

Faculty of Engineering and Information Technology
University of Technology Sydney

**Application of Information Theory to
RNA-sequencing Data Sets for Better
Understanding of Human Cancers**

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Chaowang Lan

February 2020

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Chaowang Lan declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: Jan, 30, 2020

Acknowledgments

Firstly and foremost, I profoundly grateful to my chief supervisor Prof. Jinyan Li and my co-supervisor Prof. Gyorgy Hutvagner. They have guided and encouraged me to carry on through these four years and allowed me to grow as a good researcher. My research would have been impossible without their assistance and support. I have been extremely lucky to have two supervisors who cared about my research and responded my questions patiently.

I would like to express my sincere gratitude to Dr. Eilleen M.McGowan in the School of Life Sciences of UTS. Thank you for offering time to revise my paper and providing me lots of information on breast cancer. Without her help, I would not have been published paper on the high quality journal. Thank you so much for your knowledge in cancer research, writing skill, and support.

I really grateful to two my team members Dr. Hui peng and Yi Zheng, not only for providing valuable feedback and developing my ideas in my research, but also their help in my life in Australia. I also grateful to Demi Truong Aung for the effect on doing the wet-lab experiment to confirmed my prediction. When you told me that the wet-lab experiment is confirmed my prediction, I realised my research was very important for biological research.

Of course, I sincere thanks to Dr. Renhua Song, Dr. Jing Ren, and Dr. Shameek Ghosh for their help to discover the research topic at the beginning of my PhD study. I would like to thank the other team members: Yuangsheng Liu, Zhixun Zhao, Xiaocai, Zhang, Tang Tao, and Xuan Zhang, not only for giving me large amount of inspiring advise to me research in group meeting,

Acknowledgments

but also for giving me four years of interesting activities. I could hardly forget the wonderful time when we participated in these interesting activities.

I thank to my PhD scholarship funding source which is provided by China Scholarship Council. I also think to all staffs in Advanced Analytics Institute and School of Computer Science for offering conveniences in my research.

Finally, I deepest think to my parents and brother. It is because of their trust, encouragement, and unconditional support to me, I could overcome the difficulties encountered during my four years study. Thank your very much!

Chaowang Lan

August 2019 @ UTS

Contents

Certificate	i
Acknowledgment	iii
List of Figures	ix
List of Tables	xv
List of Publications	xvii
Abstract	xix
Chapter 1 Introduction	1
1.1 Background	1
1.1.1 Information theory and its application in biological research	1
1.1.2 The regulatory mechanisms of long noncoding RNA, microRNA (isomiR), and mRNA in cancers	3
1.1.3 The introduction of RNA-sequencing data	8
1.2 Research aims and objectives	10
1.3 Research contribution	11
1.4 Thesis Structure	14
Chapter 2 Literature Review and Background	15
2.1 Technologies for miRNA (isomiR) target Prediction	15
2.1.1 Traditional methods for predicting miRNA target mRNA	16
2.1.2 High-throughput based methods for predicting miRNA target mRNA	17
2.1.3 Methods for predicting miRNA target lncRNA	18

2.2	Algorithms for constructing RNA regulatory network	19
2.2.1	Target-based method for constructing ceRNA network	20
2.2.2	Expression-based method for constructing ceRNA network	22
2.3	Methods for discovering biomarker in cancers	26
2.4	Applied mathematical methods and bioinformatics tool	34
2.4.1	Methods for selecting threshold	34
2.4.2	Kyoto Encyclopedia of Genes and Genomes pathway	35
2.4.3	Support vector machine	36
2.5	Summary	37
Chapter 3 Construction of Competing Endogenous RNA Networks from Paired RNA-seq Data Sets by Pointwise Mutual Information		
3.1	Introduction	40
3.2	Method	41
3.2.1	Definitions and Data Preprocessing	43
3.2.2	Constructing a candidate ceRNA network.	44
3.2.3	Computing the competition score	46
3.2.4	Selecting a crosstalk which has a significant competition score	48
3.3	Results	49
3.3.1	Two important ceRNA networks related to breast cancer	50
3.3.2	Characteristics of the two ceRNA networks	52
3.3.3	CeRNA networks and breast cancer treatment	55
3.3.4	Roles of ceRNA networks in KEGG pathways	57
3.3.5	A ceRNA which may be an efficient drug target for breast cancer treatment	60
3.3.6	Comprehensive Comparison with Other Methods	61
3.4	Conclusion and Discussion on Future Work	63

