

Reconstruction of Gene Regulatory Networks Using Biological Domain Knowledge

A thesis submitted in fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Akther Shermin

November 2012

Master of Science (Dhaka University) 2001

Bachelor of Science (Dhaka University) 1999



Department of Computing

Faculty of Science

Macquarie University

NSW 2109, Australia

© Akther Shermin, 2012

DECLARATION

I declare that this thesis was composed by myself and the work contained herein is the result of original research except where otherwise indicated. Some parts of the thesis include revised versions of published papers. This work has not been submitted for a higher degree to any other university or institution.

Signed: _____

Date: _____

ACKNOWLEDGMENTS

First and foremost I offer my sincerest gratitude to my supervisor, Professor Mehmet A. Orgun, who has supported me throughout my studies with his patience and knowledge whilst allowing me the room to work in my own way. Without his encouragement, effort and enthusiasm, this thesis would not have been completed or written.

I am also grateful to my co-supervisor, Dr. Abhaya Nayak, for his guidance, inspiration, encouragement and suggestions.

Apart from my supervisors I would like to thank all the researchers who are working hard to conduct biological experiments and make the data publicly available. Special thanks go to the research group of Linda Breeden Lab for giving me their unpublished data. I also thankfully acknowledge Dr. Ulrik de Lichtenberg and Dr. Min Zou for providing me their programs. I am also indebted to the many countless contributors to the "Open Source" programming community for providing the BNT tool that I have used to produce my experimental results in this thesis.

I would like to thank Dr. Hasan Jamil for his invaluable remarks on some of the topics in my thesis work. Especially, I gratefully acknowledge his kindness in presenting a research paper on my behalf in an international conference.

I am deeply indebted to my husband Afifur and my beautiful daughter Samah. Sharing with you the hours that I did not work on this thesis, made the long hours at work bearable. Thank you for your support, understanding and love!

In my daily work I have been blessed with a friendly and cheerful group of fellow students and staff members. I thank all of them for providing a friendly and enjoyable environment during my time in the department of Computing, Macquarie University.

LIST OF PUBLICATIONS

- Shermin, A., and Orgun, M.A. (2007). Learning Dynamic Bayesian Networks from cDNA Microarray Gene Expression Data. *CD-ROM Proc. of International Conference on Modelling Decisions for Artificial Intelligence (MDAI 2007)*, pp. 1-11. Kitakyushu, Japan.
- Shermin, A., and Orgun, M.A. (2009). Using Dynamic Bayesian Networks to Infer Gene Regulatory Networks from Expression Profiles. *Proc. of the 24th Annual ACM Symposium on Applied Computing (SAC 2009)*, pp.799-803. Honolulu, USA.
- Shermin, A., and Orgun, M.A. (2009). A 2-Stage Approach for Inferring Gene Regulatory Networks Using Dynamic Bayesian Networks. *Proc of the 3rd IEEE International Conference on Bioinformatics & Biomedicine (BIBM09)*, pp.166-169. Washington DC, USA.
- Shermin, A., and Orgun, M.A. (2010). Analysis of Microarray Data to Infer Transcription Regulation in the Yeast Cell Cycle. *International Journal of Functional Informatics and Personalised Medicine*, 3(1): 73-88.
- Shermin, A., Jamil, H., and Orgun, M.A. (2011). A Scalable Approach for Inferring Transcriptional Regulation in the Yeast Cell Cycle. *Proc of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp. 345-349. Chicago, USA.

ABSTRACT

High throughput technologies such as microarrays generate an enormous amount of genomic data at the cellular level. The computational reconstruction of gene regulatory networks (GRN) from this abundance of data has become a major area of research in systems and computational biology. However, the reconstruction task suffers from two major challenges: the excessive computational complexity and the low accuracy of the estimated networks. Literature of related works suggests the utilization of domain knowledge in addressing these challenges. The main interest of this thesis is to study the effectiveness of incorporating biological knowledge and other sources of biological data in the computational reconstruction of GRN.

