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

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 Gottgtroy, 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.

1. Introduction

This chapter presents the introduction to this thesis and describes some of the molecular biology key questions that we undertook in this study and explains our approach to address the important issues in this area of research. Later, we talk about the overview and the importance of this project. It is followed by a section on our original research contribution and in the same section we also describe the complete structure in which this thesis is organized. We conclude the chapter by revisiting some of the critical aspects of the problem.

1.1 Key question, Objective and Motivation

Core question of our research is how to integrate knowledge for understanding the complex gene regulation phenomenon. As an answer 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 (refer figure 1.1). 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).

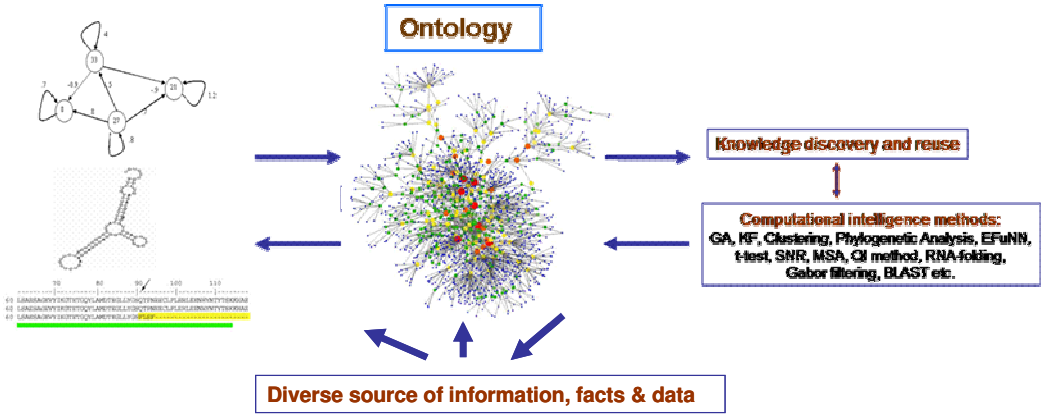


Figure 1.1: Integrative approach for information fusion and knowledge discovery in bioinformatics

The broader view of this doctoral research is the development and the use of novel computational and information science methods and systems that facilitate the understanding of the complex molecular interactions through which a gene is regulated in a molecular cell or in an organism. 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. If specifically we want to investigate this problem, there is a need to examine each regulatory stage of a cell and later integration is required to make the possible reuse of this knowledge. Through this research we have attempted to examine the knowledge at different regulatory steps, which has not only

resulted in the development of novel bioinformatics methods, but also in the modelling of such knowledge that has not been revealed in system biology so far. For such a problem, we have considered different case studies and have gained knowledge expertise on different domains. The overall investigation was done by considering data at different levels, such as the molecular sequence level, the macromolecular structure level, microarray gene expression level. The integration of knowledge was implemented through the use of ontology and demonstrated on the case study of brain gene data in brain gene ontology – "BGO" system. The developed system will keep evolving and one can always input the latest database knowledge and experimental results, thus simultaneously knowledge can be accessed and reused. This is an emerging area where researchers are starting to combine the knowledge from different but interlinked domains. More specifically this research work has contributed towards the development of a novel computational and information science method, knowledge discovery and development of a brain-gene ontology system. We believe that our research has established the significant fact that integrative use of various computational intelligence methods is critical to reveal new aspects and finally knowledge integration is also a must. The details of the different case studies, data types and applied methodologies are discussed later in the respective chapters of this thesis. During this doctoral research, we have presented our work at several conferences, produced book chapters and also published papers in international journals.

1.2 Issues to be addressed

For understanding molecular interactions (i.e. gene regulation) traditional methods have not been very successful (see detailed explanation later in the literature review of this thesis). As an example, for the problem of inferring gene regulatory networks (GRNs) most of the current methods demand large amounts of training samples and use the common remedy of increasing the number of samples within a time interval by using interpolation methods. However, interpolated points can be erroneous especially when the observed samples are sparse and noisy. Moreover, there has always been a lack of knowledge integration and the problem has not been addressed from the perspective of each regulatory step. Therefore, considering these facts, our goal is to develop and use novel intelligent methods to solve this problem in system biology (see respective chapters of this thesis in which we have discussed different domains of this study). The application of computational methods is a bioinformatics attempt that allows biologists to identify important genes and proteins related to a particular disease and to build information models, a concept generally known as molecular modelling. In terms of significance, modelling of interacting networks enables scientists to model complex dynamic processes involving large numbers of interacting variables and to extract inherent variable relationship networks. Various kinds of data

analyses using methods of computational intelligence have provided significant results for biomedical researchers [Kasabov 2005].

As we know that in biological systems everything is interconnected and ostensibly unrelated fields are related — the separation of biology into smaller subject areas is artificial. Therefore our aim is not only to represent the integrated knowledge but also to link the molecular information and findings with respective phenotypes, such as disease and to answer questions such as "Which genes are related to the occurrence of a particular disease like epilepsy and when?" In this respect, ontology encloses the development of a semantic repository of systematically ordered relevant concepts in molecular biology. Thus we have used ontology to provide the conceptual framework and the factual knowledge that is necessary to deal critically with the rapidly changing science of biology. In our case study on brain genes data with the BGO system, one can represent uncertain knowledge and evaluate the quality of the knowledge discovered through different referenced measures, such as the quality of the annotation, number of related publications and so on. For knowledge elicitation and inference, a system may require specific operations like intelligent querying using a specialised interface. This area of research is quite demanding and we hope that our obtained results are of great scientific value for biomedical researchers and the integrated ontological framework will

definitely allow other researchers to better understand the biological processes and further explore research possibilities.

1.3 Study overview, Rationale and Significance

We have studied the gene regulation area (see above section on issues to be addressed) from a different viewpoint by integrating multiple aspects and have developed and applied novel computational intelligence methods to infer gene regulatory networks and classified the microRNAs. More importantly, the knowledge, facts and the data (related to genes, diseases and interactions) which reside in different databases has been brought together as a part of a knowledge integration process. Further, from a broader point of view to understand the significance of this research work, let us imagine a cell in which, suddenly, some specific interactions between gene-gene, gene-protein and protein-protein would change. Such changes may result in tissue disintegration, because these specific interactions are involved in almost any physiological process and controls the gene regulation. Therefore, predicting the interactions between genes and proteins involved in common cellular functions is a way to get a broader view of how they work cooperatively in a cell.

It has been said that understanding the gene regulation at the system level is one of the most interesting and challenging problems in biology [De Jong 2002]. It is not surprising that organismal development involves complex regulatory mechanisms that target every aspect of gene expression [De Jong 2002]. Major cell functions are dependent on interactions between DNA, RNA and proteins. A single gene interacts with many other molecules in the cell, inhibiting or promoting, directly or indirectly, the expression of some of them at the same time. Gene interaction may control whether and how much that gene will produce RNA with the help of a group of important proteins known as transcription factors. All enzymes that take part in various catalytic processes and other signalling pathways are proteins and always synthesized from genes. But on the other hand, several small molecules in the genome known as microRNAs are actually produced by the transcription of small genes that are not translated to produce proteins. They function in controlling the process of gene regulation usually by binding to specific messenger RNAs to block the translation or decoding process [Carter Richard et al. 2001]. It is hypothesized that the secondary structure of RNA – the 2-D geometrical layout obtained by folding its sequence into a stable structure – has significant implication in its functionality. It indicates that functional classification of micro-RNA is another step towards understanding gene regulation through molecular interactions [Carter Richard et al. 2001]. Therefore as we see, the whole of this “central dogma” process (from DNA to protein production) is very interactive and

complicated and in a cell the interplay of all the interactions between DNA, RNA, proteins leads to genetic regulatory networks (GRN) [De Jong 2002 and Carter Richard et al. 2001]. Thus it is not realistic to understand the gene regulation by considering the interaction only at any one level, i.e. either gene-gene or gene-protein or protein-protein.

We have learnt by now that biological processes are not realized by a single molecule, but rather by complex interactions of proteins with their environment, including genes, ions, lipids, membranes and, of course, other proteins [Endy D 2001 and Carter Richard et al. 2001]. The expression of a gene by which dynamic biological information becomes available in the cell is actually regulated at several stages and a stage of regulation may vary (detail of this topic have been discussed by us in the literature review i.e. chapter 2 of this thesis) [Arnone et al. 1997, Hofestadt 1995 and De Jong 2002]. To address this issue one needs to study the complete cell system and specifically approach the problem by looking at each regulatory step and at their interaction. The major objectives in understanding gene regulation involves discovering the architecture, dynamics, and function of regulatory networks, making useful computational models of them, learning how to design and adapt them; and finally coming out with some knowledge discovery of biological mechanisms [De Jong 2002]. So in conclusion, we can say that in order to understand these complex molecular interactions we need to

understand and analyse the concept from the gene sequence level in the DNA to gene and protein expression level. A detailed study of an organism, such as the analysis of gene expression levels to know which crucial genes are expressed, when and where in the organism they are expressed, and to which extent, can help the scientific community to better understand the functioning of the organism [De Jong 2002]. Simultaneously, the sequence and structure data analysis to study and classify microRNAs can be an excellent input towards understanding gene regulatory networks. At last, we need a system that can combine or fuse the vast amount of experimental knowledge about genes, proteins and so on, and in turn researchers and students can make the use of it to reveal new facts or hypotheses that have not been demonstrated earlier. This has been discussed in thesis chapter 3 where we deal with the issue of data integration, information fusion and reusing knowledge to enable discoveries with the use of ontology.

1.4 Original contribution and research discoveries

I have used the word “we” throughout the thesis as this research has been done under the supervision of my designated mentors. Further, on several occasions other talented scientists inspired me with novel ideas and some of my publications are multiple authored with my colleagues who have participated in certain research projects partially. In this section we will discuss some of our original contribution to the science area through the research that I

have undertaken for this doctoral degree. After learning that the problem of understanding gene regulation, i.e. molecular interactions – gene regulatory networks (GRN) is very complicated, we strongly believe that my thesis has taken an original approach to demonstrate that problem of GRN must be studied at different levels using a range of datasets (gene expression, sequence and structure data etc.) including the study of microRNAs. We further emphasize that application of several novel computational intelligence (possibly hybrid) methods may reveal some new aspects of the problem (possibly new knowledge) and to infer the GRNs it is more suitable to adopt such integrative approach as ours that we have suggested by means of proposed integrative framework in chapter 3. Also, we have promoted the vision of knowledge integration by making the use of ontologies in current biological knowledge management. We state that machine learning tools may be applied to elicit knowledge that can be interpreted by domain experts and may be reused for various kinds of research discoveries. Below with the explanation of the thesis structure, we will summarize few of the important discoveries that we have made and the journals into which these findings were published. In general, my doctoral research has resulted in seven journal publications and over ten conference papers/book chapters. In this respect, the detailed list of publications may be seen in the section of references.

Chapter 2 is devoted to a detailed literature survey and provides the reader with a foundation of problems in molecular biology. I am the sole person responsible for doing this literature survey and defining the research questions and other problem in molecular biology. The next, chapter 3 explains the proposed integrated framework that we have later used to address the problem of understanding molecular interactions. It is the original idea that I and my primary supervisor (Prof. Kasabov) thought and proposed in the first instance. In [Kasabov, Jain and Benuskova 2008] we explain more on the approach taken and use the developed integrated system to make novel discoveries on a cancer dataset. In this publication, I have developed the central nervous system (CNS) cancer data ontology (within BGO) and also used it for all results interpretation and validation. My primary supervisor used the ECOS and wrote that section within publication. I wrote integrative framework related parts in the paper also acted as a contact person and managed the whole publication process. Dr. Benuskova supervised me during the framework development and was involved in the proof reading of paper. In summary, a set of 10 genes was found to give 85% correct prediction on drug response of a CNS cancer tumour in children (72% for the class of non-responding, and 92% for the responding to drugs class), i.e.: HMG-I(Y), NBL1, UBPY, Dynein, APC, TARBP2, hPGT, LTC4S, NTRK3, and Gps2. More detailed results are discussed in chapter 8 in which we also describe through the use of BGO system that these genes were found to be involved in

other brain functions and diseases, and are dynamically interacting with other brain-related genes.

Chapter 4 presents our scientific approach of integrating Genetic algorithm (GA) with Kalman filter (KF) for extracting GRN from gene expression leukaemia data. It reports very interesting results in this respect. For example, our method was able to identify the genes that co-regulate telomerase activity (crucial in leukaemia). An outstanding gene identified was TCF-1. Such important results were published in [Kasabov, Chan, Jain et al. 2004] and as a chapter in [Kasabov, Chan, Jain et al. 2005]. In these publications, Prof. Kasabov suggested the idea of using GA with KF and Dr. Chan was involved in programming part. I have contributed in gene selection, interpretation of GRNs and writing these parts in the paper. I am also the GnetXP software manager that was developed based on the suggested method of GA with KF. Data was provided by Dr. Sidirov and Dr. Dimitrov (National Cancer Institute, USA) and they also tested one of our finding i.e. interaction of TCF-1 gene with the telomerase gene in their laboratory. Further, I acted as a contact person and managed the whole publication process.

In chapter 5 we explain and implement the integrative use of several methods: Least Angle Regression (LARS), Expectation Maximization (EM) with Kalman Filter (KF), and Evolving Fuzzy Neural Network (EFuNN). We

took the widely used benchmark gene expression data of the yeast cell cycle and report our findings in [Chan, Havukkala and Jain et al. 2008]. In this publication, Dr. Chan did some LARS related programming and Dr. Havukkala supervised LARS related experimentation. I was involved in interpretation of GRNs, application of EM-KF, suggesting new hypothesis and writing these parts of the paper. I worked with another doctoral student for using EFuNN. Prof. Kasabov suggested the ideas and supervised the project as a leader. Also, I managed the whole publication process as a corresponding author. Each method revealed some new aspects of the problem and it is agreed that to infer the GRN and to understand the processes behind gene regulation it is more suitable to adopt such integrative approach through which some new knowledge is discovered. For example, using LARS we hypothesize – first, an exoglucanase gene exg1 is now implicated to be tied with MCB cluster regulation and second, 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. Using the method of EM with KF, 25 cell-cycle regulated key genes were successfully clustered into three functionally co-regulated groups. We have also identified two genes namely Cdc22 and Suc22 that indeed interact with each other and are the potential candidates as a control in Ribonucleotide reductase (RNR) activity. Based on the EFuNN results and integrating knowledge from the EM-KF method, we hypothesize that interaction between Suc22, Cdc22 and

Mrc1 may be mediated by two other genes namely Cds1 and Spd1. All these discoveries could not have been made with the use of a single method, which demonstrates the power of our integrative approach.

In chapter 6, we have studied Long Term Potentiation (LTP) related GRNs in the brain using Quantum Inspired Algorithm (QiEA) and clustering approach. In this respect we state that functional analysis of gene functions and temporal patterns of their expression have revealed that 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. There are many more important findings resulted from this research also discussed in chapter 6. QiEA was first developed in [Defoin-Platel, Schliebs and Kasabov 2007] and I have been mainly responsible for gene selection procedure, promoter prediction, clustering (genes, proteins and promoters) and result interpretation. Dr. Benuskova was responsible for suggesting all experiments and managing the publication which is now submitted to an international journal.

In chapter 7 we have studied human microRNAs and have classified them using the novel approach based on Gabor filtering, RNA Vienna package, CLUSTALW and BLAST etc. Two human miRNA families (let-7 and mir-30) were successfully identified using our suggested approach. It is hypothesized

that using visual information from bitmap images of 2D structures of microRNA precursors, one can extract potentially novel and useful information that can be used for discovery and classification of related molecules. We published these important findings in [Havukkala, Pang, Jain and Kasabov 2005]. In this publication, Dr. Pang did the programming and I have done the miRNA feature selection (identification), data selection, applied RNA Vienna package, CLUSTALW and BLAST. Prof. Kasabov was the project leader and Dr. Havukkala supervised all the experiments, helped me in interpreting the results and managed the publication process.

In continuation, chapter 8 of this thesis is devoted to knowledge integration where we propose integrative models and the Brain Gene Ontology (BGO) system. We explain some of the useful results that were obtained by analysing the multiple types of BGO data using computational intelligence (CI) tools, for example, Multiple Alignment Approach (MSA). The most interesting investigation was the consistent conservation of phenylalanine (F at position 269) and leucine (L at position 353) in all 20 proteins (of AMPA, GABRA and NMDA receptors) taken into account with no mutations. We expect these residues to play some role as a binding centre for interaction of these proteins with several other genes/proteins such as c-jun, mGluR3, Jerky, BDNF, FGF-2, IGF-1, GALR1, NOS and S100beta that are also believed to have a regulatory effect upon these receptors. Based on such observations we

assume that the expression of these individual subunits should be coordinated within one gene group. In addition, these regions can be the basis for mutual interactions. Mutual interactions between subunits of different receptors have been recently confirmed experimentally. Some of these important discoveries were published in [Benuskova, Jain et al. 2006]. I was responsible for doing all experiments, suggesting novel hypothesis and writing this part in the paper. Dr. Benuskova and Prof. Kasabov also did some other experiments on same data using Computational Neurogenetic Modelling (CNGM) approach and that was published within the same paper. As a whole, the BGO system was officially presented in [Kasabov, Jain et al. 2007] and during its development progression, different results were published in [Kasabov, Jain et al. 2007 and 2006]. Prof. Kasabov has been the project leader and I have contributed the main development of BGO in terms of structure, data and information, running experiments and using it to teach the postgraduate bioinformatics paper at AUT. The data export plugins were developed by another doctoral student and the animations were developed by one of our master's student. Dr. Benuskova has supervised all the experimentation and written papers with me. Further, I have been responsible for managing the system and liaising with publishers as a corresponding author.

I conclude the thesis with chapter 9 by discussing some implications, potential applications and future directions. These were identified by me in

consultation with my supervisors and I have expressed our opinions within the chapter. As this research is a “never ending topic”, I have also put a section for an open discussion. At last the appendices are presented on some methods and systems, such as GA, KF, EFuNN, ECF, Neucom, Siftware, GnetXP, BGO etc.

1.5 Conclusion

As discussed above, we will see in the next chapter of this thesis that molecular biology central dogma is a very complicated process. Direct or indirect, positive and negative molecular interactions are known as a “gene regulatory networks” and it is one of the very important problems to study in molecular biology. MicroRNAs also play a valuable role in gene regulation and have recently come into the highlight. Earlier studies in this area have some limitations in the sense that either the applied methods were not fully appropriate, or only one domain of the problem was considered. We have studied this problem from a different viewpoint by integrating multiple aspects and have developed and applied novel computational intelligence methods to infer gene regulatory networks and classified the microRNAs. More importantly, last but not least, the knowledge, facts and the data (related to genes, diseases and interactions) which reside in different databases need to be brought together as a part of a knowledge integration process. We have done so with the use of ontology and have demonstrated it by taking a case

study on brain gene data and developing a system called Brain Gene Ontology (BGO). Its evolving nature can be used to further store any kind of knowledge related to brain. Biomedical researchers may share this knowledge to derive new hypotheses and it can also be used to facilitate research path education. Overall, in this chapter, we have discussed the structure of this thesis and explicitly stated our original contribution to the science area. The research discoveries and methodologies have been published in journals, conferences and books.

2. Foundation and problems in molecular biology

This chapter presents the background on the central dogma theme, gene regulation process and molecular interactions in a cell. Next, we discuss the microRNAs, their biogenesis and emphasize on their role in the process of gene regulation. This section is followed by a discussion on the case studies that we undertake in this doctoral research. Finally, we conclude the literature survery by summarizing the facts of molecular biology.

2.1 Central dogma of molecular biology

In this section, we will highlight the main biological processes that are directly related with my overall research by providing the background focus on the dogma theme of molecular biology. It is appropriate to mention here that theories of molecular biology central dogma are well established scientific facts and there has been contribution of various prominent researchers for several years to agree upon such hypotheses. We will not discuss this well known concept from a biologist perspective but here our intention is to provide the knowledge recap on central dogma to a computer scientist or to a bioinformatician. The central dogma states that information is always transferred from DNA to RNA to protein. When DNA prepares its duplicates this phenomenon is known as replication. Process by which DNA is

transformed into RNA is called as transcription. Mature RNA transcripts form the proteins through the process of translation. In general each of these three processes (i.e. replication, transcription and translation) involves the participation of various types of proteins usually called enzymes (remember all enzymes are proteins but not all the proteins are enzymes) at different steps. Readers are requested to refer to the figure 2.1 to understand the much generalized meaning of dogma theme. Below we will discuss each of these processes one by one.

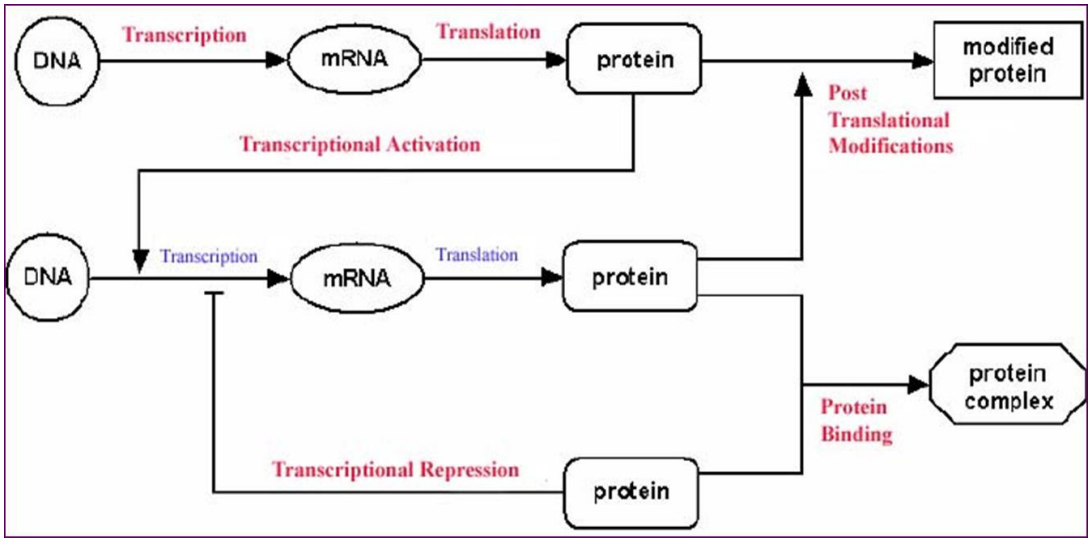


Figure 2.1: The Dogma theme in molecular biology

DNA and its replication

DNA is a double stranded helical structured molecule as described in [Watson and Crick 1953]. It is our genetic material and has been regarded as

the blue print of life. It is the basis for modern forensic science. Two strands of DNA run *antiparallel* such that one strand runs 5' -> 3' while the other one runs 3' -> 5'. Each strand has polarity, such that the 5'-phospho group of the first nucleotide begins the strand and the 3'-hydroxyl group of the final nucleotide ends the strand; accordingly, we say that this strand runs 5' to 3' ("*Five prime to three prime*"). At each nucleotide residue along the double-stranded DNA molecule, the nucleotides are complementary. That is, **A** (Adenine) forms two hydrogen-bonds with **T** (Thymine); **C** (Cytosine) forms three hydrogen bonds with **G** (Guanine). Refer to the chemical structures of A, T, C and G in the figure 2.2.

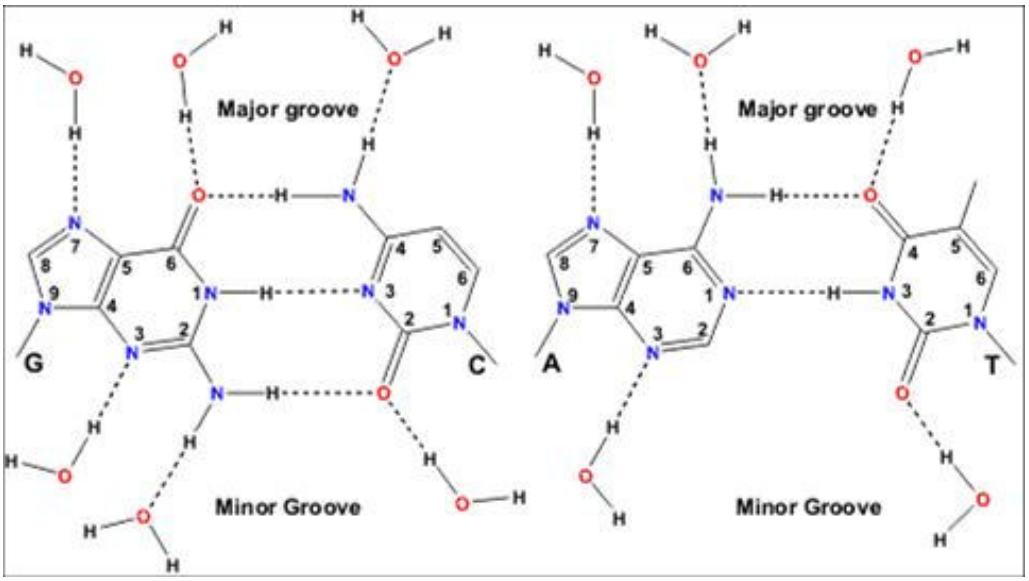


Figure 2.2: Chemical structure of DNA bases. Figure is taken from Encyclopaedia Britannica, http://www.encyclopedia.com/

DNA replication (DNA to DNA, refer to figure 2.3) begins with a partial unwinding of the double helix at an area known as the replication fork. Enzymes such as topoisomerase and DNA helicase accomplish helical unwinding [Lewin 1999]. However both of the strands have the potential to act as the template but only one of the two DNA strands of the DNA double helix carries the biological information, and is called the template strand. As the two DNA strands unzip and the bases are exposed, the enzyme DNA polymerase moves into position at the point where synthesis will begin. There are certain other enzymes also involved in the complete synthesis of DNA but as we stated earlier the goal of this thesis chapter is not to discuss the minute details of whole process here.

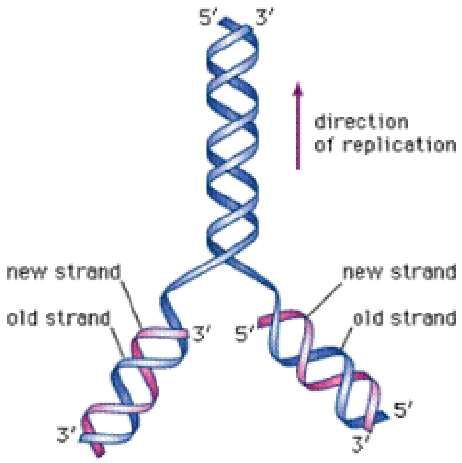


Figure 2.3: Simplified model of DNA replication according to Watson and Crick. Figure is taken from Encyclopaedia Britannica, http://www.encyclopedia.com/

RNA transcription and the role of promoters

RNA transcription (DNA to RNA) is a polymerization reaction in which individual ribonucleotides link sequentially into a chain. Not all open reading frames (ORFs) are transcribed into genes. The transcription depends on regulatory regions that control the transcription rate [Dow 1996]. The expression of a gene is regulated by a segment of DNA upstream of the coding sequence called the promoter, which binds to RNA polymerase and associated transcription factor proteins and initiates synthesis of an RNA molecule. Transcription regulatory proteins (initiation factors) affect the binding of the RNA polymerase (RNAP - the enzyme that synthesizes the RNA from the coding region of the gene) to the promoter, that is, the binding sites of RNAP located upstream of genes. A promoter is a regulatory region of DNA located upstream (towards the 5' region) of a gene providing a control point for regulated gene transcription [Lewin 1999].

In prokaryotes (lower organisms), the promoter is recognized by RNA polymerase and an associated sigma factor, which in turn are brought to the promoter DNA by an activator protein binding to its own DNA sequence nearby. In eukaryotes, the process is more complicated, and at least seven different factors are necessary for the transcription of an RNA polymerase II promoter [Lewin 1999]. Promoters represent critical elements that can work in

concert with other regulatory regions (enhancers, silencers, boundary elements/insulators) to direct the level of transcription of a given gene. It is worth noting here that eukaryotic (higher organisms) promoters are extremely diverse and are difficult to characterize. They typically lie upstream of the gene and can have regulatory elements several kilobases away from the transcriptional start site. Many eukaryotic promoters, but by no means all, contain a TATA box (sequence TATAAA), which in turn binds a TATA binding protein which assists in the formation of the RNA polymerase transcriptional complex [Levine M and Tjian R 2003]. The TATA box typically lies very close to the transcriptional start site (often within 50 bases). Eukaryotic promoter regulatory sequences typically bind proteins called transcription factors which are involved in the formation of the transcriptional complex. As promoters are typically immediately adjacent to the gene in question, positions in the promoter are designated relative to the transcriptional start site (TSS), where transcription of RNA begins for a particular gene (i.e., positions upstream are negative numbers counting back from -1, for example -100 is a position 100 base pairs upstream).

Once the transcription has been started, the first base becomes copied into an RNA Transcript. As indicated above, for mRNA genes in particular, a number of regulatory proteins participate in initiation by indicating which of the many protein-encoding genes are to be copied. The template is not always the

same nucleotide chain of the DNA double helix; different genes may have their template chains on either side of the DNA helix. By means of various mechanisms, regulatory proteins can alter any step in the transition involving the binding and the subsequent steps leading to the initiation of a stable elongation complex (Collado-Vides and Hofestadt 2002). Once the first base is added, elongation begins and RNA nucleotides add sequentially until the polymerase reaches the end of the template. In eukaryotic mRNA genes, termination may be coupled to processing reactions rather than occurring in response to specific signal sequences in the DNA. Termination factors contribute to separation of RNA polymerase and elongating factors from the template and release of the transcript in prokaryotes and possibly also in eukaryotes. These primary transcripts must undergo processing events such as capping, tailing, splicing (mainly in eukaryotes) and transport to the cytoplasm in order to produce mature, functional RNA molecules. The coding sequence of a gene is split into a series of segments called exons that are separated by non-coding sequences called introns (refer to the figure 2.4), which usually account for most of the gene sequence. The number and sizes of the introns vary between genes. Introns are removed from RNA transcripts by a process called splicing prior to protein synthesis. It is worth mentioning here that introns are not usually present in prokaryotes [Dow 1996 and Lewin 1999].

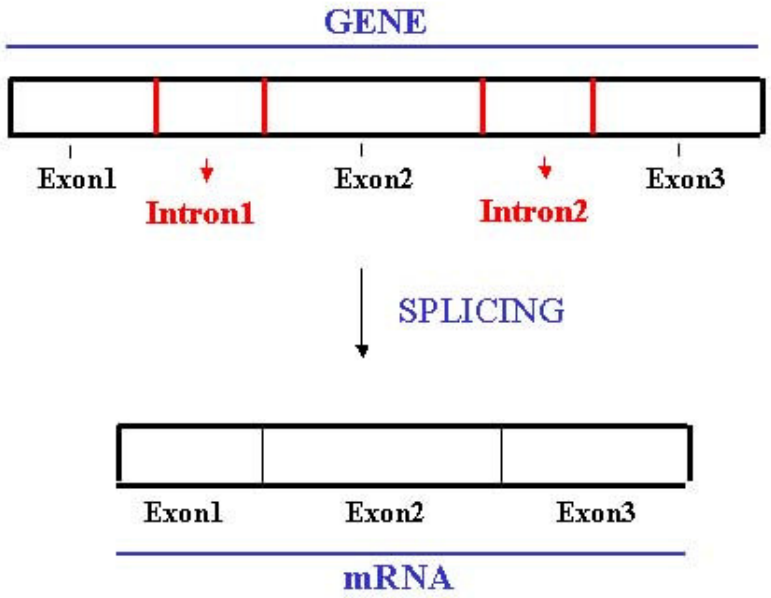


Figure 2.4: Brief description about the splicing of intervening sequences in eukaryotes

Protein translation and genetic code

Translation (RNA to Protein) involves taking the message that's in the messenger RNA and in a sense decoding the message from the language of nucleic acids to the language of proteins or polypeptides. The genetic code (refer to the figure 2.5) is the set of rules by which information encoded in genetic material (DNA or RNA sequences) is translated into the proteins (amino acid sequences) by living cells. Specifically, the code defines a mapping between tri-nucleotide sequences called codons and amino acids; every triplet of nucleotides in a nucleic acid sequence specifies a single amino acid. The genetic code consists of 64 triplets of nucleotides. With three exceptions, each codon encodes for one of the 20 amino acids (refer to their

chemical structure in the figure 2.6) used in the synthesis of proteins [Dow 1996 and Lewin 1999]. That produces some redundancy in the code: most of the amino acids being encoded by more than one codon. The genetic code is almost universal. It is known as "universal", because it is used by all known organisms as a code for DNA, mRNA, and tRNA. The universality of the genetic code encompasses animals (including humans), plants, fungi, archaea, bacteria, and viruses. However, all rules have their exceptions, and such is the case with the genetic code; small variations in the code exist in mitochondria and certain microbes. Nonetheless, it should be emphasized that these variances represent only a small fraction of known cases, and that the genetic code applies quite broadly, certainly to the all known nuclear genes.

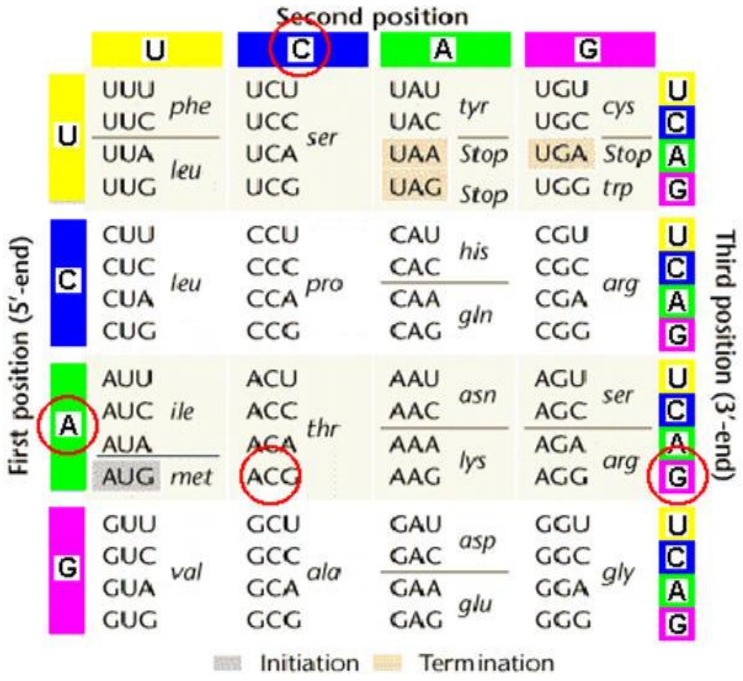


Figure 2.5: The genetic code. Figure is taken from Encyclopaedia Britannica, http://www.encyclopedia.com/

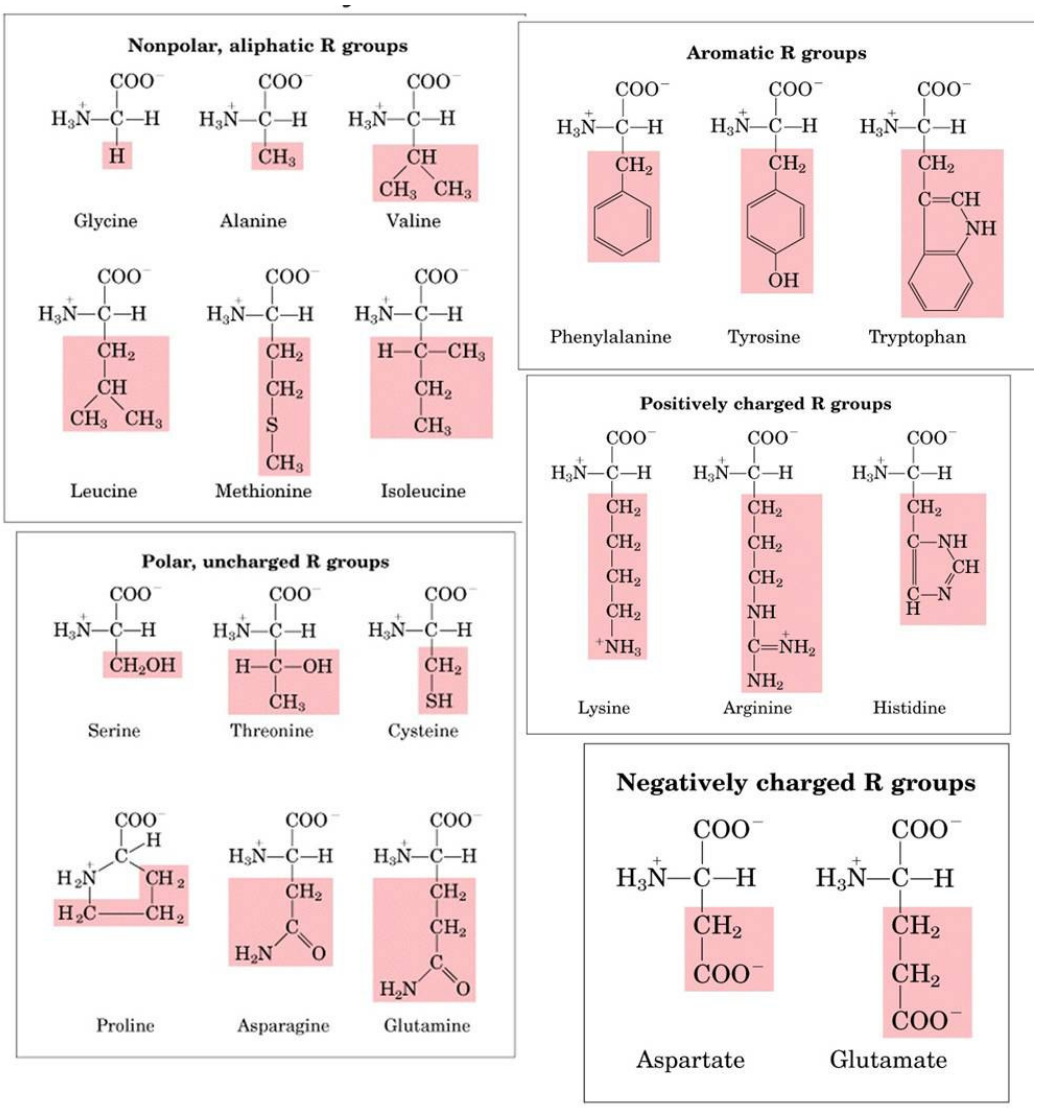


Figure 2.6: Chemical structures of naturally occurring amino acids. Figure is taken from Encyclopaedia Britannica, http://www.encyclopedia.com/

Translation of messenger RNAs occurs on ribosomes, which are specific complexes of RNAs and proteins, formed into large and small subunits [Dow 1996 and Lewin 1999]. The ribosome binds to the mRNA at the start codon (AUG - codes for methionine) that is recognized only by the initiator tRNA. Unlike stop codons, the start codon alone is not sufficient to begin the process,

therefore nearby sequences and initiation factors are also required to start translation. The ribosome proceeds to the elongation phase of protein synthesis [Kahn 1995 and Mann 1999]. During this stage, complexes, composed of an amino acid linked to tRNA, sequentially bind to the appropriate codon in mRNA by forming complementary base pairs with the tRNA anticodon. The ribosome moves from codon to codon along the mRNA. Amino acids are added one by one, translated into polypeptide sequences dictated by DNA and represented by mRNA. Protein synthesis uses not free amino acids but amino acids covalently joined to specific tRNAs, the aminoacyl-tRNAs. There are three termination or stop codons namely: UAG, UGA and UAA. At the end, a release factor binds to the stop codon, terminating translation and releasing the nascent polypeptide from the ribosome [Dow 1996 and Lewin 1999]. The key difference between prokaryotes and eukaryotes is that eukaryotes have a nucleus and prokaryotes do not. Other differences are in the complexity of the transcription complex, mRNA splicing, and the role of chromatin [Pandey and Mann 2000]. So the message here is eukaryotic genetic and molecular mechanisms are even more complicated to understand.

