

**Integrative approaches to modelling and
knowledge discovery of molecular interactions in
bioinformatics**

by

**Vishal Jain
(2008)**

*A thesis submitted to Auckland University of Technology in partial fulfilment for
the degree of "Doctor of Philosophy – PhD"*

School of Computing and Mathematical Sciences

Supervisors:

Prof Nikola Kasabov, Dr. Lubica Benuskova and Dr. Paul Pang

KEDRI, Auckland University of Technology

Content

Attestation of authorship.....	iv
Acknowledgements.....	v
Abstract.....	viii
1. Introduction.....	1
1.1 Key question, Objective and Motivation.....	1
1.2 Issues to be addressed.....	4
1.3 Study overview, Rationale and Significance.....	6
1.4 Original contribution and research discoveries.....	9
1.5 Conclusion.....	17
2. Foundation and problems in molecular biology.....	19
2.1 Central dogma of molecular biology.....	19
2.2 Background on gene regulation processes and molecular interactions..	30
2.3 MicroRNAs as a gene regulatory element.....	36
2.4 Case studies throughout the thesis.....	44
2.5 Conclusion.....	48
3. An integrative ontology-based framework for modelling and knowledge discovery in bioinformatics.....	50
3.1 Introduction and problem specification	50
3.2 Computational intelligence methods within integrated framework for knowledge discovery.....	52
3.3 Concept of knowledge integration and information fusion.....	54
3.4 The ontology approach to integrate and reuse knowledge.....	56
3.5 Application of machine learning tools in integrated framework.....	61
3.6 Conclusion.....	62
4. Modelling and discovery of Gene regulatory networks (GRNs): An integrative Kalman filter (KF) Genetic Algorithm (GA) method.....	64
4.1 Introduction and problem specification.....	65
4.2 Existing methods for modelling of GRN.....	67
4.3 Proposed integrative approach: Kalman filter (KF) with Genetic Algorithm (GA).....	69

4.4 Case study: GRN modelling from Leukaemia gene expression time series microarray dataset.....	78
4.5 Results and discoveries.....	80
4.5.1 Biological validation of results.....	86
4.6 Conclusion.....	88
5. An integrative Least Angle Regression (LARS) and machine learning approach for GRN extraction	90
5.1 Introduction and problem specification.....	91
5.2 Proposed integrative approach: Least Angle Regression (LARS), Expectation Maximization (EM) with Kalman Filter (KF) and Evolving Fuzzy Neural Network (EFuNN).....	93
5.3 Case study on Yeast cell-cycle time series microarray dataset to infer GRNs.....	102
5.4 Results, discoveries and biological validation.....	103
5.5 Application of EM with KF and important findings.....	113
5.5.1 Biological interpretation of results.....	117
5.6 Application of EFuNN and important findings.....	120
5.7 Conclusion.....	125
6. Studying LTP related GRNs using quantum inspired evolutionary algorithm (QiEA) and clustering analysis.....	128
6.1 Introduction and problem specification.....	128
6.2 Case study: Mouse LTP time series microarray dataset analysis to infer GRNs.....	131
6.2.1 Method for gene Selection.....	133
6.2.2 Gene clustering and functional analysis.....	143
6.2.3 Application of QiEA to predict GRNs and gene knock-in mice experiments.....	150
6.3 Results, discoveries and biological validation.....	154
6.4 Conclusion and discussion	160
7. Computational methods to discover novel microRNAs using 2-D structures.....	168
7.1 Introduction and problem specification.....	169
7.2 Existing methods for microRNA classification.....	170
7.3 Proposed integrative method of Gabor Filter, BLAST and CLUSTALW.....	173
7.4 Case study on human microRNAs dataset.....	176
7.5 Results, discoveries and biological validation.....	184
7.6 Conclusion and discussion.....	191