The thesis starts with the identification of several key features of gene regulation that are used by the transcriptional regulators and employs that knowledge to restrict the number of possible regulators for each gene. We choose Dynamic Bayesian Network for the computational reconstruction of the GRN. The thesis then explores the co-regulation of genes and the potency of integrating multiple sources of biological data in the reconstruction task. Through the analysis of both real and synthetic data, this thesis also quantifies to what extent the computation time and reconstruction accuracy of the model has been improved.

The comprehensive performance and scalability analysis of various GRN models demonstrate that the employment of biological features can convincingly reduce the computational complexity of the model. Moreover, the integration of other sources of biological data makes the model computationally efficient and estimates networks with improved accuracy. Most importantly, such integration results in a scalable model; that is, the model estimates networks including thousand genes while preserving its level of performance.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	v
LIST OF PUBLICATIONS	vii
ABSTRACT.....	ix
LIST OF TABLES	xvii
LIST OF FIGURES	xix
CHAPTER 1. INTRODUCTION	1
1.1 Issues in Reconstructing GRN from Experimental Data	3
1.2 Thesis Motivation	7
1.3 Thesis Statement	8
1.4 Contributions	9
1.5 Thesis Outline	12
CHAPTER 2. EUKARYOTIC GENE TRANSCRIPTION	15
2.1 Introduction.....	16
2.2 Biological Aspect of Transcriptional Regulation	17
2.2.1 Gene expression.....	17
2.2.2 Genomic regulation system	19
2.2.2.1 Transcription regulation.....	20
2.2.2.2 Gene regulatory network	22
2.3 Quantification of Gene Expression.....	24
2.3.1 cDNA Microarrays	24
2.4 Our Model Organism and the Cellular Process	26
2.4.1 Cell Division in yeast cell cycle	27
2.5 Experimental Micorarray Data	28
2.6 Conclusion	34

CHAPTER 3. CURRENT APPROACHES TO MODELING GENE

REGULATORY NETWORKS 35

3.1 Introduction.....	36
3.2. Challenges of Analyzing Microarray Data	37
3.3. Current Approaches for Modeling GRN from Gene Expression Data	39
3.3.1 Models for discrete variables.....	40
3.3.1.1 Boolean Networks	41
3.3.1.2 Probabilistic Boolean Networks (PBN).....	43
3.3.1.3 Bayesian Networks (BN).....	44
3.3.1.4 Dynamic Bayesian Networks (DBN)	46
3.3.2 Models for continuous variables.....	48
3.3.2.1 Ordinary Differential equations (ODEs)	49
3.3.2.2 S-System model	50
3.3.2.3 Neural network models.....	51
3.4. Reconstruction of GRN by Incorporating Biological Information.....	53
3.5 Discussion.....	56
3.6 Conclusion	59

CHAPTER 4. PHASE-SPECIFIC REGULATION IN THE YEAST CELL CYCLE 63

4.1 Introduction.....	64
4.2 Related Work.....	65
4.3 Background.....	67
4.3.1 Regulation of transcription factors in yeast cell cycle.....	67
4.3.2 Learning dynamic Bayesian network from data.....	68
4.4 Methods	72
4.4.1 Data discretization	73
4.4.2 Peak time calculator of genes	73
4.4.3 Assigning genes in phase specific clusters	74
4.4.4 Inference of GRN with DBN algorithm	75
4.5 Experiments and Results.....	76

