

Faculty of Engineering and Information Technology
University of Technology, Sydney

**Rule Mining on MicroRNA
Expression Profiles for Human
Disease Understanding**

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

by

Renhua Song

July 2016

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

Acknowledgments

First, and foremost, I would like to express my gratitude to my chief supervisor, Assoc. Prof. Jinyan Li, and to my co-supervisors Assoc. Prof. Paul Kennedy and Assoc. Prof. Daniel Catchpoole. I am extremely grateful for all the advice and guidance so unselfishly given to me over the last three and half years by these three distinguished academics. This research would not have been possible without their high order supervision, support, assistance and leadership.

The wonderful support and assistance provided by many people during this research is very much appreciated by me and my family. I am very grateful for the help that I received from all of these people but there are some special individuals that I must thank by name.

A very sincere thank you is definitely owed to my loving husband Jing Liu, not only for his tremendous support during the last three and half years, but also his unflinching and often sorely tested patience. My deepest thanks is also extended to every member of my family, especially my little son Boao Liu, for his love, understanding and obedience. Special thanks are also owed to my parents Congchun Song and Guiyun Zhang for all their tremendous assistance and unflinching support.

My sincere appreciation and gratitude goes to Dr. Qian Liu and Dr. Yun Zheng, for all their freely given invaluable advice and insightful discussion throughout my research. My heartfelt admiration and thanks is also extended to Dr. Gyorgy Hutvagner and Dr. Hung Nguyen in the Centre for Health Technologies of UTS, Dr. Kotagiri Ramamohanarao in the University of

Acknowledgments

Melbourne, Dr. Limsoon Wong in the National University of Singapore, Dr. Tao Liu in the Children's Cancer Institute Australia. Their respective assistance in terms of biological knowledge was invaluable.

I would like to thank of my colleagues at the Advanced Analytics Institute (AAI), for their selfless support over the course of my PhD candidature. and for all the fun that we shared in the last three and half years.

Thanks and praises are also due to Almighty God, my Lord and Saviour, who has bestowed great blessing upon me throughout my life. During the last three and half years, while conducting my research and preparing to write this dissertation, I have come to realise the great relevance of the words of the apostle Saint Paul who proclaimed, "I can do all things through Christ who strengthens me".

Finally, to anyone whom I have not mentioned, please forgive me. I can most definitely assure you that you have occupied a unique and special place in my thoughts. Thank you!

Renhua Song

July 2016 @ UTS

Contents

Certificate	i
Acknowledgment	iii
List of Figures	xi
List of Tables	xix
List of Publications	xxi
Abstract	xxv
Chapter 1 Introduction	1
1.1 Background	1
1.1.1 Rule Mining	1
1.1.2 microRNAs, mRNAs and TFs, and their Relationships	2
1.1.3 Human Disease Studied by this Work	7
1.2 Research Questions	8
1.3 Research Contributions	11
1.4 Thesis Structure	14
Chapter 2 Related Work	17
2.1 miRNA Biomarkers	17
2.1.1 Identification of miRNA Biomarkers by qRT-PCR	19
2.1.2 Microarray Analysis of miRNA Biomarkers	21
2.2 miRNA-mRNA Regulatory Modules	24
2.2.1 Existing Databases Based on Sequence Data	24
2.2.2 Disadvantages of miRNA Target Prediction	32

2.2.3	Computational Methods for Discovering miRNA-mRNA Regulation	32
2.3	Co-regulation miRNA Network	39
2.4	miRNA-TF Co-regulatory Networks	41
2.4.1	Sequence-Based Methods	42
2.4.2	Methods Using Expression Data and Other Data	45
2.5	Limitations of Existing Methods	47
2.6	Summary	49
Chapter 3	Research Methodology	50
3.1	Definitions for Information Gain Ratio, Euclidean Distance, 10-Fold Cross Validation and Pearson's Correlation Coefficient	50
3.1.1	Information Gain Ratio	50
3.1.2	Euclidean Distance	52
3.1.3	10-Fold Cross Validation	52
3.1.4	Pearson's Correlation Coefficient	52
3.2	Data Mining Methods	53
3.2.1	A committee of decision trees	53
3.2.2	Naive Bayes Classifier	53
3.2.3	K-nearest Neighbors Algorithm	54
3.3	Our Proposed Rule Mining Methods	55
3.3.1	Rule Discovery	55
3.3.2	Strong Discriminatory Rules	56
3.4	Bioinformatics Tools	57
3.4.1	GO Term Enrichment Analysis	57
3.4.2	KEGG Pathway Enrichment Analysis	58
3.4.3	PPI Network Construction	59
3.5	Performance Measurement	60
Chapter 4	Rule Discovery and Distance Separation to Detect Reliable miRNA Biomarkers for the Diagnosis of Lung Squamous Cell Carcinoma	61

