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.

Signature: Production Note:
Signature removed prior to publication.

Date: Jan. 30. 2020

Acknowledgments

Firstly and formost, 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,

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
Augest 2019 @ UTS

Contents

List of Figures

List of Tables

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 enable 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 flitter 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.