In the next section of this chapter, we will provide background on how the genes are regulated in a cell and the importance of molecular interactions to determine the fate of any cell.

2.2 Background on gene regulation processes and molecular interactions

Understanding the gene regulation at the system level is one of the most interesting and challenging problems in biology. As we have noticed, major cell functions are dependent on interactions between DNA, RNA and proteins. The interplay of interactions between DNA, RNA and proteins leads to genetic regulatory networks (GRN) and in turn controls the gene regulation. Underlying cause of evolution is mutation which in turns affect gene regulatory network through change in protein sequence which may give rise to alterations in function [Arnone and Davidson 1997].

Gene regulatory network are systems controlling the fundamental mechanisms that govern biological systems [De Jong 2002]. A single gene interacts with many other genes in the cell, inhibiting or promoting directly or indirectly, the expression of some of them at the same time. Gene interaction may control whether and how vigorously that gene will produce RNA with the help of a group of important proteins known as transcription factors. When these active transcription factors associate with the target gene sequence (DNA bases), they can function to specifically suppress or activate synthesis of the corresponding RNA [Lander 1996]. As we have described in the section of molecular biology central dogma, each RNA transcript then functions as the

template for synthesis of a specific protein. Thus the gene, transcription factor and other proteins may interact in a manner that is very important for determination of cell function. Gene regulatory networks govern which genes are expressed in a cell at any given time, how much product is made from each one, and the cell's responses to diverse environmental cues and intracellular signals. Much less is known about the functioning of the regulatory systems of which the individual genes and interaction form a part [Collado-Vides 1989, Fields 1999, Loomis 1995, O P B 1997 and Thieffry 1999]. Transcription factors provide a feedback pathway by which genes can regulate one another's expression as mRNA and then as protein [Bolouri 2001]. An example of a simple genetic regulatory network is represented below in the Figure 2.7 modified from [Brownstein 1998 and Kolchanov 2003].

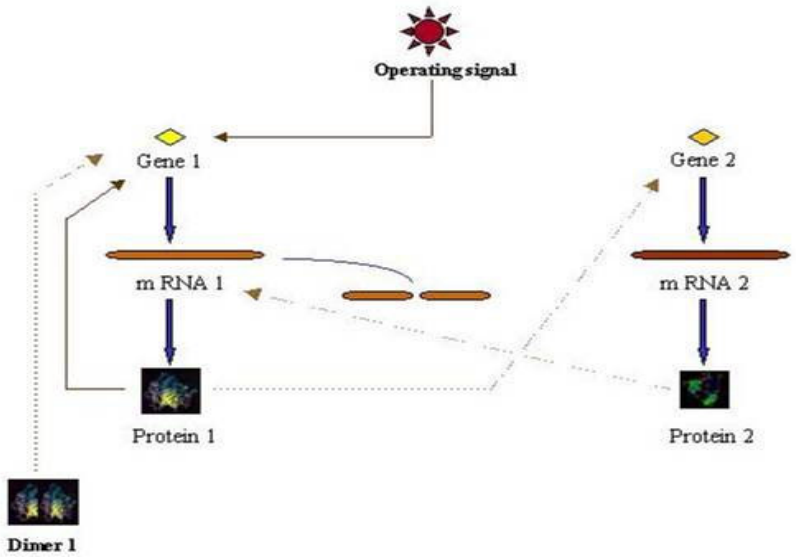


Figure 2.7: Formal description of a gene regulatory network to illustrate biological processes to be considered for computational modelling. Figure is modified from [Brownstein 1998 and Kolchanov 2003].

In the above example (figure 2.7), the system includes two genes, gene1 and gene2, coding for individual mRNAs, mRNA1 and mRNA2, wherefrom proteins protein1 and protein2 are translated. Monomer of protein1 forms dimer1. Protein2 is a specific factor responsible for degrading mRNA1; protein1 inhibits the transcription of gene2 and simultaneously activates transcription of its own gene1; while dimer1 of protein1 inhibits transcription from its own gene1. Generally gene1 is inactive and its primary activation depends on the outside operating signal. Gene2 in norm synthesizes mRNA2 with a certain nonzero activity. Both mRNA and proteins have limited life spans. Direct and indirect feedbacks typically are important. More realistic networks often feature multiple tiers of regulation, with first-tier gene products regulating expression of another group of genes, and so on.

Data in molecular biology for understanding molecular interactions

All the cells in an organism contain the same genetic material (except few exceptions). This implies that in order to understand these complex molecular processes one needs to understand the concept from the sequence level to gene expression level. Understanding sequence data, and the ability to utilize this hidden knowledge, creates a significant impact on many aspects of the problem that we are discussing in this thesis chapter. Nature and relevance of molecular sequence information, computer-based protein and DNA sequence

analysis are crucial directions. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at National Centre for Biotechnology Information (NCBI). These three organizations exchange data on a daily basis. Sequence analysis encompasses the use of various bioinformatics methods to determine the biological function and/or structure of genes and the proteins they code for. The regulated expression of thousands of genes controls the mechanisms by which morphological form and cell functions are spatially organized during development [Kahn 1995, Mann 1999 and Pandey 2000]. Important molecular biology techniques such as gel electrophoresis and spectrophotometry allows the state of a cell to be characterized, further, in addition, the recent advances in large-scale assay technologies are providing experimentalists with an almost overwhelming quantity of data that defy description or characterization by conventional means [Bolouri 2001]. A large number of genes have been discovered along with their regulatory sites after the completion of several sequencing projects. It has resulted in the availability of huge amounts of datasets for analyses. A number of new microarray-based technologies have been developed over the last few years, and the technological development in this area is likely to continue at a brisk pace. These technologies include DNA hybridization arrays (gene expression arrays, oligonucleotide arrays for sequencing and polymorphism), protein arrays, tissue arrays and

combinatorial chemistry arrays. In these arrays total RNA is reverse-transcribed to create either radioactive or fluorescent-labelled cDNA that is hybridized with a large DNA library of gene fragments attached to a glass or membrane support. Phosphorimaging or other imaging techniques are used to produce expression measurements for thousands of genes under various experimental conditions. Use of these arrays is producing large amounts of data, potentially capable of providing fundamental insights into biological processes ranging from gene function to development, cancer, and ageing and pharmacology. DNA gene expression microarrays datasets allow biologists to study genome-wide patterns of gene expression in any given cell type, at any given time, and under any given set of conditions. The detailed study of organisms at the molecular level such as to analyze gene expression levels to know which crucial genes are expressed, when and where in the organism, and to which extent can help scientific community to better understand the functioning of organism. Measuring the expression rate of each gene over time gives a temporal profile of its expression level [Kasabov 2003, Kasabov and Dimitrov 2002] thus time ordered gene expression profiles are sequential snapshots of genome expression pattern [Ben-Dor 1999]. Analysis of large-scale gene expression is motivated by the premise that the information on the functional state of an organism is largely determined by the information on gene expression (based on central dogma). Gene expression array data

can be analysed on at least three levels of increasing complexity [Baldi and Brunak 2001]:

1) First level is that of single genes, where one seeks to establish whether each gene in isolation behaves differently in a control versus a treatment situation.

2) Second level is multiple genes, where clustering of genes are analysed in terms of common functionalities, interactions and co-regulation etc.

3) The third level attempts to infer the underlying gene and protein networks that ultimately are responsible for the patterns observed.

The ultimate goal of analysis of expression data is the complete reverse engineering of a genetic network (functional inference of direct causal gene interactions). Each gene in a cell may express differently over time this makes gene expression analysis based on static data ("one shot") not a very reliable mechanism therefore in order to draw meaningful inferences from gene expression data, it is important that each gene be surveyed under a variety of conditions, preferably in the form of expression time series in response to perturbations. Such datasets may be analysed using a range of methods with increasing depth of inference. We will discuss about these existing methods for such analysis in the next few chapters of this thesis. Also the case studies that we undertake in this research are introduced in this chapter and are well described in rest of this thesis.

2.3 MicroRNAs as a gene regulatory element

It is not surprising that organismal development involves complex regulatory mechanisms that target every aspect of gene expression [Amy et al. 2005]. Traditionally, the role of RNA in the cell was considered mostly in the context of the gene to protein translation, limiting RNA to its function as mRNA, tRNA, and rRNA. Although the analysis of sequenced genomes to date has focused most heavily on the protein-coding set of genes however a new paradigm of gene expression regulation has emerged recently with the discovery of microRNAs (miRNAs). Most of the miRNAs are thought to control gene expression, mostly by base pairing with miRNA-recognition elements (MREs) found in their messenger RNA (mRNA) targets whereas other non-coding RNA (ncRNA) participate in gene expression indirectly [Marianthi et al. 2004]. Therefore, the discovery of a diverse array of transcripts that are not translated to proteins but rather function as RNAs has changed this view profoundly. A comprehensive understanding of cellular processes is impossible without considering RNAs as key players. Efficient identification of functional RNAs (ncRNAs as well as *cis*-acting elements) in genomic sequences is, therefore, one of the major goals of current bioinformatics. Their discovery adds a new dimension to our understanding of complex gene regulatory networks. In this review, we will revisit the history of miRNAs, discuss their biogenesis and summarize recent findings in the miRNA research

and biological function etc. We conclude by highlighting the specific limitations of experimental approaches in this area and focus on important bioinformatics methods for classification of miRNAs.

Some genes produce transcripts (miRNAs) that function directly in regulatory, catalytic, or structural roles in the cell. These microRNAs are prevalent in all living organisms, and methods that aid the understanding of their functional roles are essential. MicroRNAs (miRNA) are ~22 nucleotide-long RNAs that function in translational repression by base pairing with their target mRNA in a variety of plants and animals. Processing of pri-miRNA into the premiRNA stem-loop occurs in the nucleus, while subsequent processing of pre-miRNA into 21-22 mers is a cytoplasmic event. They originate from long precursors (pri-miRNA) that, in animals, are cleaved by the Drosha endonuclease in the nucleus [John and Philip 2004] to give ~70 nucleotide long miRNA precursors (pre-miRNAs) with a characteristic hairpin structure. In the plants, excision of pre-miRNAs is performed by DCL1, a Dicer homologue [Michel J Weber 2005]. In plants, most microRNA (miRNA) genes bind sequences perfectly and lead to mRNA degradation [Harlan 2005]. However, in animals, with a notable exception, they function by preventing translation without mRNA degradation [Harlan 2005]. The precise molecular mechanism by which the bound miRNA down-regulates translation of its target

mRNA remains unknown. Biogenesis of miRNAs is discussed in detail in the next section of this thesis chapter.

Biogenesis of miRNAs

Although the mechanism of miRNA action remains elusive, their biogenesis is rapidly becoming clear. In general, it is thought that miRNA biogenesis proceeds via intermediate precursor transcripts of more than 70 nucleotides that have the capacity to form an extended stem-loop structure (pre-miRNA), although at least some pre-miRNAs are further derived from even longer transcripts (primary miRNA transcripts, or pri-miRNAs). These can exist as long individual pre-miRNA precursor transcripts, as operon-like multiple pre-miRNA precursors, or even as part of primary mRNA transcripts.

Two processing events lead to mature miRNA formation in animals. In the first, the nascent miRNA transcripts (pri-miRNA) are processed into ~70-nucleotide precursors (pre-miRNA); in the second event that follows, this precursor is cleaved to generate ~21–25-nucleotide mature miRNAs. Little is known about the transcriptional regulation of pri-miRNAs, except that certain primiRNAs are located within introns of host genes, including both protein-coding genes and non-coding genes, and might therefore be transcriptionally regulated through their host-gene promoters [Lin He and

Gregory J.Hannon 2004]. In addition, certain miRNAs are clustered in polycistronic transcripts, indicating that these miRNAs are co-ordinately regulated during development.

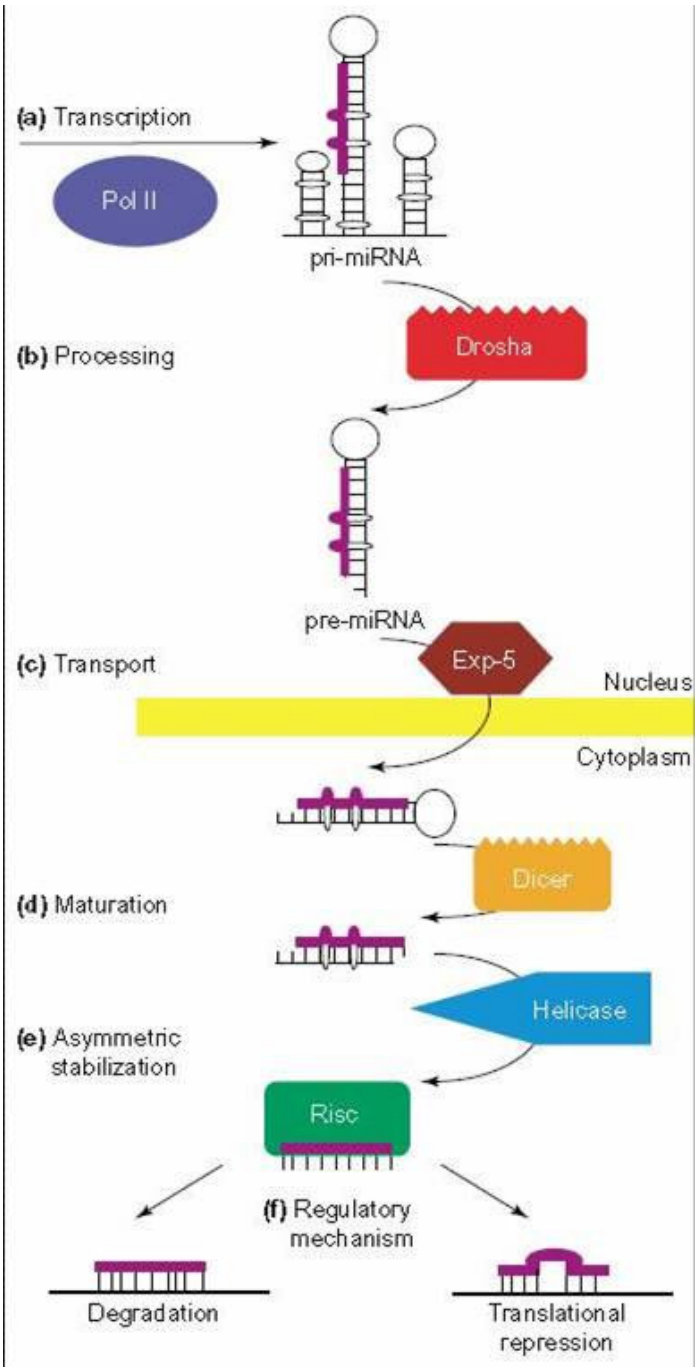


Figure 2.8: Potential regulatory points during expression and function of animal miRNAs.

[Figure is taken from Lin He and Gregory J.Hannon 2004]

The sequential cleavages of miRNA maturation are catalysed by two RNase-III enzymes, Drosha and Dicer [refer to figure 2.8]. Both are dsRNA-specific endonucleases that generate 2-nucleotide-long 3′□overhangs at the cleavage site. Drosha is predominantly localized in the nucleus and contains two tandem RNase-III domains, a dsRNA binding domain and an amino-terminal segment of unknown function [Lin He and Gregory J Hannon 2004]. Regardless of the diverse primary sequences and structures of pri-miRNAs, Drosha cleaves these into ~70-bp pre-miRNAs that consist of an imperfect stem-loop structure. Although the precise mechanisms that Drosha uses to discriminate miRNA precursors remain unknown, several studies have addressed the features of pri-miRNAs that contribute to Drosha cleavage both *in vitro* and *in vivo* [Lin He and Gregory J Hannon 2004]. The efficiency of Drosha processing depends on the terminal loop size, the stem structure and the flanking sequence of the Drosha cleavage site, because shortening of the terminal loop, disruption of complementarity within the stem and removal or mutation of sequences that flank the Drosha cleavage site significantly decrease, if not abolish, the Drosha processing of pri-miRNAs.

After the initial cleavage by Drosha, pre-miRNAs are exported from the nucleus into the cytoplasm by Exportin 5 (Exp5), a Ran-GTP dependent nucleo/cytoplasmic cargo transporter [refer to figure 2.8]. Once inside the cytoplasm, these hairpin precursors are cleaved by Dicer endonuclease into a

small, imperfect dsRNA duplex (miRNA: miRNA*) that contains both the mature miRNA strand and its complementary strand (miRNA*) a strand of which, corresponding to the mature miRNA, is predominantly incorporated in the RNA-induced silencing complex (RISC). Dicer contains a putative helicase domain, a DUF283 domain, a PAZ (Piwi–Argonaute–Zwille) domain, two tandem RNase-III domains and a dsRNA-binding domain (dsRBD). The RISC complex either inhibits translation elongation or triggers mRNA degradation, depending upon the degree of complementarity of the miRNA with its target [Mattick John S and Makunin Igor V 2005].

Techniques for microRNA identification and available data

Several hundred miRNAs have been experimentally determined in mammals, fish, flies, worms and plants; this number roughly corresponds to only 1% of the protein-coding genes in each organism [Sam Grifths-Jones 2004]. In addition to the experimental approaches, bioinformatics predictions have helped to identify novel miRNAs in various organisms, mostly on the basis of pre-miRNA hairpin structures and sequence conservation throughout evolution [Xiu-JieWang et al. 2004]. With the growing number of experimentally confirmed miRNAs, refined bioinformatics approaches will undoubtedly increase in power for identifying miRNAs that have escaped

experimental searches [Jia-Fu Wang et al. 2004, Stefan Washietl 2006 and Washietl et al. 2005].

It is quite challenging to use molecular biology techniques to confirm microRNAs existence in the cell. Also, there is a shortage of lab technologies available for the identification of microRNAs. The two main biological methods for cloning microRNAs to date are random cloning and sequencing, and PCR based methods providing partial sequence data. Although random cloning and sequencing of microRNAs is considered the gold standard for validating microRNAs, it is slow, expensive and lacks enough sensitivity to consistently detect microRNAs. The PCR-based methods use primers covering at least half of the predicted microRNAs and thus provide only limited independent sequence data. Due to the common sequence similarity between many microRNA families, this approach does not provide enough validation evidence for the presence of specific microRNAs. Other approaches for microRNA expression validation, such as Northern blot analysis, RNase protection, and laser-based nano-technologies only provide expression data on validated microRNAs, but do not offer validation of the existence of a specific predicted microRNA. If one uses northern blotting and probing it can be expensive and time consuming, so in summary we can say experimental miRNA identification is technically challenging and incomplete for the following reasons: miRNAs tend to have highly constrained tissue- and time specific expression patterns;

degradation products from mRNAs and other endogenous non-coding RNAs coexist with miRNAs and are sometimes dominant in small RNA molecule samples extracted from cells. Despite such issues scientists have developed and published some publicly available databases in field. So, nowadays there are few such repositories available for example, the miRBase database [Sam Griffiths-Jones et al. 2006] aims to provide integrated interfaces to comprehensive microRNA sequence data, annotation and predicted gene targets. miRNAMap [Paul WC Hsu et al. 2006] is a recently developed integrated database to store the known miRNA genes, the putative miRNA genes, the known miRNA targets and the putative miRNA targets. Thus these database lists out many known candidates but unknown miRNAs are still to be identified and needs to be submitted to such repositories.

Considering the above constraints of experimental approaches and limited access to the biological laboratory, for our research and case study on human microRNAs, we use the known data of human microRNAs and search within them to be able to identify some of them using our novel approach. We used a comprehensive and up-to-date repository (searchable database) of published miRNA sequences and annotation for the development and application of computational approach for predicting miRNA (for details see chapter 7). We discuss the potential and applicability of our method in this thesis chapter 7. Readers who wish to continue reading on the details of miRNA classification

method that we have employed and result interpretation etc. are advised to go directly on the chapter 7. However those who wish to take overall thesis idea on the integrative approach to understand molecular interactions in molecular biology may continue reading from chapter 3 onwards.

2.4 Case studies throughout the thesis

After reading the review on gene regulation and learning about the role of microRNAs in understanding the molecular interactions within a cell, researchers would surely agree with us that the expression of a gene by which biological information becomes available in the cell is actually regulated at several stages and the stage of regulation may vary (see detailed explanation in literature review). Thus in order to understand these complex molecular interactions we need to understand and analyse the concept from the gene sequence level to gene and protein expression level.

The studies on understanding molecular interactions of a cell can be addressed by considering genes and proteins from cells of any kind, as for example leukaemia blood cells or neuronal brain cells. Therefore, study may involve consideration of specific diseases like leukaemia, and neurogenetic disorders like epilepsy. Simultaneously, the sequence and structure data analysis to study and classify microRNAs can be an excellent input towards

understanding gene regulatory networks. We selected the individually complex case studies based on following criteria (reader needs to link here with the literature survey that makes obvious to select wide variety of data to understand gene regulation):

(1) In the above provided literature survey, we learnt that it is important to analyse the time-series microarray gene expression data for addressing the issue of gene regulation (GRN).

(a) For our first case study, we obtained the time-series gene expression data of Leukaemia U937 cell line from the National Cancer Institute (NCI), USA. The institute is a collaborator of KEDRI and we have also published jointly. The method applied was the integrative approach of Kalman filter (KF) with Genetic Algorithm (GA). More details can be found in the chapter 4 of this thesis.

(b) As our second case study, recently released publicly available data of *Schizosaccharomyces pombe* (yeast) was analysed and we hypothesize our novel findings. Integrative approaches of Least Angle Regression (LARS), Expectation Maximization (EM) with Kalman Filter (KF) and Evolving Fuzzy Neural Network (EFuNN) were used to analyse this time-series gene expression dataset of yeast. Chapter 5 has further details on this case study.

(c) For our third case study, we were provided with the gene expression time series dataset on the long term potentiation (LTP) related genes by [Chen

J Park et al. 2006]. They published their preliminary findings on this dataset and later this dataset was given to us as a courtesy for doing further analysis. We studied the LTP related GRNs using the integrative approach of quantum inspired evolutionary algorithm (QiEA) and clustering approach. This area has been described in the chapter 6 of this thesis.

(2) Following literature survey, we also learnt that microRNAs play a very crucial role in the regulation of genes. Therefore as our next case study we obtained data on all available known human microRNAs (till 2005) from RFAM database - miRNA Registry and dealt with functional classification and prediction using integrative approach of Gabor Filter, BLAST and CLUSTALW. Chapter 7 deals with this area of case study.

(3) As discussed earlier in the introduction of this thesis, for proper understanding of molecular interactions, in addition to analysing the gene expression data, sequence data, structure data of microRNA etc., it is also very important to link the research findings, data, facts about genes and proteins etc. in a way so that this centralized information may be shared among researchers and can be reused for further discoveries. Therefore, later we try to prove the concept of knowledge integration and information fusion on brain gene data through the development of brain-gene ontology (BGO). We have used most of the public data for developing this system and for

knowledge discovery we have shown the potential use of this system in discovering novel relationship among certain genes/proteins. We have used BGO data to perform gene-protein sequence analysis, studied clustering relationships between the subunits of crucial neuronal proteins and analysed the BGO data using computational intelligence methods (like ECOS) that lead to novel discoveries. In turn, this kind of analysis also aids in supporting the biological validation of models obtained from Computational Neurogenetic Modelling project [Kasabov and Benuskova 2004]. Chapter 8 of this thesis is dedicated to this area of doctoral study.

Therefore, by considering the above mentioned case studies in this research we had a chance to investigate molecular data from sequence level to gene and protein expression level. In addition, we also wanted to learn more about the microRNAs therefore analysis on human microRNA data and classifying their obtained structures was a good step. Further, ontology helped us in integrating the facts, data and knowledge etc. and using brain gene data as a case study allowed us to discover the novel findings in genetic neuroscience. In general, we can say that the undertaken case studies are very much illustrative and have potential applications in cancer prognosis, drug discovery and brain disease modelling etc. Further, the overall investigation on the respective case studies has allowed us to develop and apply various

generalized integrative approaches that may be used to analyse the different kinds of datasets as well (with certain restrictions of course).

2.5 Conclusion

It is well understood that the molecular biology central dogma is a very complex process. It includes the production of proteins from the genetic material DNA via an intermediate transcript molecule called RNA. The molecular interactions between DNA, RNA and proteins are called genetic regulatory networks (GRN). GRNs are so central to understanding and this area is a very important subject to study in bioinformatics. Major objectives in genetic regulatory studies should involve discovering the architecture, dynamics, and function of regulatory networks; make useful computational models of them; learn how to adapt and design them and finally coming out with some knowledge discovery of biological mechanisms. A new paradigm of gene expression regulation has emerged recently with the discovery of microRNAs (miRNAs). Most, if not all, miRNAs are thought to control gene expression, mostly by base pairing with miRNA-recognition elements (MREs) found in their messenger RNA (mRNA) targets. Their discovery adds a new dimension to our understanding of complex gene regulatory networks. Therefore, a comprehensive understanding of cellular processes is impossible without considering RNAs as key players. We have done the extensive

literature survey (part of which has been discussed in this chapter 2) to understand the generalized problem of understanding molecular interactions and gene regulation in the field of the bioinformatics. It is seen from literature review that various kinds of datasets are emerging from the experimental biology. For proper understanding of such cellular regulatory events, it is worth studying/analysing different sets of datasets that are available in the field of bioinformatics. Above mentioned areas of case studies in this PhD research has given us an opportunity to learn at different levels (a) the microarray gene expression (b) microRNA structure (c) the molecular sequence. We further discuss on specific problem description, methodologies and undertaken research in coming chapters of this thesis. For example, chapter 3 describes our integrated framework for modelling and knowledge discovery. Chapters 4, 5 and 6 are dedicated to analysing the various time-series gene expression datasets using different integrative approaches. Next, chapter 7 addresses the issue of microRNA classification in understanding gene regulatory networks. Finally we discuss the importance of ontology in current knowledge integration and describe the brain gene ontology (BGO) that was developed on brain gene data as an example of information fusion in chapter 8 of this thesis. It has been used in conjunction with the computational intelligence machine learning methods like ECOS and CLUSTALW etc. to report novel findings in the genetic neuroscience. This thesis is concluded with chapter 9 in which we discuss potential applications of this research and provide future directions etc.

3. An integrative ontology-based framework for modelling and knowledge discovery in bioinformatics

This chapter discusses and presents our proposed novel integrated framework for modelling and knowledge discovery. In this respect we provide a quick recap of the central dogma problem and explain our proposed approach. We later focus on the importance of ontologies in current data management and emphasize that machine learning tools play a valuable role in extracting/analysing the ontology data and thus to make discoveries. We conclude the chapter by giving a short summary of our proposed framework.

3.1 Introduction and problem specification

With the help of a proposed integrated framework, we describe our approach to study the gene regulation problem and fulfil the desire of scientific community that molecular interactions should be studied at different checkpoints of the central dogma and range of data should actually be taken into account for analysis using novel computational intelligence methods. We have learnt in the previous chapter of this doctoral thesis, the central dogma theme states that proteins are produced from the DNA (gene) via intermediate transcript called as RNA. Later these proteins play the role of enzyme to perform the checkpoints as a gene expression control. Also, according to the recently emerged paradigm sometimes genes don't code for proteins but results in small molecules of microRNAs which in turn control the gene regulation. The idea is that, such very complicated molecular biology process

results in the production of wide variety of data that can be used by computer scientists for modelling and to enable discoveries. Researchers can apply the novel or widely accepted bioinformatics and machine learning methods for such kind of data analysis. We further emphasize that the proper maintenance of diverse sources of data, structures and, in particular, their adaptation to new knowledge is one of the most challenging problems but one of the crucial tasks towards the knowledge integration vision is the efficient encoding of human knowledge in ontologies. Therefore in addition to the novel bioinformatics methods, our proposed integrated framework involves the use of ontology (in our case brain gene ontology) that helps the researchers to store the facts, results, data etc. The system can further act as a database so that new computational methods may be applied to reveal further hidden knowledge in this area. One can also make the use of several protégé based plug-ins that itself helps in discovering new knowledge. This novel knowledge (or discoveries) can then become the part of system in a way that it is new input data. One can again apply the methods of computational intelligence to make further discoveries and this process may be continued iteratively. During the whole process, at any stage the knowledge can be extracted, be examined and reused (refer to figure 3.1).

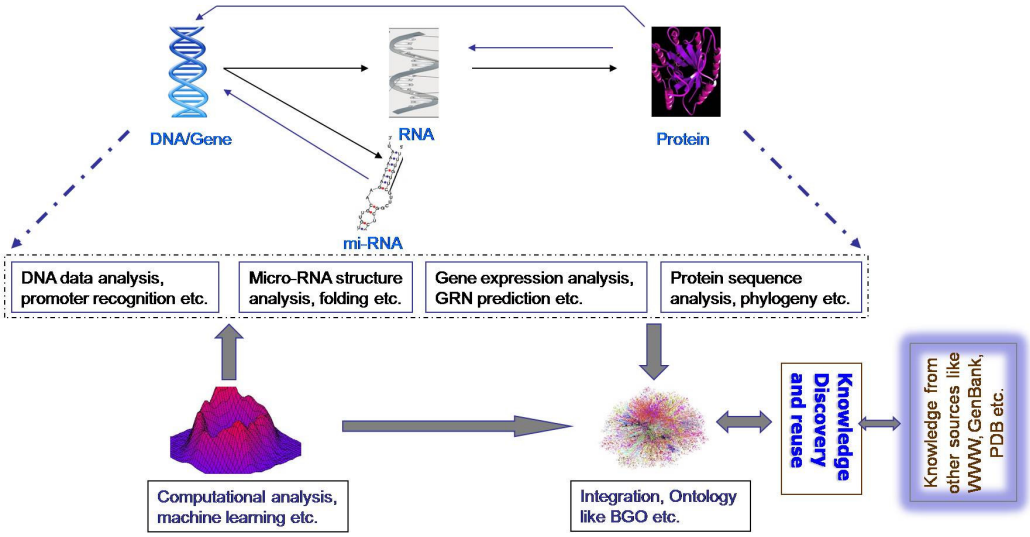


Figure 3.1: Proposed integrated framework for modelling and knowledge discovery. Small figures are taken from Encyclopaedia Britannica (http://www.encyclopedia.com/) to generate above figure

3.2 Computational intelligence methods within integrated framework for knowledge discovery

The framework that we have proposed in this chapter involves the use of several novel computational intelligence methods. As mentioned earlier, it is not so easy to understand the process of gene regulation (i.e. prediction of gene regulatory networks – GRNs) using bioinformatics approach. The problem is very complex and currently computational algorithms have not been highly 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. In our proposed framework to study GRN area, we have suggested the possible use of multiple data types, for example, gene expression time series, sequence and literature etc. We further suggest in this

framework that microRNAs prediction should also be studied simultaneously as they have recently emerged as a key player in controlling gene regulation. Finally, after examining each regulatory stage of a cell – we propose that integration is required to make the possible reuse of the obtained knowledge. Rest of the chapters in this thesis are directly related to the figure 3.1 in which we have represented the design of the proposed framework. For example, in chapters 4, 5 and 6, three separate gene expression time series datasets are analysed with the aim of inferring the meaningful GRNs. For this purpose, we didn't rely upon only one source of data (i.e. gene expression) and we have potentially used the valuable inputs from diverse sources of information, like gene and protein sequence data analysis, promoter prediction and analysis, functional analysis, clustering, and literature data and so on. Similarly, different integrative computational intelligence methods were applied on above mentioned data (information) types, for example, Genetic algorithm (GA) with Kalman filter (KF), Least Angle Regression (LARS), Expectation Maximization (EM) with Kalman Filter (KF), Evolving Fuzzy Neural Network (EFuNN) and Quantum Inspired Algorithm (QiEA) etc. Next, in chapter 7 we have dealt with the miRNA structure analysis (folding) based on our novel approach of the Gabor filtering. Again, while doing the miRNA analysis we have used multiple data information, for example, miRNA sequence analysis (BLAST) and clustering (CLUSTALW). By doing this thorough investigation (while working on different case studies) in several years of time, we have learnt that multiple sources of data and hybrid methods reveals some new aspects of the problem and we suggest that to infer the GRNs it is more suitable to adopt such integrative approach as ours through which some new

knowledge is discovered. Later, in accordance to the figure 3.1 we focus on knowledge integration and information fusion (using ontology) and discuss the utility of the machine learning tools in integrated framework. It has been described later in this chapter.

3.3 Concept of knowledge integration and information fusion

Although each cell within an organism will usually contain the same set of genes; there are significant differences in how the genes are utilized between them. Research from structural genomics towards functional genomics is leading for availability of new biological knowledge [Pandey and Mann 2000]. Opportunities arise by the simple act of connecting different facts and points of view that have been created for one purpose, but in light of subsequent information, they can be reused in a quite different context, to form new concepts or hypothesis.

The explosion of biomedical data and the growing number of disparate data sources are exposing researchers to a new challenge - how to acquire, represent, maintain and share knowledge from large and distributed databases in the context of rapidly evolving research [Chandrasekaran, Josephson & Benjamins 1999]. It is not wrong to say that biological knowledge is evolving so rapidly that it is difficult for most scientists to assimilate and integrate the new information with their existing knowledge. Therefore with the growing amounts of medical knowledge available it is

becoming impossible to contemplate successful biomedical research without good sharable knowledge bases and data structures. By integrating the domain of discovery outputs from different experimentation techniques with extracted appropriate biological and medical knowledge, there becomes a scope to discover new metabolic pathways and modelling of the metabolic and regulatory networks in living organisms, and ultimately to understand the pathogenesis of various diseases.

