

**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**

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

<b>Certificate</b>	<b>i</b>
<b>Acknowledgements</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>List of Figures</b>	<b>xv</b>
<b>List of Tables</b>	<b>xxv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Literature Review . . . . .	3
1.3 Evaluation of the Quality of Predictions made by Existing Tools . . . . .	9

---

1.4	Research Agenda . . . . .	11
1.5	Review of Nath's Approach . . . . .	14
1.5.1	Generation of Nodes and Graph . . . . .	14
1.5.2	Nath's Analysis Framework . . . . .	17
1.5.3	Dissecting Nath's Model . . . . .	19
1.5.4	Energy Rules . . . . .	22
1.5.5	Validity Rules . . . . .	22
1.6	Conclusion . . . . .	23
<b>2</b>	<b>Representing a Genomic String by a Graph</b>	<b>25</b>
2.1	Representing All Folds of a Genomic Sequence within a Specified Range by a Graph . . . . .	26
2.1.1	Folds of a Genomic Sequence . . . . .	26
2.1.2	Representing a Fold by a Graph . . . . .	30
2.1.3	Representing Different Folds of a Sequence by a Graph . . . . .	33
2.1.4	Representing All Folds with End Points in a Given Range . . . . .	35
2.1.5	Building the Graph Incrementally . . . . .	36

---

2.2	Graph Encoding Hairpin Folds . . . . .	39
2.3	Representing Specific Types of Folds . . . . .	45
<b>3</b>	<b>Incremental Suffix Tree</b>	<b>51</b>
3.1	Motivation . . . . .	51
3.2	Literature Review . . . . .	54
3.3	Articulate Suffix Tree . . . . .	55
3.4	Generalized Articulate Suffix Tree . . . . .	56
3.5	Forward sliding . . . . .	57
3.5.1	Deletion at the Beginning . . . . .	58
3.5.2	Insertion at the End . . . . .	59
3.6	Reverse Sliding . . . . .	65
3.6.1	Deletion at the End . . . . .	66
3.6.2	Insertion at the Beginning . . . . .	70
<b>4</b>	<b>Incremental Graph Generation</b>	<b>75</b>
4.1	Node and Edge Generation . . . . .	75

---

4.2	Incrementally Updating the Graph . . . . .	76
4.3	Dynamic Programming for General Folds . . . . .	79
4.4	Dynamic Programming on the Incremental Graph . . . . .	82
4.5	Validity Rules . . . . .	87
<b>5</b>	<b>Graph Visualization</b>	<b>99</b>
5.1	Mapping Strategy . . . . .	101
5.2	Image Generation for DNN based Classification . . . . .	102
<b>6</b>	<b>Deep Learning Based Analysis of Hairpin Fold Images</b>	<b>109</b>
6.1	Dataset Generation . . . . .	109
6.2	Training a 2-Layer Network on Resnet50 Features . . . . .	111
6.3	Combining 41 Probability Values of the Trained Model . . . . .	113
6.4	Testing the Model on Larger Sequences . . . . .	115
6.5	Experiments . . . . .	119
6.6	Some Running Time Experiments . . . . .	134
6.7	Conclusion . . . . .	137

<b>CONTENTS</b>	<b>xiii</b>
<b>7 Conclusions</b>	<b>139</b>
<b>Bibliography</b>	<b>143</b>
<b>Appendices</b>	<b>155</b>
<b>A Suffix Tree</b>	<b>157</b>
A.1 Data Structure . . . . .	157
A.2 Proof: Node to be Merged Does Not Have an Incoming Suffix Link in an Articulate Suffix Tree . . . . .	158
A.3 An Articulate Suffix Ends at an Internal Node . . . . .	158
A.4 Converting the Largest Articulate Suffix . . . . .	159
A.5 Recomputing General Edge Label . . . . .	159
A.6 Edge Split . . . . .	160
A.7 Edge Merge . . . . .	160
A.8 Maintain Suffix Links and Indicator and Link Vectors . . . . .	160
A.9 Maintaining First and Last Leaves . . . . .	161
A.10 Generalization . . . . .	161