Chapter 4	An isomiR Expression Panel Based Novel Breast Cancer Classification Approach using Improved Mutual Information	65
4.1	Introduction	65
4.2	Method	69
4.2.1	Data Source and Definitions	69
4.2.2	Removal of lowly expressed isomiR	71
4.2.3	Calculating the weight of isomiR by improved mutual information	73
4.2.4	Identification of isomiR biomarkers that classify breast cancer subtypes	75
4.3	Results and Discussion	76
4.3.1	Characterization of isomiRs identified in different subtypes of breast cancer	76
4.3.2	Identification of isomiRs that classify breast cancer subtypes	81
4.3.3	Comparing the performance of improved mutual information to other feature selection methods	83
4.3.4	IsomiRs are superior biomarkers compared to protein coding gene expression-based approaches for the classification of different subtypes of breast cancer	84
4.3.5	IsomiRs may play important regulatory roles in different subtypes of breast cancer	86
4.3.6	Assessing the role of individual isomiRs in the regulation of breast cancer specific pathways	89
4.4	Conclusion	91
Chapter 5	Identification of Glioma Subtypes Biomarkers through Information Gain	92
5.1	Introduction	92
5.2	Definition and Materials	93
5.3	Methods	95

5.3.1	Threshold selection	95
5.3.2	Information gain	95
5.4	Result	97
5.4.1	IsomiRs are highly expressed in gliomas	97
5.4.2	Selecting highly expressed isomiRs	100
5.4.3	IsomiRs could be biomarkers for classifying different glioma subtypes	102
5.4.4	The role of isomiR in glioma cancer subtypes	105
5.4.5	Predicting molecular pathways of isomiRs in glioma subtypes	107
5.4.6	Predicting general pathways that miss regulated due to elevated of 3' isomiR expression	108
5.4.7	Predicting the subtype specific changes of individual targets of miRNAs based on isomiRs	109
5.5	Conclusion	109
Chapter 6	Summary and Future Work	111
6.1	Summary	111
6.2	Future Work	112
Chapter A	Appendix: Long Table	115
Chapter B	Appendix: List of Symbols	119
Bibliography	123

List of Figures

2.1	The one-versus-all method for classified multiple classes in SVM. If the dataset contains n classes. This method is that train a classifier which distinguish one class and the rest classes. Repeat this process until all the classes are distinguished.	38
3.1	The framework of our method	42
3.2	The examples of ceRNA crosstalk and ceRNA network. (a) A ceRNA crosstalk; (b) A ceRNA network	42
3.3	A ceRNA network mediated by hsa-miR-451a. The rectangle and oval boxes contain the names of lncRNAs and mRNAs, respectively	51
3.4	The ceRNA network formed from the top 50 candidate ceRNA crosstalks mediated by hsa-miR-375. Text words in the rectangle boxes are the names of the lncRNAs and text words in the oval boxes are the names of the mRNAs.	53
3.5	The binding sites of lncRNA, miRNA, and mRNA. . .	53
3.6	The ceRNA networks involved in the chemokine signaling pathway.	58
3.7	A ceRNA network cross-regulates two mRNAs through three miRNAs.	61
3.8	The common and unique ceRNA crosstalks predicted by various methods.	62

4.1 **IsomiR biomarker subtyping methodology.** The framework of the novel methodology designed for breast cancer biomarker subtyping is composed of three discrete steps from isomiR expression profiling to identification of key isomiRs used as novel biomarkers. 70

4.2 **The distribution of total expression levels of isomiRs.** The x-axis presents the total expression level. The ratio of the isomiRs was calculated using the number of the isomiRs in the bin divided the total number of isomiRs. For example, the ratio of the expression level isomiRs that lower than 1 is about 0.65. This implies that 65% of the isomiRs total expression level is lower than 1 72

4.3 **The distributions of 3' isomiR and their wild type miRNAs across different breast cancer subtypes.** The x-axis is the variant symbol. The variant symbol is divided into two parts by the sign (-). The left part of the sign (-) is the variate type at 3' position. The right part of sign (-) is the number of nucleotide added or trimmed at the 3' position. 77

