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

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 Special thanks go to the research group of Linda Breeden Lab for giving me their unpublished data.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 Litchtenberg 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

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page