In the above context, the biggest problem today is that we are faced with an overwhelming array of nomenclature for genes, proteins, drugs and even diseases and thus biomedical community is suffering from a communication problem and the ability to use resources to search the vast sources of information more effectively, to extract appropriate meanings. Many researchers and databases use (at least partially) idiosyncratic terms and concepts for representing biological information. It has been observed that often, terms and definitions differ between groups, with different groups not infrequently using identical terms with different meanings. The concept "gene", for example, is used with different semantics by the major international genomic databases [Ashburner et al. 2000]. The quick aim in this respect may be to produce a dynamic, controlled vocabulary and to provide the knowledge on genes and proteins can be used for the analysis by various researchers.

As stated above, with the accumulation of both data and knowledge in the biomedical area and bioinformatics, it becomes necessary that these data and

knowledge need to be organized in a more global knowledge repository and used in their complexity and richness for an efficient novel discovery for domain specialist in our scientific community. To illustrate the problem above, let us take for example the brain in its multiple aspects of functioning and disease. The brain evolves in its structure and functionality at different levels from genetic (gene/protein level) to the evolutionary and the processes at each intermediate level are very complex and difficult to understand, but much more difficult to understand is the interaction between the different levels, e.g. gene- brain function-disease [Kasabov, Jain and Benuskova 2007]. The ultimate and broader goal in this respect is to use the stored data to discover novel knowledge and this has led to the development of ontologies [Chandrasekaran, Josephson & Benjamins 1999].

3.4 The ontology approach to integrate and reuse knowledge

Today, the key role of ontologies in information management has resulted in the rapid development of a large number of ontologies. Ontologies are an application-independent way to represent knowledge about the entities of an enterprise. To integrate genetic, proteomic and brain activity data and to perform data analysis, modelling, prognosis and knowledge extraction that reveals relationships between brain functions and genetic information, we need to build new global data and knowledge repositories and new mathematical and computational models.

In recent years, ontologies have been adopted in many business and scientific communities as a way to share, reuse and process domain knowledge [Fensel 2004 and Pisanelli 2004]. As a database technology, ontologies are commonly coded as triple stores (subject, relationship, object), where a network of objects is formed by relationship linkages, as a way of storing semantic information [Gruber 1993]. The term ontology has its origin in philosophy. In philosophy, ontology is the study of being or existence and usually describes the basic categories and relationships of being or existence to define entities and types of entities. For ontologies, many definitions exist like (1) It is specification of a conceptualization of a knowledge domain [Gruber 1993] (2) Ontology is a description (like a formal specification of a program) of the concepts and relationships between them to support the sharing and reuse of formally represented knowledge among AI systems [Fensel 2004 and Pisanelli 2004]. Ontology captures the intrinsic conceptual structure of a domain. Ontology can be said to study conceptions of reality. In modern computer science and information science, ontology is a data model that represents a set of concepts within a domain and the relationships between those concepts. Ontology is used to reason about the objects within that domain. Ontology specifies at a higher level the classes of concepts that are relevant to the domain and the relations that exist between these classes. Ontology captures the intrinsic conceptual structure of a domain. For any given domain, its ontology forms the heart of the knowledge representation. According to Gruber, the meaning of ontology in the context of computer science is the description of concepts and relationships that can exist for an agent or a community of agents [Gruber 1993]. By agent(s) we mean a

database, software tool, or any computational system. So in general we can say that ontology defines a common vocabulary and a shared understanding about a shared, formal, explicit and partial account of a conceptualization. In recent years, ontologies have been adopted in many business and scientific communities as a way to share, reuse and process domain knowledge [Fensel 2004 and Pisanelli 2004]. One advantage of ontologies over terminological systems is to support reasoning. Ontologies range from taxonomies and classifications, database schemas, to fully axiomatized theories.

Ontologies are now central to many applications such as scientific knowledge portals, information management and integration systems, electronic commerce, and semantic web services. For experimental purposes the medical ontologies, biomedical ontology (http://www.bioontology.org/) and the gene ontology (GO) have been created (http://www.geneontology.org/). The Gene Ontology (GO) project is a collaborative effort to provide a controlled vocabulary to describe gene and gene product attributes in any organism. The project began as a collaboration between three model organism databases, FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Database (MGD), in 1998. Since then, the GO Consortium has grown to include many databases, including several of the world's major repositories for plant, animal (mouse, rat), human and microbial genomes. But this is still not all. According to the 2008 update of the world-wide molecular database collection, there are 1078 freely available gene/protein related databases [Galperin 2008], 110 more than the

previous one. Since 2004, a total of 110-170 databases have been added each year. Therefore intelligent integration of relevant knowledge needs to be embodied in any biological data ontology that deals with novel knowledge discovery. The GO Consortium (http://www.geneontology.org/) is the set of about 20 bodies, i.e. model organism and protein databases and biological research communities actively involved in the development and application of the Gene Ontology. Disease Ontology is a controlled medical vocabulary designed to facilitate the mapping of diseases and associated conditions to particular medical codes such as ICD9, SNOMED and others (http://diseaseontology.sourceforge.net/). The goal of a Biomedical Ontology is to allow scientists to create, disseminate, and manage biomedical information and knowledge in a machine-processable form for accessing and using this biomedical information in research. The Gene Ontology (GO) project provides a controlled vocabulary to describe gene and gene product attributes in any organism addressing the need for consistent descriptions of gene products in different databases [Ashburner et al. 2000].

Simultaneously with the emerging need for standardized nomenclatures and concept ontologies for biosciences, the new science of systems biology has emerged. It is needed for the grand unification of biological and medical knowledge for basic and applied research. Importantly, systems biology is the ultimate tool for describing metabolic and genetic networks interacting with environmental variables to produce phenotypes of all organisms, including health and disease in individuals. In such systems biological knowledge needs to be represented, stored and analysed in a standardized ontological

framework, so that data from different domains of biology and medicine can be properly integrated.

Ontology can provide conceptual framework and factual knowledge that is necessary to deal critically with the rapidly changing science of biology and is necessary to explore the relationships between genes involved in brain disorders. Ontology is one way to provide a semantic repository of systematically ordered relevant concepts in molecular biology [Fensel 2004 and Pisanelli 2004]. Such repositories can be used, for example, to bridge the different notions in various databases by explicitly specifying the meaning of and relation between fundamental concepts. Thus in general, ontology permits researchers to define and share domain-specific vocabularies. Ontologies are not tied to any kind of particular formalism for their representation, nor are they concerned, in principle, with issues of computer tractability. To enable novel discoveries and to make the best possible reuse of available knowledge, ontologies should be regularly updated with new data, facts, information and knowledge (we all are aware that same is the case with any other database as well). Depending upon the type of ontology and available resources such updates can be consistently done on a daily to monthly basis. The new facts can be added manually or through an automated way. There is more information available on open and evolving ontologies from earlier published research paper [Gottgtroy, Kasabov and MacDonell 2006].

3.5 Application of machine learning tools in integrated framework

One of the novel ideas of this thesis is the integration between ontology and machine learning tools in relation to feature selection, classification and prognostic modelling with results incorporated back into the ontology. Machine learning is concerned with the design and development of algorithms and techniques that allow computers to "learn". The major focus of this integrated framework is to extract information from developed ontology data, by computational and statistical methods using a machine learning approach. For this purpose we have adopted a collaborative approach between human and machine. Human intuition cannot be entirely eliminated since the designer of the system must specify how the data is to be represented and what mechanisms will be used to search for a characterization of the data. In this respect, desired data from the ontology system (like our BGO) may be selected and exported in text file or in NEUCOM (neuro-computing decision support enviroment, www.theneucom.com/) format file for range a of analyses. This integrates the power of ontology with other novel machine learning softwares developed locally at KEDRI (Auckland University of Technology). For example, few of the top scoring genes out of thousands may be selected using a t-test or signal to noise ratio (SNR) method in a software environment Siftware [Kasabov, Jain and Benuskova 2007]. One can also apply the WEKA (data mining software developed in Java, http://www.cs.waikato.ac.nz/ml/weka/, which is developed by University of Waikato) and NeuCom to train prediction or classification models and to

visualize relationship information. Evolving Connectionist System (ECOS) [Kasabov 2003 and 2007a] can be used for building adaptive classification or prognostic systems and for extracting rules (profiles) that characterize data in local clusters. Analysis may also be done in a different manner by standard bioinformatics software like BLAST and FASTA for revealing homology patterns for those genes/proteins of interest, etc. [Benuskova, Jain et al. 2006]. To use the ontology data in conjunction with other computational intelligence or machine learning tools like NEUCOM (see appendix D), SIFTWARE (see appendix D), WEKA and CLUSTALW etc. one may have to obtain the licence permission from the respective organization. Chapter 8 of this thesis describes more on our approach and discusses a few of the discoveries that we have made and published internationally in this area.

3.6 Conclusion

This thesis chapter has discussed the integrated framework that we have developed for modelling and knowledge discovery. We have proposed for the first time that the discovery of molecular interactions should be made at different control points of central dogma and therefore one should consider the wide variety of data obtained from laboratory experiments. We have suggested that integrative use of computational intelligence methods (like GA, KF, EM, LARS, EFuNN, QiEA, BLAST, CLUSTALW etc.) is must within our proposed framework as each method may reveal new aspect of the problem and can potentially reveal novel knowledge. If one wants to view from a broader vision of understanding the phenomenon of gene regulation then we claim that it is very crucial to integrate the obtained knowledge, data, facts etc.

Currently we propose the approach of ontology for the knowledge integration and discuss the role of machine learning tools to make further discoveries from the proposed integrated framework. It is also emphasized that reuse of the knowledge plays a vital role in this area of research. In relation to our proposed framework, next three chapters 4, 5 and 6 are dedicated to analysing the various time-series gene expression datasets (along with other multiple information and/or data) using different computational intelligence integrative approaches. Chapter 7 is dedicated to the analysis of microRNAs and in chapter 8 we have discussed the concept of knowledge integration. At last, in chapter 9 implications and future directions of this research has been discussed.

4. Modelling and discovery of Gene regulatory networks (GRNs): An integrative Kalman filter (KF) Genetic Algorithm (GA) method

In chapter 2, we have described the problems in molecular biology and in particular, the discovery of molecular interactions i.e. gene regulatory networks (GRN). We have discussed the proposed integrated framework in the previous chapter where analysis of the gene expression data is the most important domain and here, now we describe one of our novel approaches to deal with the mentioned problem. In the initial two sections we quickly revisit the importance of studying GRNs and focus on the already existing approaches for their modelling. In the next section, we describe our integrative method of Kalman filter (KF) and Genetic Algorithm (GA). It is then followed by a section on the case study on GRN modelling from Leukaemia gene expression time series microarray dataset (which is one of the hot topics in current research world) and later we provide a summary of important results and discoveries that we have obtained and published in lecture notes in computer science and as a book chapter. This thesis chapter is then concluded with a highlight on the usefulness of the generic applicability of our method and the obtained results.

4.1 Introduction and problem specification

Detailed biological background on molecular interactions (so called Gene regulatory network – GRN) and description of the problem has been already provided in the chapter 2 of this thesis. As most regulatory systems of interest involve many genes connected through interlocking positive and negative feedback loops, an intuitive understanding of their dynamics is hard to obtain. Gene regulatory network is one of the two main targets in biological systems because they are systems controlling the fundamental mechanisms that govern biological systems. As we stated earlier, in order to draw meaningful inferences from gene expression data, it is important that each gene be surveyed under a variety of conditions, preferably in the form of expression time series in response to perturbations. GRN modelling will enable scientists to model complex dynamic processes involving large numbers of interacting variables and to extract inherent variable relationship networks (VRN). The discovery of gene regulatory networks (GRN) from time series of gene expression observations can be used to: (1) Identify important genes in relation to a disease or a biological function, (2) Gain an understanding on the dynamic interaction between genes, (3) Predict gene expression values at future time points. Gene expression datasets may be analysed using a range of computational intelligence methods with increasing depth of inference.

Current progress in the study of both naturally occurring and synthetic genetic networks indicates that computational modelling should have an important role in the description and manipulation of the dynamics that underlie cellular control. Computational and theoretical approaches will lead to testable predictions regarding the current understanding of complex biological networks. Models are usually made up of abstractions that are easier to manipulate than the actual system. Models that are developed upon knowable quantities; amount of different entities (e.g. proteins, transcripts or regulatory sites) and the rates at which it reacts can be regarded as somewhat more reliable. Genetic networks that include many genes and many signal pathways are rapidly becoming defined in prokaryotes and eukaryotes [Shapiro 1995]. However these past efforts to model the behaviour of gene regulatory network interactions were qualitatively or quantitatively incomplete. There are two major difficulties hampering the modelling and simulation of genetic regulatory networks. First, the biochemical reaction mechanisms underlying regulatory interactions are usually not known or are incompletely known. This means that detailed kinetic models cannot be built and more appropriate models are needed. Second, quantitative information on kinetic parameters and molecular concentrations is only seldom available. As a consequence, traditional models for numerical analysis are difficult to apply. Computationally predicted model may be compared with the observed gene expression profiles and thus one can get the estimation of its accuracy level. If the predicted and observed

behaviour do not match, and the experimental data is considered reliable, the model must be revised [De Jong 2002].

4.2 Existing methods for modelling of GRN

An important problem with inferring GRN is the large problem of dimension (thousands of genes) relative to the small number of observations (several to tens of time points). Often even after filtering of noisy signals, hundreds or thousands of candidate genes still remain. For this reason, many clustering algorithms are developed to reduce the problem dimension. Beginning with cluster analysis and determination of mutual information content, one can capture control processes shared among genes. A variety of clustering algorithms have been used to group together genes with similar temporal expression patterns [Ben-Dor et al. 1999, Brown et al. 2000, Cho et al. 1998, Eisen et al. 1998, Holter et al. 2001, Spellman et al. 1998]. Traditional clustering methods that perform hill-climbing from randomly initialized cluster centres are prone to produce inconsistent and sub-optimal cluster solutions over different runs. A crucial problem is to infer an accurate model for interactions between important genes in a cell. The major approaches that deals with the modelling of gene regulatory networks involve differential equations (mathematical equation for an unknown function of one or several variables that relates the values of the function itself and of its derivatives of various orders) [Likhoshvai et al. 2000], stochastic models (randomness is

present, and variable states are not described by unique values, but rather by probability distributions) [Mc Adams et al. 1997], evolving connectionist systems (genes are represented as neurons and the interaction between them – as weighted connections) [Kasabov 2003 and 2007a], boolean networks (where boolean vectors represent the state of the genes at every time point, i.e. values of 1 or 0; this representation is too simplistic and is imprecise) [Sanchez et al. 1997], generalized logical equations [Thieffry 1995], threshold models (multifactorial causality but a dichotomous phenotype) [Tchuraev 1991], evolutionary algorithms (GRN are evolved from gene data based on a fitness function) [Goldberg 1989 and Baeck 1995], petri nets (consists of places, transitions, and arcs; graphically depicts the structure of a distributed system as a directed bipartite graph with annotations) [Hofestadt 1995], bayesian networks (transitional probabilities are represented in the model) [Friedman et al. 2000], directed (edge is replaced by a directed graph edge) and undirected graphs (nodes are connected by undirected arcs). But there are certain problems with these existing methods like missing values in data, diverse data sources, multiple model and data integration, comparing GRN derived from different models. Such methods usually requires large amount of training samples. A common remedy to this is to increase the number of samples by interpolation methods but we all know that the interpolated points can be erroneous especially when the observed samples are sparse and noisy. So in general we can say that despite the existence of these methods,

the problem of the genetic network discovery has not been solved so far. One of the reasons is that the processes are too complex for the existing computational models. Generally speaking, we believe that modelling genetic networks requires the combination of two or more techniques together and if possible integrating the knowledge obtained from the application of such methods can be even better idea. It may involve careful gene selection or clustering, deriving smaller gene networks of clusters or of genes and then obtaining a globally optimized gene interaction network. In addition, some biological inputs to the model like checkpoint controls or validating at least a few relations within gene network are required to obtain meaningful models. In this thesis chapter we will describe our novel approach that we take to address this problem and the results on Leukaemia case study will be discussed in this direction.

4.3 Proposed integrative approach: Kalman filter (KF) with Genetic Algorithm (GA)

Introduction

We propose here a novel method that integrates Kalman Filter [Brown 1983] and Genetic Algorithm (GA) [Goldberg 1989]. For more details on KF and GA, readers are advised to refer to the appendices A and B. The GA is used to select a small number of genes, and the Kalman filter method is used to derive the GRN of these genes. After GRNs of smaller number of genes are

obtained, these GRNs may be integrated in order to create the GRN of a larger group of genes of interest. The primary goal of this research is develop a method for GRN discovery from multiple and short time series data of a large number of genes and thus to understand the molecular interactions within cell. The secondary goal is to apply the method as to identify the genes that co-regulate telomerase activity. Telomerase is one extra gene taken into account apart from our 32 selected genes (it is important to study leukaemia disease; results are discussed later in the case study). The integrated method can be easily generalized to extract GRN from any time series gene expression data. In coming few sections of this thesis chapter we report on our methodology (using small examples to illustrate parts of the method) and the experimental findings.

Modelling GRN with first-order differential equations, state-space representation and Kalman Filter

Here, GRN is modelled with the discrete time approximation of first-order differential equations, given by:

$$\mathbf{x}_{t+1} = \mathbf{F}\mathbf{x}_t + \boldsymbol{\varepsilon}_t \qquad (1)$$

where $\mathbf{x}_t = (x_1, x_2, \ldots x_n)'$ is the gene expression vector at the t-th time interval and n is the number of genes modelled, ε_t is a noise component with covariance E=cov (ε_t), and F=(f_{ij}) i=1,n, j=1,n is the transition matrix relating x_t

to x_{t+1}. It is related to the continuous first-order differential equations $d\mathbf{x}/dt = \mathbf{\Psi x} + \mathbf{e}$ by $\mathbf{F} = e^{\tau\mathbf{\Psi}} + \mathbf{I}$ and $\varepsilon_t = \pi\mathbf{e}$ where τ is the time interval {note the subscript notation (t+k) is actually the common abbreviation for (t+kτ)} [Fields 1999]. We work here with a discrete approximation instead of a continuous model for the ease of modelling and processing the irregular time–course data (with Kalman filter). Besides being a tool widely used for modelling biological processes, there are two advantages in using first-order differential equations.

First, gene relations can be elucidated from the transition matrix **F** through choosing a threshold value ($\zeta; 1>\zeta>0$). If $|f_{ij}|$ is larger than the threshold value ζ, $x_{t,\,j}$ is assumed to have significant influence on $x_{t+1,i}$. A positive value of f_{ij} indicates a positive influence and vice-versa. Second, they can be easily manipulated with KF to handle irregularly sampled data, which allow parameter estimation, likelihood evaluation and model simulation and prediction.

The main drawback of using differential equations is that it requires the estimation of n^2 parameters for the transition matrix **F** and n (n-1)/2 parameters for the noise covariance **E**. To minimize the number of model parameters, we estimate only **F** and fix **E** to a small value. Since both series contain only four samples, we avoid over-parameterization by setting n to 4, which is the maximum number of n before the number of parameters exceeds the number of training data {It matches the number of model parameters (the size of **F** is

n^2=16) to the number of training data (n×4 samples =16)}. Since in our case study one of the n genes must be telomerase, we can search for a subset of size K=3 other genes to form a GRN.

To handle irregularly sampled data, we employ the state-space methodology and the KF (see appendix A). We treat the true trajectories as a set of unobserved or hidden variables called the *state variables*, and then apply the KF to compute their optimal estimates based on the observed data. The state variables that are regular/complete can now be applied to perform model functions like prediction, parameter estimations instead of the observed data that are irregular/incomplete. This approach is superior to interpolation methods as it prevents false modelling by trusting a fixed set of interpolated points that may be erroneous.

State-Space Representation

To apply the state-space methodology, a model must be expressed in the following format called the *discrete-time state space representation*

$$\mathbf{x}_{t+1} = \mathbf{\Phi}\mathbf{x}_t + \mathbf{w}_t \tag{2}$$

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \mathbf{v}_t \tag{3}$$

$$\mathrm{cov}(\mathbf{v}_t) = \mathbf{R} \qquad \mathrm{cov}(\mathbf{w}_t) = \mathbf{Q} \tag{4}$$

where, $\mathbf{x}_t$ is the system state; $\mathbf{y}_t$ is the observed data; $\mathbf{\Phi}$ is the state transition matrix that relates $\mathbf{x}_t$ to $\mathbf{x}_{t+1;}$ $\mathbf{A}$ is the linear connection matrix that

relates $\mathbf{x}_t$ to $\mathbf{y}_t$; $\mathbf{w}_t$ and $\mathbf{v}_t$ are uncorrelated white noise sequences whose covariance matrices are **Q** and **R** respectively. The first equation is called the *state equation* that describes the dynamics of the state variables. The second equation is called the *observation equation* that relates the states to the observation.

To represent the discrete-time model in the state-space format, we simply substitute the discrete-time equation (1) into the state equation (2) by setting $\mathbf{\Phi}=\mathbf{F}$, $\mathbf{w}_t=\boldsymbol{\varepsilon}_t$ and $\mathbf{Q}=\mathbf{E}$ and form a direct mapping between states and observations by setting $\mathbf{A}=\mathbf{I}$. The state transition matrix $\mathbf{\Phi}$ (functional equivalent to **F**) is the parameter of interest as it relates the future response of the system to the present state and governs the dynamics of the entire system. The covariance matrices **Q** and **R** are of secondary interest and are fixed to small values to reduce the number of model parameters.

Using GA for the selection of a gene subset for a GRN

Genetic Algorithm (GA) as a general optimization procedure is described in appendix B. The task is to search for a set of genes that form the most probable GRN models, using the model likelihood computed by the KF as an objective function. Given *N* the number of candidates and *K* the size of the subset, the number of different gene combinations is *N!/K!(N-K)!*. In our case study, *N*=32 is small enough for an exhaustive search. However, as more

candidates are identified in the future, the search space grows exponentially in size and exhaustive search will soon become unfeasible. For this reason a method based on GA is proposed. The strength of GA is twofold:

1. Unlike most classical gradient methods or greedy algorithms that search along a single hill-climbing path, a GA searches with multiple points and generates new points through applying genetic operators that are stochastic in nature. These properties allow for the search to escape local optima in a multi-modal environment. GA is therefore useful for optimizing high dimensional functions and noisy functions whose search space contains many local optima points.

2. A GA is more effective than a random search method as it focuses its search in the promising regions of the problem space.

GA Design for Gene Subset Selection

In the GA-based method for gene subset selection proposed here, each solution is coded as a binary string of N bits. A "1" in the ith bit position denotes that the ith gene is selected and a "0" otherwise. Each solution must have exactly K "1"s and a repair operator is included to add or delete "1"s when this is violated. The genetic operators used for crossover, mutation and selection are respectively the standard crossover, the binary mutation and the (μ, λ) selection operators. Since there are two series – the plus and the minus series

(these two series means data was generated from the two different clones) of time-course gene expression observations in our case study, a new fitness function is designed to incorporate the model likelihood in both series. For each solution, the ranking of its model likelihood in the plus series and in the minus series are obtained and then summed to obtain a joint fitness ranking. This favours convergence towards solutions that are consistently good in both the plus and the minus series. The approach is applicable to multiple time series data.

Procedures of the GA-based method for gene subset selection

Population Initialization. Create a population of μ random individuals (genes from the initial gene set, e.g. of 32) as the first generation parents.

Reproduction. The goal of reproduction is to create λ offspring from μ parents. This process involves three steps: crossover, mutation and repair.

- *Crossover*. The crossover operator transfers parental traits to the offspring. We use the uniform crossover that samples the value of each bit position from the first parent at the crossover probability p_c and from the second parent otherwise. In general, performance of GA is not sensitive to the crossover probability and it is set to a large value in the range of [0.5, 0.9] [Baeck and Fogel et al. 2000]. Here we set it to 0.7.

- *Mutation*. The mutation operator induces diversity to the population by injecting new genetic material into the offspring. For each bit position of the offspring, mutation inverts the value at a small mutation rate p_m. Performance of GA is very sensitive to the mutation probability and it usually adapts a very small value to avoid disrupting convergence. Here we use p_m=1/N, which has been shown to be both the lower bound value and the optimal value for many test functions [Muhlenbein 1992 and Baeck et al. 2000], providing an average of one mutation in every offspring.
- *Repair*. The function of the repair operator is to ensure that each offspring solution has exactly K "1" to present the indices of the K selected genes in the subset. If the number of "1"s is greater than K, invert a "1" at random; and vice-versa. Repeat the process until the number of "1"s matches the subset size K.

Fitness Evaluation. Here λ offspring individuals (solutions) are evaluated for their fitness. For each offspring solution, we obtain the model likelihood in both the plus and the minus series and compute their ranking (lower the rank, higher the likelihood) within the population. Next, we sum the rankings and use the negated sum as fitness estimation so that the lower the joint ranking, the higher the fitness.

Selection. The selection operator determines which offspring or parents will become the next generation parents based on their fitness function. We use the (μ, λ) scheme that selects the fittest μ of λ offspring to be the next generation parents. It is worth comparing this scheme to another popular selection scheme $(\mu+\lambda)$ that selects the fittest μ of the joint pool of μ parents and λ offspring to be the next generation parents, in which the best-fitness individuals found are always maintained in the population, convergence is therefore faster. We use the (μ, λ) scheme because it offers a slower but more diversified search that is less likely to be trapped in local optima.

Test for termination. Stop the procedure if the maximum number of generations is reached. Otherwise go back to the reproduction phase.

Upon completion, GA returns the highest likelihood GRNs found in both the plus and the minus series of gene expression observations. The proposed method includes running the GA-based procedure over many iterations (e.g. 50) thus obtaining different GRN that include possibly different genes. Then we summarize the significance of the genes based on their frequency of occurrence in these GRNs and if necessary we put together all these GRNs thus creating a global GRN on the whole gene set.

4.4 Case study: GRN modelling from Leukaemia gene expression time series microarray dataset

For all of our experiments in this research we use the time series gene expression microarray dataset of the leukaemia extracts of U937 cell line of plus and minus clones (series) that was actually obtained in the National Cancer Institute (NCI), National Institute of Health (NIH), USA. This institute is a KEDRI collaborator and we received this dataset to study molecular interactions using our novel computational methods. Within dataset each series contains the time-series expression of 12,625 genes of which we deal with only 32 pre-selected candidate genes that have been found potentially relevant (based on molecular biology facts), as well as the expression of the telomerase (it is one of the expressed gene in this dataset). Both the plus series and minus series contains four samples recorded at the $(0, 6, 24, 48)^{th}$ hour. Discovering GRN from these two series is challenging in two aspects: first, both series are sampled at irregular time intervals; second, the number of samples is scarce (only four samples). A third potential problem is that the search space grows exponentially in size as more candidate genes are identified in the future. Several GRNs of three genes most related to the telomerase gene are discovered, analysed and integrated. The results and their interpretation confirm the validity and the applicability of the proposed method.

The integrated GA-KF method introduced above is applied to identify genes that regulate telomerase in a GRN from a set of 32 pre-selected genes. Since the search space is small (only C_3^{32}=4960 combinations), we apply exhaustive search as well as GA for validation and comparative analysis. The experimental settings are as follows. The expression values of each gene in the plus and minus series are jointly normalized in the interval [-1, 1]. The purpose of the joint normalization is to preserve the information on the difference between the two series in the mean. For each subset of *n* genes defined by the GA, we apply KF for parameter estimation and likelihood evaluation of the GRN model. Each GRN is trained for at least 50 epochs (which is usually sufficient) until the likelihood value increases by less than 0.1. During training, the model is tested for stability by computing the eigenvalues of $(\Phi\text{-}I)$ [Bay 1999 and Dorf et al. 1998]. If any of the real part of eigenvalues is positive, the model is unstable and is abandoned.

For the experiments reported in this chapter relatively low resource settings are used. Parent and offspring population sizes (μ, λ) are set to (20, 40) and maximum number of generations is set to 50. These values are empirically found to yield consistent results over different runs. We run it for 20 times from different initial populations to obtain the cumulated results. The results are interpreted from the list of 50 most probable GRNs found in each series (we can lower this number to narrow down the shortlist of significant

genes). The frequencies of each gene being part of the highest likelihood GRNs in the plus and in the minus series are recorded. Next, a joint frequency is calculated by summing the two frequencies. The genes that have a high joint frequency are considered to be significant in both minus and plus series.

For exhaustive search, we simply run through all gene combinations of three genes plus the telomerase; then evolve through KF a GRN for each combination and record the likelihood of each model in both the plus and minus series. A similar scoring system as GA's fitness function is employed. We obtain a joint ranking by summing the model likelihood rankings in the plus series and the minus series, and then count the frequency of the genes that belong to the best 50 GRNs in the joint ranking.

4.5 Results and discoveries

The top ten highest scoring genes obtained by GA and exhaustive search are tabulated in Table 4.1 and a comprehensive list of genes with their function is provided in Table 4.2.

Table 4.1: Significant genes extracted by GA and through an exhaustive search from 32 selected genes

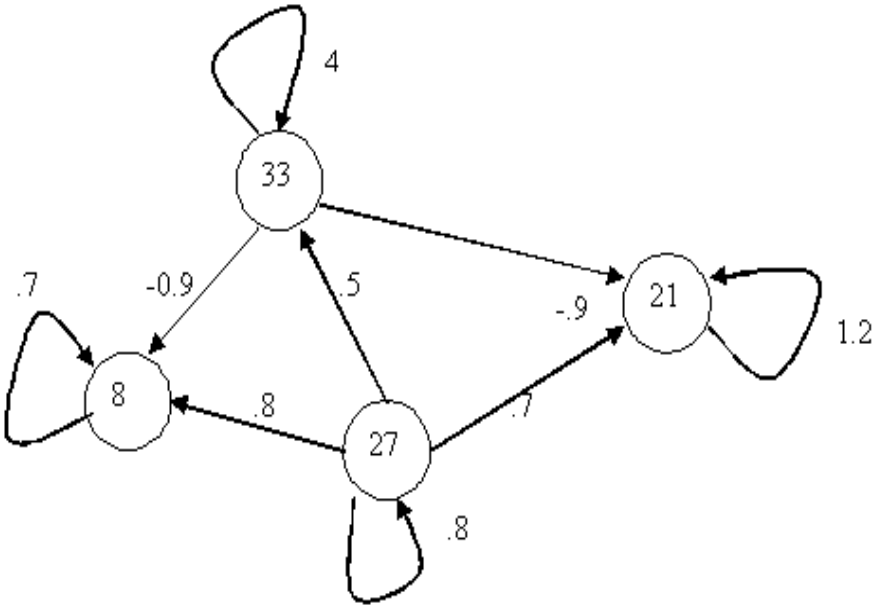
Rank	Indices of significant genes found by GA (Freq. of occurrence in Minus GRNs, Freq. of occurrence in Plus GRNs) and their accession numbers in Genbank	Indices of significant genes found by exhaustive search (gene Index)
1	27 (179,185) X59871	20 M98833
2	21 (261,0) U15655	27 X59871
3	12 (146, 48) J04101	32 X79067
4	32 (64, 118) X79067	12 J04101
5	20 (0, 159) M98833	6 AL021154
6	22 (118, 24) U25435	29 X66867
7	11 (0, 126) HG3523-HT4899	5 D50692
8	5 (111, 0) D50692	22 U25435
9	18 (0, 105) D89667	10 HG3521-HT3715
10	6 (75, 0) AL021154	13 J04102

The identified GRNs can be used for model simulation and prediction. The GRN dynamics can also be visualized with a network diagram using the influential information extracted from the state transition matrix. As an example, we examine one of the discovered GRN of genes (33, 8, 27, 21) for both the plus and minus series, shown in figure 4.1 and figure 4.2 respectively. The network diagram shows only the components of Φ whose absolute values are above the threshold value ζ=0.3.

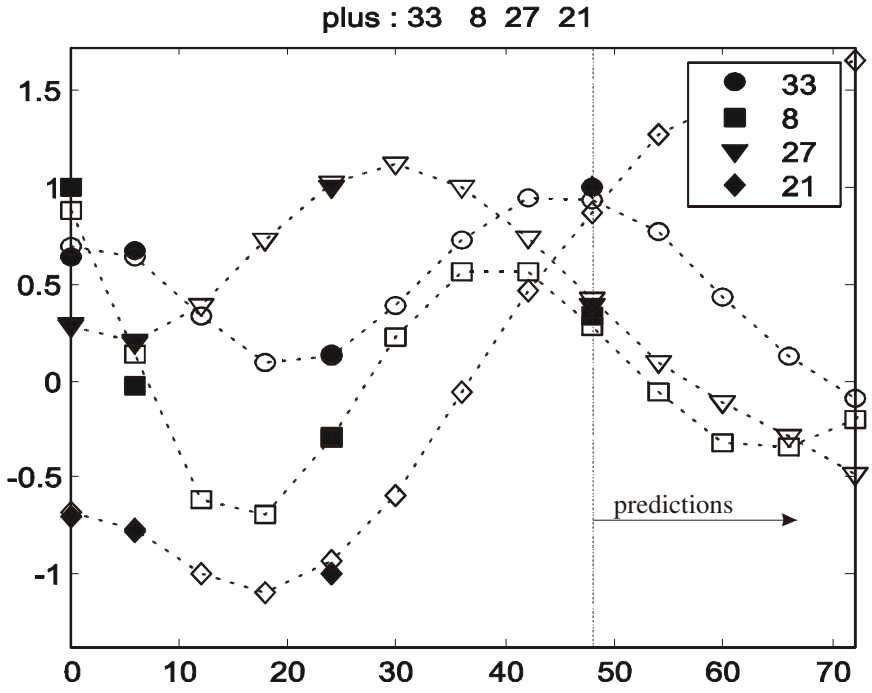
Table 4.2: Corresponding gene name and description/function of each gene index

Index	Name	Description/function	Index	Name	Description/function
1	AB012124	Homo sapiens TCFL5 mRNA for transcription factor-like 5	18	M55914	Human c-myc binding protein (MBP-1) mRNA
2	AB016194	Homo sapiens elk1 oncogene	19	M64788	Human GTPase activating protein (rap1GAP) mRNA
3	AF002999	Homo sapiens TTAGGG repeat binding factor 2 (hTRF2) mRNA	20	M98833	Human ERGB transcription factor (FLI-1 homolog) mRNA
4	Y11306	Homo sapiens mRNA for hTCF-4	21	U15655	Human ets domain protein ERF mRNA
5	AF058696	Homo sapiens cell cycle regulatory protein p95 (NBS1) mRNA	22	U25435	Human transcriptional repressor (CTCF) mRNA
6	AL021154	AL021154: dJ150O5.2 (Inhibitor of DNA binding 3 (dominant negative helix-loop-helix protein, 1R21, HEIR-1))	23	U32645	Human myeloid elf-1 like factor (MEF) mRNA
7	D13891	Human mRNA for Id-2H	24	U40705	Homo sapiens telomeric repeat binding factor (TRF1) mRNA
8	D50692	Homo sapiens mRNA for c-myc binding protein	25	U43189	Human Ets transcription factors NERF-1a and NERF-1b (NERF-1a, b) mRNA
9	D89667	Homo sapiens mRNA for c-myc binding protein	26	V00568	1973_s_at _DDD, RNA, 0, v00568, V00568
10	HG3521-HT3715	1903_at _UDU, mRNA, 0, ras-related, protein	27	X59871	Human TCF-1 mRNA for T cell factor 1 (splice form C)
11	HG3523-HT4899	Proto-oncogene c-myc	28	X60287	1981_s_at _DDU, RNA, 0, x60287, X60287
12	J04101	Human erythroblastosis virus oncogene homolog 1 (ets-1) mRNA	29	X66867	422_s_at _DUU, Gene, 0, x66867, X66867
13	J04102	Human erythroblastosis virus oncogene homolog 2 (ets-2) mRNA	30	X77956	H. Sapiens Id1 mRNA
14	J04977	Human Ku autoimmune antigen gene	31	X78992	H. Sapiens ERF-2 mRNA
15	M13929	Human c-myc-P64 mRNA, initiating from promoter P0, (HLmyc2.5) partial cds	32	X79067	38740_at _DDD, mRNA, 0, x79067, Cluster
16	M25269	Homo sapiens tyrosine kinase (ELK1) oncogene mRNA	33	AF015950	Homo sapiens telomerase reverse transcriptase (hTRT) mRNA
17	M30938	Human Ku (p70/p80) subunit mRNA			

For the plus series, the network diagram in figure 4.1 (a) shows that gene 27 has the most significant role regulating all other genes (note that gene 27 has all its arrows out-going). The network simulation, shown in figure 4.1 (b) fits the true observations well and the predicted values appear stable, suggesting that the model is accurate and robust. For the minus series, the network diagram in figure 4.2 (a) shows a different network from that of the plus series. The role of gene 27 is not as prominent. The relationship between genes is no more causal but interdependent, with genes 27, 33 and 21 simultaneously affecting each other. The difference between the plus and minus models is expected. Again, the network simulation result shown in figure 4.2 (b) shows that the model fits the data well and the prediction appears reasonable. Thus, in general the figures 4.1 (a) and 4.2 (a) are simply the matrix representation that we have shown as a visualization diagram (or network). The value on lines (arrows) shows the affect of one gene on the other with certain impact (positive or negative). Telomerase gene (number 33) is shown in both the figures. Diagrams are actually self-explanatory and curve as such doesn't indicate any additional information.