B Dynamic Programming	163
Biography	168

# List of Figures

1.1	A hairpin fold of precursor miRNA mi0001519 in human genome. The sequence of this precursor miRNA is <i>AGU ACCAAAGUGCUC AU AGUGC AGGUAGUUUUGGCAU GACUCU ACUGUAG UAUGGGCACUUC CAGUACU</i> . . . . .	4
1.2	Sample fold with multiloop. . . . .	10
1.3	Three local minima folds of precursor miRNA mi0001519, as computed by our tool. . . . .	16
1.4	A graph constructed from the folds shown in Figure 1.1, 1.3a, 1.3b and 1.3c .	17
1.5	2-Dimensional representation of global and local minima folds for a precursor miRNA [56]. . . . .	18
1.6	Competing substructures of a graph by Nath [56] (page 60 chapter 4). . . . .	19
1.7	Different nodes for wobble and non wobble as created by Nath [56]. . . . .	19

---

2.1	Sample fold with all possible types of loops. . . . .	27
2.2	A graph created from a fold. . . . .	29
2.3	A graph created from another possible fold of subsequence 3 to 58 of the genomic sequence. . . . .	33
2.4	Graph representing the folds in Figure 2.2 and Figure 2.3. . . . .	34
2.5	A graph created from a fold of subsequence 42 to 57 of the genomic sequence.	35
2.6	A graph created from a fold of subsequence 6 to 23 of the genomic sequence.	36
2.7	Graph generated after combining the graphs in Figure 2.4, Figure 2.5 and Figure 2.6 . . . . .	37
2.8	Graph after adding nodes and edges due to incrementing the end index of the subsequence from 58 to 59. . . . .	38
2.9	Fold for the subgraph of the traversal that starts and ends at $n_{s4}$ . . . . .	39
2.10	A graph pattern for the clover leaf fold. . . . .	40
2.11	A graph pattern for the hairpin fold. . . . .	40
2.12	Graph created from the hairpin fold. . . . .	41
2.13	A hairpin fold from index 1 to 28 of sequence <i>GAGGGAUAGGUAGAAA CUACCCU AAUUUU</i> . . . . .	43

---

2.14	Graph generated for structure represented in Figure 2.13. . . . .	43
2.15	Graph generated after representing a pair of subsequences by an edge. . . . .	43
2.16	Graph generated for the considered sequence. . . . .	45
2.17	Histogram of distribution of length of precursor miRNA sequences as reported in mirBase v22 [23]. . . . .	46
2.18	Distribution of average approximate version of number of local minima folds of precursor miRNA and random sequences. The average number of folds are computed by taking an energy interval of 0.5 kcal/mol. . . . .	48
2.19	Distribution of average number of local minima folds of tRNA and random sequences as shown by Higgs [30]. The average number of folds are computed by taking an energy interval of 0.5 kcal/mol. The solid line represents the distribution for tRNA and the dashed line represents the distribution for random. . . . .	49
3.1	Explicit suffix tree of the sequence $S_1 = ACUCAG$ and $S_2 = CUGAGU$ . A \$ has been appended at the end of $S_1$ and # has been appended at the end of $S_2$ . . . . .	52
3.2	Common substring of the sequence $S_1 = ACUCAG$ and $S_2 = CUGAGU$ . A \$ has been appended at the end of $S_1$ and # has been appended at the end of $S_2$ . . . . .	52

---

3.3	Mapping of the common substring in $S_2$ , back to $S_1$ . The matches created in $S_1$ are shown by red curves. . . . .	53
3.4	Node created for the matches shown in Figure 3.3. . . . .	53
3.5	Implicit suffix tree of the sequence $ACGA$ . . . . .	55
3.6	Articulate suffix tree of the sequence $ACGA$ . . . . .	56
3.7	Generalized articulate suffix tree for the window $ACGA$ from forward string $S$ and window $UCGU$ from the reverse complemented string $S^{rc}$ . $S$ is denoted by $S_1$ and $S^{rc}$ is denoted by $S_2$ in the image. . . . .	57
3.8	Generalized articulate suffix tree for the window $ACGA$ from forward string $S$ and window $UCGU$ from the reverse complemented string $S^{rc}$ . The edge to be deleted, for performing delete at beginning of window in forward string, has been crossed by red edges. . . . .	58
3.9	Generalized articulate suffix tree after performing delete at beginning in window over forward string. The tree now represents the suffixes for the window $CGA$ from forward string $S$ and window $UCGU$ from the reverse complemented string $S^{rc}$ . . . . .	59
3.10	Generalized articulate suffix tree after showing the extension of articulate suffixes when the character to be inserted does not exist already on any outgoing edge of the node where the articulate suffix ends. . . . .	61