4.1	Introduction	61
4.2	Materials and Methods	63
4.2.1	Data Sets of miRNA Expressions in SCC Patients	63
4.2.2	Rule Discovery within Top-Ranked miRNAs	64
4.2.3	Rule Discovery across the Whole Feature Space	68
4.3	Results	69
4.3.1	Prediction Performance by Rules	69
4.3.2	Distance Separation in 2D and 3D Spaces to Identify Reliable Biomarkers	75
4.3.3	The Reliability of Identified Best 2D and 3D Biomarkers	78
4.3.4	The Genomic Location of Biomarker miRNAs	78
4.3.5	Target Genes of Biomarker miRNAs and Their Associated Diseases	80
4.4	Discussion	82
4.5	Conclusion	83

**Chapter 5 Connecting Rules from Paired miRNA and mRNA
Expression Data Sets of HCV Patients to Detect
both Inverse and Positive Regulations 85**

5.1	Introduction	85
5.2	Methods	89
5.2.1	miRNA and mRNA Expression Data Sets	89
5.2.2	Rule-based Identification of miRNA-mRNA Regulatory Modules	89
5.3	Results	93
5.3.1	2-miRNA Discriminatory Rules from the miRNA Expression Data	93
5.3.2	Rules from the mRNA Expression Data	94
5.3.3	A miRNA-mRNA Regulatory Interaction Network	96
5.3.4	Many-to-Many miRNA-mRNA Regulatory Modules	101
5.3.5	Negatively and Positively Regulated mRNAs by Multiple miRNAs	103

5.4	Conclusion	110
Chapter 6 Identification of Lung Cancer miRNA-miRNA Co-regulation Networks Through a Progressive Data Refining Approach 113		
6.1	Introduction	113
6.2	Materials	116
6.3	Methods	117
6.3.1	Preprocessing of miRNA Expression Data	117
6.3.2	Network Construction for Co-regulating miRNAs	119
6.3.3	Validation of the Co-regulating miRNAs	121
6.4	Results	122
6.4.1	Co-regulating miRNA Pairs and Their Big Network	122
6.4.2	Topological Characteristics of the Co-regulation Network and the Functional Modules	125
6.4.3	Lung Cancer Related miRNAs Have More Functional Synergism	128
6.4.4	Lung Cancer Related miRNAs and Their Functional Modules	128
6.4.5	KEGG Pathway Analysis Results	130
6.4.6	Transcription Factors Related to Lung Cancer	133
6.5	Summary and Conclusion	134
Chapter 7 A Novel Framework for Inferring Self-regulation miRNAs 136		
7.1	Introduction	136
7.2	Materials	138
7.2.1	Construction of TF-miRNA Relationships and miRNA- target Relationships in the Post-transcriptional Regulatory Network:	138
7.2.2	Construction of TF-target Relationships in the Transcriptional Regulatory network:	140

7.2.3	Identification of Important Network Motifs: miRNA-mediated Feed Forward Loops	140
7.2.4	Identification of Experimentally Validated Regulatory Interactions	142
7.3	Results	143
7.3.1	Self-Regulated miRNAs in the Human Regulatory Network	143
7.3.2	Identification of Self-regulated Transcription Factors	144
7.3.3	Identification of miRNA-mediate Feed-Forward Loops	145
7.3.4	Validation of miRNAs Self-Regulations	149
7.4	Discussion	150
7.5	Conclusion	152
 Chapter 8 Summary and Conclusion		153
8.1	Main Achievements	153
8.2	Direction for Future Research	157
8.3	Closing Summary	160
 Chapter A Appendix: Long Table		161
 Chapter B Appendix: Algorithm of Prim Code		163
 Chapter C Appendix: List of Symbols		165
 Bibliography		176