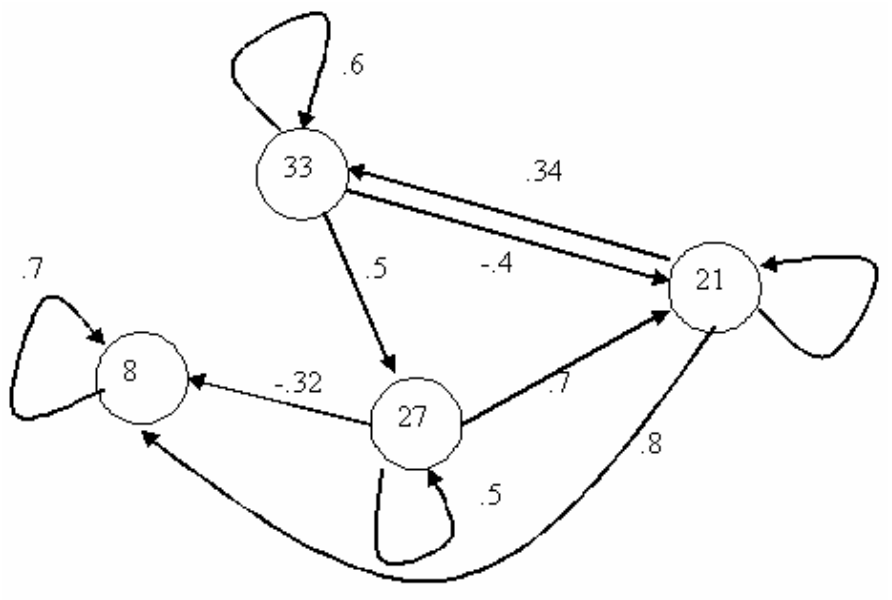


(a)

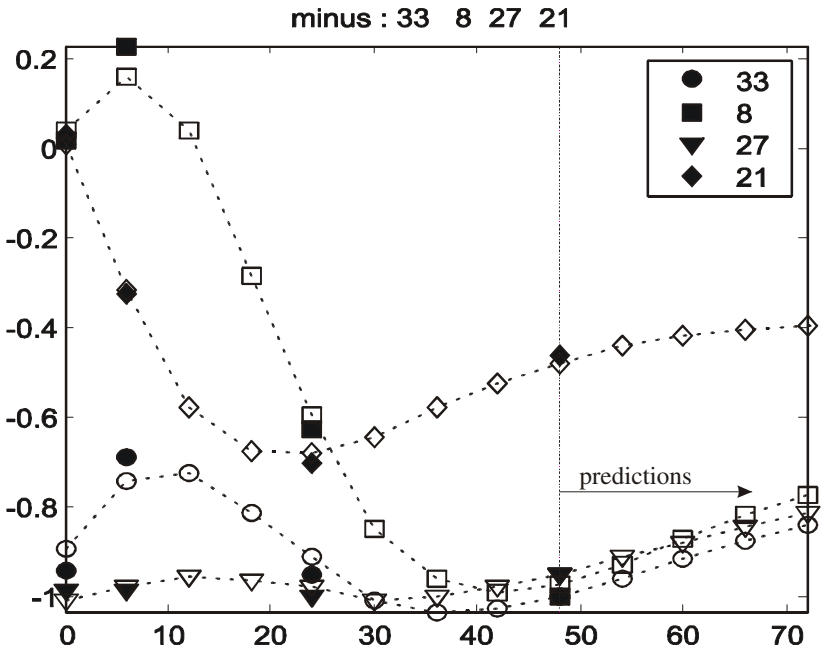


(b)

Figure 4.1: The identified best GRN of gene 33 (telomerase) and genes 8, 27 and 21 for the plus series - (a) the network diagram (b) the network simulation and gene expression prediction over future time; X axis: time points (hours) and Y axis: expression values. Solid markers represent observations.



(a)



(b)

Figure 4.2: The identified best GRN of gene 33 (telomerase) and genes 8, 27 and 21 for the minus series - (a) the network diagram (b) the network simulation and gene expression prediction over future time; X axis: time points (hours) and Y axis: expression values. Solid markers represent observations.

4.5.1 Biological validation of results

We can see that the results obtained by GA and exhaustive search (table 4.1) are strikingly similar. In both lists, seven out of top ten genes are common (genes 27, 12, 32, 20, 22, 5, 6) and four out of top five genes are the same (genes 27, 12, 32 and 20). The similarity in the results supports the applicability of a GA-based method in this search problem and in particular, when the search space is too large for an exhaustive search. An outstanding gene identified is gene 27, TCF-1. Validation of our obtained results (i.e. interacting genes that we have found in relation to telomerase gene) is indeed done through biological experimentation carried out at National Cancer Institute (NCI). For example, it was verified successfully that gene 33 (telomerase) interact with the gene 27 (TCF-1). The biological implications other high scoring genes may be the work of future investigation.

A simple method for building a global GRN - putting the pieces of the puzzle together

After many GRNs of smaller number of genes are discovered, each involving different genes (with a different frequency of occurring), these GRNs can be put together to create a GRN of the whole gene set (out of the GRNs of smaller number of genes). Representation and illustration for the top five (fittest) GRNs from our experiment are shown in figure 4.3 and table 4.3

respectively. In this particular example, one can see that gene number 33 (*Homo sapiens* telomerase reverse transcriptase, hTRT, mRNA, AF015950) is crucial for our case study (discussed earlier) and we are interested in finding the GRNs related to this gene. Therefore, here we have tried to visualize the most likely genes that interact with this central gene (i.e. gene 33). Such a simple yet powerful method may be easily applied to any interacting genes/proteins of a given dataset to infer global GRNs. Some other examples on computational modelling strategies for gene regulatory network reconstruction may be found in [Sehgal et al. 2008]. Some implications and future directions to the methods suggested here may be found in chapter 9 of this thesis.

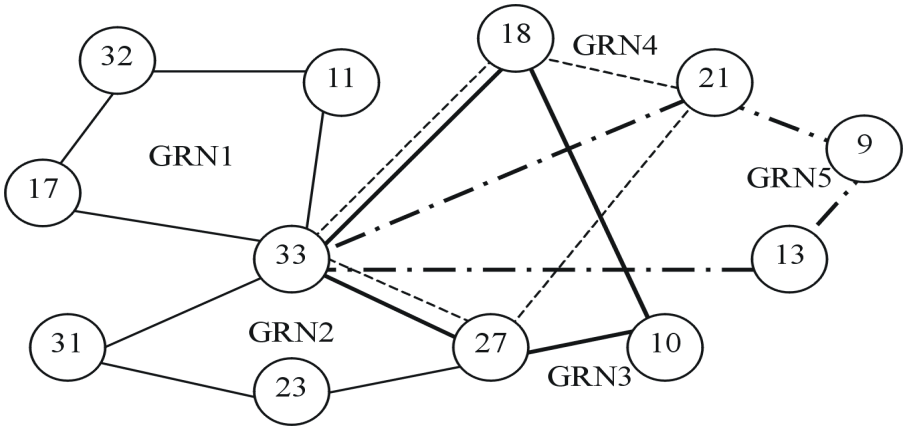


Figure 4.3: The five highest likelihood GRN models found by GA in the plus series are put together.

Table 4.3: Illustration of top five fittest GRNs (plus series)

GRN Number	GRN identified
1	(33 32 17 11)
2	(33 31 27 23)
3	(33 27 18 10)
4	(33 27 21 18)
5	(33 21 13 9)

4.5 Conclusion

In this chapter, we have proposed a novel method that integrates Kalman Filter and Genetic Algorithm for the discovery of GRN from gene expression observations of several time series (in this case they are two) of small number of observations. As a case study we have applied the method for the discovery of GRN of genes that regulate telomerase in two sub-clones of the human leukemic cell line U937. The time-series contain 12,625 genes, each of which sampled four times at irregular time intervals, but only 32 genes (and telomerase) of interest are dealt within this chapter. The method is designed to deal effectively with irregular and scarce data collected from a large number of variables (genes). GRNs are modelled as discrete-time approximations of first-order differential equations and Kalman Filter is applied to estimate the true gene trajectories from the irregular observations and to evaluate the likelihood of the GRN models. GA is applied to search for smaller subset of genes that are probable in forming GRN using the model likelihood as an optimization objective. After several runs of the GA-based method the genes

that occur with the highest frequency in the optimal GRNs are considered significant. The obtained GRNs may be put together to form a global GRN of the whole gene set. This approach reduces the size of GRN to be modelled from the number of candidate genes to the size of the smaller subsets, thus reducing the dependence on the amount of data. One of our finding i.e. interaction of TCF-1 with telomerase was validated through biological experimentation and confirmed as indeed correct. It is worth mentioning that the proposed method in this chapter for GRN extraction is a generic approach that can be easily applied on any time series gene expression dataset to infer gene regulatory networks. In the next two chapters of this thesis, we will discuss the potential integrative use of some other methods to infer GRNs.

5. An integrative Least Angle Regression (LARS) and machine learning approach for GRN extraction

In relation to the proposed framework to study gene regulation (refer to chapter 3), in this chapter we present another novel computational intelligence integrated approach to tackle the problem of inferring gene regulatory networks (GRNs). A brief introduction is followed by the outline description of our proposed integrative approach: Least Angle Regression (LARS), Expectation Maximization (EM) with Kalman Filter (KF) and Evolving Fuzzy Neural Network (EFuNN). Then we summarize the LARS algorithm and discuss its application by undertaking the case study on yeast cell-cycle time series microarray dataset [Peng et al. 2005] to infer GRNs and report important results and other discoveries (they were validated and found to be biologically significant). Further down we show the clusters and gene regulatory network obtained using the method of EM with KF and the next section is dedicated to the EFuNN application on this problem. The biological significance of our results has been discussed in each of these sections. We conclude by discussing the possible hypothesis that may be derived based on our results using each of these methods.

5.1 Introduction and problem specification

As mentioned earlier in this thesis, meaningful inferences may be drawn from the gene expression data and it is important that each gene be surveyed under a variety of conditions, preferably in the form of time series expression in response to perturbations. Range of the computational intelligence methods may be applied for analyzing the gene expression datasets to infer the gene regulatory networks (GRNs). It has been discussed in the previous chapter that traditional models for numerical analysis are difficult to apply. In the same chapter we have also dedicated one section on the existing methods for the modelling of GRN, highlighting some of the drawback they have. Here, in this chapter we have explored the idea of determining whether or not an integrated approach would be more suitable to reveal about the controls of gene regulation (i.e. understanding molecular interactions) and what knowledge can be derived from different models. To select the smaller number of genes (i.e. reducing problem dimension) in addition to apply the appropriate clustering methods, here we have also considered the valuable inputs (i.e. data) from the biological experiments (based on literature). We have analysed the Yeast cell-cycle time series gene expression microarray dataset [Peng et al. 2005] using Least Angle Regression (LARS), Expectation Maximization (EM) with Kalman Filter (KF) and Evolving Fuzzy Neural Network (EFuNN) and report the important results in the individual sections of this chapter. The discovered

knowledge and other findings have been realized so important that we have proposed some of new hypotheses in yeast cell-cycle control direction. For example, using LARS we hypothesize firstly that the exoglucanase gene exg1 is now implicated to be tied with Mlu1 cell-cycle box (MCB) cluster regulation and secondly, a mannosidase gene with histone linked mannoses cluster regulation. A new possible quantitative prediction or hypothesis is that the time delay of the interaction between the two genes seems to be approximately 30 minutes, or 0.17 cell cycles. Using the method of EM with KF, 25 cell-cycle regulated key genes were successfully clustered into three functionally co-regulated groups. We have also identified two genes namely, ribonucleoside-diphosphate reductase large chain (Cdc22) and ribonucleoside-diphosphate reductase small chain (Suc22) that indeed interact with each other and are the potential candidates as a control in Ribonucleotide reductase (RNR) activity. Based on the EFuNN results and integrating knowledge from EM-KF method, we hypothesize that interaction between Suc22, Cdc22 and hypothetical coiled-coil protein (Mrc1) may be mediated by two other genes namely, CDP-diacylglycerol Synthase (Cds1) and RNR inhibitory protein (Spd1). These results were published very recently in an international journal [Chan, Havukkala, Jain, Hu and Kasabov 2008]. We conclude this chapter with a discussion on adopting similar integrative approaches for discovering GRNs and also summarize the crucial results.

5.2 Proposed integrative approach: Least Angle Regression (LARS), Expectation Maximization (EM) with Kalman Filter (KF) and Evolving Fuzzy Neural Network (EFuNN)

To unravel the controlling mechanisms of gene regulation and thus to understand the molecular interactions, in this thesis chapter we undertook the application of sophisticated soft computing methods for inferring gene regulatory networks (GRN) from time series gene expression microarray data. To infer the GRN we have applied three computational intelligence methods - Least Angle Regression (LARS), Expectation Maximization (EM) with Kalman Filter (KF), and an Evolving Fuzzy Neural Network (EFuNN). The methods were applied on time series microarray data of *Schizosaccharomyces pombe* yeast cell cycle genes [Peng et al. 2005]. Each method revealed some new aspects of the problem and it was strongly believed by us that to infer the GRN and to understand the molecular interactions in molecular biology it is more suitable to adopt such integrative approach as ours through which some new knowledge is discovered. The methods discussed and applied here can be used to analyze any kind of short time series of many interacting variables for inferring the regulatory network. Researchers should take such integrative computational intelligence approach seriously to understand the complex

phenomenon of gene regulation and thus to simulate the development of the cell.

As we have discussed earlier, a model that is capable of efficient feature selection and sparse parameter learning – i.e. selecting only a few important explanatory genes or connections – is of particular advantage. By reducing the number of predictors and model parameters, we improve model parsimony and learning efficiency. Moreover, it is known that biological regulatory networks exhibit limited connectivity, proving that sparse models are biologically more plausible [Davidson et al. 2003, Yeung et al. 2002 and Han et al. 2004]. However we have discussed earlier in this chapter but just to remind the readers that currently, only few classes of GRN models are capable of feature selection or sparse learning. They are predominantly graphical models, such as Bayesian Networks [Pe'er et al. 2001, Friedman et al. 2000 and Murphy et al. 1999], Gaussian Networks [Wille et al. 2004] and linear system-based methods such as singular value decomposition (SVD) [Yeung et al. 2002], Lasso developed by Tibshirani [Tibshirani 1996] used for microarray data and their variants. Notably, [Someren et al. 2003] used Lasso successfully to analyze a small set of eight genes over 18 timepoints of yeast cell cycle data. Lasso is one of the new well-performing methods [Huang et al. 2005] and was used successfully for large-scale inference about indegrees and outdegrees of internet data network connections [Gustafsson et al. 2005],

indicating that linear methods work well. Previously, LARS was recommended by [Segal et al. 2003] for gene selection in microarray data for phenotype classification. Recently proposed in [Efron et al. 2004], LARS is a linear system-based method, closely related to the classical automatic model-building methods of Forward Stagewise linear regression [Hastie et al. 2001] and Lasso [Tibshirani 1996]. LARS is capable of producing similar or identical solutions to LASSO, yet it has the advantage that the full path of *p* solutions can be obtained at significantly reduced computational effort, which is important in handling datasets with large *p*.

Explanation and background on LARS

LARS is related to three statistical automatic model-building methods, the Forward Stepwise, Forward Stagewise linear regression [Hastie et al. 2001] and the Lasso [Tibshirani 1996]. A brief review is given on these three methods.

Let **y** be *n*-vector representing the response, $X=[\mathbf{x}_1,..,\mathbf{x}_p]$ be the collection of *p* possible *n*-vector predictors. Both the covariates and the response are normalized to zero mean and unit variance. They are assumed to be related through the linear equation $\mathbf{y} = X\boldsymbol{\beta}$. The problem is to build the model using only a few of the most significant covariates, which can also be viewed as learning a sparse representation of the coefficient vector **β**, so that only few coefficients

are non-zero.

Both Forward Stepwise and Forward Stagewise begin with an empty set of covariates and then add to the set in a greedy manner. In Forward Stepwise, we select the predictor $\mathbf{x}_{j1}$ that has the largest absolute correlation with $\mathbf{y}$, and then perform a linear regression of $\mathbf{y}$ on $\mathbf{x}_{j1}$. This leaves the residual vector $\mathbf{r}$ that is orthogonal to $\mathbf{x}_{j1}$, which is now considered to be the next response. Next, we select the predictor $\mathbf{x}_{j2}$ having the largest absolute correlation with the new response, and so on. After k steps, we obtain k most significant predictors $\mathbf{x}_{j1},...,\mathbf{x}_{jk}$ from the covariate set, which can then be used to construct a k-parameter model with the usual least square method. While computationally efficient, the greedy selection process may be overly coarse, eliminating useful predictors that may be correlated to the already selected predictors.

Forward Stagewise is similar to Forward Stepwise, except that the estimates move in many tiny steps instead of one big step in order to achieve a higher precision in the greedy selection process. Model building is performed over many iterations. In each iteration, we update the estimate in the direction of the predictor having the largest absolute correlation by a small stepsize ε. Its main drawback is that the stepsize ε must be small enough to prevent oscillation, which in turn raises the computational cost. An example of Forward Stagewise is shown in figure 5.1 (top).

Lasso is essentially an ordinary least squares algorithm constrained by L1-penalty.

$$\min \|\mathbf{y} - X\boldsymbol{\beta}\|^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \le t$$

Unlike the more common L2-penalty which suppresses the overall coefficient sizes, L1-penalty drives all but a few to zero, leaving a sparse coefficient vector. However, a major problem is that the control parameter *t* must be adapted for each target gene. An example of Lasso over a range of *t* is shown in figure 5.1 (middle), as analysed in [Efron et al. 2004]. A recently proposed method called the "ElasticNet" [Zou and Hastie 2003] combines both L1- and L2- penalty to achieve grouping effect as well as sparse coefficients, which is an attractive feature for GRN inference. Such application could be investigated in future studies.

While Forward Stagewise and Lasso are effective tools for automatic model-building, their applications have been limited by the cumbersome task of parameter choice – i.e. the choice of stepsize ε for Stagewise and of *t* for Lasso. LARS eliminates these problems, by computing efficiently the full path of the solution (figure 5.1, bottom) and by being parameter-free. In the figure 5.1, Stagewise is evaluated at stepsize ε =2, a rather high resolution to avoid oscillation. Lasso is evaluated at every 200 increments of *t*. LARS is even more efficient, evaluated only at branch points of the solution path.

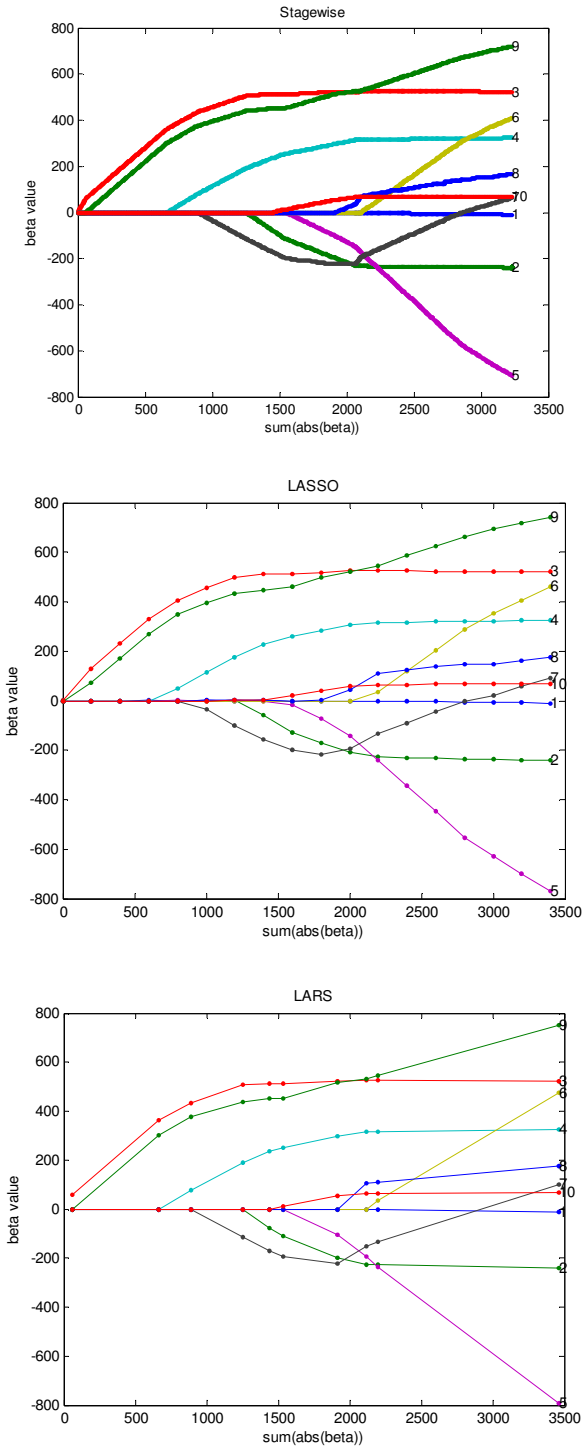


Figure 5.1: Path of the solutions identified by Forward Stagewise (top), Lasso (middle) and LARS (bottom) for the diabetes data, as analysed in [Efron et al. 2004]; the dataset consists of 442 instances and 10 covariates. Each node represents single iteration of the underlying algorithm.

LARS algorithm and its application to GRN inference

For the LARS experiments five genes are omitted (see details later in case study) because more than 50% of their values are missing, leaving 742 genes in the dataset. As with Forward selection methods, we begin with zero coefficients and select the covariate $\mathbf{x}_{j1}$ having the greatest absolute correlation with the response. Next, we move the estimate in the direction of the predictor $\mathbf{x}_{j1}$, with the largest possible stepsize until the next most correlated predictor $\mathbf{x}_{j2}$ has the same correlation as the current estimate. Now having two predictors already selected, we move the estimate in the direction equiangular between $\mathbf{x}_{j1}$ and $\mathbf{x}_{j2}$, with the largest possible stepsize until the next most correlated predictor $\mathbf{x}_{j3}$ has the same correlation as the current estimate. As more predictors are added, we keep proceeding in the direction equiangular between all existing predictors until the next most correlated predictor, and so on. Since one predictor is added in each step, LARS takes a total of only p steps to compute the full set of solutions. The complete description of the procedures of LARS and the formula can be referred in the original paper [Hastie et al. 2001]. Computationally, LARS is very suitable to the task of GRN inference, mainly because of its computational efficiency and capability of evaluating the full path of solutions. For the following experiment on yeast time series data, it takes only 20 seconds with our MATLAB code on a Pentium 2.8GHz to compute the first ten predictors for all 742 genes (32

samples per gene). To achieve the same results, it would take much longer time for Stagewise to iterate in small steps and for Lasso to trial different values of t. By evaluating the full path of solutions, beginning from the most significant predictor to the least, we can evaluate the optimality of different model building criteria, such as selecting only the first ten best predictors, or the best set that achieve a certain optimality index (e.g. prediction error, BIC or AIC).

We apply LARS using two models: one static model and the other, AutoRegressive (AR) model. Static models, such as those used in Bayesian Networks [Pe'er et al. 2001, Friedman et al. 2000 and Murphy et al. 1999] and Gaussian Networks [Wille et al. 2004] are usually used for data that are collected irrespective of time. We assume that the expression of gene *j* at time *t* is a function of the expression of all other genes *i≠j* at the same time instance. The response $\mathbf{y}^{(static)}$ and the covariates $X^{(static)}$ of the model are defined as:

$$\mathbf{y}^{(static)} = \mathbf{x}_j^{(t)}$$
$$X^{(static)} = [\mathbf{x}_1^{(t)},...,\mathbf{x}_{j-1}^{(t)},\mathbf{x}_{j+1}^{(t)},...,\mathbf{x}_p^{(t)}]$$

AR models, such as those used in Dynamic Bayesian Network [Murphy et al. 1999], are used only for time series data. We assume that the expression of gene *j* at time *t* is a function of the expression of all genes inclusive of itself at the previous time instance. The response $\mathbf{y}^{(AR)}$ and the covariates $X^{(AR)}$ of the model are defined as:

$$\mathbf{y}^{(AR)} = \mathbf{x}_j^{(t)}$$
$$X^{(AR)} = [\mathbf{x}_1^{(t-1)},...,\mathbf{x}_p^{(t-1)}]$$

For each model, we identify the top 10 predictors and their coefficients for each gene and investigate their biological implications. For visualization of networks, we use the software WebInterViewer [Han et al. 2004]. In the plots we have shown (figure 5.2), edges are drawn between the target gene and its predictors. Sometimes we constrain the search by plotting only those that rank within the top 20% in terms of the absolute value of the coefficients. Since it is well-known that microarray data are noisy and the GRN problem is ill-poised due to the lack of training samples, identification of the exact marker gene for the regulation of a particular target gene is very difficult (*e.g.* distinguishing the marker gene from its co-regulated genes). Therefore, we focus on analyzing examples of well-known gene groups and overall connections rather than on individual genes. We would like to mention here that rest of the two parts of our integrative approach i.e. Expectation Maximization (EM) with Kalman Filter (KF) and Evolving Fuzzy Neural Network (EFuNN) are discussed later in this chapter and currently in the next section we will continue with the detailed explanation on the case study (experiments with LARS) on yeast cell cycle.

5.3 Case study on Yeast cell-cycle time series microarray dataset to infer GRNs

Yeast is an important model for general biology and also for the drug development - crucial cell cycle component identification is important for example novel antibiotic target selection. This fact is related to the high degree of conservation in evolution between the primitive eukaryote and mammals and also the yeast has almost no non-coding introns. The cell cycle, or cell-division cycle, is the series of events that take place in a eukaryotic cell leading to its replication. These events can be divided in two brief periods: interphase (G1 phase, S phase, G2 phase) - during which the cell grows, accumulating nutrients needed for mitosis and duplicating its DNA and the mitotic (M) phase, during which the cell splits itself into two distinct cells, often called "daughter cells". One may obtain further information on cell cycle by visiting http://en.wikipedia.org/wiki/Cell_cycle. Therefore, microarray gene expression data on the cell cycle of *Schizosaccharomyces pombe* (yeast) [Peng et al. 2005] was used as a case study in this chapter. This dataset is a carefully made time series experiment with ten minutes interval sampling of mRNA for six hours (~two cell cycles) from yeast cells with synchronized cell division. The set of 747 genes selected as the cell cycle related genes is provided in this dataset within the system controlling cell division through M, G1, S, and G2 stages. We have treated any missing values in the same way

as reported in Peng et al. (2005) using the Gaussian smoothing interpolation method. The dataset we have used have fewer biological replicates but we use this data because comparing to the other datasets as analysed in Rustici et al.(2004) and Spellman et al. (1998), it has the most frequently sampled time points which can reduce interpolation errors between time points.

In summary, LARS generated clusters of interactions, that seem to correspond to known gene interactions, and can identify new interesting genes, like the exoglucanase gene linked to MCB motif related regulatory cluster. There are some caveats, though. Very recently, Marguerat et al. (2006) combined data of three *S. pombe* comprising 10 experiments, to show that there are at most about 500 clearly periodic genes. We have analysed the 742 genes from Peng et al. (2005), thus possibly including some non-cycling genes and missing some. Repeating the present LARS analyses with the combined dataset of Marguerat et al. (2006) could further confirm our results.

5.4 Results, discoveries and biological validation

Several strongly linked clusters were examined for their biological verification (interpretation) and were found to be indeed highly likely to exist in nature under the biological conditions tested. Two examples are discussed below. Literature mining for data on orthologous genes from *Saccharomyces cerevisiae* (much better studied yeast than *S. pombe*) was greatly facilitated by

the yeast orthologous groups resource developed by Valerie Wood [Wood 2006] and the YOGY search tool which is a web-based resource for retrieving orthologous proteins [Penkett et al. 2006]. Evaluation of the possible false positives is more difficult, as little negative data has been published on yeast gene interactions, and utililization of for example cellular location data to suggest unlikely interaction partners is not a trivial task. Comprehensive evaluation of false negatives (not reported by LARS, but biologically real interactions) in our regression coefficients is beyond the scope of the present chapter.

Peng et al. (2005) identified 8 of the 9 known histones in an expression cluster. The static LARS model result (refer to figure 5.2 - top) shows an improvement, clustering together all nine histone genes. Few genes seem to be co-regulated simultaneously and multiply linked with the nine histones, which are essential components of chromosomes. The AR-LARS model result (refer to figure 5.2 - bottom) shows that in addition to all nine histones being very well connected, several interconnecting genes are also apparent. The central regulator seems to be *cyp4*, an orthologue of human cyclophilin B (gene 345 = SPBP8B7.25), connecting directly to all nine histones. This is validated by Rustici et al. (2004) also finding *cyp4* in the histone cluster. Gene *cyp4* functions in chaperoning plasma membrane proteins through the secretory pathway [Pemberton et al. 2005]. Relevance to cell division is

plausible, as according to Arevalo-Rodriguez et al. (2005) and Wang and Heitman (2005), a similar gene, cyclophilin A is meiosis related, linking to chromosome duplication and chromatin formation. However in Joseph (1999) *Aspergillus nidulans* cyclophilin B gene is linked to the growth in high stress response.

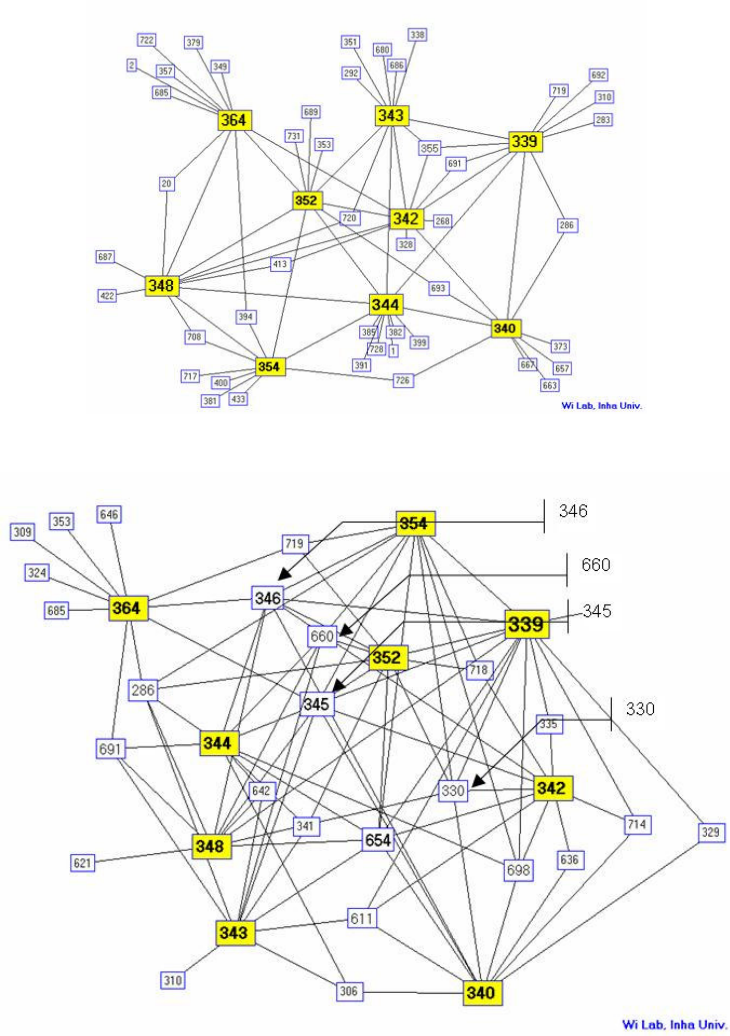


Figure 5.2: The nine histone genes (shaded/yellow boxes with numbers in bold) are interlinked by the LARS algorithm in the static model (top) and the AR model (bottom). Only genes affecting the histone are shown (not those affected by histone).

The proteinase inhibitor gene 346 (SPCC338.12) is similar to the *S. cerevisiae* (YNL015W) and in our results it connects to the eight histones. This is in agreement with Rustici et al. (2004) results. It is reported to be a core environmental stress response (CESR) gene [Chen D. et al. 2003], so its role in the cell cycle is unclear; perhaps certain stages in the cell cycle can be considered as "stress" events. A histone binding protein SPBC577.15c (gene 330, in refer to figure 5.2 - bottom) links to six histones as expected by its function, but it is not among the top 500 cycling genes [Marguerat et al. 2006]. A mannosidase SPAC2E1P5.01c (gene 660 in refer to figure 5.2 - bottom) is presumably related to control of mannoses linked (bound) with histones and may be under same transcriptional control [Clare and King 2002]. The *S. cerevisiae* homologue YJR131w of this mannosidase is also involved in the cell cycle progression by genetic evidence and interacts with yeast transcriptional inhibitor (*MET30)* gene known to control cell cycle progression [Guldener et al. 2005].

Turning to an examination of another gene interaction cluster, Peng et al. (2005) identified 20 genes in an expression cluster, all containing the MCB sequence motif. The MCB motif is a well-known sequence bound by DNA synthesis control (DSC) transcriptional complexes called MCB-binding factors, which are required for the transition from G1 phase to S phase in the cell cycle. Our AR model also connects all the 20 genes very closely, plus two further

genes belonging to the same cluster (refer to figure 5.3 - top).

These two genes seem to be master regulators. The first: a beta-transducin homologue (CDC20 i.e. gene 82) is validated as a well-known cell cycle regulator affecting nuclear movement and chromosome separation. The second, an exo-beta-glucanase (gene 162 - SPBC1105.05, exg1), similar to *S. cerevisiae* (YLR300W, YOR190W and YDR261C) is a novel finding. Corroborating our result, Rustici et al. (2004) also found that exg1 gene is co-expressed with MCB motif genes; the gene however contains a new motif (not MCB), suggesting that it may be regulated by another transcription factor [Marguerat et al. 2006]. Rustici et al. (2004) also mention an endoglucanase, SPAC821.09 (eng1) and SPAC14C4.09 (a putative glucanase) being "activator of yeast metallothionein expression" (Ace2) dependent. The gene eng1 (endo-glucanase) is known to help to digest the division septum and separate the cells after cell division [Martın-Cuadrado et al. 2003]. The exoglucanase might be related to similar functions, or cell wall assembly/growth, but this would have to be verified experimentally. In summary for the MCB-cluster, LARS identified known and plausible regulators, and discovered a novel connection of an exoglucanase gene to MCB-regulated genes and cell wall restructuring.

Furthermore, when all the 90 genes affecting any of these 20 genes were analyzed for their Gene Ontology annotation bias [Berriz et al. 2003], it was found that they include a set of genes statistically significant for cell division and cytokinesis (Table 5.1). This further confirms that the AR model extracted biologically relevant subsets of linked genes. Verification of any of these links could be done by yeast knockout models or specific gene pertubation methods. When only 20% of the strongest connections included in the analysis for the network of the MCB motifs genes (AR model), 17 out of the 22 MCB motif genes were remained in the network diagram (comparing top and bottom diagram of figure 5.3). This gives credence to the significance of these links as the major correlated genes. This is further supported by noting that the central linkages to genes 82 (CDC20) and 162 (exg1) are also retained in the top 20% strongest links.

Table 5.1: MCB motif genes: GO annotation bias compared to the complete set of 742 genes in Peng et al. (2005). Dataset is based on FuncAssociate analysis [Berriz et al. 2003].