4.4 **The distributions of 5' isomiR and their wild type miRNAs across different breast cancer subtypes.** The Y-axis is the total expression level of isomiR (or wild type miRNA). The x-axis is the variant symbol. The variant symbol is divided into two parts by the sign (-). The left part of the sign (-) is the variate type at 5' position. The right part of sign (-) is the number of nucleotide added or trimmed at the 5' position. 78

4.5 **The distributions of miRNA has-let-7d-5p and its isomiRs across different breast cancer subtypes.** The 3' (5') isomiR could have different lengths. The total expression level of 3' (5') isomiR is the sum of the expression level of different length of 3' (5') isomiR. 79

- 4.6 **The performance of classification by using different number of isomiR.** The x-axis is the number of isomiRs that are used to classify the breast cancer subtype. The Y-axis is the performance of the classification. 80
- 4.7 **Comparison of our isomiR panel based novel method classification with other feature selection methods.** The x-axis is the number of isomiRs that are used to classify the breast cancer subtype. Y-axis is the performance of the classification. The higher the AUC, the better the classification. Legend: the star represents the novel method described in this chapter. The circle, and cross sign are the Filter method, and the Hellinger method, respectively. 84
- 4.8 **Comparison of isomiR and gene classification for breast cancer subtyping.** The x-axis is the number of isomiR that are used to classify the breast cancer subtype. Y-axis is the performance of the classification. The higher the AUC, the better the classification. Legend: the star and circle present the classification using mRNA and isomiR, respectively. 86
- 5.1 **The number of isomiR in each sample.** More than twenty thousand forms of isomiR are identified in each glioma patient. The x-axis presents the name of the label of glioma sample. GA: patients with astrocytoma subtype, EPE: patients with ependymoma subtype, and cell line is the patient with cell subtype. 98

5.2 **The expression level of different types of isomiRs and miRNA.** We can found that the expression level of 3' trimming isomiR is higher than wild type miRNA and the 3' untemplated additional isomiR has comparable expression level to the wild type miRNA. The 3' trimming and untemplated isomiRs may play important roles in regulating the gene pathway of glioma. The expression level of polymorphic, 5' isomiR, and 3' templated additional isomiR are related lowly expressed compare to other types of isomiRs and wild type miRNA. 99

5.3 **The expression level distribution of 5' isomiRs in different glioma subtypes.**The positive value in the x-axis means the isomiR is 5' trimming isomiR and the negative value implies that the isomiR is 5' added isomiR. 100

5.4 **The expression level distribution of 3' trimming or templated additional isomiRs in different glioma subtypes.** The negative value in x-axis indicates that the isomiR is 3' templated addition isomiR otherwise is 3' trimming isomiR. . . 101

5.5 **The expression level distribution of 3' untemplated addition isomiRs in different glioma subtypes.** The x-axis is the number of nucleotide added at the 3' position. . . . 102

5.6 **The expression level distribution of polymorphic isomiRs in different glioma subtypes.** The y-axis is the total expression level of the isomiR. The x-axis is the nucleotide substitution position of the miRNA. 103

5.7 **The PDF of the total expression level of isomiRs.** The expression level of the isomiR is continuous data. A histogram of which the ‘bin’ of the bar graph equaled 1 was applied. X-axis is the bin number. Bin 1 is the number of the isomiR which its expression level is below 1 and bin 2 is the isomiR which its expression level is between 1 and 2. Since the total expression level of isomiR was wide ranging, this histogram proved to be very large and therefore, the complete histogram could not be displayed: the distribution of the total expression level less than 40. 104

5.8 **The expression distribution of has-miR-138-5p|3’g-1 (A) and isomiR has-miR-4510|ms-6G/U (B) in different glioma subtypes.** The x-axis in this figure is the glioma subtype and the y-axis is the expression level of the isomiR. 106