List of Figures

1.1	Gene regulation of miRNAs and TFs. miRNAs regulate biological processes in proliferation, metabolism, differentiation, development, apoptosis, cellular signaling and even cancer development and progression. TFs are fundamental players of gene expression regulation at the transcriptional level. miRNAs regulate target gene expression at the post-transcriptional level.	9
1.2	A graphical description framework of research plan. ①: to identify paired biomarkers instead of individual miRNA biomarkers. ②: to discover both positive and negative miRNA-mRNA regulatory modules, and miRNA-miRNA co-regulation modules. ③: to study the relationships among miRNAs, mRNAs and TFs.	11

1.3	Thesis Structure. Ch. 1 introduces the research background. Ch. 2 provides the literature review. Ch. 3 describes research methodology. Ch. 4 presents a novel method to miRNA biomarkers for SCC diagnosis. Ch. 5 presents a ‘change-to-change’ method to detect both inverse and positive regulatory relationships from a paired miRNA and mRNA expression data set of HCV patients. Ch. 6 presents a novel integrative approach to the discovery of miRNA-miRNA co-regulating networks. Ch. 7 presents a study of self-regulating miRNAs through combining the relationships among miRNAs, TFs and target genes. Ch. 8 provides a final summary of this research and also suggests some future directions.	15
4.1	Heatmap representation of the expression levels of the 19 miRNAs. A single miRNA is unable to distinguish cancer samples from normal samples, while the combination of 2 or 3 miRNAs can faultlessly identify cancer (or normal) samples from normal (or cancer) samples.	66
4.2	Distance separation by 100%-frequency rules in 2D space. The left panel shows a shorter distance separation between the cancer and normal samples than the separation shown in the right panel.	67
4.3	The procedure of rule discovery with 19 miRNAs. The up panel is the dataset processing phase, and 19 miRNAs are obtained. The down panel is the discovery phase to get biomarkers.	68
4.4	Decision trees. The left panel is a decision tree made of miR-205 and miR-98. The right panel is a decision tree made of miR-205 and miR-451.	72

4.5 **Expression data on 2D planes.** The left panel is the plane co-ordinated by miR-205 and miR-98. The right panel is coordinated by miR-205 and miR-451. The blue rectangles indicate the expression ranges of all of the normal samples. 73

4.6 **Examples of 2D rules.** The left panel describes two miRNAs whose class-label is related to normal. The right panel shows two miRNAs whose class-label is related to cancer. 76

4.7 **Examples of 3D rules.** The left panel contains miR-100, miR-199a and miR-200c. The right panel contains miR-133a, miR-21 and miR-520a-AS. 76

5.1 **Our approach in comparison to previous approaches.** We construct miRNA-mRNA regulatory modules using rule-based methods as shown in the right panel where the mRNA data set is narrowed down by the identified miRNA rules which are derived at the first step. 88

5.2 **Computational steps for the identification of miRNA-mRNA regulatory modules.** 1) Collection of miRNA expression profile data set. 2) Discovery of discriminatory rules from the miRNA expression data set using our rule discovery algorithm. 3) Construction of a selected and relevant mRNA expression data set. 4) Discovery of discriminatory rules from the relevant mRNA data set. 5) Identification of candidate miRNA-mRNA regulatory modules by combining the miRNAs and mRNAs in the discovered rules. 91

5.3 **A miRNA-mRNA regulatory interaction network.**
There is an edge between two miRNAs if they are components of a miRNA rule. The edge between a miRNA and a mRNA represents a regulation of the miRNA for its target. Six miRNAs (miR-214, miR-34a, miR-129, miR-765 and miR-210) and 9 mRNAs (ACVR1C, RAB43, FNDC5, WDR33, ALDH4A1, ANKRD12, KCTD9, ARMC1 and DICER1) all in red are confirmed by literature work. 98