- 
- 3.11 Generalized articulate suffix tree after showing the extension of articulate suffixes when one character is inserted at the end. . . . . 62
- 3.12 Generalized articulate suffix tree after performing insertion at the end in window over the forward string. The tree now represents the suffixes for the window  $CGAC$  from forward string  $S$  and window  $UCGU$  from the reverse complemented string  $S^{rc}$ . . . . . 64
- 3.13 Generalized articulate suffix tree after showing the deletion of a character from an explicit suffix. In this case, at least one character remains on the leaf edge and the suffix remains explicit after deletion of the character. . . 67
- 3.14 Generalized articulate suffix tree after showing the deletion of a character from an explicit suffix. In this case, the leaf edge is not left with any character and the explicit suffix gets converted to an articulate suffix. . . . 67
- 3.15 Generalized articulate suffix tree showing the deletion of a character from the end of an articulate suffix. . . . . 68
- 3.16 Generalized articulate suffix tree after performing deletion at the end in the window over the reverse complemented string. The tree now represents the suffixes for the window  $CGAC$  from forward string  $S$  and window  $UCG$  from the reverse complemented string  $S^{rc}$ . . . . . 70

3.17	Generalized articulate suffix tree after performing insertion at the beginning in the window over the reverse complemented string. The tree now represents the suffixes for the window <i>CGAC</i> from the forward string <i>S</i> and window <i>GUCG</i> from the reverse complemented string $S^{rc}$ . . . . .	72
4.1	Adjacent matched regions for Watson Crick and wobble pairs. . . . .	76
4.2	Matches created when the window is slid over the genomic sequence. . . .	77
4.3	An example of the case when a match is added to a node with 2 matches. The last match of both the nodes is the same. . . . .	77
4.4	An example of the case when the child nodes of a node with 3 matches are copied to the child adjacency of a node with 2 matches and having the same first match. . . . .	78
4.5	Various types of folds that can be constructed given that the character at index $i$ matches the character at index $j$ . . . . .	80
4.6	Various types of folds that can be constructed if the characters at index $i$ and $j$ may enclose a multi loop. . . . .	81
4.7	Graph generated for the sequence <i>GAGGGAUAGGUAGAAACUACCCUAAUUUU</i> . . . . .	84
4.8	Graph after computing minimum path energy at $n_2$ . . . . .	85
4.9	Graph after computing minimum path energy at $n_5$ . . . . .	86

---

4.10	Graph after computing minimum path energy at $n_1$ . . . . .	87
4.11	Graph after computing minimum path energy at $n_4$ . . . . .	88
4.12	Distribution of loop length in miRNA folds present in mirBase v22 [23]. . . . .	89
4.13	Distribution of asymmetry in any bulge in miRNA folds. . . . .	90
4.14	Distribution of fold length in miRNA folds present in mirBase v22 [23]. The dangling ends are excluded in computing this length. . . . .	91
4.15	Distribution of total asymmetry in miRNA folds. . . . .	92
4.16	Distribution of maximum bulge size in miRNA folds. . . . .	93
4.17	Distribution of maximum lower bulge size in miRNA folds. . . . .	93
4.18	Distribution of maximum upper bulge size in miRNA folds. . . . .	94
4.19	Distribution of total bulge length in miRNA folds. . . . .	95
4.20	Distribution of total upper bulge length in miRNA folds. . . . .	96
4.21	Distribution of total lower bulge length in miRNA folds. . . . .	96
4.22	Distribution of number of matches in miRNA folds computed by our im- plementation. . . . .	97
5.1	An image constructed for the graph shown in Figure 1.4. This is a scaled version of mapping the graph to an image for illustration. . . . .	100

