ADVANCED INFORMATION THEORY FOR THE ANALYSIS OF BIOLOGICAL DATA

NITHYA RAMAKRISHNAN

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

DECEMBER 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-इनफॉरमैटिक्स के क्षेत्रों में अनुसंधान ने हाल के वर्षों में बहुत तरक़्क़ी 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।

एंट्रोपी 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 “एंट्रॉपियों का एंट्रोपी" के रूप में समझा जा सकता है और एंट्रोपी की तुलना में अधिक उपयुक्त जानकारी सैद्धांतिक माप साबित हुई है। डीएनए के दोहरे और एकल गुथे हुए रूपों में हम द्विध्रुवीय आघूर्ण की विशिष्ट सूचना की खोज करते हैं और इंट्रॉन-एक्सॉन विभाजन समस्या में इस का प्रयोग करते हैं।

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

आधारित नहीं है, सी.पी.जी द्वीपों में मेथिलैशन 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 मेथिलैशन वितरण" पर आधारित हैं जिसकी गणना स्वस्थ नमूनों के मेथिलैशन वितरण का उपयोग करके की जाती है। औसत स्वस्थ मेथिलैशन वितरण से विशिष्ट नमूने के मेथिलैशन 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 इंट्रॉन-एक्सॉन के ज्ञानक्षेत्र के साथ-साथ कम्प्यूटेशनल एपिजेनेटिकस में आगे के शोध के लिए सुझाव प्रदान करते हैं।

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation /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