5.4 **The regulatory module inferred from the first miRNA rule and its corresponding mRNAs.** miR-557 and miR-214, the miRNAs of the first HCV+ miRNA rule are placed in the up panel. Four mRNA rules are identified and their mRNAs are placed in the middle and bottom panels. The edges linking miR-214 and its mRNA targets are in solid lines, while the edges linking miR-557 and its mRNA targets are in dashed lines. The confirmed target mRNAs are also highlighted with an underline. 101

5.5 **The regulatory module inferred from the first HCV-rule consisting of miR-129 and miR-765.** In this module, miR-765 targets 6 mRNAs and miR-129 regulates 4 mRNAs. ANKRD12, a target of miR-765, is validated to be associated with chronic liver disease by existing works. 102

5.6 **The many-to-many relationship between some mRNAs and miRNAs identified in our modules** (i.e., one mRNA is targeted by many miRNAs and one miRNA can regulate many mRNAs). 102

- 5.7 **The positive expression relationship between *GFRA2* mRNA and miR-557, miR-765, and miR-17-3p.** The expression levels of the three miRNAs are preprocessed in the log scale, and the expression levels of *GFRA2* are expanded by 10 times. The three miRNAs all have a high gain ratio, separating the HCV+ and HCV- samples very well. 105
- 5.8 **The partial complementary sequence pairing between the 5' UTRs of *GFRA2* and the seed sites of miR-557, miR-765 and miR-17-3p.** The mismatched base pairs are shown in smaller font. 107
- 5.9 **The negative expression relationship between the *QKI* mRNA and miR-493-3p, miR-129, and miR-765.** The expression levels of the three miRNAs are preprocessed in the log scale. The three miRNAs all have a good gain ratio, separating the HCV+ and HCV- samples very well. 109
- 5.10 **The positive expression relationship between the *ALDH4A1* mRNA and miR-184.** The expression levels of miR-184 are preprocessed by dividing 10 and this has a good gain ratio, classifying the HCV+ and HCV- samples very well. 111

6.1 **The flowchart of constructing a miRNA-miRNA co-regulation network starting from three lung cancer data sets (①: DS1, DS2 and DS3). Preprocessing:** ②: using DS1', DS2' and DS3' to represent the common miRNAs of DS1, DS2 and DS3; ③ : using data sets PAIRS-1, PAIRS-2 and PAIRS-3 for storing miRNA pairs and their common targets by selecting highly correlated miRNA pairs containing no fewer than 10 common targets. **Identification:** ④: Identifying a miRNA pair co-regulating the same function modules by performing GO function and protein interaction analyses; ⑤: repeating the procedure for every miRNA pair in PAIRS-1, PAIRS-2 or PAIRS-3; ⑥: identifying the functional modules and constructing a common miRNA-miRNA co-regulation network by assembling all the miRNA pairs with miRNAs which are detected at least twice from the three data sets. **Verification:** ⑦: using existing databases (KEGG pathway, miR2Disease and OMIM) and graph theoretical methods to validate this co-regulation network and functional modules. 118

6.2 **A miRNA-miRNA co-regulation network. There are 41 connections and 43 nodes in this network.** A node stands for a miRNA, and an edge connecting two nodes represents a co-regulation. The miRNAs with red circles (points) are confirmed to be associated with lung cancer from the miR2Disease database. The verified co-regulating miRNA pairs are highlighted in the blue dashed circles and let-7a/b/c/d/f/g are in red. 124

6.3 **Degree distribution of the miRNAs in the co-regulation network.** The X axis stands for the degrees of each miRNA and the Y axis represents the proportion of each degree category in the miRNA-miRNA co-regulation network. There are 19 nodes having a degree of 1, while there is only one miRNA with a degree of 6. 126

6.4 **An example of co-regulation between miRNAs (let-7a, b, c and f) and one of their functional modules (SMAD2, ACVR1B, ACVR2A and ACVR2B).** There is a co-regulation between let-7c/f, let-7c/a, let-7c/b, let-7a/b, let-7a/f, and let-7a/b. SMAD2, ACVR1B, ACVR2B and ACVR2A define a functional module, and they are directly connected to each other in the protein-protein interaction network. 129

6.5 **KEGG pathway enrichment analysis for the target subsets of each miRNA pair in the co-regulation miRNA network.** The X axis shows the existing numbers of the corresponding pathways and the Y axis describes the pathways' names (P-value < 1.0e-4) in the three data sets. . . 132

