COMPUTER VISION BASED DEEP LEARNING TO PREDICT PRECURSOR MIRNA IN HUMAN GENOME

ANKIT SINGHAL

AMAR NATH AND SHASHI KHOSLA
SCHOOL OF INFORMATION TECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI

MAY 2021

COMPUTER VISION BASED DEEP LEARNING TO PREDICT PRECURSOR MIRNA IN HUMAN GENOME

by

ANKIT SINGHAL

Amar Nath and Shashi Khosla School of Information Technology

Submitted

in fulfilment of the requirements of the degree of Doctor of Philosophy

to the

Indian Institute of Technology Delhi

MAY 2021

Certificate

This is to certify that the thesis titled **Computer Vision based Deep Learning to predict precursor miRNA in Human genome** being submitted by **Mr Ankit Singhal** for the award of **Doctor of Philosophy** in **Information Technology** and applications of Computer Science is a record of *bona fide* work carried out by him under our guidance and supervision at the Amar Nath and Shashi Khosla School of Information Technology, Indian Institute of Technology Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

S. N. Maheshwari Sanjiva Prasad
Honorary Professor Professor
Department of Computer Science and Engineering
Indian Institute of Technology Delhi
New Delhi 110016

Acknowledgements

I am greatly indebted to Prof. S. N. Maheshwari and Prof. Chetan Arora for making this thesis a success. I am also grateful to Prof Sanjiva Prasad for helping me during the course of the Ph. D. I am also thankful to the HPC facilities at IITD, which allowed me to do a variety of experiments with large CPU, GPU and Disk requirements.

I am also grateful to my parents. Without their support, it would not have been possible for the day to come that I become able to write a thesis. I wish to thank my friends Happy Mittal, Ankit Anand, Dinesh Khandelwal, Vijay, Saurabh for making my stay memorable at IITD. I am also thankful to the lab staff for their cooperation.

Ankit Singhal

Abstract

This thesis reports an investigation of an *ab initio* method to predict precursor miRNAs in the human genome. The research program followed is based on a computational study done from a thermodynamic perspective by Higgs on transfer RNAs (tRNA). Higgs reported that in terms of free energy most of the tRNAs have a significant gap between the ground state and the next excited state. In contrast, the minimum energy conformation of random sequences is separated from the next (excited) state by a much smaller amount. Higgs also observed that, in general, the number of alternative competing states (*i.e.*, neighbouring local minima) are smaller for random sequences. That is, although the total number of local minima states is larger for tRNA, the number of local minima states close to the ground state is larger for random sequences. In effect, the ground state is less stable in random sequences because there are more alternative secondary structures with similar Gibbs energies. Since miRNA secondary structures are under similar evolutionary pressures as tRNAs, it is expected that Higgs's observations for tRNAs would apply equally to miRNA. Exploration of local minima around a global minimum as suggested by Higgs's observations requires working with a collection of hairpin-shaped folds.

A straightforward way of exploring all possible alternate folds of a sequence and its immediate periphery by repeatedly folding various subsequences becomes computationally expensive. We have developed a graph-based model to capture the dynamics of alternate folds along the lines of an earlier investigation carried out as a part of this research paradigm. The graph consists of nodes which represent complementary matched regions. Edges represent bulges in the hairpin folds of a precursor miRNA. A path in such a graph represents a hairpin fold corresponding to a subsequence. The challenge lies in analysing the graph structure as a whole. It is our thesis that modern image processing techniques, which are now effective for scene analysis, face recognition, etc., can be used to analyse images of graphs that have encoded in them context-dependent information such as distribution of local minima.

We have attempted such an analysis using the following steps:

We have developed a graphical representation for the local minima folds of a sequence.

We have developed an algorithm to dynamically maintain a generalized suffix tree where one of the strings being maintained is a reverse complement of the other. Because of this reverse complement relationship the challenge is in devising an algorithm that dynamically updates a suffix tree to simultaneously slide a window in the forward direction over one string and in the reverse direction over the other string. This suffix tree is then used to incrementally build the graph for the whole genome.

We have developed a technique to represent the graph and its corresponding sequence by an image. We have then applied Deep Learning on the images generated, to predict

whether the given sequence of genome has a precursor miRNA.

We have developed a technique to aggregate the results obtained over images that are shifted by some fixed amount to increase the confidence level in declaring a genomic site as containing precursor miRNA.