List of Tables

3.1	A matrix of expression levels of RNAs	44
3.2	The binary expression matrix of RNAs transformed from Table 3.1	45
3.3	Expression fold change ratios and p-values of the lncRNAs involved in the ceRNA networks mediated by hsa-miR-451a and hsa-miR-375	54
3.4	Top-5 competition scores in the ceRNA crosstalks mediated by <i>hsa - miR - 375</i> and <i>hsa - miR - 451a</i>	55
3.5	KEGG pathways which can be regulated by ceRNA networks	59
4.1	Breast cancer subtype reclassification for isomiR identification.	71
4.2	The 20 isomiR biomarkers, their weights, and their ratios . . .	82
4.3	Five KEGG pathways which are relative to breast cancer progresses and subtype classification	88
4.4	The average expression level of isomiRs and miRNA in each breast cancer subtype.	90
4.5	5' variant isomiRs' predicted target genes	90
5.1	The type symbol and the variation form detail of isomiR . . .	94
5.2	3' isomiRs influence the KEGG pathways that relative to glioma	109
5.3	5' isomiRs and polymorphic isomiRs effect the KEGG pathways that relative to glioma	110

List of Publications

Below is the list of journal and conference papers associated with my PhD research:

Journal Papers Published

- **Lan, C.**, Peng, H, McGowan, E.M., Hutvagner, G., and Li, J., 2018. An isomiR expression panel based novel breast cancer classification approach using improved mutual information. *BMC medical genomics*, 11(6), pp.118
- Zhao, Z, Peng, H, **Lan, C.**, Zheng, Y, Fang, L, and Li, J., 2018. Imbalance learning for the prediction of N 6-Methylation sites in mRNAs. *BMC genomics*, 19(1), pp.574
- Liu, Y, **Lan, C.**, Blumenstein, M, and Li, J., 2017. Bi-level Error Correction for PacBio Long Reads. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, <https://doi.org/10.1109/TCBB.2017.2780832>
- Peng, H, **Lan, C.**, Zheng, Y, Hutvagner, G., Tao, D, and Li, J., 2016. Cross disease analysis of co-functional microRNA pairs on a reconstructed network of disease-gene-microRNA tripartite. *BMC Bioinformatics*, 18(1), 193:1-193:7
- **Lan, C.**, Chen, Q, and Li, J., 2016. Grouping miRNAs of similar functions via weighted information content of gene ontology. *BMC*

Bioinformatics, 17(S-19), pp.159-170

- Zheng, Y, **Lan, C.**, Peng, H, and Li, J., 2016. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.2460-2463
- **Lan, C.**, Peng, H, Hutvagner, G., and Li, J., Construction of Competing Endogenous RNA Networks from Paired RNA-seq Data Sets by Pointwise Mutual Information. (accepted by BMC genomics)

Abstract

This research utilizes information theory to study the regulatory roles of non-coding RNAs in human cancers. microRNAs (miRNA) are small non-coding RNAs binding to mRNAs to suppress protein expression. Long non-coding RNAs (lncRNA) can act as competing endogenous RNAs (ceRNAs) to compete with mRNAs to bind to miRNAs. LncRNAs, miRNAs, and mRNAs form the ceRNA networks, which play a vital role in regulating molecular pathways of human cancers. Furthermore, miRNA isoforms, which are called isomiRs, are also able to regulate the gene expression and could be used to distinguish cancer subtypes. Therefore, constructing ceRNA regulatory networks and identifying isomiRs as cancer subtype biomarkers are very important for understanding the regulatory role of non-coding RNAs in cancers.

Current methods for constructing ceRNA networks and discovering biomarkers that faithfully classify different cancer subtypes have some limitations. Information theory is a powerful tool for better understanding the regulatory role of non-coding RNAs in human cancer. This thesis utilizes information theory for constructing ceRNA network and discovering human cancer subtype biomarkers in cancers. The novel contributions to the research field by this thesis are enlisted below:

- A competition rule-based pointwise mutual information is proposed to construct ceRNA networks.
- An improved mutual information and an information gain are developed to identify isomiRs as biomarkers for classifying different cancer

subtypes.

- A distribution-based method is proposed to filter out the noisy data in RNA-seq data.

Three case studies have been performed to study the regulatory roles of non-coding RNAs in human cancers. (1) The first case study is to construct the competition relationships between lncRNA, miRNA, and mRNA in breast cancer by using pointwise mutual information. (2) The second case study is to utilize the improved mutual information to discover isomiR biomarkers for classifying different breast cancer subtypes. (3) The third case study applies the improved information gain to detect isomiR based biomarkers to classify different glioma subtypes.