MCB motif genes (20 genes)			
Genes found	Total genes	*P* value	GO annotation
8	25	<0.001	0006260: DNA replication
6	21	0.002	0051320: S phase
90 genes linked to MCB motif genes			
Genes found	Total genes	*P* value	GO annotation
15	42	0.007	0000910: cytokinesis / cell division

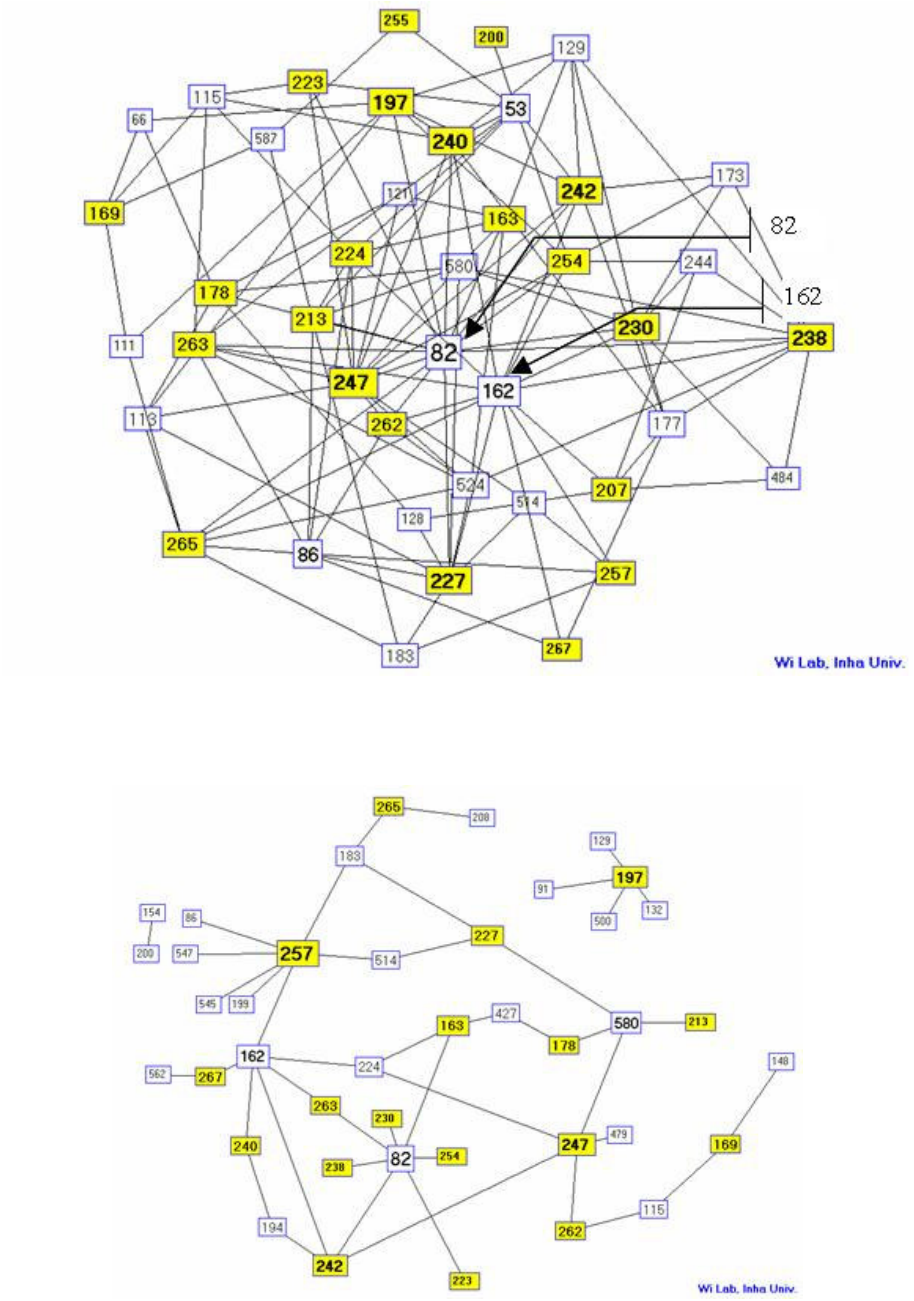


Figure 5.3: The MCB motif containing genes are linked to each other (AR model), and to many other genes of similar Gene Ontology annotation. MCB motif genes are the shaded/yellow boxes. In top – “all ten strongest links included” and all 20 MCB motif genes are closely linked. In bottom - only top 20% strongest links included, still 17 of 22 MCB motif genes are closely linked.

Evidence of Serial Regulation in the Yeast Cycle

Finally, the original dataset from Peng et al. [Peng et al. 2005] was sorted sequentially according to the peak expression time along the cell cycle. Therefore, the links between genes co-expressed or co-varying around the same time should have correlating identification numbers. Top graph in figure 5.4 shows that this is clearly the case between the gene number and its strongest linked gene (only regression coefficients >0.4 in the AR model results included). The points along the diagonal are mostly linked genes with positive coefficient. Genes that are slightly below the diagonal are indicating a short time delay in the co-variation and suggesting co-regulation. The two other groups of parallel points to the diagonal indicate genes that are negatively regressed to the genes with peak expression approximately half a cell cycle later, at maximum phase difference, as expected. Thus the LARS-calculated regulatory network revealed two distinct main groups of genes: one group of upregulating genes (positive feedback) at the same stage of cell cycle, the other group of downregulating genes (negative feedback) at half cell cycle away. This control seems continuous, so that at any phase of cell cycle there are both types of genes being expressed.

The bottom graph in figure 5.4 reveals further the following predictions (that could be verified experimentally later): 1) the number of strong positive

feedback interactions is 50% more (114 vs 78) than the number of strong negative feedback interactions, 2) the strongest positive interactions occur about 0.17 cell cycles later, suggesting 30 minutes (= 0.17×180 minutes, the duration of the cell cycle in the original experiment [Peng et al. 2005]) time lag between peak of regulator and peak of regulated gene, 3) the majority of strong negative interactions occur between genes that are half a cell cycle separated, possibly keeping the cell cycle in synchrony for all relevant genes throughout the cell cycle, 4) some negative interactions have a time lag of 30 minutes, representing perhaps genes with more complex multimodal expression patterns. It is interesting to note that several CDK/cyclin regulators and regulatory transcription factors also show interactions with genes about half a cell cycle ahead [Simon et al. 2001]. The number of such interacting genes that show peaks half a cell cycle apart may be larger than previously suspected. Further functional analysis of the subset of genes in figure 5.4 might clarify this.

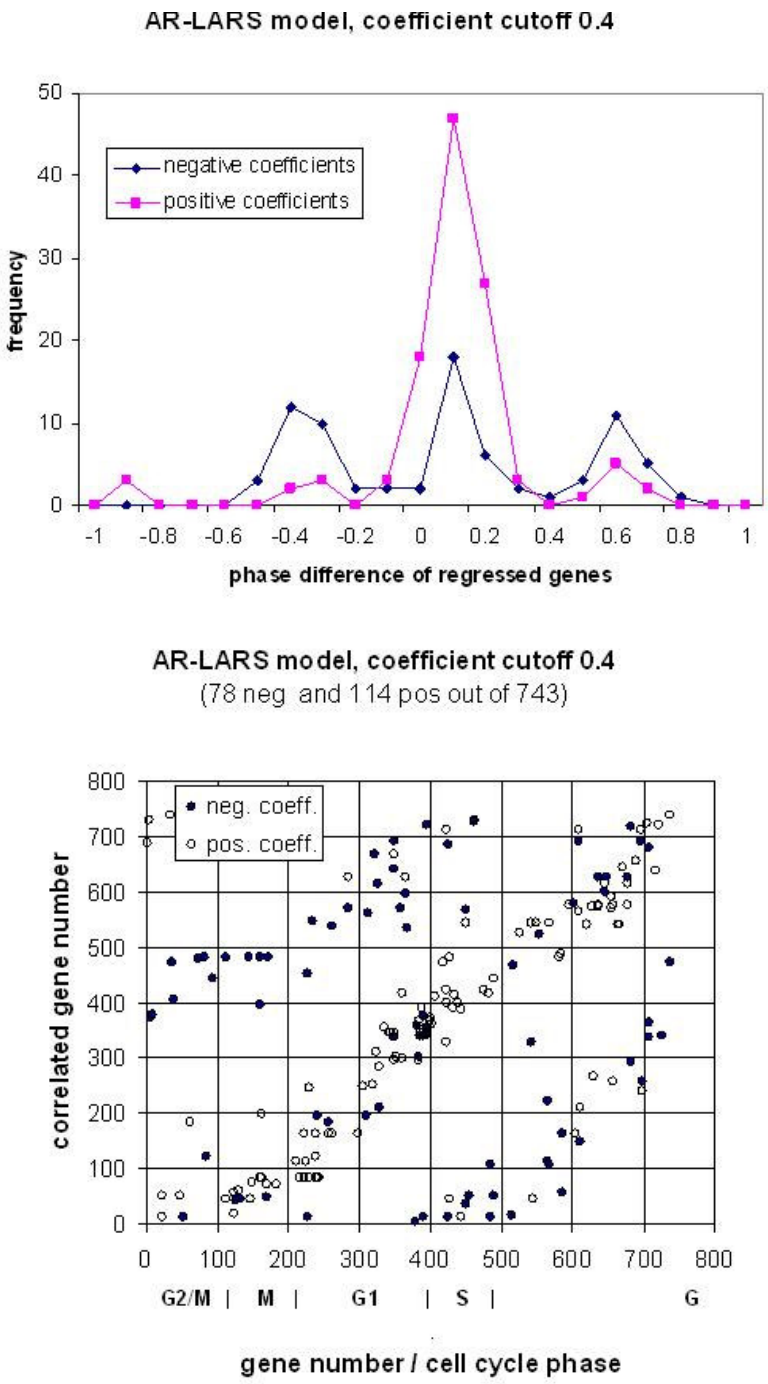


Figure 5.4: Original gene ordering by cell cycle phase is recognized by the strongest links discovered by AR-LARS model. Top is the graph for phase difference in regressed genes. Bottom is the graph for regressed genes (on x-axis: gene identification number and on y-axis: regressed gene strongest linked gene are shown).

5.5 Application of EM with KF and important findings

Next, we consider some of the checkpoint controls as a part of model reliability and its validation. Therefore to observe the regulatory interactions at a closer level and to reduce the search space for the clustering solutions, we selected only the 25 genes controlling cell cycle-regulated transcription from a pool of 742 genes that was analysed by LARS. However we do not claim this list is exhaustive but this subset of cell cycle-regulated transcripts is in accordance with the published experimental evidence of *S. pombe* genome as reported in [Wood et al. 2002]. Many of these genes also contain Mlu1 cell-cycle box (MCB) consensus sequences in their promoters. As discussed earlier, MCB motif is a well-known sequence bound by DNA synthesis control (DSC) transcriptional complexes called MCB-binding factors, which are required for the transition from G1 phase to S phase in the cell cycle.

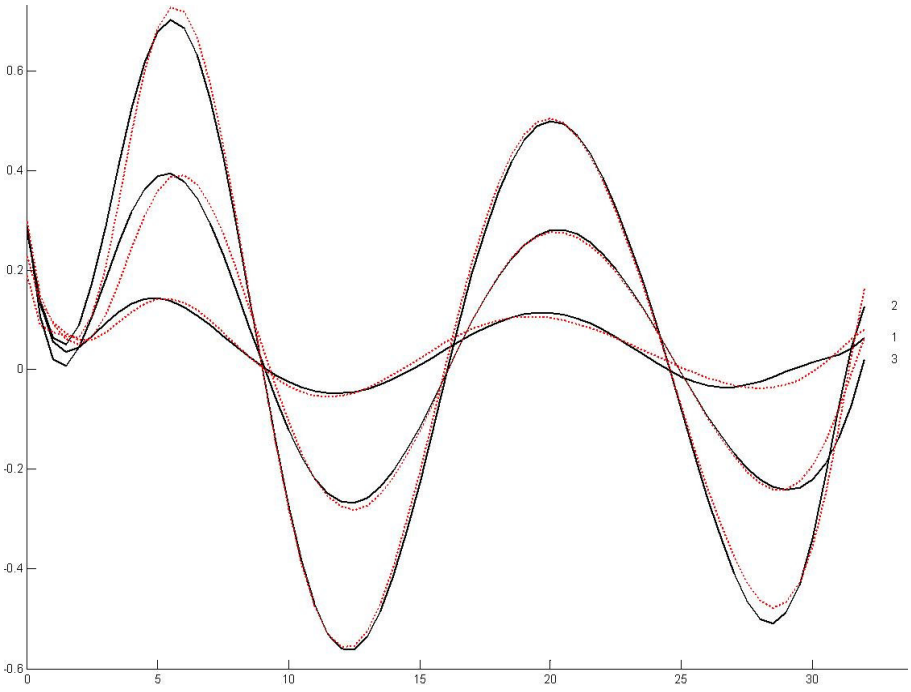
Expectation Maximization (EM) algorithm and Kalman Filter for GRN Modelling

After applying LARS, we also apply a two-stage methodology that is implemented in our software "Gene Network Explorer (GNetXP)" for extracting GRNs from gene trajectory data [Chan, Kasabov and Collins 2005]. However, GNetXP is also equipped with Hybrid Algorithm (power of Genetic Algorithm -

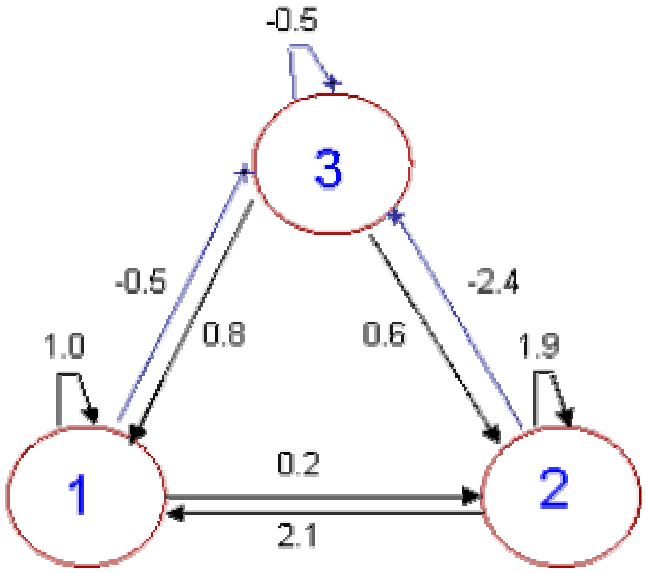
GA with Expectation Maximization - EM) algorithm), but for our experiments we avoided its use mainly because (a) after some attempts we actually found that EM is sufficient for this small subset of 25 genes to model trajectories (b) it can significantly reduce computation time. Therefore, in the first stage, we apply EM on clustering the number of gene trajectories using the mixture of multiple linear regression models for fitting the trajectory data. In the second stage, we apply the Kalman Filter to identify a set of first-order differential equations that describe the dynamics of the representative trajectories, and use these equations for discovering important gene interactions. GNetXP has been developed in KEDRI based on the method that we have described for the understanding of molecular interactions from Leukaemia dataset earlier in this chapter, more specific details about the method implemented in this software are available from [Chan, Kasabov and Collins 2005]. Gene relations can be elucidated from the obtained transition matrix. Significant gene interactions can be identified as those elements whose absolute value is greater than a pre-defined threshold. Such information may also be expressed in a network diagram. Table 5.2 lists all the 25 genes we selected for our experiments and the grouping results show that they are distributed into three different clusters, their molecular functions are also listed along with indication of their potential role in certain cell-cycle phase.

Table 5.2: Gene clustering obtained in ten runs of the standard EM

No.	Gene Symbol	Cluster	Molecular Function	Phase
1	SPAC24H6.05 | cdc25	1	M-phase inducer phosphatase	G2/M
2	SPAC23C11.16 | plo1	1	serine/threonine-protein kinase plo1 (EC 2.7.1.-)	M
3	SPAC821.08c | slp1	2	wd-domain protein; CDC20/p55CDC/Fizzy homolog	M
4	SPAC20G8.05c | cdc15	3	phosphoprotein; subcellular localization of GFP fusion-Cytoplasm dot and septum	M
5	SPBC2F12.11c | rep2	1	transcriptional activator, zinc finger	M
6	SPAC24B11.11c | sid2	1	putative serine/threonine protein kinase	G1
7	SPBC14C8.07c | cdc18	2	cell division control protein 18	G1
8	SPBC887.14c | rph1	1	rrm3-pif1 helicase homolog	G1
9	SPBC32F12.09 | rum1	1	rum1 protein	G1
10	SPBC428.18 | cdt1	3	cell division cycle protein cdt1; replication factor	G1
11	SPCC338.17c | rad21	1	double-strand-break repair protein rad21	G1
12	SPCC550.13 | dfp1	1	dbf4 homolog, subunit of Hsk1 protein kinase	G1
13	SPAC17H9.19c | cdt2	3	target of Cdc10 transcription factor: coupling START with cytokinesis; WD domain	G1
14	SPAC694.06c | Mrc1	3	hypothetical coiled-coil protein	G1
15	SPCC1442.01 | ste6	1	guanine-nucleotide releasing factor, Ste6p	G1
16	SPAC144.13c | srw1	1	WD domain containing srw1 protein	G1
17	SPCC290.04	3	putative transcriptional regulator; zinc finger	G1
18	SPAC19E9.02 | fin1	1	putative G2-specific serine/threonine specific protein kinase (EC 2.7.1.-); promoter of chromatin condensation	G1
19	SPAPB2B4.03 | cig2	2	g2/mitotic-specific cyclin cig2/cyc17	G1
20	SPBC660.14 | mik1	3	mitosis inhibitor protein kinase mik1	G1
21	SPAC1F7.05 | cdc22	2	ribonucleoside-diphosphate reductase large chain	G1
22	SPBC1105.17 | cnp1	1	probable histone h3 variant	G1
23	SPBC25D12.04 | suc22	1	ribonucleoside-diphosphate reductase small chain	G1
24	SPBC4C3.12 | sep1	1	hnf-3/fork head transcription factor homologue.	G2
25	SPCC4B3.15 | dmf1	1	septum positioning protein Dmf1p	G2



(a)



(b)

Figure 5.5: (a) Actual (black) and KF simulated (red – dotted) trajectories for three clusters; X axis: time points and Y axis: expression values (b) Network diagram obtained using the gene interaction information available from transition matrix (within circle: gene clusters and numbers on lines shows positive or negative impact b/w clusters)

After limited trials, we found that the interactions between clusters are most easily elucidated using three clusters. We use the given settings of GnetXP for modelling gene trajectories, which are as follows: desired number of three clusters, ten as the number of coefficient (the higher the number of coefficients, the higher the modelling precision and the lesser the curve smoothness) with a value of 0.1 as a stopping threshold (amount of log likelihood increase between iterations below which the EM algorithm exists). For our experiments each EM evaluation required less than ten seconds (running in MATLAB on a Pentium IV 2.8GHz, 2GB RAM) and algorithm was run with a total number of ten times. First of all as shown in figure 5.5 (a), note that the model tracks closely to the actual trajectories showing that the first-order differential equations are sufficient even for the complex trajectories in this case. Network information is represented as a diagram in figure 5.5 (b) where we have displayed three clusters (each cluster can have more than one gene) and the influence (positive or inhibitory) of one cluster over the other.

5.5.1 Biological interpretation of results

For the biological interpretation and validation of our results, we have carefully examined some examples of gene groupings and the interaction between these clusters. These groupings are based on the findings from *S. pombe* gene expression studies [Wood et al. 2002], which show that certain key cell cycle-regulated transcripts are expressed in a similar fashion during a

time-course experiment and should thus be clustered together upon microarray analysis.

When we carefully analysed cluster one it was found that out of its total 15 genes, ten were having the assigned function of protein binding in the gene ontology (GO) annotation. Two genes were found to be involved in the DNA binding process and one act as a transcription factor for cell cycle phenomenon. Gene - WD domain containing srw1 protein “SPAC144.13c” was not properly annotated and only one gene - guanine-nucleotide releasing factor, Ste6p “SPCC1442.01” was found to have the guanyl-nucleotide exchange factor activity which is a bit different from rest of the genes of this group. Therefore, considering these biological evidences, it was obvious for most of the genes from this group to be clustered together. Turning to the examination of another cluster, we found that all the four genes from cluster two were involved in the very similar processes like cyclin-dependent protein kinase regulator activity, mitotic cell cycle spindle assembly checkpoint or ATP binding that are essential for the maintenance of the cell-cycle process which is again a very good biological support for them to be grouped with each other. Cluster three has the six genes and is comparatively more heterogeneous group. Phosphoprotein; subcellular localization of GFP fusion-Cytoplasm dot and septum “SPAC20G8.05c” and target of Cdc10 transcription factor: coupling START with cytokinesis; WD domain “SPAC17H9.19c” are involved

in the protein anchoring and bridging functions respectively. Another two genes - hypothetical coiled-coil protein “SPAC694.06c” and mitosis inhibitor protein kinase mik1 “SPBC660.14” are involved in the protein serine/threonine kinase activity. Putative transcriptional regulator; zinc finger “SPCC290.04” has zinc ion binding property and cell division cycle protein cdt1; replication factor “SPBC428.18” is not properly annotated in the GO.

When we looked into sophisticated details about the information on the interactions between the clusters of these genes, it was found from [Håkansson et al. 2006] that S phase delayed “Spd1p” inhibits fission yeast ribonucleotide reductase (RNR) activity by interacting with (binding to) the Cdc22 (“SPAC1F7.05” in cluster two) which is the large subunit of RNR and Spd1 also inhibits ribonucleotide reductase (RNR) activity by anchoring the small RNR subunit Suc22 (“SPBC25D12.04” in cluster one) in the nucleus [Liu C et al. 2003]. This is very clear suggestive evidence that genes from the cluster one and cluster two interact with each other as a control of ribonucleotide reductase activity, however involvement of some intermediate genes is not surprising in this process. It justifies up to much extent our findings of gene regulatory network as shown in figure 5.5 (b). Elucidation of the mechanisms of these processes may have important implications for our understanding of the fundamental mechanisms of *S. pombe* cell cycle. It is understood from this analysis that progression of cell cycle is regulated by a

wide range of mechanisms such as protein binding, cyclin-dependent protein kinase regulator activity, and mitotic cell cycle spindle assembly checkpoint etc. Key genes have been found to play a pivotal role as a controlling switch.

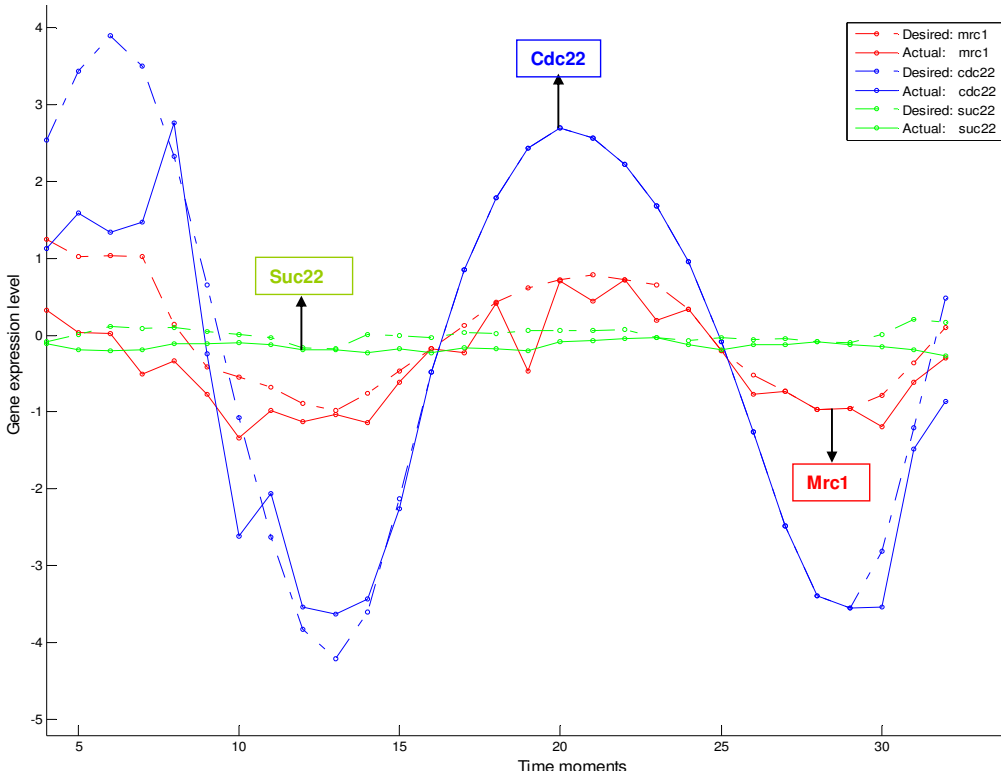
5.6 Application of EFuNN and important findings

We have seen above that the results of Expectation Maximization (EM) algorithm using GnetXP on selected subset of 25 genes are quite promising and are biologically significant. Considering the fact that the biological processes are too complex for the existing computational models, we thought to take an one-step-ahead approach to understand gene regulation of *S. pombe* (and specifically the phenomenon of ribonucleotide reductase regulation) by inferring even smaller gene regulatory networks of individual genes (rather than cluster of genes). We aimed at obtaining a very small gene network that requires the model to evolve both its structure and functionality in time. To do so, we decided to pick only one appropriate key gene from each of the three clusters obtained using EM-KF above. It was quite obvious to select Suc22 ("SPBC25D12.04" from cluster one) and Cdc22 ("SPAC1F7.05" from cluster two) as they are potential GRN candidates according to [Håkansson et al. 2006 and Liu et al. 2003] in ribonucleotide reductase regulation. As we stated earlier that cluster three was little heterogeneous so we chose – Mrc1 (SPAC694.06c) from this group as based on the literature survey this gene is

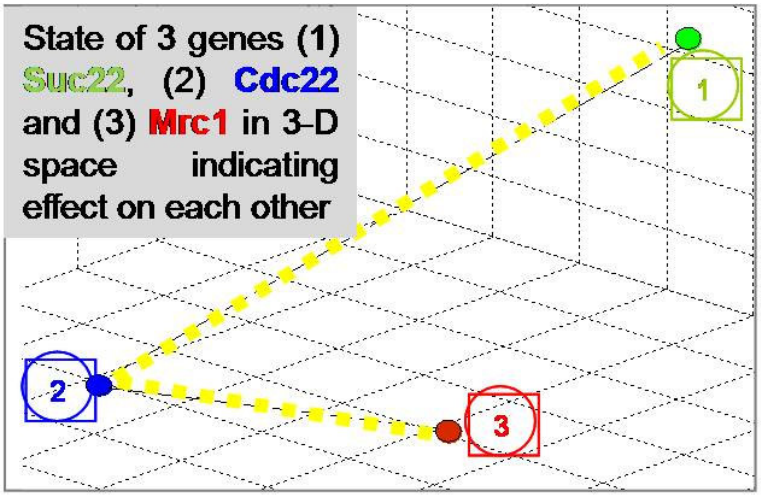
found to be involved in wide range of functions like protein serine/threonine kinase activator activity and DNA binding etc.

To obtain a model that evolves both its structure and function in time, we applied a approach of the *evolving connectionist systems (ECOS)* model Evolving Fuzzy Neural Network (EFuNN) for this task. The method has been developed and patented by Kasabov. It is described in [Kasabov 2006, 2005, 2003, 2004, 2001a,b,c,d; Kasabov and song 2002; Kasabov and Dimitrov 2002]. To discuss the details of EFuNN is beyond the scope of this thesis chapter but in general, an evolving connectionist system is a neural network that operates continuously in time and adapts its structure and functionality through a continuous interaction with the environment and with other systems according to: (a) a set of parameters that are subject to change during the system operation; (b) an incoming continuous flow of information with unknown distribution; (c) a goal (rationale) criteria (also subject to modification) that is applied to optimise the performance of the system over time. Our task is to find the gene regulatory network of three genes: g1= ribonucleoside-diphosphate reductase small chain “Suc22” (belongs to cluster one from KF results), g2= ribonucleoside-diphosphate reductase large chain “Cdc22” (present within cluster two from KF results), g3= hypothetical coiled-coil protein “Mrc1” (grouped in cluster three from KF results) while taking into account the integrated influence of expression values over time.

After the EFuNN was trained on a selected subset of above three genes of *S. pombe* time course gene expression data, rules that express transitions of gene states over time were extracted from it. The EFuNN model for predicting the gene expression level at the next time moment was trained for *n (n = 26)* runs in our experiment. During each run, value of the expression level of all the three genes at three time moments (t-3, t-2 and t-1) was used as the input vector to train the EFuNN model for predicting their expression levels at the next time moment (t). Once the training was successfully finished, an EFuNN model was obtained that was used for the next run of training. Then, the time lag was shifted to next three time moments. In this manner, after 26 runs of training, an optimized EFuNN model was finally obtained for the independent testing of expression level of the above mentioned genes at last three time moments. For comparison, we have also used the final trained EFuNN model to test all the expression values of these genes from time moment four to the time moment 32. It is shown in figure 5.6 (a) that describes the actual and predicted expression patterns of three genes over time using the EFuNN model.



(a)



(b)

Figure 5.6: (a) Gene expression level (mRNA expression) over time (in minutes) when all the 3 genes were tested using EFuNN model (dotted line: predicted; continuous line: actual) (b) EFuNN model showing the effect of three key genes upon each other. Each point represents a state of the 3 genes used in the model and the lines are representing (rules) transitions of the states (example is shown at the time moment 31).

The rule nodes in an EFuNN capture information of input genes that are related to the output genes at next time moment. We extracted rules from the trained structure that describe the transition between the gene states in the problem space. We will not talk in detail about these rules because the aim of this thesis chapter is to represent the final knowledge on molecular interactions that we have obtained using different methods but readers may obtain more information on the kind of rules that may be extracted using EFuNN from [Kasabov and Dimitrov 2002; Kasabov 2001 and Kasabov 2003]. In summary, these rules are linked to each other in terms of time links of their creation, thus representing the GRN. Figure 5.6 (b) is representing the knowledge that we have derived from these rules by showing how these important genes may affect each other in terms of their expression at a certain time moment (here as for example we have shown this phenomenon at the time moment 31). It is clearly shown here using EFuNN model that these key genes have some sort of impact on each other. Functional states and interaction events of Suc22 and Cdc22 have already been discussed by us in the previous section of this chapter and using Pubmed-Entrez we further verified the biological properties of Mrc1 which showed us that this gene also have interacting properties with some other genes like Cds1. Thus considering the EM-KF results discussed earlier and GRN obtained using EFuNN model, we hypothesize that the genes like Cds1 and Spd1 may act as intermediate molecules for the interaction between our selected three key genes namely Suc22, Cdc22 and Mrc1.

Biological laboratory experiments may help further to verify such results and thus to further understand the controlling points of yeast cell-cycle.

5.7 Conclusion

A more integrative approach is suggested for the analysis of any microarray time-series gene expression dataset. For the first time, LARS is applied to the GRN inference of *S. pombe* cell cycle microarray data. The algorithm successfully extracted regulatory networks that seem to be biologically relevant and functionally correct. New knowledge was gained, as novel genes were linked to histone and MCB related gene clusters. The static and autoregressive (AR) LARS model gives consistent results, but, as expected, AR model is better and provides more accurate known regulatory genes related to cell cycle. The LARS method seems an attractive alternative to traditional regression methods, like Lasso. As applied to inference of gene regulatory networks, LARS could be developed further by internal validation and weighting of the regression coefficients, as proposed in another context in [Tikka 2004]. Different timelags could be added to improve the model fitting. The time series smoothing could be improved by using cubic splines [Bar-Joseph 2004], instead of Gaussian smoothing. Incremental on-line and multipass methods could be derived as suggested in [Balakrishnan and Madigan 2006]. Further, several microarray datasets could be integrated by using LARS regression models, as done by [Gilks et al. 2005] for standard

multivariate regression models.

Using our integrative approach some new knowledge is discovered on yeast cell cycle, such as: using LARS we could produce biologically relevant known gene regulatory networks and we hypothesize – first, an exoglucanase gene exg1 is now implicated to be tied with MCB cluster regulation and second, 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. Gene Ontology functional annotation supported the relevance of linked subsets of genes. Global analysis discovered that co-varying genes were of two major types: putative up-regulators, correlated positively to genes at around the same time in the cell cycle, and putative down-regulators, correlated negatively to genes about half a cell cycle later. These down-regulators may include the far ahead looking inhibitors that effectively control and synchronize the cell cycle continuously in succession, keeping the process in track. Next, using the method of EM with KF 25 cell-cycle regulated key genes were successfully clustered into three functionally co-regulated groups. We have also identified two genes namely Cdc22 and Suc22 that indeed interact with each other and are the potential candidates as a control in Ribonucleotide reductase (RNR) activity. This phenomenon was further studied with the application of EFuNN algorithm. Our model was trained and tested on three key genes (Suc22,

Cdc22 and Mrc1). Upon testing, the performance was found very good and based on integrating the different aspects of obtained knowledge we hypothesize that the interaction between these three genes may be mediated by two other genes namely Cds1 and Spd1.

It seems that the proposed integrated approach for inferring gene regulatory network makes it suitable to be applied on proteomics and metabolomics time-series to derive possible regulatory networks between genes, proteins and various metabolites. Ultimately, information fusion methods (like our BGO, discussed later in the chapter 8) could be used to link all these interaction networks for a global integrated analysis. Such approach can be easily extended by incorporating other novel computational intelligence methods (based on their computing efficiency and parameter-free operation like LARS) and comparative analysis may also be accounted. As per our proposed framework (refer to chapter 3), in the next chapter we have extended the investigation by implementing another integrative approach for obtaining meaningful GRNs.

6. Studying LTP related GRNs using quantum inspired evolutionary algorithm (QiEA) and clustering analysis

As per our previously described framework (refer to chapter 3), in this chapter we have extended the GRN inference investigation by implementing another integrative approach for obtaining meaningful results. Here, we will discuss the application of our integrative approach of (1) clustering based on gene expression profiles and genes, promoters and proteins sequence data and (2) quantum inspired evolutionary algorithm (QiEA) to analyse the mouse time series gene expression dataset [Park, Gong and Tang 2006] for studying long term potentiation (LTP) related GRNs. In this respect, we will describe in detail, the process of gene selection and discuss the obtained results through clustering etc. Next, we describe the application of QiEA to predict GRNs and perform some virtual gene knock-in mice experiments. We conclude the chapter by summarizing and discussing some important results.

6.1 Introduction and problem specification

In the previous chapter we have determined whether or not an integrated approach would be more suitable to reveal about the controls of gene regulation (i.e. understanding molecular interactions) and what knowledge can be derived from the different models. We learnt that each computational intelligence method may reveal some new aspects of the problem and we understood that to infer the GRN and to understand the interactions in molecular biology it is more suitable to adopt the integrative approaches

through which some new knowledge may be discovered. Here, we have applied a different computational intelligence integrative approach to understand the molecular interactions (GRN) on another gene expression data (time course of mouse DNA microarray) and our findings were supported by rigorous literature and database knowledge analysis. We have selected the genes purely based on literature data, thus considered the valuable inputs from the biological experiments (based on pubmed). We investigated the selected subset of 79 genes out of 12,488 activity regulated genes and results unravel the genetic and molecular mechanisms of synaptic plasticity, which are the basis of learning and memory. The genes were selected based on their demonstrated link to long-term potentiation and/or learning&memory disorders (Alzheimer's disease, Fronto-temporal dementia, X-linked mental retardation, Rett syndrome and schizophrenia), however we do not claim this list is exhaustive. First, we clustered our selected 79 genes based on their temporal profile of gene expression to obtain 14 clusters with co-expressed genes. We have performed the functional analysis of genes within clusters based on NCBI Entrez Gene Database. Based on the functional analysis and temporal profile of gene expressions within each cluster and available literature, we have made a link between gene functions and different stages of induction, stabilization and maintenance of long-term potentiation. Functional analysis was further backed up with the clustering analysis based on the gene and protein sequences to identify similarity of genes and their proteins. To investigate whether genes within clusters can be co-regulated, we have also investigated the similarity of promoter regions of genes within clusters. Results of this analysis show that some genes within clusters can be indeed

co-regulated based on the similarity of their promoter regions. By using a novel optimization method called quantum inspired evolutionary algorithm (developed at Centre for Neurocomputing and Computational Intelligence, KEDRI - AUT), we have inferred an abstract matrix of interactions between temporal clusters of genes, which we have validated using documented interactions. The same optimization method was used to run virtual transgenic and gene knock-in mice experiments. Predictions on which genes are up- or down-regulated as a result of the other gene(s) mutations were verified based on available literature. Our approach (it is one of the method in the proposed framework that has been described previously) that we have suggested in this thesis chapter has a general application to suit the analysis and model development scenario on any time-series gene expression dataset accompanying any cellular process to obtain further insights into the underlying genetic and molecular mechanisms. In addition, the abstract interaction network can serve to simulate virtual gene knock-out and knock-in experiments to predict the effect of mutations upon the rest of interacting genes. Below in this chapter, we will discuss the gene selection process that we have adopted for this research, which is followed by a section on clustering of genes and their functional analysis. Then the application of QiEA to predict GRNs and gene knock-in mice experiments are presented and we conclude the chapter by discussing and summarizing some of our important results that we have obtained.