7.1 **The topological model of miRNA-mediate FFLs.** A TF regulates a miRNA, and they both regulate the target mRNA, and miRNA regulation of the target gene is negative. S and D represents synthesis and degradation respectively. 142

7.2 **The relationships between miRNAs and the self-edge TFs.** The number of TFs is no less than 8. 145

7.3 **The relationships between miRNAs and the self-edge TFs.** The number of TFs is no less than 8. 145

7.4 **Self-regulated miRNAs involved in self-edge TFs.** These relationships exist in both the miRNA-target and miRNA-TF regulations. 146

7.5	Results of FANMOD trial with a subgraph size of 3. only the ID of 38 and 6 are shown in this figure.	149
8.1	Main achievements of my PhD study	155

List of Tables

2.1	miRNAs with altered expression in malignancy.	23
4.1	Projection of 5 important miRNAs onto a prioritised list of 328 miRNAs, resulting in 19 miRNAs ranked as high as these 5 miRNAs.	65
4.2	Comparisons of three classifiers on four data sets	70
4.3	Multiple 100%-frequency rules derived from the 19-miRNA data set through our iterative decision tree method.	71
4.4	The performance comparison of three datasets.	74
4.5	Shortest pair-wise Euclidean distance between the cancer and normal samples in 2D and 3D biomarker spaces.	77
4.6	The probability of different AUC values in the 1000 randomization tests.	79
4.7	The chromosomal location of the 13 miRNAs in our 2D and 3D biomarker rules	79
4.8	The targets and associated disease of our biomarkers	81
5.1	The top-ranked miRNAs with a gain ratio larger than 0.5	94
5.2	The target mRNAs and their rules for each miRNA rule.	95
5.3	The number of predicted mRNA targets in the TargetScan database and those targets common in our mRNA data set	96
5.4	All target mRNAs of miRNAs in HCV+ and HCV- modules.	97

5.5	Pearson's correlation coefficients between the miRNAs and mRNAs in the many-to-many regulatory module. '-' indicates the mRNA (in a column) is not the target of the miRNA (in a row).	104
5.6	Transition probability of two adjacent bases in the 5' UTRs of <i>GFRA2</i>	107
6.1	Progress of the miRNA pairs numbers when our analysis and requirements were getting refined.	125
6.2	GO functional analysis of a functional module with seven genes in the TP63 interaction network (p-value <1.0e-04).	127
6.3	Target genes in the functional modules	131
7.1	All of the identified motifs using a FANMOD trial with a subgraph size of 3.	147
A.1	Categorisation of miRNA target prediction tools (Categorisation was taken from a survey of computational algorithms for miRNA target prediction)	162

List of Publications

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

Journal Papers Published

- Zheng, Y., Ji, B., **Song, R.**, Wang, S., Li, T., Zhang, X., Chen, K., Li, T. and Li, J., 2016. Accurate detection for a wide range of mutation and editing sites of microRNAs from small RNA high-throughput sequencing profiles. *Nucleic acids research*, p.gkw471. <http://dx.doi.org/10.1093/nar/gkw471>
- Tee, A.E., Liu, B., **Song, R.**, Li, J., Pasquier, E., Cheung, B.B., Jiang, C., Marshall, G.M., Haber, M., Norris, M.D. and Fletcher, J.I., 2016. The long noncoding RNA MALAT1 promotes tumor-driven angiogenesis by up-regulating pro-angiogenic gene expression. *Oncotarget*, <http://dx.doi.org/10.18632/oncotarget.6675>
- Tee, A.E., Liu, P., Maag, J., **Song, R.**, Li, J., Cheung, B.B., Haber, M., Norris, M.D., Marshall, G.M., Dinger, M. and Liu, T., 2015. The long noncoding RNA MALAT1 promotes hypoxia-driven angiogenesis by upregulating pro-angiogenic gene expression in neuroblastoma cells. *Cancer Research*, 75(Supplement 15), pp.146-146
- Liu, Q., **Song, R.** and Li, J., 2015. Inference of gene interaction networks using conserved subsequential patterns from multiple time

