                                                                                             | 71  |
| 5.1: Figure illustrating the algorithm for the computation of modified methylation entropies that can be used for the classification process. ....                                                                                                                                                            | 83  |
| 5.2: Figure showing how the transformations influence the methylation entropy of the samples for the case of tumor suppressor genes. ....                                                                                                                                                                     | 90  |
| 5.3: Figure showing how the transformations influence the methylation entropy of the samples for the case of oncogenes. ....                                                                                                                                                                                  | 92  |
| 6.1: Figure showing the plots of the PDFs of the (a) Kullback-Leibler and (b) Jensen-Shannon distance measures of the healthy and tumor methylation distributions from the average healthy methylation distribution. ....                                                                                     | 106 |

|                                                                                                                                                                                                                                                                   |     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 6.2: Figure showing the plots of the PDFs of the (a) Kullback-Leibler (KL) distance and (b) Jensen-Shannon (JS) distance measures of the healthy and tumor methylation distributions from the average healthy methylation distribution, for a bin size of 10..... | 111 |
| 6.3: Figure showing the plots of the PDFs of the (a) Kullback-Leibler and (b) Jensen-Shannon distance measures of the healthy and tumor methylation distributions from the average healthy methylation distribution for the global case. ....                     | 113 |

# List of Tables

|     |                                                                                                                                                                    |    |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1 | Magnitudes and Angles of the dipole moments of the DNA bases .....                                                                                                 | 31 |
| 3.2 | Parameters defining the geometry of the DNA Helix: the roll angle, slide length and twist angle in the B, A and C forms of DNA .....                               | 36 |
| 3.3 | Mean values of dipole angular entropies ( $H_{ang}$ ) and dipole angular spectral entropies ( $H_{as}$ ).....                                                      | 45 |
| 3.4 | Values of accuracy of separation for the PDFs of the dipole moment entropies and dipole moment spectral entropies of the introns and exons.....                    | 49 |
| 4.1 | Spatial and angular co-ordinates for the dipole moment vectors of the base-pairs in double-stranded form of the DNA. ....                                          | 59 |
| 4.2 | Calculation of the dipole moment vector angles of the base-pairs in an imaginary sequence G-C C-G A-T T-A. ....                                                    | 60 |
| 4.3 | Sample computation of the dipole moment angles for an imaginary single-stranded sequence (GCAT). ....                                                              | 62 |
| 4.4 | Mean values of the dipole base angular superinformation ( $H_{bas}$ ) and dipole nucleotide angular superinformation ( $H_{nas}$ ) for the exons and introns ..... | 66 |
| 5.1 | Table containing the range of methylation intensity levels and the corresponding methylation alphabet used in our experiments. ....                                | 79 |
| 5.2 | Parameters used in the computation of classification efficiency.....                                                                                               | 83 |
| 5.3 | List of Tumor Suppressor Genes and Oncogenes for the case of KIRC .....                                                                                            | 87 |

|                                                                                                                                                                                                                                                                       |     |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 5.4 Table listing the sensitivity and specificity values of classification into healthy and tumor phenotypes for different probability transformations, using the Naïve Bayes classifier.....                                                                         | 91  |
| 5.5 Table listing the sensitivity and specificity values of classification into healthy and tumor phenotypes for different probability transformations, using the Naïve Bayes classifier. ....                                                                        | 93  |
| 6.1 Definitions of the parameters in the computation of sensitivity and specificity measures. ....                                                                                                                                                                    | 103 |
| 6.2 Values of the mean KL distance and JS distance measures obtained for the case of tumor suppressor genes for a bin-size of 10.....                                                                                                                                 | 105 |
| 6.3 Table listing the values of sensitivity and specificity for the KL distance and JS distance measures for various bin-sizes using the Naïve Bayes classifier for the case of Tumor Suppressor Genes.....                                                           | 109 |
| 6.4 Table listing the values of sensitivity and specificity for the KL distance and JS distance measures for various bin-sizes using the Naïve Bayes classifier for the case of oncogenes. The data pertains to KIRC cancer type obtained from the TCGA database..... | 110 |
| 6.5 Table listing the values of sensitivity and specificity for the KL and JS distance quantities for various bin-sizes using the Naïve Bayes classifier for the case of the global set of genes in a sample. ....                                                    | 112 |

# List of Abbreviations

| Abbreviation<br>/Acronym | Expansion                         |
|--------------------------|-----------------------------------|
| DNA                      | Deoxy-ribo Nucleic Acid           |
| CpG                      | Cytosine-phosphorous-Guanine      |
| RNA                      | Ribo Nucleic Acid                 |
| mRNA                     | messenger RNA                     |
| cDNA                     | complementary DNA                 |
| A in DNA                 | Adenine                           |
| C in DNA                 | Cytosine                          |
| T in DNA                 | Thymine                           |
| G in DNA                 | Guanine in DNA                    |
| 5mC                      | 5 methyl Cytosine                 |
| DMR                      | Differentially Methylated Regions |
| KIRC                     | Kidney Renal Clear Cell Carcinoma |
| KL distance              | Kullback Leibler distance         |
| JS distance              | Jensen Shannon distance           |
| TSGs                     | Tumor Suppressor Genes            |

|        |                                |
|--------|--------------------------------|
| SVM    | Support Vector Machines        |
| RNAseq | RNA sequencing                 |
| TCGA   | The Cancer Genomic Atlas       |
| PDF    | Probability Density Function   |
| tp     | True positives                 |
| tn     | True negatives                 |
| fp     | False positives                |
| fn     | False negatives                |
| SNP    | Single Nucleotide Polymorphism |

---