6.2 Case study: Mouse LTP time series microarray dataset analysis to infer GRNs

Activity-dependent synaptic plasticity is a process in which synapses (connections between neurons) change their efficacy as a consequence of their previous activity. At present, changes in the efficacy of excitatory synapses are thought to be fundamental to information storage within neuronal networks of the brain. In hippocampus and cerebral cortex, long-term synaptic potentiation (LTP), a long-lasting increase in synaptic efficacy, is produced by high-frequency stimulation (HFS) of presynaptic afferents [Bliss and Lomo 1973]. To unravel the genetic and molecular mechanisms of synaptic plasticity, we have analysed time course DNA microarray data of a selected subset of 79 genes out of 12,488 activity regulated genes (ARGs) [Park, Gong and Tang 2006]. ARGs are defined as being up-regulated or down-regulated by various experimental stimuli including LTP-inducing tetanic stimulations, electroconvulsive seizures, KCl-mediated membrane depolarization, N-methyl-D-aspartate stimulations, and learning tasks. Park et al. (2006) induced a stable LTP lasting for 2 h in the mini slices made from the mice hippocampal dentate gyrus. Total RNA was extracted from frozen dentate gyrus mini slices following the electrophysiology. Time course DNA microarray analyses were performed to determine the temporal expression profiles of ARGs in response to LTP-inducing tetanic stimulation. The authors have clustered genes according to their function. They have also discovered that these ARGs are clustered on chromosomes, and these ARG clusters are conserved during evolution. In addition they have discovered that ARGs

specific clusters have different molecular properties, but they are functionally corregulated by the cAMP-response element-binding protein (full details of the study can be found in Park et. al. 2006). We were kindly provided with original microarray data by Dr. Chang Sin Park from University of California at Irvine to perform further analysis by bioinformatics and computational intelligence means to dig out more knowledge about genetic and molecular mechanisms of LTP. The genes for our analysis were selected from Park et al's data based on their documented link to LTP and/or learning&memory disorders (Alzheimer's disease, Fronto-temporal dementia, X-linked mental retardation, Rett syndrome and schizophrenia), however we do not claim this list is exhaustive. First, we have clustered this subset of genes based on their temporal profile of gene expression over time to obtain 14 clusters with co-expressed and thus supposedly co-regulated genes. Functional analysis of genes revealed that all these temporal clusters are functionally heterogenous. In other words they contain genes which code for proteins with different functions. Based on the functional analysis and temporal profile of gene expressions and available literature, we have made a link between gene functions and different stages of induction, stabilization and maintenance of long-term potentiation. We have done the clustering analysis of genes and their coded proteins within each temporal cluster based on the gene sequences to identify evolutionary closeness of genes/proteins within each cluster. Not all genes / proteins within the same clusters were related but half of them were. A stronger indicator of co-regulation is the similarity of promoter regions of the genes within clusters. By this kind of analysis, we have

identified certain genes within each cluster that can be indeed co-regulated based on the similarity of their promoter regions.

Next, we have focused on regulatory interactions between identified temporal gene clusters. To infer the regulatory interaction coefficients between gene clusters, we have employed a new method of quantum inspired evolutionary algorithm (developed at Centre for Neurocomputing and Computational Intelligence, KEDRI - AUT), in which each solution represents all the possible solutions with a certain probability. To provide more details on methodology is beyond the scope of this chapter but it can be referred in [Defoin-Platel, Schliebs and Kasabov 2007]. These probabilities evolve according to the fitness function based on regulatory interaction coefficients leading eventually to the optimal match of model to expression profiles of gene clusters. Inferred regulatory coefficients have been validated using the information about interactions between individual genes within and between clusters from NCBI Gene Entrez database. We used optimized regulatory interaction matrices to simulate virtual transgenic and gene knock-in mice experiments with genes that are within our subset of 79 selected genes. Based on these virtual experiments we make predictions about the effect of mutated gene upon other gene's expression levels, some of which can be validated based on available literature and the rest can serve as a basis for future experimental testing.

6.2.1 Method for gene Selection

The main criterion for gene selection was the documented role in the LTP induction and/or learning&memory disorders (Alzheimer's disease,

Fronto-temporal dementia, X-linked mental retardation, Rett syndrome and schizophrenia) by additional works. In the following text the selected genes are indicated in bold, and the full list is in table 6.1. The induction of LTP in dentate gyrus requires Ca^{2+} influx through the *N*-methyl-D-aspartate (NMDA) glutamate receptors (subunits **Grin1, Grin2a, Grin2b**). The short-lasting form of LTP or early E-LTP, which precedes the lasting LTP, requires the participation of a Ca^{2+}/calmodulin-depenendent protein kinase II (CaMKII, **Camk2a** - subunit α) [Miller et al. 2002]. CaMKII in the basal state is completely dependent upon Ca^{2+}/calmodulin for its activity, but upon activation can rapidly convert to a Ca^{2+}-independent kinase by autophosphorylation, which is necessary for the induction of E-LTP. A molecular basis of E-LTP is the insertion of new "alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors" (AMPARs) into the postsynaptic membrane [Shi et al. 1999] (α subunit **Gria1**). AMPARs mediate most of excitatory postsynaptic response in glutamatergic synapses; therefore changing their number (and/or properties) is a powerful way to control the strength of synaptic transmission. Insertion of new AMPAR occurs through endocytosis, i.e. fusion of storage vesicles with new AMPARs with the postsynaptic membrane [Lledo, Zhang, Sudhof, Malenka and Nicoll 1998]. Ca^{2+} and CaMKII play a key role in this process [Sudhof 1995]. Localization of excitatory synapses on spines in turn leads to induction of strong intra-spine electric field that can electrophoretically drive vesicles for fusion with the postsynaptic membrane to bring new membrane and receptors for postsynaptic density and spine enlargements [Benuskova 2000].

The long-lasting form of LTP or L-LTP requires the activation of the cAMP signalling pathway through **PKA** (cAMP-dependent protein kinase, protein kinase A), ERK/MAPK (extracellular signal-regulated protein kinase/ mitogen-activated protein kinase, **MAPK**/**Mapk1**/**Erk2**/**ERK** and **Mapk14**/**p38**), and RSK2 (ribosomal S6 kinase 2) (the latter not found in the Park's data). Activation of gene transcription via **CREB** (cAMP-responsive transcription factor) pathway follows [Mayford and Kandel 1999]. Constitutively active CaMKII (**Camk2a**) activates a related kinase, CaMKIV (**Camk4**) that contributes to early Ca^{2+}-stimulated CREB phosphorylation [Wu, Deisseroth and Tsien 2001]. CAMP is activated through Ca^{2+}/**calmodulin**-dependent adenylyl cyclase (calmodulin, **Calm3**). (Alternatively, a neurotransmitter or neuromodulator like dopamine binds to G-protein-coupled receptors. G-protein activates adenylyl cyclase, which catalyzes production of cAMP). CAMP then activates PKA. Ca^{2+} also activates ERK/MAPK. Facilitated by cAMP, both CaMKII and CaMKIV translocate to the cell nucleus along with PKA and ERK/MAPK to activate gene transcription via phosphorylation of CREB. Translocation of ERK to the nucleus requires activation of PKA [Poser and Storm 2001]. Activation of the ERK/MAPK pathway may trigger nuclear translocation of RSK2 and thus phosporylation and transactivation of CREB [Impey et al. 1998].

Another transcription factor interacting with CREB/Creb1 is the homeodomain protein **Barx2**, which in turn interacts with activating transcription factor 2 (**Atf2**/**CREB-2**) [Edelman, Meech and Jones 2000]. We have included in our analysis also the cAMP responsive element modulator, **Crem**/**ICER,** which is also a transcription factor. Induction of LTP in the

dentate gyrus *in vivo* leads to rapid phosphorylation of the two downstream transcriptional targets of MAPK/ERK, i.e. already mentioned CREB and the ternary complex factor **Elk-1**, a key transcriptional-regulator of serum response element (SRE)-driven gene expression, and the immediate early gene **zif268**/**Egr1** is upregulated [Davis, Vanhoutte, Pagès, Caboche and Laro 2000]. MAPK/ERK, CREB and the immediate early gene (IEG) zif268 are essential components of a signalling cascade required for the expression of lasting phase LTP and of certain forms of long-term memory, because pharmacological blockade of MAPK/ERK phosphorylation, functional inactivation of CREB in an inducible transgenic mouse and inactivation of zif268 in a mutant mouse result in a similar deficit in long-term recognition memory [Bozon et al. 2003]. **ATF4**/**CREB-2** & **C**/**EBP**/**Cebpa**: memory storage and synaptic plasticity in transgenic mice expressing an inhibitor of ATF4 (CREB-2) and C/EBP proteins CCAAT/enhancer binding protein (C/EBP) were both impaired [Chen et al. 2003]. We have not found the CREB binding protein, Crebbp/CBP in Park et al's data. However we have found the signal transducer and activator of transcription, **Stat1** that can directly interact with the CREB-binding protein (CBP)/p300 family of transcriptional coactivators [Zhang et al. 1996].

Brain-derived neurotrophic factor (**BDNF)** induces LTP in intact adult hippocampus [Ying et al. 2002]. In addition, BDNF infusion leads to rapid phosphorylation of the MAPK/ERK and p38/Mapk14 but not **JNK**/**Mapk8** (c-Jun N-terminal protein kinase). BDNF-LTP was further coupled to ERK-dependent phosphorylation of the transcription factor CREB. BDNF infusion resulted in selective upregulation of mRNA and protein for Arc (not found in

the Park et al's data). MEK (**MEK2**/**Map2k2**) inhibitor blocked Arc upregulation in parallel with BDNF-LTP. The neuroplasticity-associated Arc gene is a direct transcriptional target of Early Growth Response (Egr) transcription factors [Li, Carter, Gao, Whitehead and Tourtellotte 2005]. Early growth response (Egr) transcription factors (**Egr1**/**Zif268**/**Krox24, Egr2**/**Krox20,** Egr3 and Egr4) are synaptic activity-inducible immediate early genes (IEGs). In particular, Egr1/Zif268 is essential for persistence of late-phase long-term potentiation (L-LTP), for hippocampus-dependent long-term memory formation, and for reconsolidation of previously established memories. Krox20 may play a key role in the stabilization of late LTP since its mRNA levels are increased 24 hours after LTP induction [Williams et al. 1995].

Tetanic stimulations induce NMDA-receptor-dependent synaptic **Wnt3a** release, nuclear **beta-catenin** accumulations, and the activation of Wnt target genes [Chen, Park and Tang 2006]. Suppression of Wnt signaling impairs LTP. Conversely, activation of Wnt signaling facilitates LTP. Stabilized beta-catenin translocates to the nuclei and binds the TCF/LEF **(Lef1)** family of transcription factors to regulate the expression of Wnt target genes. Recent studies suggest that aberrant Wnt signaling is implicated in multiple abnormal brain conditions, including Alzheimer disease and schizophrenia. Using a knock-out mouse, it was found that **c-rel**/**Rel**, one of the transcription factors identified in our bioinformatics study, is necessary for hippocampus-dependent long-term memory formation in dentate gyrus [Levenson et al. 2004].

Role of Tissue Plasminogen Activator (**tPA**) Receptor **LRP**/**Lrp1** in hippocampal LTP has been shown [Zhuo et al. 2000]. Both tPA and LRP are synthesized by hippocampal neurons. LRP is the major cell surface receptor that binds tPA. tPA&LRP have been also implicated in the pathogenesis of Alzheimer's disease. LTP is significantly decreased in mice lacking tPA. It was found that tPA binding to LRP in hippocampal neurons enhances the activity of cyclic AMP-dependent protein kinase, a key molecule that is known to be involved in L-LTP. Biochemical and molecular biology experiments indicate that the expression and secretion of tPA and BDNF are enhanced by strong tetanic stimulation that induces L-LTP as well as by training in hippocampal-dependent memory tasks [Pang and Lu 2004]. Inhibition of either tPA or BDNF by gene knockout and specific inhibitors results in a significant impairments in L-LTP and long-term memory.

It is widely thought that Alzheimer's disease (AD) begins as a malfunction of synapses, eventually leading to cognitive impairment and dementia. Double knockin (2xKI) mice carrying human mutations in the genes for amyloid precursor protein (APP) and presenilin-1 exhibit age-related downscaling of AMPAR-mediated evoked currents and spontaneous, miniature currents [Chang et al. 2006]. Electron microscopic analysis further corroborates the synaptic AMPAR decrease. Additionally, 2xKI mice show age-related deficits in bidirectional plasticity (LTP and LTD) and memory flexibility. These results suggest that AMPARs are important synaptic targets for AD and provide evidence that cognitive impairment may involve downscaling of postsynaptic AMPAR function (subunit **Gria1**).

Apolipoprotein E (**apoE**/**Apoe**) and low density lipoprotein receptor-related protein (LRP) facilitate intraneuronal Abeta42 accumulation in amyloid model mice of AD [Zerbinatti et al. 2006]. LRP binds and endocytoses Abeta42 both directly and via apoE but endocytosed Abeta42 is not completely degraded and accumulates in intraneuronal lysosomes. Amyloid-beta (**Abeta**/**App**), a peptide thought to play a crucial role in Alzheimer's disease (AD), has many targets that, in turn, activate different second-messenger cascades. Amyloid-beta peptide inhibits activation of the nitric oxide(NO)/ soluble guanylyl cyclase (sGC)/cyclic GMP (**cGMP)**/CREB pathway during hippocampal synaptic plasticity and has been found to markedly impair hippocampal LTP [Puzzo et al. 2005]. Therefore we include the genes for soluble guanylyl cyclase **Gucy1b3** and the NO synthetase **nNOS**. Amyloid beta-peptide treatment of cultured hippocampal neurons leads to the inactivation of protein kinase A (PKA) and CREB phosphorylation in response to glutamate is decreased [Vitolo et al. 2002]. Although there is eventual loss of synapses in both AD and animal models of AD, deficits in spatial memory and inhibition of LTP precede morphological alterations in the models, suggesting earlier biochemical changes in the disease. Abeta is generated by presenilin-dependent gamma-secretase cleavage of beta-amyloid precursor protein (**betaAPP**). In addition, presenilins (**PS1**/**Psen1** and **PS2**/**Psen2**) also regulate Abeta degradation [Pardossi-Piquard et al. 2005]. Presenilin-deficient cells fail to degrade Abeta and have drastic reductions in the transcription, expression, and activity of neprilysin, a key Abeta-degrading enzyme. Neprilysin gene promoters are transactivated by AICDs from APP-like proteins (APP, **APLP1**/**Aplp1**, and **APLP2**/**Aplp2**). PS1 mutations also

increase levels of the pre-apoptotic transcription factor **Gadd153/chop**, however without affecting its mRNA levels [Milhavet et al. 2002]. Another protein that has been implicated in the development of AD is hydroxyacyl-Coenzyme A dehydrogenase type II or hydroxysteroid (17-beta) dehydrogenase 10, **Hsd17b10** [Yan et al. 1997].

Fragile X syndrome, the most common inherited form of human mental retardation, is caused by mutations of the **Fmr1/FMRP** gene that encodes the fragile X mental retardation protein (FMRP). Silenced expression of the FMR1 gene is responsible for the fragile X syndrome. The FMR1 gene codes for an RNA binding protein (FMRP), which can shuttle between the nucleus and the cytoplasm and is found associated to polysomes in the cytoplasm. Biochemical evidence indicates that FMRP binds a subset of mRNAs and acts as a regulator of translation. A novel protein interacting with FMRP: nuclear FMRP interacting protein (**NUFIP/Nufip1**) has been identified [Bardoni, Schenck and Mandel 1999]. FRAXE mental retardation results from expansion of a CCG trinucleotide repeat located in exon 1 of the **Fmr2/Aff2**, which results in transcriptional silencing. FMR2 is hypothesized to be a transcriptional activator. LTP was enhanced in hippocampal slices of Fmr2 knock-out compared with wildtype mice [Gu et al. 2002]. Mutations of aristaless related homeobox gene, **Arx**, have been found in X-linked mental retardation, as well as of guanosine diphosphate (GDP)-dissociation inhibitor **Gdi1**, p21-activated kinase **Pak3**, and angiotensin II receptor **Agtr2** have been implicated in non-syndromic X-linked mental retardation [Ropers et al. 2003].

Loss-of-function mutations or abnormal expression of the X-linked gene encoding methyl CpG binding protein 2 (**MeCP2**) cause a spectrum of postnatal neurodevelopmental disorders including Rett syndrome (RTT), nonsyndromic mental retardation, learning disability, and autism. Mice expressing a truncated allele of Mecp2 (Mecp2308) reproduce the motor and social behavior abnormalities of RTT. Hippocampus-dependent spatial memory, contextual fear memory, and social memory are significantly impaired in Mecp2308 mutant males (Mecp2308/Y). The morphology of dendritic arborizations, the biochemical composition of synaptosomes and postsynaptic densities, and BDNF expression were not altered in these mice but LTP was impaired [Moretti et al. 2006]. LTP was also reduced in the motor and sensory regions of the neocortex. These data demonstrate a requirement for MeCP2 in learning and memory and suggest that functional and ultrastructural synaptic dysfunction is an early event in the pathogenesis of RTT. MeCP2 functions as a global repressor of transcription. Neuronal activity and subsequent calcium influx trigger the de novo phosphorylation of MeCP2 at serine 421 (S421) by a CaMKII-dependent mechanism. It was found that S421 phosphorylation controls the ability of MeCP2 to regulate dendritic patterning, spine morphogenesis, and the activity-dependent induction of Bdnf/BDNF transcription [Zhou et al. 2006]. These findings suggest that, by triggering MeCP2 phosphorylation, neuronal activity regulates a program of gene expression that mediates nervous system maturation and that disruption of this process in individuals with mutations in MeCP2 may underlie the neural-specific pathology of RTT.

Schizophrenia is also accompanied with memory disorders [Aleman, Hijman, Haan and Kahn 1999]. Complexin 2 (**Cplx2**) playing role in membrane exocytosis, **QKI/qk** playing role in splicing and expression of myelin-related genes, COMT/Comt involved in dopaminergic metabolism, extracellular matrix serine protease (**Reln/reelin**), and the *α*7 subunit of the nicotinic receptor for Acetylcholine (**Chrna7**) were implicated in the neuropathology of the disease [Cloninger, 2002; Egan et al. 2001; Sugai et al. 2004]. Micrutubule assicated protein tau gene (**MAPT/Mapt**) has been implicated in the fronto-temporal dementia [Bertram and Tanzi 2005].

Transglutaminase-3/Tgm3 is involved in stabilization of synapses, synapse formation, modulation of adenylyl cyclase and CREB. Inhibition of Tgm3 impairs both the early and late phases of LTP in dentate gyrus [Park et al. 2006]. **Cdc25b** encodes a tyrosine phosphatase that controls the activity of CDC2/cyclin B kinase. An abnormal up-regulation of Cdc25b has been suggested to be involved in the development of AD. Inhibitor of Cdc25b blocked both the early and late phases of LTP in dentate gyrus [Park et al. 2006]. **Ptpns1/SHS-1** is a synaptic adhesive Ig-like transmembrane glycoprotein and a signal regulatory protein involved in receptor tyrosine signalling and can regulate cell-cell interactions; can initiate MAPK pathway. Blockade of Ptpns1 impaired early but not late LTP in dentate gyrus [Park et al. 2006]. LTP in the dentate gyrus depends on the NCAM glycoprotein, the neural cell adhesion molecule, **Ncam1** [Stoenica et al. 2006]. The growth factor, neurotrophin 3, **Ntf3**, also plays role in LTP [Shimazu et al. 2006].

The following genes are up-regulated with different peaks following LTP induction at perforant path synapses *in vivo* [Abraham and Williams 2003], i.e.: Transcription Factors (TF): **Egr2/Krox20** reaches the over-expression peak at around 2 h post-induction of LTP [Williams et al. 1995]**; AP-1 family (Fosb ; Fosc; Junb; Junc,** increase 20 min after LTP induction [Abraham et al. 1993])**; COX-2/Ptgs2;** Receptor function: **Homer1a/vesl, TkrB/Ntrk2 and TkrC/Ntrk3** (LTP induction triggers a rapid 2 h elevation in TrkB and TrkC gene expression [Bramham, Southard, Sarvey, Herkenham and Brady 1996]. Synaptic structure: **Synaptopodin/Synpo; Synapsin 1/Syn 1; Stx1bl/syntaxin; Syp/synaptophysin; Syt1/ synaptotagmin.** An increased level of synaptopodin/Synpo mRNA was observed at 75 min and 3.5 h after the onset of LTP in the dentate gyrus [Yamazaki, Matsuo, Fukazawa, Ozawa, and Inokuchi 2001]. There was an increase in the levels of three proteins (synapsin, synaptotagmin and synaptophysin) at 3 h, but not at 45 min after induction of LTP [Lynch, Voss, Rodriguez and Bliss 1994]. Syntaxin level was increased at 2 h and 5 h after the LTP induction [Hicks et al. 1997]. Enzymes: **tPA; PKC/Prkca** (increase at 1h but not 2h [Meberg, Valcourt and Routtenberg 1995])**;** Structural proteins: **AKAP/Akap1; PSD95**.

6.2.2 Gene clustering and functional analysis

We have done a simple clustering of selected 79 genes by aligning temporal profiles of their temporal expression data. Most comprehensive list of selected genes (into the respective clusters) along with the functions of their coded proteins as found by means of NCBI Entrez Gene database is provided in the table 6.1.

Table 6.1: Genes with respective IDs, temporal cluster (group with multiple genes and is shown in different colours), process they are involved in and their functions

Gene common name	Genebank ID	Cluster	Process	Transcription Factor / DNA binding	RNA binding /translation regulation	nucleic acid binding	synaptic structure/plasticity	cell matrix and/or adhesion	Cell cycle regulation	cytoskeleton; structure-related	intracellular signalling	membrane fusion	retrograde signal	IEG	receptor-associated binding	ATP binding	receptor activity	ion-channel activity	protein kinase activity	transferase activity	hydrolase activity	phosphatase activity	growth factor activity	beta-amyloid/tau binding	antioxidant
Abeta; App	U82624	1	AD/LTP	x			x	x																	
Aplp2	M97216	1	AD/LTP	x				x																	
Apoe; AI255918	D00466	1	AD																					x	x
Arx	AB006103	1	X-MR	x					x																
Camk2a_X87142	X87142	1	LTP			x										x			x	x					
Cdc25b	AI849132	1	LTP/AD																		x	x			
Fmr2; Aff2	AJ001549	1	XMR/LTI	x																					
Gria1	X57497	1	LTP														x	x		x					
Grin2b	D10651	1	LTP														x	x							
Gucy1b3	AF020339	1	LTP								x								x	x					
Homer1-pending_A	AF093257	1	LTP								x				x										
JNK; Mapk8	AB005663	1	LTP/AD			x										x			x	x					
Junb	U20735	1	LTP						x					x											
PKA; Prkaca	M12303	1	LTP			x										x			x	x					
PSD95; SAP90	D50621	1	LTP				x			x		x			x										
Ptpns1; SHPS-1	AV317524	1	LTP					x		x	x						x								
Atf4, CREB-2	M94087	2	LTP	x																					
Calm3	M19380	2	LTP								x														
Camk2a_X14836	X14836	2	LTP			x						x				x			x	x					
Cebpa; CBF-A; CC	M62362	2	LTP	x					x																
Cplx2	D38613	2	SCH									x													
c-Rel; Rel	X60271	2	LTP	x					x																
Grin1	D10028	2	LTP														x	x							
Grin2a	D10217	2	LTP														x	x							
Hsd17b10	U96116	2	AD																						x
Mapt	M18776	2	FTD							x															
Ncam1_X15050	X15050	2	LTP					x	x							x	x						x		
Stx1bl-pending; sy	D45207	2	LTP									x													
Akap1_U95146	U95146	3	LTP		x	x													x						
Atf2; CREB-1; Crel	U46026	3	LTP	x		x																			
nNOS; Nos1; NO	D14552	3	LTP										x												x
qk	U44940	3	SCH		x	x																			
beta-catenin; Catnb	M90364	4	LTP	x						x															
COX-2; Ptgs2	M88242	4	LTP																					x	x
Fos; c-fos	V00727	4	LTP	x										x											
Gdi1	U07950	4	XMR																						
Krox20; Egr2	M24377	4	LTP	x		x								x											
Psen2	U57325	4	AD																		x				

Gene common name	Genebank ID	Cluster	Process	Transcription Factor / DNA binding	RNA binding /translation regulation	nucleic acid binding	synaptic structure/plasticity	cell matrix and/or adhesion	Cell cycle regulation	cytoskeleton; structure-related	intracellular signalling	membrane fusion	retrograde signal	IEG	receptor-associated binding	ATP binding	receptor activity	ion-channel activity	protein kinase activity	transferase activity	hydrolase activity	phosphatase activity	growth factor activity	beta-amyloid/tau binding	antioxidant
Agtr2	U04828	5	XMR								x						x								
Bdnf	X55573	5	LTP				x																x		
Camk4	J03057	5	LTP	x		x										x			x	x					
Comt	AF076156	5	SCH																	x					
Homer1_AB01947	AB019479	5	LTP											x											
MEK2; Map2k2	AW123542	5	LTP			x										x			x	x					
Ntf3; NT3; NT-3; N	X53257	5	LTP												x								x		
Syn1; synapsin 1	AF085809	5	LTP							x		x				x									
Tgm3; transglutam	L10385	5	LTP						x											x					
Elk1	X87257	6	LTP	x										x											
Fosb	X14897	6	LTP	x										x											
Jun; AP-1; Junc; c-	X12761	6	LTP	x										x											
Lef1	D16503	6	LTP	x																					
Mapk1; ERK; Erk2	D87271	6	LTP			x	x									x			x	x					
Mapk14; p38	D83073	6	LTP/AD			x										x			x	x					
Ncam1_X15052	X15052	6	LTP				x									x							x		
Reln, reelin	U24703	6	SCH					x													x				
Syp; synaptophysir	X95818	6	LTP				x																		
Akap1_U95145	U95145	7	LTP		x	x													x						
CREB	X67719	7	LTP	x																					
Fmr1; FMRP	L23971	7	XMR/LTD		x		x																		
ICER,Crem_10509	M60285	7	LTP	x																					
Nufip1_AA681274	AA681274	7	X-MR		x	x	x																		
PKC; Prkca	M25811	7	LTP			x										x			x	x					
Syt1; synaptotagmi	D37792	7	LTP									x			x										
Aplp1	L04538	8	AD/LTP					x																	
cGK; Prkg2	L12460	8	LTP			x										x			x						
Synaptopodin; Syn	AW046661	8	LTP				x			x															
Wnt3a	X56842	8	LTP								x				x										
Tkrb; Ntrk2	M33385	9	LTP													x	x		x	x					
Trkc; Ntrk3	AF035400	9	LTP			x										x	x		x	x					
Zif268; Egr1; Krox2	M28845	9	LTP	x		x	x			x				x											
Gadd153; chop	X67083	10	LTP	x		x																			
Psen1	L42177	10	AD																		x			x	
Mecp2	AJ132922	11	RTT/LTP/LTD/XMR						x																
Pak3	U39738	11	XMR/LTP			x	x									x			x	x					
Stat1	U06924	11	LTP	x							x												x		
Lrp1	X67469	12	LTP/AD														x							x	
tPA; Plat	J03520	12	LTP/AD																		x		x		
Barx2	L77900	13	LTP	x																					
Chrna7	L37663	14	SCH														x	x							

Distribution of functions of genes within each cluster showed that these temporal clusters are heterogenous, that is genes within clusters have multiple functions. Analysis of the functions of each gene within each cluster, based on NCBI Entrez Gene database, has given us the opportunity to match gene expressions with various subcellular processes that underlie LTP induction and stabilization over the course of 2 hours post-tetanization. For that purpose we have organized the genes according to the time they are over-expressed, i.e. all the genes that are over-expressed at 30 min, 60 min, 90 min and 120 min, plus the 5^{th} group of always under-expressed genes (during the period of 2 hrs), see appendix F. At each time point during the first 2 hours some genes that are involved in multiple cellular functions were over-expressed while others involved in the same functions were under-expressed. These particular molecular functions include: regulation of transcription and translation, immediately early genes, cell cycle regulation, cell matrix and cell adhesion molecules, growth factors, genes related to synapse structure and plasticity, genes related to cytoskeleton, membrane fusion (exocytosis), intracellular and retrograde signalling, receptor-associated binding, ion-channel and receptor activity, protein-kinase activity, transferase activity, hydrolase activity and genes that are beta-amyloid or tau related. This complex scenario points to the fact that induction and maintenance of LTP consists of series of parallel and overlapping events leading from signal transduction from stimulated synapses to the nucleus, where gene expression is initiated to result in new protein synthesis, which in turn changes the structure and composition of stimulated synapses.

Clustering analysis based on sequence data

We retrieved the genes and their respective protein sequences from the NCBI database using their accession numbers. Then all the sequences were converted into FASTA format. For our investigation, we employed the standard operating bioinformatics procedure, i.e., comparative analysis from similarity measures that is a widely accepted approach by most of bioinformaticians. The method is familiarly known as multiple sequence alignments (MSA). Scope of this bioinformatics technique is three fold: (a) gains understanding to identify the shared regions of homology; (b) determines the consensus sequence of several aligned sequences; (c) identifies the evolutionary phylogenetic relationship between different molecules. For such strategies, there are however many software available free to the academic users, but here in our case we have used the CLUSTALX package (version clustalx1.81). It provides a window-based user interface to the ClustalW multiple alignment program and uses the vibrant multi-platform user interface development library developed by the National Center for Biotechnology Information (NCBI), for details see [Jeanmougin, Thompson, Gouy, Higgins and Gibson 1998]. The program was installed on a Windows XP and all the experiments were run on a Pentium 4 machine having 2.8GHZ processor with 2GB RAM installed. A few specific reasons for using this software were that (a) it is mostly cited; (b) secondly it produces the result output that are easy to interpret than others in the field and (c) flexibility of using its locally installed version without being connected to the internet. Sequence similarity between the different molecules was visualized and later the clustering trees (dendograms) were obtained using the minute information

available from this multiple sequence alignment file. Clustering tree was then saved and displayed in the neighbour joining (NJ) plot software (it also measures the distance between taxa) [Perrière and Gouy 1996]. Later we manually assigned the colours to the molecules in accordance with their clustering results as obtained from QiEA. It is worth mentioning here that we have ignored the protein analysis in some cases where short sub-sequence of a transcribed spliced nucleotide sequence called expressed sequence tags (EST) was present. However we notice there were not many such cases but these genes were Nufip1, MEK2, Synaptopodin, Cdc25b and Ptpns1.

In order to understand the regulatory behaviour of our selected set of 79 genes and to analyze co-regulation of genes within temporal clusters, we performed a specific analysis on the genes to predict their potential promoter sites. All the high quality promoters sequences for the respective genes were extracted (predicted) using the “promoser”, which is a large-scale promoter and transcription start site (TSS) identification program mainly for *Homo sapiens* (human), *Mus musculus* (mouse) *and Rattus norvegicus* (rat) genes [Halees, Leyfer and Weng 2003]. The system utilizes the most recent genome assemblies of each organism and the mRNA have been kept updated frequently to keep pace with an expanding GenBank mRNA collection. We predict the promoters that are 1000 bases upstream and 50 bases downstream starting from the TSS and only the promoter nearest upstream to the 5' end of the transcript. Also, we ignored the upstream overlapping clusters and any genomic gaps in the sequence and the promoters that we found were filtered for TSS's with a minimum quality of two and support of at least one high quality sequences. In general, we consider only those

promoters that could be confidently positioned and in very few cases din't ignored the result that is a best guess to its location. Some accessions could not be mapped very well to the mouse genomic loci or they were below the specified quality criteria that we have mentioned here for promoter selection, therefore, in our clustering analysis we were unable to include the promoters from each of the 79 genes. The sequence of the predicted promoters were assigned appropriate name and were put into FASTA format and then they were analysed using usual MSA method as described above.

Each dendogram may represent the related closeness between the genes, promoters and the proteins. In our experiments we use the trees for clustering purpose and it can be noted from different trees that genes, proteins or promoters that are more closely related to each other (in terms of functions and course of evolution) are readily aligned one after another. We are interested to see which genes /proteins and promoter regions are close to each other (based on the sequence data) and also which of these close sequences belong to the same temporal cluster of gene expression. All the dendograms may be referred in the appendix F and for simplicity here we have interpreted the essence of the important findings in table 6.2.

Table 6.2: Related proteins, genes and promoters (based on sequence); gene, protein and promoter name is followed by a cluster number within bracket (based on gene expression)

Phylogenetically close proteins	Phylogenetically close genes	Phylogenetically close promoters
Gria1 (1) Grin2b (1)	Gria1 (1) Abeta (1)	Fmr2 (1) Cdc25b (1)
PKA (1) JNK (1)	ApoE (1) Junb (1)	Atf4 (2) Grin2a (2)
Homer1AF093257 (1) Abeta (1)	Homer1AF093277 (1) JNK (1)	Bdnf (5) Tgm3 (5)
Grin1 (2) Grin2a (2)	Cplx2 (2) Mapt (2)	
Calm3 (2) syntaxin (2)	Grin1 (2) Grin2a (2)	
Ntf3 (5) Bdnf (5)	c-fos (3) Egr2/ Krox20(3)	
Syn1 (5) Comt (5)	Ntf3 (5) Bdnf (5)	
Ncam (6) Jun (6)	Agtr2 (5) Homer1 AB019479 (5)	
Fosb (6) Syp (6)	Mapk1 (6) Mapk14 (6)	
Mapk1 (6) Mapk14 (6)	Ncam (X15052) (6) Lef1 (6)	
CREB (7) ICER (7)		
Trkb (9) Trkc (9)		
Atf2 (3) Fmr1 (7)	Ncam1X15052 (6) Trkc (9)	nNOS (9) Nufip1 (7)
Akap1U95146 (3) Akap1U95145 (7)	Homer1AF093257 (1) Homer1AB019479 (5)	cGK (6) Mapt (2)
Syp (6) Syn1 (5)	JNK (1) Mapk14 (6)	Syt1 (7) Pak3 (11)
Gadd153 (10) Fmr2 (1)	Mapk1 (6) Agtr2 (5)	COX2 (4) Camk4 (5)
Psen2 (4) Psen1 (10)	Elk1 (6) Zif268 (9)	Grin2b (1) Elk1 (6)
Homer1AF093257 (1) Homer1AB019479 (5)	Akap1U95146 (3) Akap1U95145 (7)	Atf2 (3) Psen2 (4)
Homer1AF093257 (1) Ncam1X15050 (2)	Grin2a (2) Grin2b (1)	Aplp2 (1) Agtr2 (5)
Camk2ax87142 (1) Camk2ax14836 (2)	Grin1 (2) Psen1 (10)	
cGK (9) PKA (1) PKC (7)	CREB (7) Camk2ax87142 (1)	
JNK (1) Mapk14 (6) Mapk1 (6)		

6.2.3 Application of QiEA to predict GRNs and gene knock-in mice experiments

For the prediction of gene regulatory networks (GRN), we have shown the application of valuable computational intelligence methods like GA, KF,

LARS, EM, EFuNN etc. in the previous chapters 4 and 5 of this thesis. In this chapter we aimed to infer coefficients of an abstract regulatory network which has 14 temporal clusters of gene expression in its nodes by using the method of Quantum Inspired Evolutionary Algorithm (QiEA, the method is a revised description of QEA originally proposed by KEDRI researchers in 2007, for more details see [Defoin-Platel, Schliebs and Kasabov 2007]. The inferred interaction coefficients were interpreted as an average influence between clusters of genes, which results from averaging individual gene influences between genes from different clusters. Optimization resulted in 189 matrices of interactions between clusters that faithfully reproduced the temporal course of gene expressions found in 14 clusters during the first 2 hours (Park et al's data for these 79 genes). We then used each solution in simulated transgenic mice experiments, whether the altered levels of some genes would lead to gene expression levels observed in transgenic and gene knock-in experiments. Since clusters are composed of more than one gene, we have simulated one gene mutation within a particular cluster with a slight increase in the average expression level of the whole cluster (if the mutation led to an enhanced expression of the gene in question). We did it by recalculating the cluster expression average for each time point with the new increased value of the expression of the mutated or knock-in gene. This manipulation reflects the following assumption: if levels of expression of other genes within that cluster remain the same, then increase in one's gene expression shifts the cluster average to higher expression values. The same assumption has been applied to other clusters that were affected through interaction weights. If the result of gene mutation in one cluster was the shifted average of another

cluster, then this change can reflect change of expression levels of any gene within that cluster.