course gene expression datasets. BMC Genomics, 16(Supplement 12): S4

- **Song, R.**, Catchpoole, D.R., Kennedy, P.J. and Li, J., 2015. Identification of lung cancer miRNA-miRNA co-regulation networks through a progressive data refining approach. Journal of Theoretical Biology, 380, pp.271-279
- **Song, R.**, Liu, Q., Liu, T. and Li, J., 2015. Connecting rules from paired miRNA and mRNA expression data sets of HCV patients to detect both inverse and positive regulatory relationships. BMC Genomics, 16(Supplement 2): S11
- **Song, R.**, Liu, Q., Hutvagner, G., Nguyen, H., Ramamohanarao, K., Wong, L. and Li, J., 2014. Rule discovery and distance separation to detect reliable miRNA biomarkers for the diagnosis of lung squamous cell carcinoma. BMC Genomics, 15(Supplement 9): S16

Conference Papers Presented

- Liu, Q., **Song, R.** and Li, J. “Inference of gene interaction networks using conserved subsequential patterns from multiple time course gene expression datasets.” www.jsbi.org/giw-incob2015/, Tokyo, Japan. **Oral Presentation**
- **Song, R.**, Liu, Q., Liu, T. and Li, J. “Connecting rules from paired miRNA and mRNA expression data sets of HCV patients to detect both inverse and positive regulatory relationships.” <http://apbc2015.mbc.nctu.edu.tw/>, HsinChu, Taiwan. **Oral Presentation.**
- **Song, R.**, Liu, Q., Hutvagner, G., Nguyen, H., Ramamohanarao, K., Wong, L. and Li, J. “Rule discovery and distance separation to detect reliable miRNA biomarkers for the diagnosis of lung squamous cell carcinoma.” <http://incob.apbionet.org/incob14/>, Sydney, Australia. **Oral and Poster Presentation.**

Papers to be Submitted/Under Review

- **Song R.**, Zheng, Y., Catchpoole, D.R., Kennedy, P.J. and Li, J.
“A Novel Framework for Inferring Self-regulation miRNAs.” **Under
Revision**

Abstract

This research employs rule mining methods to study the important roles of miRNAs in human diseases. From past experience and from reviewing the literature, rule mining is a widely used data mining technique for the discovery of interesting relationships in large data sets. MicroRNAs (miRNAs) are endogenous and highly conserved non-coding RNA molecules. They can inhibit and/or promote the post-transcriptional expression of target messenger RNAs (mRNAs). miRNAs thus play a pivotal role in a cell's differentiation, proliferation, growth, mobility, and apoptosis, as well as in viral replication and proliferation. This has inspired many research works aimed at detecting miRNAs' functions in human disease. However, with the current deluge of miRNA data, previous works have suffered from limitations in terms of handling the relationship between various molecules. Firstly, they usually identify single miRNAs as biomarkers, and always produce low sensitivity and specificity. Secondly, intensive research largely depends on the inverse expression relationships between miRNAs and mRNAs to discover miRNA-mRNA regulatory modules. Finally, the miRNA-miRNA co-regulations and miRNA self-regulations have not been well investigated. As a result, rule mining is a powerful new technology with great potential to help researchers focus on the most important miRNAs for understanding human diseases. This thesis reports our past and current research outcomes in this area. The contributions of the thesis are as follows:

- A novel rule mining method is proposed to detect the significant miRNA biomarkers.

- A “change to change” method is proposed to mine both positive and negative regulatory relationships from paired miRNA and mRNA expression data sets.
- A progressive data refining approach is proposed to identify the lung cancer miRNA-miRNA co-regulation network.
- A novel framework is proposed to detect the self-regulation miRNAs.

The research was conducted through four case studies. (1) The first case study was on lung squamous cell carcinoma for accurate diagnosis of this disease through the reliable miRNA biomarkers identified by a novel rule discovery method. (2) The second case study was on paired miRNA and mRNA expression data of HCV patients to detect both positive and negative regulatory modules. (3) The third case study was on lung cancer data sets for the computational methods to identify miRNA-miRNA co-regulation networks and miRNA-miRNA co-regulatory relationships. (4) The fourth case study was on multiple data types to infer self-regulation miRNAs in humans through an integrative rule mining framework and approach. All the results have been verified by the existing literature and databases.