5.2	A graph generated for a 224 length sequence obtained after appending genomic contextual sequence on both sides of precursor miRNA mi0001519. This is a 224x224 image as generated by our tool. . . . .	103
5.3	An image encoding path energy gradation for a graph of 224 length genomic sequence containing precursor miRNA mi0001519. The graph is shown in Figure 5.2. . . . .	104
5.4	Minimum energy distribution for the folds of precursor miRNA and random sequences as computed by our tool. . . . .	105
5.5	An image encoding the graded path energy at scale 2 for the graph shown in Figure 5.2. The resulting image is of size 448*448. . . . .	106
5.6	A portion of the 448*448 image, from Figure 5.5, that contains most of the information about folds of the given sequence. . . . .	107
6.1	Model <i>m1</i> trained on Resnet50 features. . . . .	111
6.2	Train accuracy for the model trained on Resnet50 features. . . . .	112
6.3	Validation accuracy for the model trained on Resnet50 features. . . . .	113
6.4	Train accuracy for the model trained on probability values generated by model <i>m1</i> . . . . .	114
6.5	Validation accuracy for the model trained on probability values generated by model <i>m1</i> . . . . .	115

---

6.6	Prediction over a 1000 length sequence which contains the precursor miRNA in the center. . . . .	115
6.7	Prediction over a 1000 length sequence which does not contain a precursor miRNA. . . . .	116
6.8	Prediction over a 47000 length sequence from the genome. . . . .	117
6.9	Images corresponding to support vectors of precursor miRNA that have low energy folds as precursor miRNA in general. . . . .	121
6.10	Images corresponding to support vectors of precursor miRNA that have higher energy folds as in the case of random sequences. . . . .	121
6.11	Distribution of support vectors w.r.t shift from the center image. A negative shift implies shift towards left and a positive shift implies shift towards right. . . . .	122
6.12	Accuracy on the model with 1024 and 512 units in the first and second layer respectively. The lreepoch was 4 in this case. . . . .	124
6.13	Probability corresponding to correct classification of precursor miRNA. . .	129
6.14	Probability corresponding to incorrect classification of precursor miRNA. .	130
6.15	Probability corresponding to correct classification of random sequences. . .	131
6.16	Probability corresponding to incorrect classification of random sequences. .	132

---

6.17 Probabilities mi0000270 largerContext. . . . .	132
6.18 Probabilities set1 135153 chr19 largerContext. . . . .	133
6.19 Time variation for the graph generation as the sequence length is increased by 10k every time. . . . .	134
6.20 Space variation for the graph generation as the sequence length is increased by 10k every time. . . . .	135
6.21 Variation of time required for running the pipeline as the sequence length is increased by 10k every time. This is the time required to run the image generation, generation of Resnet50 features, normalizing Resnet50 features, passing the Resnet50 features through first model that has (1024,512) as the number of units in the hidden layers. This generates the probability values that whether the given sequence has a precursor miRNA. After normalizing, these probability values are passed through the second model that has (30,15) as the number of units in the hidden layers. . . . .	136

# List of Tables

6.1	The number of false positives when the prediction is run for many sequences of length near 40k in chromosome 2. The start position, length and number of false positive for each sequence are shown. . . . .	118
6.2	5-fold accuracy when a linear SVM was trained over Resnet50 features. . .	119
6.3	Comparison of mean accuracy in 5 fold validation when shift -40 to 40 and -25 to 25 were used to train a linear SVM over Resnet50 features . . . . .	122
6.4	Comparison of accuracy when different number of layers were tried in the neural network. The different values that were tried for lreproch have also been shown. . . . .	123
6.5	Accuracy when batch normalization and dropout has been applied to the network. . . . .	125
6.6	Accuracy when lower lrdf was used. . . . .	126

6.7	Accuracy when higher dropout was used . . . . .	126
6.8	Accuracy when less lrdf and more lreepoch was used. . . . .	126
6.9	Combining the predictions by simple voting . . . . .	127
6.10	Accuracy on test set when different types of aggregation methods are used.	133