4.5.1 Experimental data	76
4.5.2 Experimental setup	77
4.5.3 Experimental results	77
4.6 Conclusion	82
CHAPTER 5. CO-REGULATION OF CO-EXPRESSED GENES	85
5.1 Introduction.....	86
5.2 Related Work	87
5.3 Background.....	89
5.3.1 Co-expression of Co-regulated Genes in yeast cell cycle.....	89
5.4 Methods	91
5.4.1 Partitioning Around the Mediods (PAM)	91
5.4.2 Identification of co-expressed genes with PAM.....	93
5.4.3 Inference of regulation among groups of co-expressed genes.....	94
5.4.4 Merge networks learned by the models	95
5.5 Experiments and Results.....	95
5.5.1 Experimental data	96
5.5.2 Experimental setup	96
5.5.3 Experimental results	96
5.6 Conclusions.....	100
CHAPTER 6. TRANSCRIPTIONAL REGULATION FROM	
MULTI-SOURCE DATA	103
6.1 Introduction.....	104
6.2. Related Work	105
6.3 Background.....	107
6.3.1 Protein-protein interaction (PPI) data	107
6.3.2 Transcription factor binding site (TFBS) data.....	109
6.4 Methods	110
6.4.1 Impute missing values	111
6.4.2 Extraction of potential regulators of genes	112

6.4.3 Learning the structure of GRN	113
6.5 Experiments and Results.....	114
6.5.1 Experimental data	114
6.5.2 Experimental setup	115
6.5.3 Experimental results	115
6.5.3.1 Experiment 1.....	115
6.5.3.1 Experiment 2.....	117
6.6 Conclusions.....	119
CHAPTER 7. PERFORMANCE ANALYSIS OF THE GRN MODELS	121
7.1 Introduction.....	122
7.2 Existing DBN-based Models for Comparison.....	123
7.3 Model Evaluation Criteria	124
7.4 Experimental Setup.....	126
7.5 Analysis of Simulated Data	126
7.6 Analysis of Experimental Data.....	134
7.6.1 Inference of small-scale networks	135
7.6.2 Comparison with synthetically generated random networks.....	139
7.6.3 Inference of large-scale networks.....	144
7.7 Conclusion	148
CHAPTER 8. SCALABILITY ANALYSIS	151
8.1 Introduction.....	152
8.2 Some Existing Scalable Approaches	153
8.3 Experimental Setup.....	154
8.4 Scalability Analysis of the Proposed GRN Models.....	155
8.4.1 Computation time	156
8.4.2 Precision	158
8.4.3 Recall	160
8.5. Extended Scalability Analysis of GRN _{Multi-sources}	164
8.6 Conclusion	166

CHAPTER 9. CONCLUSION AND FUTURE WORK	167
9.1 Summary of Contributions.....	168
9.2 Future Work.....	170
9.3 Final Remarks	172
 BIBLIOGRAPHY.....	 173

LIST OF TABLES

Table	Page
4.1: Pseudo code of DBN structure learning	71
4.2: Number of nodes vs. number of possible network structures.....	71
4.3: Dataset alpha38, includes transcription levels of 150 genes with a sampling interval of 5 minutes and a total of 22 time points.	80
4.4: Dataset alpha30, includes transcription levels of 150 genes with a sampling interval of 5 minutes and a total of 22 time points.	81
5.1: Dataset alpha30, includes transcription levels of 200 genes with a sampling interval of 5 minutes and a total of 22 time points.	99
5.2: Dataset alpha38, includes transcription level of 200 genes with a sampling interval of 5 minutes and a total of 22 time points.	99
6.1: The Dataset includes transcription levels of 250 genes with a sampling interval of 7 minutes and a total of 18 time points.	118
7.1: Comparison of performance among the four different DBN-based GRN models. The models have been applied on 5 different simulated datasets of 20 genes over 21 time points.....	128
7.2: Comparison of performance among the four different DBN-based GRN models. The models have been applied on 5 different simulated datasets of 50 genes over 21 time points.....	130
7.3: Comparison of performance among the four different DBN-based GRN models. The models have been applied on 5 different simulated datasets of 100 genes over 21 time points.....	132
7.4: Comparison of performance among the five different DBN-based GRN models	145

8.1: Dataset alpha30. Comparison of performance among the five different DBN-based GRN models.....	162
8.2: Dataset alpha38. Comparison of performance among the five different DBN-based GRN models.....	163
8.3: Comparison of the performance of GRNMulti-sources in analyzing medium to large-scale networks including 100, 500 and 1000 genes respectively.....	164