8. Integrative Brain-Gene Ontology (BGO) and simulation system.....	193
8.1 Introduction.....	193
8.2 BGO: An overview, aims and goals.....	197
8.3 Implementation of brain-gene ontology system.....	200
8.4 Knowledge reuse, elicitation and discoveries with BGO.....	207
8.4.1 Biological validation and interpretation of results.....	214
8.5 Using BGO data for neuronal gene-protein sequence and clustering analysis.....	218
8.6 Facilitating education with BGO.....	227
8.7 Conclusion and System availability.....	229
9. Implications and Future Directions.....	232
9.1 Introduction.....	232
9.2 Implications and future directions of the suggested approach:	
9.2.1 Gene regulatory networks.....	234
9.2.2 MicroRNA regulations.....	235
9.2.3 Brain gene ontology.....	238
9.3 Ethical considerations.....	241
9.4 Potential applications of the developed methods and systems.....	242
9.5 Conclusion and Open discussion.....	250
References.....	253
Appendices.....	269
A. Kalman filter (KF).....	269
B. Evolutionary Computation and Genetic Algorithm (GA).....	271
C. EFuNN and ECF.....	274
D. Neucom and Siftware.....	281
E. GnetXP – description and user manual.....	285
F. Supplementary information for chapter 6.....	288
G. Snapshots from animations of BGO.....	296

I hereby declare that this submission is my own work carried under the guidance of my designated supervisors and that, to the best of my knowledge and belief, it contains no material previously published that was accepted for the award of any other degree or diploma of a university or other institution of higher learning, otherwise due acknowledgement is made in the references.

(VISHAL JAIN)

Acknowledgements

To the extent that I have accomplished anything in my life, I credit the encouragement, support, blessings and inspiration of my entire family. They all believe in my potential to achieve whatever goals I envision in my mind. My gratitude to them cannot be expressed with the confines of this thesis. This long journey is dedicated to them.

On the professional front, first I wish to express my sincere appreciation to my respected supervisor Prof. Nikola Kasabov for his invaluable guidance, encouragement and support during my doctoral program as well as the critical reviews and comments on my publications and this dissertation. I found him indeed an insight supervisor who has provided me the scientific vision, stimulating guidance and freedom of creativity to develop as an independent researcher. His original approach to scientific problems spurred me to redefine questions and to seek novel answers.

Next, I am deeply indebted to my secondary supervisor Dr. Lubica Benuskova. It is privilege to have such wonderful mentor and to enjoy their advices, friendship and research discussion during these years. I greatly appreciate your time and effort. I am also very thankful to my third supervisor Dr. Paul Shaoning Pang. It has been a real pleasure for me to work with him in the past few years. At the same time I am also thankful to all of my examiners.

Many other individuals have contributed in their unique ways towards this research. I am happy to meet and to work together with them in the past years. The work in this thesis would never have taken its current shape had it not been for the support, inspiration and advice that I received from a bunch of people. Therefore, I would like to extend my sincere appreciation to my advisory committee members Dr. Zeke Chan, Dr. Ilkka Havukkala, Dr. Dimiter Dimitrov and Dr. Igor Sidirov.

During my research career in Auckland I have been surrounded by an amazing group of talented, intelligent, supportive and thoughtful peoples. In this respect I also extend my thanks to Prof. Ajit Narayanan, Dr. Colleen Higgins, Associate Prof. Frances Joseph and Prof. Stephen MacDonell for their constructive professional suggestions and certain other helps during this coursework.

My study at Auckland University of Technology (AUT) in New Zealand has been a great experience and I would like to thank all of my colleagues. I would like to acknowledge my dear friends for their friendship, discussion and help to make the research life easier and interesting, in particular, Dr. Liang Goh, David Zhang, Simei Gomes Wysoski, Paulo Gottgroy, Dougal Greer, Maggie (Tian Min) Ma, Anju Verma, Richard Walton, Stefan Schliebs, Raphael (Yingjie) Hu, and Scott Heappey.

Post graduate office has also supported me by performing most of the final stage administrative tasks related to my PhD, so please accept my thanks. I also

wish to thank Joyce D'Mello and Peter Hwang for any support they have provided during my work at KEDRI.

Last but not least, I would also like to thank all the members of AUT Technology Park for providing me certain kind of official support and any other arrangements I always needed throughout my study.