There are several transgenic mouse models that are genetically modified in such a way that they produce more amyloid precursor protein (APP). These transgenic mice are considered to be animal models of Alzheimer's disease (AD) that is characterized by accumulation of the amyloid β-protein in the so-called senile plaques. Researchers have profiled gene expression in hippocampi and cerebral cortici of these transgenic mice by microarray techniques (there are also results obtained with other techniques but we consider only the microarray results since our abstract regulatory network has been inferred based on microarray data). We can simulate these experiments by artificially up-regulating clusters containing APP and/or PS1 in our abstract regulatory network, and calculate levels of expression for other clusters. After the dynamics stabilizes we can compare the predicted cluster expression values (which linearly reflect gene expression values within clusters) with data available from the transgenic mice experiments.

In the first experiment, Dickey et al. [Dickey et al. 2003] used the amyloid precursor protein + presenilin-1 (APP+PS1) transgenic mouse as a model for amyloid deposition, and like in AD, the mice develop memory deficits as amyloid deposits accumulate. At the age when these animals developed cognitive dysfunction, they had reduced mRNA expression of several genes essential for long-term potentiation and memory formation, in particular Zif268 (cluster 9) and Homer-1a (cluster 5), which are on our list. These changes appeared to be related to amyloid deposition, because mRNA expression was

unchanged in the regions that did not accumulate amyloid (nontransgenic mice served as control). According to microarray data, these changes were accompanied by changes in mRNAs of genes like synapsin (cluster 5, down), synaptophysin (cluster 6, up), synaptotagmin (cluster 7, no change). Thus, authors concluded that the memory loss in APP+PS1 transgenic mice may model the early memory dysfunction in AD before the degeneration of synapses and neurons.

Three mouse models of AD were used to assess changes in gene expression [Wu et al. 2006]. One mouse model harboured homozygous familial AD (FAD) knock-in mutations in both, APP and presenilin 1 (PS-1) genes (APP/PS-$1^{P264L/P264L}$), the other two models harboured APP over-expression of FAD mutations (Tg2576) with the PS-1 knock-in mutation at either one or two alleles. To assess changes in gene expression associated with Abeta accumulation, the Affymetrix murine genome array U74A was used to survey gene expression in the cortex of these three models both prior to and following Abeta deposition. Altered genes were identified by comparing the AD models with age-matched control littermates. Thirty four gene changes were identified in common among the three models in mice with Abeta deposition. Down-regulated genes of note included BDNF (cluster 5) in APP/PS-$1^{P264L/P264L}$ and Tg2576/PS-$1^{P264L/P264L}$, and Mapt (cluster 2) in Tg2576/PS-$1^{P264L/P264L}$ and Tg2576/PS-$1^{P264L/+}$.

In the work of Jee et al. [Jee et al. 2007] cDNA microarray was used with the large-scale screening of the brain mRNA from transgenic and normal mice of 18 months of age. The authors produced transgenic mice expressing

neuron-specific enolase promoter-controlled APPsw leading to APP overexpression. It was demonstrated that cognitive deficit along with Aβ-42 depositions were shown at 12 months of age in these mice. Among the 48 differentially expressed genes in this study, only one occurs also on our list, i.e. Psen2 (cluster 4), and it was significantly down-regulated.

6.3 Results, discoveries and biological validation

After simulating virtual gene knock-in experiment APP + PS1 with each of 189 cluster interaction matrices that were optimized to reproduce temporal behaviour of 14 LTP clusters up to two hours post LTP induction, only one interaction matrix has led to an asymptotic expression levels that matched data according to Table 6.3. This is illustrated in figure 6.1, where we can see actual cluster expressions, their predictions into the future and for comparison results of simulated virtual APP + PS1 knock-in experiment. When running only APP knock-in, results were qualitatively similar to figure 6.1, so we can verify that cluster 4 (containing gene Psen2) is indeed lower in expression than it would be if APP would not be increased.

We can also validate predictions made by this resulting abstract interaction model by means of the published data on gene expression following the induction and maintenance of LTP in normal wild type dentate gyrus granule cells at perforant path synapses according to dotted curves in figure 6.1. For instance Egr1/Krox24 (cluster 9) reaches the over-expression peak at around 2 h post-induction of LTP and at 4 h its levels decrease in comparison with the 2 h peak [Williams et al. 1995]. This temporal course of

expression is in accordance with the prediction for cluster 9 by our optimized matrix (see figure 6.1 dotted curve for cluster 9).

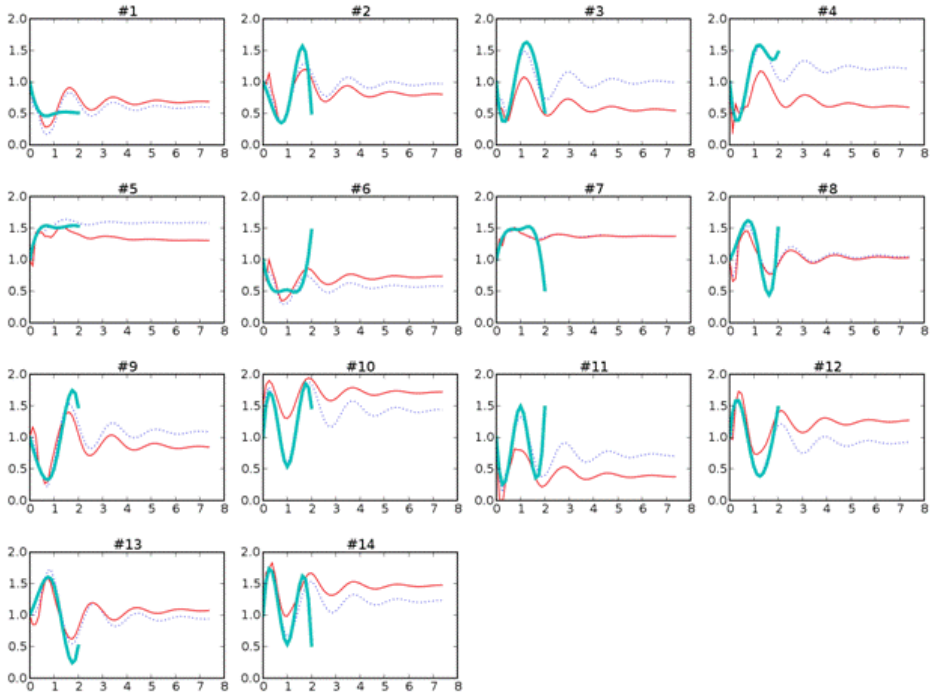


Figure 6.1: Normal and virtual gene knock-in dynamics of abstract regulatory network involving 14 temporal clusters; X axis: time points and Y axis: expression values. Thick cyan curves denote actual normalized temporal expression of clusters up to two hours post LTP induction. Dotted curves express prediction of expression in time to the future. Red solid curves express results of APP + PS1 knock-in experiments by artificially increasing the average expression levels in clusters 1 and 10, and consequences this change has upon expression of other clusters in GRN.

In another study, the following time courses of mRNA levels following LTP induction in the dentate gyrus were recorded over discrete time points 2, 6, 24 and 120 h [Bramham et al. 1996]: TrkB (Cluster 9) – increase at 2 h, and at control levels at $t \geq 6$ hrs. In the Park et al's data TrkB's mRNA is increased at 1.5 and 2 h. Prediction of our GRN model for Cluster 9 is that the levels of gene decrease back to control levels at about 6 h (figure 6.1), which is in accordance with the data. BDNF and NT3 (both in Cluster 5) are increased in

Park et al's data at all four time points, starting with 0.5 up to 2 h. In [Bramham et al. 1996], there is a significant increase in BDNF at 6 and 24 h. NT3 however is increased at all measured time points, albeit significantly only at 6 h. Prediction of our model is that there is a lasting up-regulation of expression of cluster 5, which is in accordance with the data on BDNF and NT3 (figure 6.1).

Based on the available literature we have partially verified our predicted time courses for some clusters based on the optimized abstract gene interaction matrix between clusters. Interaction coefficients that have led to the network dynamics both for simulation of wild type and transgenic mice gene expression in 14 clusters are in table 6.3.

Table 6.3: Coefficients of an abstract interaction matrix, yellow/light gray correspond to direct or indirect interactions (i.e. via other gene/s) between genes that are in our list as confirmed by means of NCBI Gene Entrez database. Pink/gray highlighted coefficients are predicted interactions in such a sense that we have found interactions of our listed genes with other genes in Park et al. (2006), which are however not in our list.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
C1	-0.07	0.37	0.27	0.05	-0.04	0.21	0.3	-0.25	-0.3	-0.14	-0.1	-0.14	-0.45	-0.1
C2	0.51	0.14	0.21	0.08	-0.2	-0.17	0.02	0.38	0.15	-0.21	-0.41	-0.34	-0.33	0.21
C3	-0.46	0.11	0.05	0.16	0.11	-0.55	-0.5	-0.04	0.26	-0.13	0.29	-0.25	0.31	0.36
C4	-0.53	0.03	-0.07	-0.17	0.6	-0.74	-0.05	0.15	0.82	-0.64	0.15	-0.08	-0.2	0.28
C5	-0.27	0.04	0.29	0.38	-0.1	-0.44	0.37	-0.11	-0.11	-0.21	-0.21	-0.03	-0.18	0.6
C6	0.57	-0.32	0.13	-0.17	0.04	0.19	0.32	-0.39	-0.12	-0.31	-0.09	-0.12	-0.27	0.35
C7	0.11	0.05	0.1	0.16	0.2	0.04	-0.24	0.14	-0.4	0.18	-0.18	-0.22	0.14	0.16
C8	-0.8	-0.14	0.2	0.04	0.34	0.46	0.1	0.02	-0.28	-0.02	-0.18	0.32	-0.18	-0.18
C9	0.84	0.11	-0.09	-0.1	0.2	-0.08	0.54	0.01	-0.13	0.18	0.05	-0.75	-0.45	-0.3
C10	0.6	-0.004	-0.05	0.1	-0.17	0.28	0.15	-0.33	0.05	0.08	-0.18	-0.02	-0.1	0.25
C11	0.06	-0.05	-0.22	0.67	0.24	-0.68	-0.05	0.07	-0.24	-0.54	-0.38	0.16	0.18	0.11
C12	-0.29	0.25	0.27	-0.07	-0.36	0.47	0.23	-0.13	-0.38	-0.15	-0.28	0.24	-0.26	0.51
C13	-0.29	0.05	-0.13	0.11	-0.12	-0.25	-0.21	-0.07	-0.23	-0.23	0.22	0.27	0.44	0.44
C14	0.08	0.24	0.44	-0.28	-0.18	0.81	0.02	0.27	-0.2	0.17	-0.36	-0.07	-0.31	-0.06

We can see that a linear model predicts nonzero regulatory coefficients between each pair of clusters. The value expresses the overall strength and sign, which theoretically is a result of many individual interactions between

genes in clusters, both positive and negative in sign. Both strong and weak interactions are vital for reproducing the temporal dynamics of all clusters. Non-highlighted values in table 6.3 represent predicted interactions between genes within particular clusters that may or may not be confirmed in the future. Interaction coefficients highlighted in yellow/light gray correspond to direct or indirect interactions (i.e. via other gene/s) between genes that are in our list as confirmed by means of NCBI Gene Entrez database. Pink/gray highlighted coefficients are predicted interactions in such a sense that we have found interactions of our listed genes with other genes in Park et al. (2006) data, which are however not in our list. These genes were then assigned to clusters based on their temporal course of gene expression and in figure 6.2 their names are written in bold.

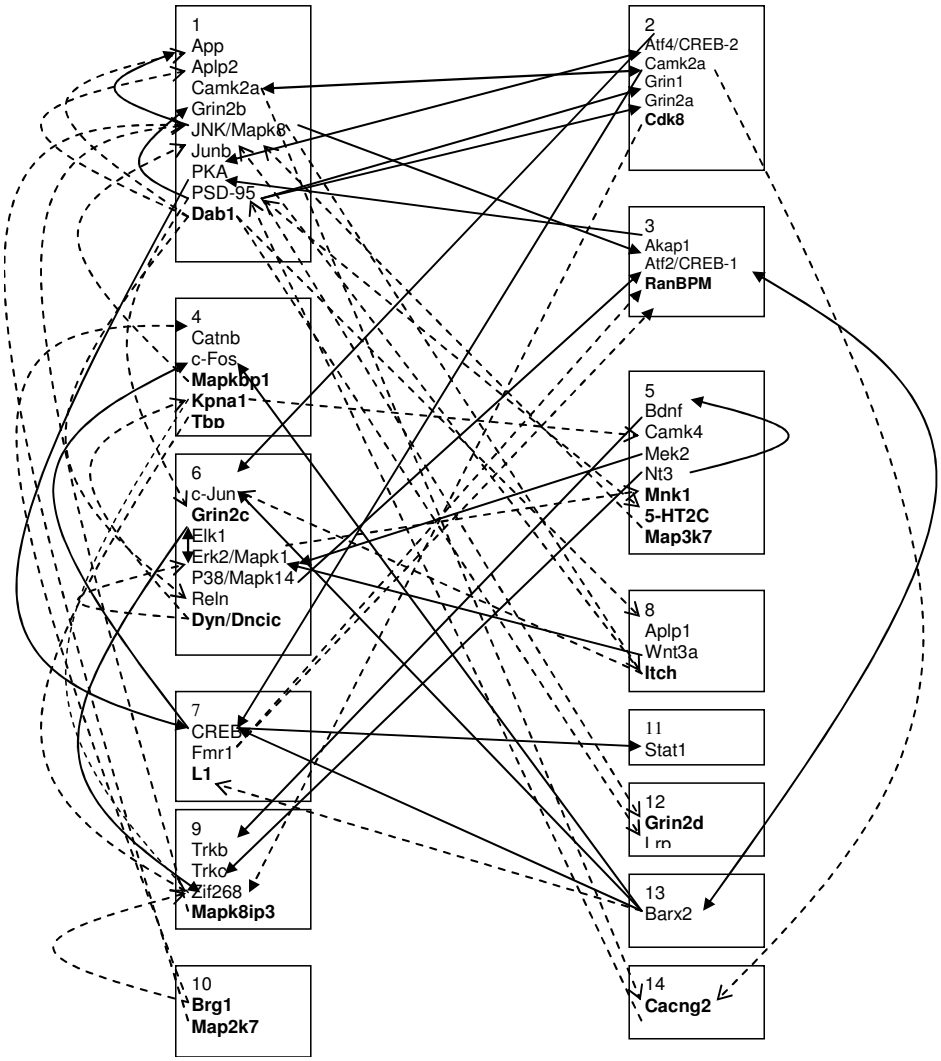


Figure 6.2: Original (solid lines/curves) and predicted interactions (dashed line/curves) between genes/proteins as found by means of NCBI Entrez gene database. Each box represents separate cluster and genes in bold are not in our list of 79 selected genes, however they are in Park et al. (2006) data and we could assign them to clusters based on their temporal profile of expression in order to infer more interactions between clusters.

The next, figure 6.3, illustrates interactions between clusters in an abstract regulatory network, in such a way that we have collapsed gene interactions between individual genes from figure 6.2 into connections between clusters, the thickness of which reflects how many genes between given clusters have confirmed interactions. We can see that the cluster

number 1 acts as hub, based on a number of connections leaving and entering it.

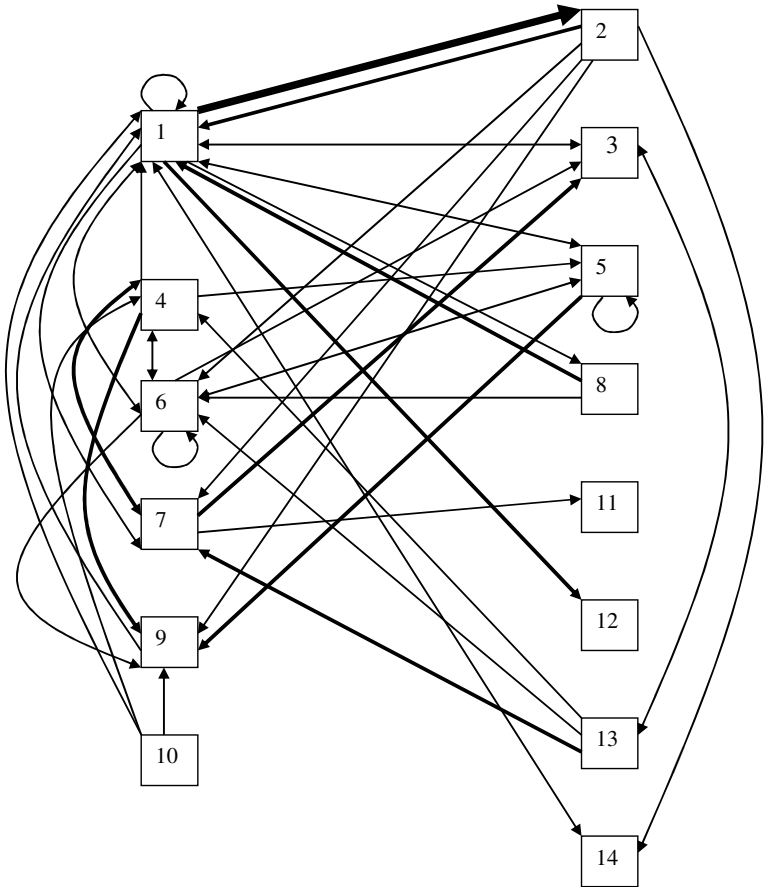


Figure 6.3: Summary of interactions between clusters in an abstract regulatory network.

However, not all interactions between genes on our list have been discovered by today so this account represents only partial illustration. There are some predictions of strong interactions for mouse model of AD (stronger than 0.75, see table 6.3), which are summarized in figure 6.4, and which can serve as a guide for further experimental testing.

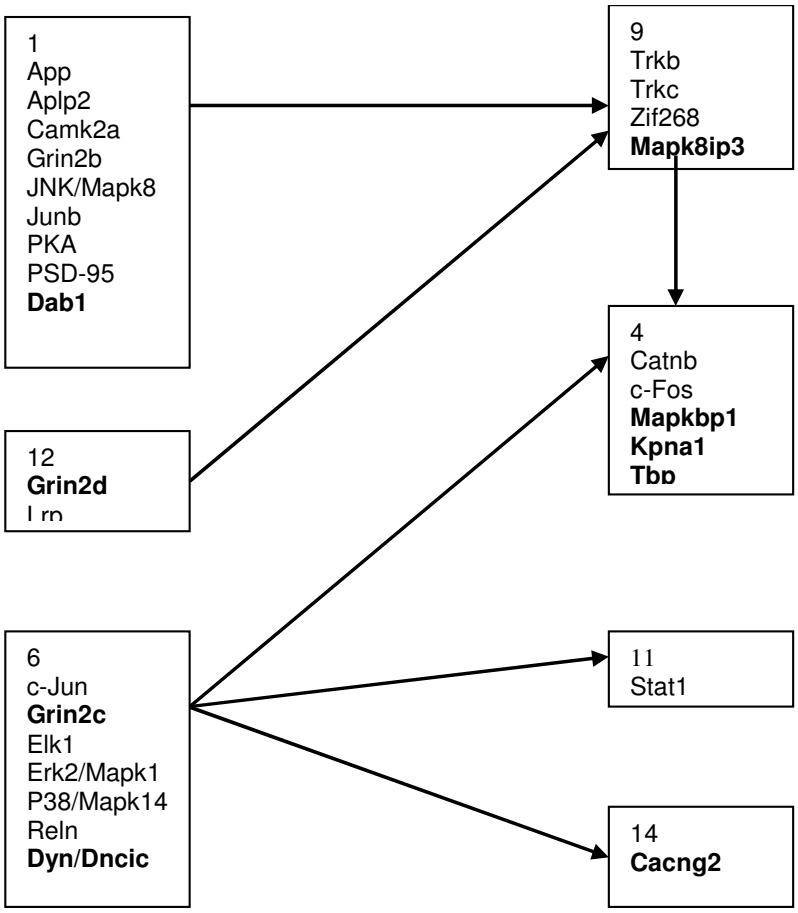


Figure 6.4: Predicted strong interactions between clusters of genes from our abstract regulatory model

6.4 Conclusion and discussion

There are many sophisticated computational intelligence methods to obtain the gene regulatory networks, in this chapter we have used the integrative approach of QiEA, clustering (based on gene expression profile and the sequence data) and functional analysis to analyse the time-series gene expression data. Understanding LTP is possible only if we take the analysis of underlying genetic mechanism into account. LTP is an example of a process that takes several hours perhaps days to develop and which has

distinct phases each of them is characterized by a different profile of expression of genes. In this thesis chapter, we have treated microarray data on temporal expression of genes involved in LTP induction and maintenance during the first 2 hours poststimulus with a variety of analytical tools to dig out more knowledge about LTP underlying mechanisms.

The genes for our analysis were selected from Park et al's 12,488 activity regulated genes [Park et al. 2006] based on their documented link to LTP and/or learning&memory disorders (Alzheimer's disease, Fronto-temporal dementia, X-linked mental retardation, Rett syndrome and schizophrenia), however we do not claim this list is exhaustive, nor that this could have been the only criterion for gene selection. Another criterion could have been the magnitude of the gene expression change after the induction of LTP or a combination of several criteria. This selection can be the topic of future study. First, we have clustered this subset of 79 genes based on their temporal profile of gene expression over time to obtain 14 clusters with co-expressed and thus supposedly co-regulated genes. Functional analysis of genes based on information in NCBI Entrez Gene Database revealed that all these temporal clusters are functionally heterogenous, i.e. genes within clusters encode proteins that have different functions. Thus, based on the functional analysis and temporal profile of gene expressions and available literature, we have made a link between gene functions and different stages of induction, stabilization and maintenance of long-term potentiation. Within the first 2 hours of LTP induction and maintenance, genes with multiple functions were either over-expressed or under-expressed. These functions include regulation of transcription and translation, immediately early genes, cell cycle

regulation, cell matrix and cell adhesion molecules, growth factors, genes related to synapse structure and plasticity, genes related to cytoskeleton, membrane fusion (exocytosis), intracellular and retrograde signalling, receptor-associated binding, ion-channel and receptor activity, protein-kinase activity, transferase activity, hydrolase activity and genes that are beta-amyloid or tau related. From the summary of functional analysis, we can see that translation factors regulating RNA are overexpressed from 30 min up to 1.5 h posttetanus. Transcription factors regulating gene transcription are up-regulated from 30 min up to 2 h. Since researchers distinguish several types of LTP based on the time of its decay and requirements of de novo protein synthesis we can infer which type of LTP was actually induced in Park et al's experiment. LTP1 decays with the average time constant of hours (~2 h), LTP2 with the average decay constant of days (1-4), and LTP3 decays over weeks (> 20 days) [Abraham et al. 1993; Abraham and Williams 2003]. It appears that LTP3 requires gene transcription and subsequent protein translation, LTP2 requires protein translation from pre-existing RNA and LTP1 requires neither. From this we can infer that LTP induced in Park et al's experiment was probably the LTP3 type, since it was accompanied by activation of genes related to synaptic structural changes from as early as 30 min post HFS.

We have performed an extensive clustering analysis of gene and their protein sequences. Twelve pairs of proteins are closely related and at the same time belong to the same temporal clusters, whereas there are 10 pairs of closely related proteins that belong to different clusters. In total 44 proteins out of 79 can be considered to be close related. 10 pairs of genes are closely

related and at the same time belong to the same temporal clusters, whereas there are 9 pairs of closely related proteins that belong to different clusters. In total 38 genes out of 79 can be considered to be closely related. Thus, our analysis indicated that about half of genes / proteins within our selected subset of genes related to LTP are in fact closely related based on the sequence data. These closely related genes / proteins can but do not have to necessarily belong to the same temporal cluster of gene expression.

Large-scale computational analysis of transcription regulation is a powerful and promising technology that should provide us with a better understanding of the nature of the intricate network of regulatory genes that provide living cells with their remarkable properties. Biologically we know that the regulation mechanism is a multilevel high complexity system that involves upstream and downstream cis-regulatory elements on the DNA but availability of the promoter sequence data has never been easy task. In recent years, the transcription start sites have been identified computationally by considering alignments of a large number of partial and full-length mRNA sequences to genomic DNA, with provision for alternative promoters. In order to understand the regulatory behaviour of our selected set of 79 genes, we have performed a specific analysis on the genes to predict their potential promoter sites. Out of 79 genes, only 10 pairs of genes have very similar promoter regions (sequence similarity). Only 3 pairs have also the same time parallel time course of their expression. This result suggests that in spite of different promoter regions, genes still can be co-regulated to have a similar time course of their expression for a quite long time period of 2 hours as is demonstrated in 14 similar profiles of 79 genes.

Further, we have focused on interactions between temporal gene clusters in an abstract regulatory network. The inferred interaction coefficients are interpreted as an average influence between clusters of genes, which are the result of influences between individual genes. To infer the interaction coefficients between gene clusters, we employed a new method of quantum inspired evolutionary algorithm (developed at KEDRI), in which each solution represents all the possible solutions with a certain probability. These probabilities evolve according to the fitness function based on abstract interaction coefficients leading eventually to the optimal fit of expression profiles of gene clusters for the 2 hours Park et al's data. We have used the method of QiEA for inference of gene regulatory network for the first time. QiEA belongs to a class of probabilistic models of promising solutions to guide the exploration of the search space, the so-called Estimation of Distribution Algorithms (EDA). QiEA was introduced in [Defoin-Platel, Schliebs and Kasabov 2007] where it was also experimentally compared against classical genetic algorithm and quantum-inspired evolutionary algorithm on such benchmark optimization problems like OneMax problem, 01-knapsack problem and NK-landscape problem, leading to better optimization results. QiEA avoids elitist optimization strategy and thus obtains better results, and therefore we have chosen this method to optimize our abstract regulatory network. In addition, the strength of QiEA is that it is a multi-model EDA, which is easy to tune, has an adaptive learning speed (requires less number of iterations), and is robust against the decision error. However, in the presented case of an abstract regulatory network inference, we have not

compared QiEA with other algorithms as this was not the goal of this research.

We used these optimized interaction matrices obtained by QiEA to simulate virtual transgenic and gene knock-in mice experiments. As a result we have selected one matrix of regulatory coefficients that reproduces both the normal data and the data from transgenic and gene knock-in experiments for APP+PS1 mice. Some interaction coefficients have been validated using the information about interactions from NCBI Gene Entrez database. The rest can serve as predictions for experimental testing. These interactions are predicted for a mouse model of Alzheimer disease. It is to be expected from the criteria we have used for choosing this particular regulatory network that another disease animal model KIO virtual experiments will lead to a different abstract regulatory network that can serve for making testable predictions about gene interactions for those diseases, should we have the appropriate data.

Analysis carried out in this thesis chapter has led to several new findings. First, clustering analysis of genes and proteins has shown that about half of the set of 79 genes / proteins related to LTP are closely related to each other. One half of these closely related genes / proteins have also similar time course of gene expression during 2 hours after induction of LTP. Analysis of closeness of their promoter regions has shown that only 20 out of 79 genes have close promoter regions (sequence similarity). In spite of that also the genes which do not have close promoter regions in the clustering sense can be coordinated in expression to lead to orchestrated gene expression

patterns. Second, functional analysis of gene functions and temporal patterns of their expression have revealed that 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. So do other genes responsible for intracellular signalling, enzymatic reactions and synapse remodelling. Yet others are suppressed during the first 2 hours, either because they may come into play later as it seems that maintenance and stabilization of LTP is a long-term process, or they may be required to be suppressed all the time, like the genes which are related to Alzheimer disease and/or mental retardation. Genes, mutations of which have been related to various disorders the main symptoms of which are cognitive impairments are integral parts of this complex scenario of activity-evoked synaptic changes. Third, we have considered and abstract regulatory network between clusters of temporally synchronized genes. The goal of this analysis was to make predictions about gene interactions in a mouse model of particular disease. Based on available data we have chosen an animal model of AD, the APP+PS1 transgenic or gene knock-in mice. Optimized regulatory matrices served to simulate virtual gene knock-in experiments to be validated based on experimental data about consequences upon other genes in the abstract regulatory network. Some of the inferred regulatory coefficients were verified based on NCBI Entrez database information about individual genes. Surprisingly, in this inferred abstract regulatory network cluster 1 acted as a hub with at least twice as many connections coming in and out as any other cluster. This cluster contains genes that are under-expressed during the first 2 hours of LTP post-induction period. It also contains several crucial genes

involved in AD, like APP, ApoE, Aplp, and several genes related to mental retardation, like Arx and Fmr2. This finding may point towards a more important role of these genes in the whole process of LTP induction and maintenance than other genes. Therefore their mutations can have more profound consequences upon the whole process of LTP than any other gene mutations. At the same time we have successfully adapted and tested a new optimization method for inference of abstract gene regulatory networks, called quantum-inspired evolutionary algorithm, in an effort to shed more light on the underlying genetic mechanisms of LTP. However, the kind of data analysis envisaged in this research work can be applied to any temporal microarray data, not just those related to LTP, and as such has a general application span.

In accordance to the earlier proposed framework (refer chapter 3), we have shown that microRNA investigation is also very important for GRN study, therefore in the next chapter we have proposed novel bioinformatics method for miRNA classification.

7. Computational methods to discover novel microRNAs using 2-D structures

In the previous three chapters we have applied the different computational intelligence methods to infer the gene regulatory networks. In accordance to the earlier proposed framework (refer to chapter 3), we have shown that microRNA investigation is also very important for GRN study therefore in this chapter we have proposed a novel integrative computational method for classifying microRNAs by Gabor Filter Features from 2D Structure Bitmap Images. We have provided the foundation of the problem in the chapter 2 of this thesis and here in this chapter we initially provide the quick specification of the problem. Then we discuss some of the existing methods in this direction proposed by other research groups. It is followed by the explanation on our integrative approach of Gabor Filter, BLAST and CLUSTALW and the details of the undertaken case study on human microRNAs. We further discuss the important obtained results that we have already published in one of the international journal and other results that were published in the lecture notes in computer science. We conclude the chapter by summarizing and discussing our contribution in this area.

7.1 Introduction and problem specification

As mentioned within the review of this thesis some genes produces transcripts (miRNAs) that function directly in regulatory, catalytic, or structural roles in the cell. These are ~22 nucleotide-long RNAs that function in translational repression by base pairing with their target mRNA in a variety of plants and animals (refer to the figure 7.1). Thus, a new paradigm of gene expression regulation has emerged recently with the discovery of microRNAs (miRNAs) that adds a new dimension to our understanding of complex gene regulatory networks. Therefore, it is wise to propose a novel bioinformatics method in this doctoral thesis where we have already addressed several aspects of studying the gene regulation.

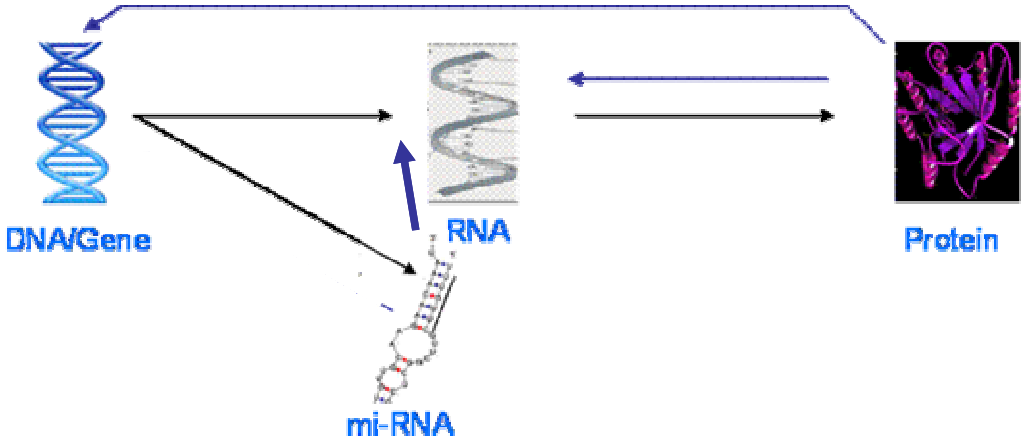


Figure 7.1: MicroRNAs as key players (base pairs with target mRNA) in gene regulation

Lack of conserved microRNA sequences or microRNA targets between animals and plants suggests that plant microRNAs evolved after the split of the plant lineage from mammalian precursor organisms. This means that the

information about plant microRNAs does not help to identify or classify most mammalian microRNAs. Also, in mammalian genomes the foldback structures are much shorter, down to only about 80 base pairs; making sequence similarity search a less effective method for finding and clustering remotely related microRNA precursors. For more details on the microRNAs, their biogenesis, potential role in the gene regulation and experimental techniques for their identification and available data readers are kindly requested to refer the chapter 2 of this thesis. In the coming sections we will talk about the current methods available in this area and then we explain our novel method to address the problem. Later we discuss the results and conclude the chapter with the justification of our approach.

7.2 Existing methods for microRNA classification

Cross-species sequence comparison is a powerful approach to identify functional genomic elements, but its sensitivity decreases with increasing phylogenetic distance, especially for short sequences [Bompfuenewerer et al. 2005]. The algorithms used for gene prediction are less efficient to predict miRNAs because miRNA sequences have a low similarity among sequences. In addition, the function of non-coding RNAs is uniquely determined by the three dimensional structure of the molecule. To reach its functional form, a single stranded RNA molecule undergoes folding – driven by GC/AU/GU base-pairing and stacking interactions – to form short helices and various

single stranded loop regions that define its secondary structure [Okazaki Y et al. 2002]. RNA secondary structure, the pattern of base-pairing, contains the critical information for determining the three dimensional structure and function of the molecule. Some RNAs also require metals or proteins to chaperone the folding process. Conservation of secondary structure, despite the sequence variability observed in the precursor sequences, suggests that the secondary structure plays a major role, presumably in the processing of the mature miRNA from the precursor. In particular, evolutionary conservation of secondary structures serves as compelling evidence for biologically relevant RNA function [Ding Y et al. 2003 and Fukushima K 1998]. These miRNAs have been discovered by various experimental methods such as northern blot and clone library etc. However, identifying miRNA by those experiments is considerably time-consuming and cost-expensive. Thus, we need a computational algorithm to efficiently predict miRNAs. This suggests that development of computational tools based on RNA secondary structure is essential for discovery of new microRNAs and classification of their functional roles. A variety of computational methods have used the secondary structure of RNA molecules to search and categorize miRNAs, but many of these methods have their own limitations.

Several sequence similarity and RNA folding based methods have been developed to find novel microRNAs. Simple BLAST similarity search identified e.g. orthologues of let-7 microRNA in several species [Pasquinelli et al. 2000].

The next approach has been screening by RNA fold prediction algorithms (best known are Mfold and RNAfold) to look for stem-loop structure candidates having a characteristically low deltaG value indicating strong hybridization of the folded molecule, followed by further screening by sequence conservation between genomes of related species. Softwares called MIRseeker [Lai et al. 2000] and MIRscan [Lim et al. 2003] have been used in this fashion for fruitfly (*Drosophila*) and human microRNA discovery, respectively. Most recently, about a thousand candidates conserved human microRNAs have been found by phylogenetic conservation based search strategy [Berezikov et al. 2005]. This method is based on careful multiple alignment of many different closely related primate species to find accurate conservation at single nucleotide resolution. The problem with all these approaches is that they require extensive sequence data and laborious sequence comparisons between many genomes as one key filtering step. Also, finding species-specific, recently evolved microRNAs by these methods is difficult, as well as evaluating the phylogenetic distance of too remotely related genes which have diverged too much in sequence.