सार

यह थीसिस मानव जीनोम में पूर्ववर्ती माइक्रो आरएनए की भविष्यवाणी करने के लिए एक प्रारम्भिक विधि की जांच की रिपोर्ट करती है।

यह शोध कार्यक्रम हिग्स द्वारा ट्रांसफर आरएनए (टीआरएनए) पर थर्मोडायनामिक परिप्रेक्ष्य से किए गए एक कम्प्यूटेशनल अध्ययन पर आधारित है।

हिग्स ने बताया कि मुक्त ऊर्जा के संदर्भ में अधिकांश (टीआरएनए में जमीनी अवस्था और अगली उत्तेजित अवस्था के बीच एक महत्वपूर्ण अंतर होता है।

इसके विपरीत, यादृच्छिक अनुक्रमों की न्यूनतम ऊर्जा संरचना को अगले (उत्तेजित) राज्य से बहुत कम मात्रा में अलग किया जाता है।

हिग्स ने यह भी देखा कि, सामान्य तौर पर, वैकल्पिक प्रतिस्पर्धी राज्यों की संख्या (यानि कि पड़ोसी स्थानीय न्यूनतम) यादृच्छिक अनुक्रमों के लिए छोटी होती है।

यही है, हालांकि टीआरएनए के लिए स्थानीय न्यूनतम अवस्थाओं की कुल संख्या बड़ी है, जमीनी अवस्था की करीबी स्थानीय न्यूनतम अवस्थाओं की संख्या यादृच्छिक अनुक्रमों के लिए बड़ी है।

वास्तव में, यादृच्छिक अनुक्रमों में जमीन की स्थिति कम स्थिर होती है क्योंकि समान गिब्स ऊर्जा के साथ अधिक वैकल्पिक माध्यमिक संरचनाएं होती हैं।

चूंकि माइक्रो आरएनए माध्यमिक संरचनाएं टीआरएनए के समान विकासवादी दबाव में हैं, इसलिए यह उम्मीद की जाती है कि टीआरएनए के लिए हिग्स के अवलोकन माइक्रो आरएनए पर समान रूप से लागू होंगे।

हिग्स के अवलोकनों द्वारा सुझाए गए वैश्विक न्यूनतम के आसपास स्थानीय न्यूनतम की खोज के लिए हेयरपिन के आकार के फोल्ड के संग्रह के साथ काम करने की आवश्यकता है।

अनुक्रम के सभी संभावित वैकल्पिक सिलवटों और उसके तत्काल परिधि को बार-बार विभिन्न अनुक्रमों को मोड़ने का एक सीधा तरीका कम्प्यूटेशनल रूप से महंगा हो जाता है।हमने इस शोध प्रतिमान के एक भाग के रूप में की गई पिछली जांच की तर्ज पर वैकल्पिक सिलवटों की गतिशीलता को पकड़ने के लिए एक ग्राफ-आधारित मॉडल विकसित किया है।ग्राफ में नोड्स होते हैं जो पूरक मिलान वाले क्षेत्रों को दर्शाते हैं।एज एक अग्रदूत miRNA के हेयरपिन सिलवटों में उभार को दर्शाते हैं।इस तरह के ग्राफ़ में पथ एक हेयरपिन फ़ोल्ड को दर्शाता है।संपूर्ण रूप से ग्राफ संरचना का विश्लेषण करने में चुनौती निहित है।

यह हमारी थीसिस है जो कि आधुनिक छवि प्रसंस्करण तकनीक, जो अब दृश्य विश्लेषण, चेहरे की पहचान, आदि के लिए प्रभावी हैं, का उपयोग उन ग्राफ़ की छवियों का विश्लेषण करने के लिए किया जा सकता है, जिनमें संदर्भ-निर्भर जानकारी जैसे स्थानीय न्यूनतम का वितरण शामिल है।

हमने निम्नलिखित चरणों का उपयोग करके इस तरह के विश्लेषण का प्रयास किया है:

हमने अनुक्रम के स्थानीय न्यूनतम फोल्ड के लिए एक ग्राफिकल प्रतिनिधित्व विकसित किया है।

हमने एक सामान्यीकृत प्रत्यय पेड़ को गतिशील रूप से बनाए रखने के लिए एक एल्गोरिदम विकसित किया है जहां एक स्ट्रिंग को बनाए रखा जा रहा है जो दूसरे के विपरीत पूरक है। इस रिवर्स पूरक संबंध के कारण चुनौती एक एल्गोरिथम तैयार करने में है जो एक प्रत्यय पेड़ को गतिशील रूप से अद्यतन करता है ताकि एक स्ट्रिंग पर आगे की दिशा में एक विंडो को स्लाइड किया जा सके और दूसरी स्ट्रिंग पर विपरीत दिशा में एक साथ स्लाइड किया जा सके। इस प्रत्यय के पेड़ का उपयोग पूरे जीनोम के ग्राफ को बढ़ाने के लिए किया जाता है।

हमने एक छवि द्वारा ग्राफ और उसके अनुरूप अनुक्रम को दर्शाने के लिए एक तकनीक विकसित की है।
हमने उत्पन्न छवियों पर डीप लर्निंग लगाई, यह अनुमान लगाने के लिए कि क्या जीनोम के दिए गए अनुक्रम में एक अग्रदूत माइक्रो आरएनए है।

हमने कुछ निश्चित मात्रा द्वारा स्थानांतरित छवियों पर प्राप्त परिणामों को एकत्रित करने के लिए एक तकनीक विकसित की है जो कि एक जीनोमिक साइट को अग्रदूत माइक्रो आरएनए युक्त घोषित करने में आत्मविश्वास के स्तर को बढ़ाती है।

Contents

List of Figures

List of Tables