LIST OF FIGURES

Figure	Page
1.1: Illustration of the process of reconstructing GRN from experimental data.....	3
2.1: The flow of biological information from DNA to RNA to protein.	19
2.2: Regulatory elements on a protein-encoding DNA module.....	21
2.3: Illustration of a simple GRN.....	23
2.4: Principles and steps of a simple microarray experiment.	25
2.5: Distinct Phases of Cell Division Cycle.....	28
2.6: Dynamics of gene expression of the 13 known cell-cycle regulated TFS over the two cell cycles.....	29
2.7: GRN of the 13 known TFS.....	30
2.8: A simple GRN among three TFs during the first two phases of the cell cycle.	31
2.9: Fluctuation in the level of gene expression of the three TFs (SWI4, HCM1, and WHI5) as measured by microarray experiments.	32
2.10: Order of expression among the three TFs (SWI4, HCM1, and WHI5) within a cell cycle.....	33
3.1: A simple GRN.	41
3.2: The logic circuit diagram representing the GRN.....	42
3.3: A Bayesian network representation of the GRN including 6 genes.	46
3.4: A Dynamic Bayesian Network representation of the GRN.....	47
4.1: Framework of the proposed GRN model.....	72
4.2: Mapping of cell cycle phases to the timeline of the microarray experiment.	75
4.3: Target network of Yeast cell cycle TFs.....	78
4.4: Estimated network structure of yeast cell cycle TFs.	79

5.1: Expression profiles of co-regulated genes.....	90
5.2: Expression profiles of co-regulated genes groups.....	91
5.3: Framework of the proposed GRN model.	92
5.4: The known network structure of 13 cell cycle TFs.	97
5.5: The estimated network structure of 13 known cell cycle TFs.....	98
6.1: Genetic interaction network among the 13 TF of yeast cell cycle.	109
6.2: Potential regulatory association among the 13 known TFs of yeast cell cycle extracted from the YEASTRACT database.	110
6.3: Framework of the proposed GRN model..	112
6.4: Known network of the 13 known TFs in the yeast cell cycle.....	116
6.5: Derived network among the 13 known TFs in the yeast cell cycle.....	116
6.6: Estimated network among the 13 known TFs in the yeast cell cycle.....	117
7.1: Comparison of performance in terms of precision (P), recall (R) and F- measure among the four DBN-based GRN models through the analysis of 5 different simulated datasets including 20 genes..	129
7.2: Comparison of performance in terms of precision (P) , recall (R) and F- measure among the four DBN-based GRN models through the analysis of 5 different simulated datasets including 50 genes.	131
7.3: Comparison of performance in terms of precision (P), recall (R) and F- measure among the four DBN-based GRN models through the analysis of 5 different simulated datasets including 100 genes.	133
7.4: Target network among the 19 CCR genes extracted from the KEGG pathway database.	136
7.5: Network estimated by the PMDL-based GRN model proposed in Chaitankar et al. (2010).....	136
7.6: Network estimated by GRN _{Multi-sources} as discussed in chapter 6.	137

7.7: Comparison of the performance of 7 GRN models in terms of precision (P), recall (R) and F-measure..	138
7.8: Performance evaluation of GRN_{Phase} against synthetically generated random networks.	141
7.8: Performance evaluation of $GRN_{Co-expressed}$ against synthetically generated random networks.	142
7.10: Performance evaluation of $GRN_{Multi-sources}$ against synthetically generated random networks..	143
7.11: Comparison of the performance of 5 GRN models in terms of precision (P), recall (R) and F-measure..	146
8.1: Comparison of the computation time of five DBN-based GRN models.....	157
8.2: Comparison of the precision of five DBN-based GRN models.	159
8.3: Comparison of the recall of five DBN-based GRN models..	161
8.4: Computation time of $GRN_{Multi-sources}$ with varying network sizes including 100, 500 and 1000 genes.	165
8.5: Comparison of the precision and recall of $GRN_{Multi-sources}$ with varying network dimensions including 100, 500 and 1000 genes.	166