This doctoral research was financially supported from the Foundation of Science Research and Technology (FRST), Knowledge Engineering and Discovery Research Institute (KEDRI), Auckland University of Technology (AUT), my sincere thank to these esteemed organizations.

Abstract

The core focus of this research lies in developing and using intelligent methods to solve biological problems and integrating the knowledge for understanding the complex gene regulatory phenomenon. We have developed an integrative framework and used it to: model molecular interactions from separate case studies on time-series gene expression microarray datasets, molecular sequences and structure data including the functional role of microRNAs; to extract knowledge; and to build reusable models for the central dogma theme. Knowledge was integrated with the use of ontology and it can be reused to facilitate new discoveries as demonstrated on one of our systems – the Brain Gene Ontology (BGO).

The central dogma theme states that proteins are produced from the DNA (gene) via an intermediate transcript called RNA. Later these proteins play the role of enzymes to perform the checkpoints as a gene expression control. Also, according to the recently emerged paradigm, sometimes genes do not code for proteins but results in small molecules of microRNAs which in turn controls the gene regulation. The idea is that such a very complicated molecular biology process (central dogma) results in production of a wide variety of data that can be used by computer scientists for modelling and to enable discoveries. We have suggested that this range of data should actually be taken into account for analysis to understand the concept of gene regulation instead of just taking one source of data and applying some standard methods to reveal facts in the system biology. The problem is very complex and, currently, computational algorithms have not been really successful because either existing methods have

certain problems or the proven results were obtained for only one domain of the central dogma of molecular biology, so there has always been a lack of knowledge integration. Proper maintenance of diverse sources of data, structures and, in particular, their adaptation to new knowledge is one of the most challenging problems and one of the crucial tasks towards the knowledge integration vision is the efficient encoding of human knowledge in ontologies.

More specifically this work has contributed towards the development of novel computational and information science methods and we have promoted the vision of knowledge integration by developing brain gene ontology (BGO) system. With the integrative use of several bioinformatics methods, this research has indeed resulted in modelling of such knowledge that has not been revealed in system biology so far. There are many discoveries made during my study and some of the findings are briefly mentioned as follows: (1) in relation to leukaemia disease we have discovered a new gene “TCF-1” that interacts with the “telomerase” gene. (2) With respect to yeast cell cycle analysis, we hypothesize that exoglucanase gene “exg1” is now implicated to be tied with “MCB cluster regulation” and a “mannosidase” with “histone linked mannoses”. A new quantitative prediction is that the time delay of the interaction between two genes seems to be approximately 30 minutes, or 0.17 cell cycles. Next, Cdc22, Suc22 and Mrc1 genes were discovered that interacts with each other as the potential candidates in controlling the Ribonucleotide reductase (RNR) activity. (3) Upon studying the phenomenon of Long Term Potentiation (LTP) it was found that the transcription factors, responsible for regulation of gene expression, begin to be elevated as soon as 30 min after induction of LTP, and remain elevated up to 2

hours. (4) Human microRNA data investigation resulted in the successful identification of two miRNA families i.e. let-7 and mir-30. (5) When we analysed the CNS cancer data, a set of 10 genes (HMG-I(Y), NBL1, UBPY, Dynein, APC, TARBP2, hPGT, LTC4S, NTRK3, and Gps2) was found to give 85% correct prediction on drug response. (6) Upon studying the AMPA, GABRA and NMDA receptors we hypothesize that phenylalanine (F at position 269) and leucine (L at position 353) in these receptors play the role of a binding centre for their interaction with several other genes/proteins such as c-jun, mGluR3, Jerky, BDNF, FGF-2, IGF-1, GALR1, NOS and S100beta.

All the developed methods that we have used to discover above mentioned findings are very generic and can be easily applied on any dataset with some constraints. We believe that this research has established the significant fact that integrative use of various computational intelligence methods is critical to reveal new aspects of the problem and finally knowledge integration is also a must. During this coursework, I have significantly published this research in reputed international journals, presented results in several conferences and also produced book chapters.