Regular-expression like pattern matching algorithms have been used to scan genome sequences for regions that fold into the canonical structures of specific families [Xiu-JieWang et al. 2004]. However, they are designed to match stringent configurations of secondary structure elements, and therefore

perform poorly on families with variations in folding. Recently, some groups defined some statistical measures of miRNA precursors to predict miRNAs with respect to those of other species' miRNA precursors [Jia-Fu Wang et al. 2004, Stefan Washietl 2006 and Washietl et al. 2005]. However, these methods need the comparative analysis among miRNA precursors of evolutionarily similar species. If the miRNA precursors of one species have been not known, through such methods it is impossible to predict putative miRNA precursors in the other similar species. Thus, it is essential to develop the general algorithm to identify putative miRNA only using the structure and sequence information of miRNA precursors. Also, the general algorithm is important to search the common structure and the conserved sequences. In this doctoral thesis chapter 7, we examine whether the basic geometric and topological properties of secondary structure are sufficient to distinguish between RNA families in a learning framework and have presented a novel, simple, and computationally efficient approach for learning RNA secondary structures that requires no complicated tuning of parameters and can be applied to a wide range of learning problems.

7.3 Proposed integrative method of Gabor Filter, BLAST and CLUSTALW

Our approach that we describe in this chapter to classify microRNAs is based on the theme that the two-dimensional (2D) structure of many

microRNAs (and non-coding RNAs in general) can give additional information useful for their discovery and classification, even with data from within only one species. This is analogous to protein three-dimensional (3D) structure showing often functional and/or evolutionary similarities between proteins that cannot easily be seen by sequence similarity methods alone. Protein 3D structural comparisons are based on accurate protein crystallization data on atomic coordinates of amino acids in the polypeptide macromolecule chain. Unfortunately, such molecular structure data is scanty for RNAs in general, and for microRNA precursors in particular. Also, RNA folding simulation in 3D is still a difficult computational problem, just as the traditional grand challenge of deducing protein folding *ab-initio* from the amino acid sequence alone. Prediction of RNA folding in 2D is more advanced, and reasonably accurate algorithms are available, which can simulate the putative most likely and thermodynamically most stable structures of self-hybridizing RNA molecules. Many of such structures have been also verified by various experimental methods in the laboratory, corroborating the general accuracy of these folding algorithms.

Our novel method to address the problem described in this chapter is based on comparing bitmap images of microRNA structures. Here we approach the problem by utilizing visual information from images of computer simulated 2D structures of macromolecules. Gabor wavelet feature method has

been widely used both in multi-resolution image processing and feature extraction for object identification [Allen E et al. 2004, Fiser and King 1997 and Altschul SF 1990]. Gabor filter method can produce rotation-invariant distinguishing features of images that can then be used to evaluate image similarity. Application of this method to RNA secondary structures is novel, and represents a potentially useful new method for comparing structures of various other kinds of biological macromolecules, e.g. proteins or complex organic compounds. The innovation here is to use suitable artificial intelligence image analysis methods on bitmap images of the 2D conformation. This is in contrast of the traditional approach of using as a starting point various extracted features, like the location and size/length of loops/stems/branches etc., which can comprise a preconceived hypothesis what are the essential features of the molecule conformation.

The procedure is to take a sample of non-coding RNA sequences, calculate their 2D thermodynamically most stable conformation, output the image of the structure to bitmap images, and use a variety of rotation invariant image analysis methods to cluster and classify the structures without preconceived hypotheses what kind of features might be important ones. While one may lose specific information about the exact location or length of loops/stems or specific sequence motifs, the image analysis could reveal novel relevant features in the image, that may not be intuitively obvious to the human eye, e.g. fractal index of

the silhouette, ratio of stem/loop areas, handedness of asymmetric configurations etc.

7.4 Case study on human microRNAs dataset

Initially, we used for our experiments as input data - the combination of a set of 222 known human (*Homo sapiens*) microRNA precursors and as an example the Gabor Filter feature extraction method was tested for this new approach. The data was extracted from RFAM database mi-RNA Registry (version 5.0, December 2004, http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml) that provides a searchable database of published microRNAs [Griffiths-Jones 2004]. Now, the most recent version of this database is also available from [Griffiths-Jones S et al. 2008]. We used the provided full sequence of the stem-loop region (precursor microRNA) which also includes some flanking sequence of the presumed primary RNA transcript.

In the figure 7.2, we have shown the examples of human microRNA secondary structures from RFAM database miRNA Registry, folded by Vienna Package RNAfold algorithm with default parameters, specifying temperature of 37 degrees Celsius. Image similarities are calculated based on features generated by the Gabor Filter, and a similarity matrix is generated for all versus all pairs of the 222 microRNA precursor sequences. Additional data for

comparative analysis is used from similarity measures based on the BLAST and CLUSTALW algorithms for traditional sequence similarity search and the Vienna Package algorithm RNApdist algorithm for structural similarity index derived from the alignment of strings representing successive structural features of the molecule from one end to the other. Results indicate bitmap image analysis of RNA molecule 2D structure can yield new and useful information for classification of discovery of microRNA genes. This kind of novel information could be used in discovery and classification based on multimodal data of microRNA gene candidates from genomic sequences of various organisms. In turn, functional classification of microRNAs and more complete understanding can lead to the modelling and knowledge discovery of molecular interactions.

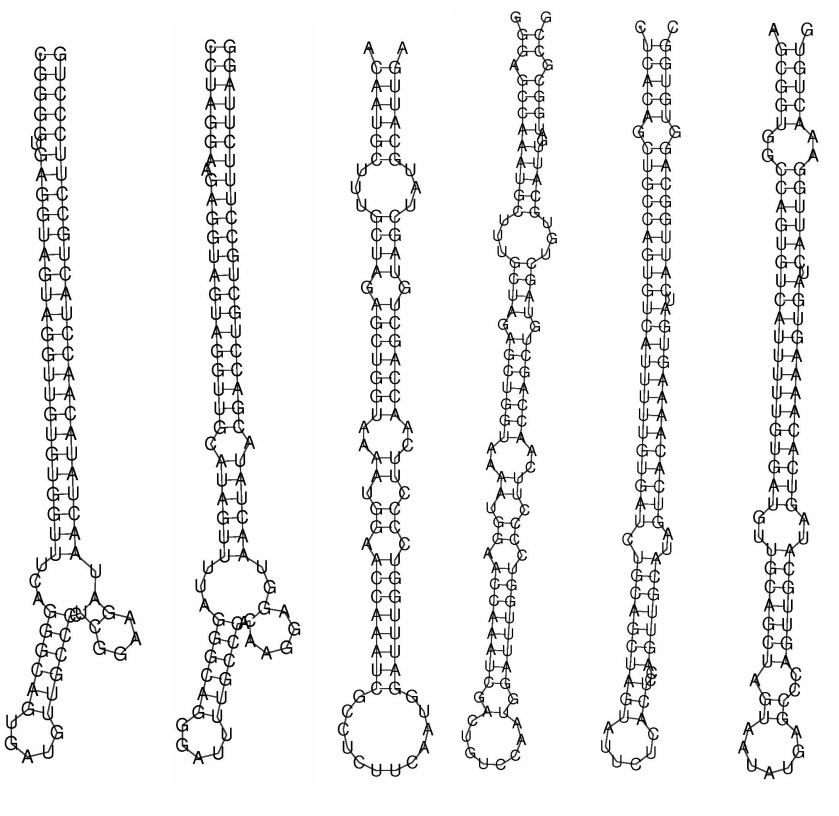


Figure 7.2: Sample pairs of microRNA structures looking similar to human visual system are shown. From left to right: microRNA sequence # 4, 6, 43, 44, 69 and 70 (numbers assigned in our dataset), corresponding to microRNA genes hsa-let-7b, hsa-let-7d, hsa-mir-133a-1, hsa-mir-133a-2, hsa-mir-153-1 and hsa-mir-153-2 in the RFAM database, respectively.

For sequence similarity searches, we used NCBI-BLAST developed by [Altschul et al. 1990], version blast-20041205-ia32-linux.tar.gz, obtained from ftp://ftp.ncbi.nih.gov/blast/executables/ installed on a RedHat Linux (version 8). Blast search was done using command line interface, with all sequences blasted against each other with BLASTN algorithm, default parameters and the following options: blastall -p blastn -m 9. The output was used to obtain the all versus all matrix of sequence similarities represented by Blast scores.

Sequence similarity between all the 222 microRNAs was visualized using multiple sequence alignment program CLUSTALW package (http://align.genome.jp/) developed by Koichi Ohkubo, (GenomeNet), for details see [Chenna et al. 2003]. Multiple alignment was run with parameters DNA Gap Open Penalty = 10.0 DNA, Gap Extension Penalty = 0.05 and using CLUSTALW DNA Matrix for all 222 sequences first, then the ordered data divided into four groups of ~ 55 sequences and circular phylogram trees were drawn using unrooted neighbour joining (NJ) algorithm.

For structural analysis of the RNA sequences we used the Vienna Package (version 1.5 beta, _http://www.tbi.univie.ac.at/~ivo/RNA/) [Hofacker 2003). For obtaining the putative 2-D structures, the algorithm RNAfold was used, with default parameters and with temperature setting T = 37 deg Celsius, The human body temperature was used, because this is the natural environment for the intracellular RNA molecules, and temperature can affect the most likely conformation an RNA molecule folds into. To calculate structural distances between different thermodynamically optimal secondary structures of all sequences, the algorithm RNApdist was used, with options -Xm and T=37 degrees Celsius. A matrix of all versus all similarities was obtained.

RNAfold produces postscript images of the most likely 2-D conformation of each RNA molecule based on its nucleotide sequence and self-hybridization.

These were converted to 8-bit greyscale bitmap images of 512 by 512 and 256 by 256 pixels using ImageMagick version 6.2.2 (http://www.imagemagick.org) installed on a Windows XP computer, run by a DOS script for batch processing. Bitmap images were used as input for image processing using MatLab (version 7.0) package running on Window XP computer. Algorithms for Gabor filter based image analysis were implemented as in-house developed MatLab scripts for processing the input data produced by BLAST, Vienna Package and the images converted to bitmaps.

Given an image with size $P \times Q$, a 2D Gabor function is a Gaussian modulated by a sinusoid. It is a non-orthogonal wavelet and it can be specified by the frequency of the sinusoid $w = 2\pi f$ and the standard deviations of Gaussian σ_x and σ_y [Amari et al. 1998]

$$g(x, y : f, \theta) = \exp\left(-\frac{1}{2}\left(\frac{x'^2}{\sigma_x^2} + \frac{y'^2}{\sigma_y^2}\right)\cos(2\pi f x')\right)$$

Where, f is the frequency of the sinusoidal plane wave along the direction θ from the x-axis, σ_x and σ_y specify the Gaussian envelope along x and y axes, respectively, which determine the bandwidth of the Gabor filter. For our experiment data (8 bits gray human microRNA structure image, with the size 512×512), 20 spatial frequencies are used, with $f = k\pi / 2^i, (i = 1,...,5)$ and $\theta = k\pi / 4, (k = 0,...,3)$.

Figure 7.4 illustrates an example of the above Gabor wavelet decomposition on a human microRNA # 182 image shown in figure 7.3 with the size 256×256, where the original image is decomposed in the frequency of 5.2 into four images that identify the energy distribution in four directions.

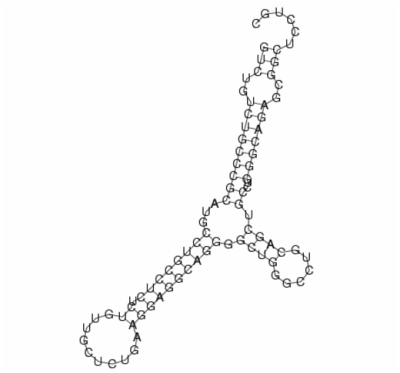


Figure 7.3: MicroRNA #182 structural image in size of 256 by 256 pixels.

Applying Gabor filters on the image with different orientation at different scales, we obtain an array of magnitudes of the filtered images:

$$E(f,\theta) = \sum_x \sum_y |\, g(x, y, f, \theta)\,|,\ f = k\pi / 2^i, (i = 1,...,5) \text{ and } \theta = k\pi / 4, (k = 0,...,3)$$

(a) $\theta = 0$ (b) $\theta = \pi / 4$

(c) $\theta = \pi / 2$ (d) $\theta = 3\pi / 4$

Figure 7.4: Gabor Wavelet Decomposition on Human microRNA # 182 of figure 7.3, $f = 5.2$,

$$\theta = k\pi / 4, (k = 0,...,3)$$

Since the main purpose of microRNA analysis is to find microRNAs with images having similar direction, shape, and similar texture of base on e.g. relative compositions of the nucleotides A, U, C and G. To this end, for each of the microRNA, we conducted the above Gabor decomposition on two resolution images (512 x 512 and 256 x 256). The higher resolution Gabor decomposition seems better tuned for the texture features extraction, and the

lower resolution emphasizes direction and shape features extraction. The lower resolution was used in the further analyses reported below.

For feature encoding, the following magnitude of the transformed coefficients is used to represent the features of microRNAs from the viewpoint of visual information in the 2D image of the molecule:

$$u_{f_i,\theta_j} = \frac{E(f_i,\theta_j)}{P \times Q}$$

Suppose *M* frequencies and *N* orientations are used in above Gabor decomposition, a feature vector for a microRNA image can be written as

$$f = [u_{f_1,\theta_1}, u_{f_1,\theta_1}, ..., u_{f_M,\theta_N}]$$

As scoring the scaling of the similarity of a pair of microRNAs , we use the City Block distance measurement, also known as the Manhattan distance measurement, to measure distance between the two feature vectors of the two microRNA Gabor features being compared:

$$d < f_a, f_b >= \sum_{i=1}^{N-1} | f_{ai} - f_{bi} |$$

Where, f_a and f_b are the query and target feature vectors, respectively. Manhattan distances were calculated for all versus all pairs of microRNA precursors to create a similarity matrix, like previously done for the BLAST similarities and the RNApdist similarities.

7.5 Results, discoveries and biological validation

Similarity measures between sequences and structures produced three symmetric similarity matrices 222 by 222 microRNA precursors, one for BLAST sequence similarity, one for RNApdist structural similarity index, and one for the bitmap image Gabor Filter feature Manhattan distance. These matrices are shown as heat colour maps in figure 7.5. In the figure colour scaling has been manually adjusted to bring out the maximum amount of information from the similarity values (rescaled to interval 0-1 before heatmap generation) throughout the matrix.

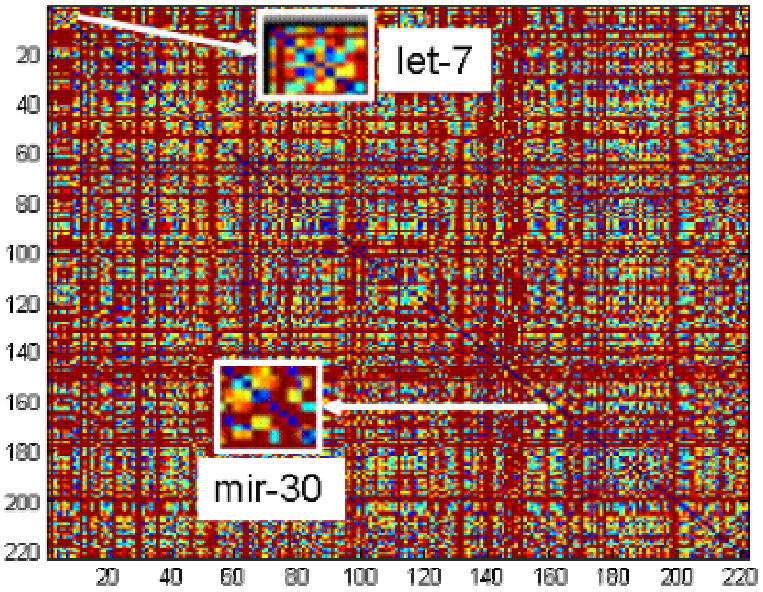
The most striking difference between the similarity matrices is the scarcity of strong BLAST similarities across most of the compared pairs in the dataset of 222 human microRNA precursors, shown by the mostly blue areas in the middle figure indicating no BLAST similarity hits (low blast score values). Both Gabor Filter visual similarity measure and RNApdist structural similarity index show much more complex pattern of similarity differences between all the pairs compared, suggesting more varied information was captured, which was detected by BLAST.

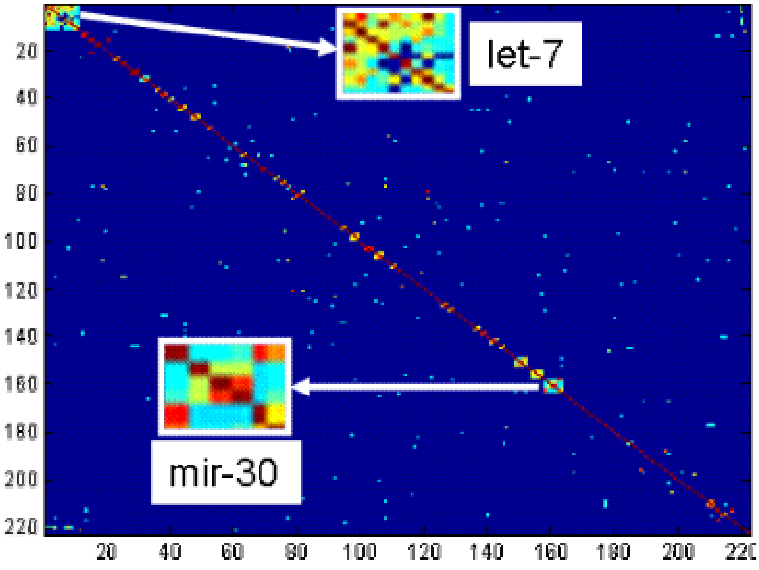
Self-similarity diagonal is evident in all matrices, but most clearly in the BLAST matrix visualization which is partly due to the favourable colour scaling for contrast in this heatmap. Heatmap magnification insets in the top left corner

show a cluster of ~10 similar sequences of the well-known let-7 family of related microRNA sequences, clearly visible as an area of higher similarity, and evident both in the Gabor Filter and the BLAST data (top and middle of figure 7.5). Similarly, the known phylogenetically related family of several mir-30 genes around sequences #160-168 is clearly visible in both BLAST and GABOR feature matrices.

The two above-mentioned clusters (let-7 family and mir-30 family) are less clearly distinguished in the RNApdist matrix. The RNApdist structural distance measure may not be performing adequately, possibly due to varying length of the microRNA precursors being compared pair-wise, or to this similarity metric not overlapping as much with BLAST similarity as the Gabor Filter measure.

Overall, the two structural similarities measured by Gabor and RNApdist seem to show more complex clustering of the microRNA precursors than the BLAST sequence similarity. There is, however, a clear partial overlap in the detected similarities among the three measures.


let-7
mir-30
20
40
60
80
100
120
140
160
180
200
220


let-7
mir-30
20
40
60
80
100
120
140
160
180
200
220

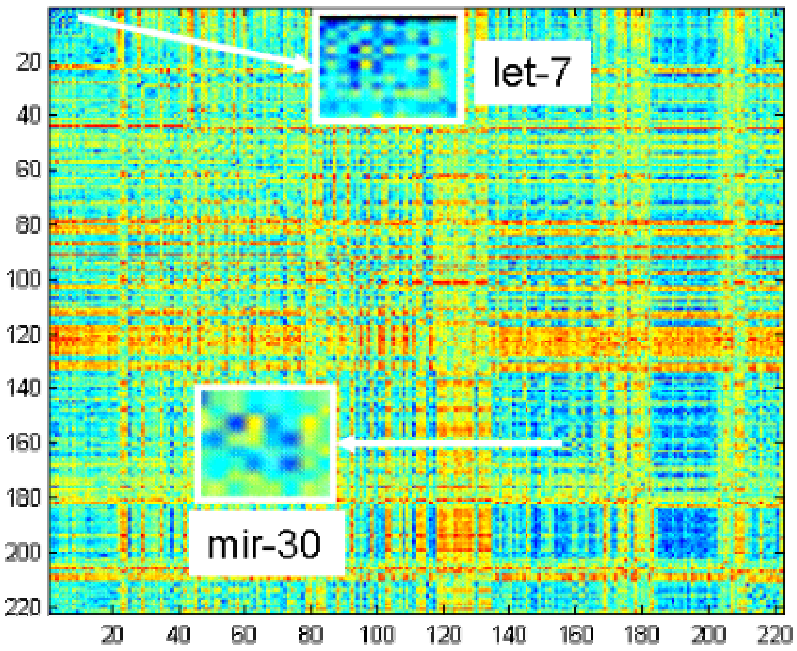


Figure 7.5: Similarity matrices as heat maps for 222 human microRNA precursors all versus all comparisons. X and Y axis: #microRNAs. Top: Gabor Filter feature vector Manhattan distances, Middle: BLAST sequence similarity based on BLAST score, Bottom: Structural similarity based on RNApdist score. Insets show magnifications of the approximate positions of microRNA families let-7 (top left) and mir-30 (lower right) in the matrices.

The relatedness of the let-7 and mir-30 genes and several other clusters in the heatmaps is corroborated by the CLUSTALW analysis, shown in figure 7.6 as a circular clustering tree diagram, with the two clusters of let-7 and mir-30 highlighted. Most of the other clusters are small, containing just a few sequences. This is as expected, as human microRNAs are known to be grouped into small independent families by sequence similarity, thought to be of either independent origin in evolution or diverged so long time ago that all detectable sequence similarities have disappeared, showing no obvious DNA sequence conservation.

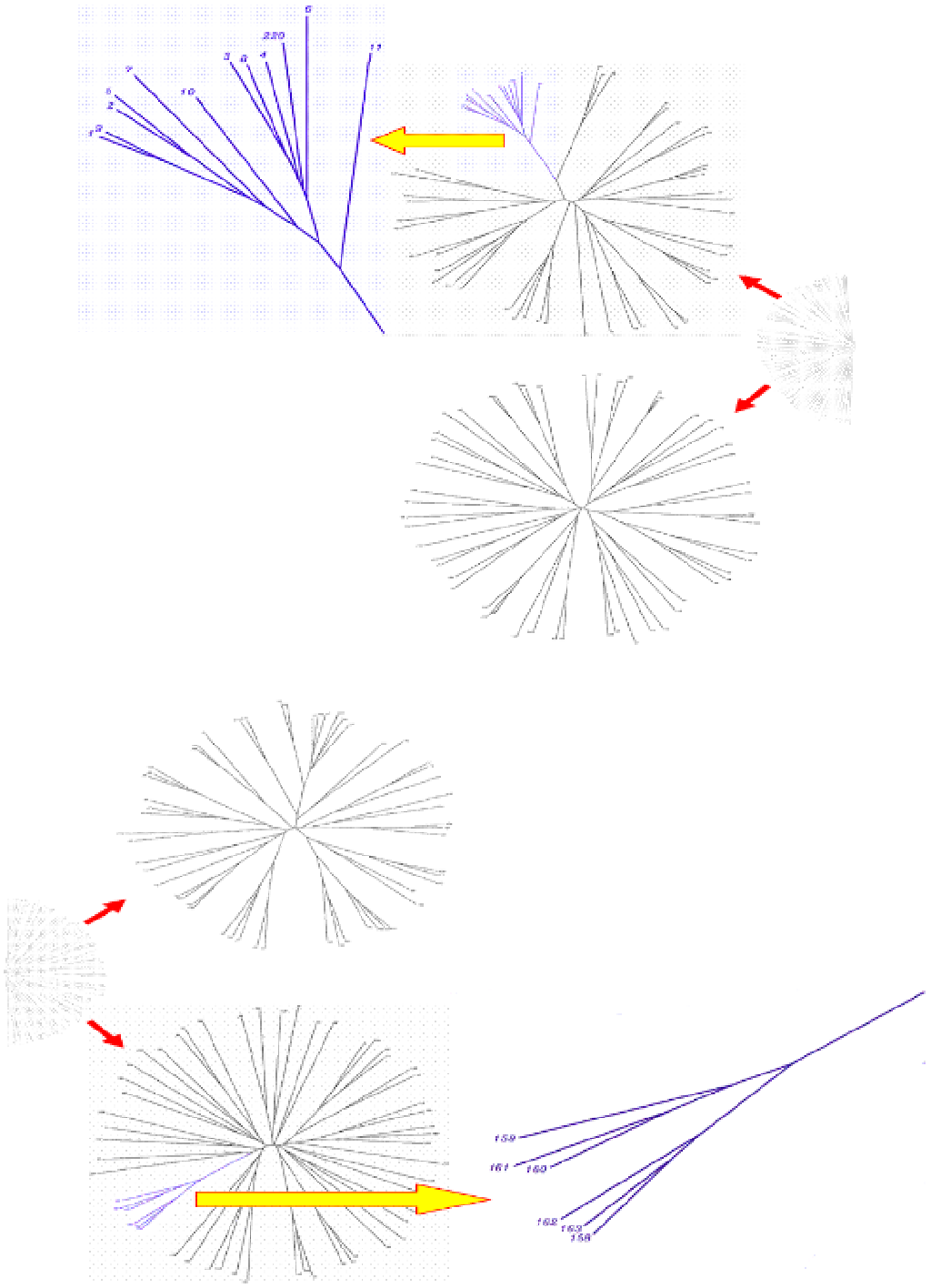


Figure 7.6: Clustering tree based on CLUSTALW algorithm showing DNA sequence similarity based relatedness among 222 human micro-RNA precursors (divided into four subsets for clarity of presentation). Clusters of the microRNA families let-7 (top) and mir-30 (bottom) are magnified in the insets as examples.

The top 10 ranked similarities for Gabor Filter, BLAST and RNApdist were then compared, as shown in table 7.1. As expected, BLAST detected best members of known microRNA families, and the overlaps to pairs detected by the other two methods are only partial. This suggests that sequence similarity measures (e.g. BLAST) extract different information about related microRNA precursors than visual information as measured by Gabor Filter features or structural similarity index RNApdist. This agrees with the very different colour patterns in the three similarity matrix representations in figure 7.5.

Table 7.1: Top ten ranked most similar pairs of human microRNA precursors as measures by BLAST sequence similarity, Gabor Filter feature vector Manhattan distance and RNApdist structural similarity index. Gene names which occur more than once in the table are in bold, pairs on microRNA precursors in known families are shaded grey.

Algorithm sort order	**Rank**	**Gabor distance ***	**BLAST score ****	**RNAdistp score ***	**pair #**			**Gene names**	
Gabor	1	0.09	145.20	20.14	17	-	18	**mir-105-1**	**mir-105-2**
	2	0.54	0.00	37.87	35	-	132	**mir-127**	**mir-221**
	3	0.54	0.00	45.94	132	-	157	**mir-221**	**mir-302d**
	4	0.60	0.00	37.09	35	-	157	**mir-127**	**mir-302d**
	5	0.66	0.00	30.44	11	-	35	let-7i	**mir-127**
	6	0.66	0.00	41.64	132	-	200	**mir-221**	**mir-380**
	7	0.76	0.00	21.81	25	-	219	**mir-1-1**	mir-96
	8	0.80	0.00	16.90	157	-	200	**mir-302d**	**mir-380**
	9	0.85	0.00	35.22	132	-	158	**mir-221**	**mir-30a**
	10	0.86	0.00	18.23	158	-	200	**mir-30a**	**mir-380**
Blast	1	0.09	145.20	20.14	17	-	18	***mir-105-1***	***mir-105-2***
	2	18.55	111.50	35.54	29	-	30	mir-124a	mir-124a-3
	3	16.59	103.60	10.03	28	-	29	**mir-124a-1**	**mir-124a-2**
	4	5.69	99.61	27.43	212	-	213	**mir-9-1**	mir-9-2
	5	11.38	91.68	24.20	1	-	9	let-7a-1	let-7f-2
	6	420.71	79.79	17.80	15	-	16	**mir-103-1**	**mir-103-2**
	7	7.77	77.80	31.60	2	-	5	let-7a-2	**let-7c**
	8	10.97	71.86	23.56	102	-	104	mir-199a-1	mir-199b
	9	12.52	69.88	19.98	212	-	217	**mir-9-1**	**mir-9-3**
	10	13.65	65.91	25.42	15	-	21	**mir-103-1**	mir-107
RNApdist	1	0.00	2.92	8.76	25	-	188	**mir-1-1**	mir-368
	2	0.00	14.84	8.77	141	-	146	mir-26a-1	mir-28
	3	0.00	25.90	9.49	187	-	198	mir-367	mir-378
	4	0.00	411.07	9.57	16	-	20	**mir-103-2**	**mir-106b**
	5	40.14	9.91	9.69	4	-	5	let-7b	**let-7c**
	6	0.00	12.91	9.70	20	-	215	**mir-106b**	mir-92-2
	7	0.00	24.26	9.88	172	-	192	**mir-106b**	mir-372
	8	103.60	16.59	10.03	28	-	29	**mir-124a-1**	**mir-124a-2**
	9	0.00	17.00	10.05	96	-	217	mir-195	**mir-9-3**
	10	0.00	22.41	10.13	63	-	186	mir-148a	**mir-9-3**

* Lower value means more similarity

** Higher value means more similarity

7.6 Conclusion and discussion

Classification of microRNAs is an important step towards understanding gene regulatory networks. Distinguishing remote evolutionary or functional relatedness by sequence similarity among the human microRNA precursors using only human data, i.e. without comparisons to other mammalian genomes) is difficult with the simple BLAST similarity search or multiple sequence alignment by CLUSTALW. This is due to extensive sequence dissimilarity, reflecting both ancient divergence of microRNAs of common origin and the possibly independent origin of many microRNA families. This explains the small number of clear similarities among the human microRNA precursors found by BLAST (middle matrix in figure 7.5) and lack of deeper branches in the clustering tree produced by CLUSTALW (figure 7.6). The functional relevance of the additional Gabor Filter based similarities uncovered by visual information extracted from bitmaps of simulated 2D structures of the microRNA precursors is a topic of further research.

The new information obtained by Gabor Filter features may reflect structural similarities persisting longer than sequence similarity (measured by BLAST) during evolution of remotely related microRNA genes of common origin. Alternatively, the structural similarities, as measured here, may be due to convergent selection leading to similar 2D structures by common constraints

for the function of the microRNA precursors in eukaryotic cells (e.g. the requirement of mostly symmetric single long hairpin structure having a straight region for the excision of the active 20-24 base-pair long microRNA).

In this chapter, we have showed that using visual information from bitmap images of 2D structures of microRNA precursors one can extract novel and useful features (up to certain extent) that can be used for discovery and classification of related molecules and ultimately to understand the process of gene regulation in a better way. Our proposed generic integrative approach may be easily applied to classify any microRNA dataset. The visual feature similarities could be used alone or preferably in combination with other sequence or structure based features for multimodal data learning by artificial intelligence methods to distinguish and discover novel microRNA precursors. For this high-throughput RNA folding could be used to generate the initial population of candidates for analysis. We have discussed some of the potential applications of the microRNA discovery for cancer and brain research in this thesis chapter 9.

Finally, in accordance to the figure 3.1 (refer to chapter 3) we have to focus on knowledge integration and information fusion and apply some machine learning tools in an integrated framework for knowledge elicitation and discovery. It is described in the next chapter of this thesis.

8. Integrative Brain-Gene Ontology (BGO) and simulation system

As we have gone through with the different chapters of this thesis, it is seen that for our investigations on understanding molecular interactions within cell we have considered different case studies and the overall analysis has been done by considering the data at the molecular sequence level, the microRNA structure level and the gene expression levels. In accordance to our proposed integrative framework, later as a proof of concept the knowledge integration on brain gene data was planned to be implemented through the ontology-assisted approach called brain-gene ontology (BGO) system [Kasabov, Jain et al. 2006, 2007, 2008]. This chapter presents and discusses the evolving brain-gene ontology (BGO) that is developed as an integrated system for storing data and knowledge and for modelling and discovery of complex interactions between genes and brain functions. In this respect we also describe some of the novel discoveries made using BGO and we conclude the chapter by summarizing some of the important facts.

8.1 Introduction and problem specification

It has been seen that the previous studies on the GRN problem demonstrated results for only one domain of the central dogma of molecular biology, so there has always been a lack of knowledge integration. For

thorough investigation of this GRN research area, there is a need to examine each regulatory stage of a cell and later integration is required to make the possible reuse of this knowledge. We have studied the gene regulation (GRN) problem from a different viewpoint by integrating multiple aspects using several integrative computational intelligence approaches and also proposed novel bioinformatics method for the microRNA classification. We have seen throughout this thesis - it is very important to build reusable models for the central dogma theme. Thus, more importantly the knowledge, facts and the data (related to genes, diseases and interactions) which reside in different databases need to be brought together as a part of a knowledge integration process. We have done so with the use of ontology and have demonstrated it by taking a case study on brain gene data and developing a system called Brain Gene Ontology (BGO). Knowledge can be accessed and reused to facilitate new discoveries and it is demonstrated later in this chapter.

We have designed the BGO to facilitate active learning and research in the areas of bioinformatics, neuroinformatics, information engineering, and knowledge management and we claim that different parts of it can be used by different users, from a school level to postgraduate and PhD student level. The BGO is concerned with the collection, presentation and use of knowledge in the form of global and shared access ontology. BGO includes various concepts, facts, data, software simulators, graphs, animations, and other

information forms, related to brain functions, brain diseases, their genetic basis and the relationship between all of them. The BGO has an open and evolving structure with knowledge and data added continuously. The current version, that includes over 400 genes/proteins and other data, related to brain functions and diseases is implemented in an ontology building environment endowed with plug-ins. The BGO allows users to: navigate through the rich information space of brain functions and brain diseases, brain related genes and their activities in certain parts of the brain and their relation to brain diseases; to run simulations; to download data that can be used in a software machine learning environment, such as WEKA and NeuCom to train prediction or classification models; to visualise relationship information; to add some new information as the BGO has an evolving structure [Kasabov, Jain et al. 2007].

For knowledge elicitation and inference, BGO system may require specific operations (e.g., querying using the interface). Using the BGO a complex interaction network was found and displayed for the GABRA1 gene. Integrating the power of BGO with the computational intelligence (CI) module, gene expression data modelling and genetic profile discovery was done for brain cancer tumour response to treatment prognosis, using an evolving connectionist model. Although there was no single gene found to accurately predict response to drugs, a set of ten genes was found to give 85% correct prediction (72% for the class of non-responding, and 92% for the responding to

drugs class), i.e.: HMG-I(Y), NBL1, UBPY, Dynein, APC, TARBP2, hPGT, LTC4S, NTRK3, and Gps2. Through the BGO these genes were found to be involved in other brain functions and diseases, and interacting with other brain-related genes [Kasabov, Jain and Benuskova 2007].

We have used the BGO data for neuronal gene-protein sequence and clustering analysis and totally new hypothesis was driven in the area of genetic neuroscience. We applied other CI tools like CLUSTALW on the BGO data i.e. subunit proteins for “amino-methylisoxazolepropionic acid receptor” (AMPAR), “gamma amino butyric acid receptor” (GABAR) and “N-methyl-D-aspartate acid receptor” (NMDAR) for clustering and sequence analysis. Result of analysis clearly showed us the extent of conservation of many amino-acid residues among all investigated receptor subunits and the most interesting investigation is the consistent conservation of phenylalanine (F at position 269) and leucine (L at position 353) in all 20 proteins taken into account with no mutations. We expect these residues to play some role as a binding centre for interaction of these proteins with several other genes/proteins such as c-jun, mGluR3, Jerky, BDNF, FGF-2, IGF-1, GALR1, NOS and S100beta that are also believed to have a regulatory effect upon these receptors. Based on such observations we assume that the expression of these individual subunits should be coordinated within one gene group [Benuskova, Jain et al. 2006]. In addition, these regions can be the basis for mutual interactions. Mutual

interaction between subunits of different receptors has been recently confirmed experimentally. So in such a way we believe that the developed BGO system is a very valuable addition to the scientific community and this integrated system will keep evolving and one can always input the experimental results and simultaneously a domain expert may access this integrated knowledge for future research or to educate peoples.

8.2 BGO: An overview, aims and goals

Our approach through this thesis chapter shows integrating various kinds of useful data and knowledge through the use of so called brain-gene ontology (BGO) and using it as knowledge and data repository to facilitate new knowledge discovery and better understanding of brain and related processes. We aim to build a multi-dimensional biomedical ontology to be able to share knowledge from different experiments undertaken across aligned research communities in order to connect areas of science seemingly unrelated to the area of immediate interest. In our investigation we have developed and represented ontology of genes and proteins that are related to specific brain related disorders like Epilepsy, Schizophrenia, Parkinson, Alzheimer disease, Rett syndrome and Mental retardation etc. The developed and presented Brain-Gene Ontology (BGO) in this thesis is focused on mammalian brain and has a broader scope than GO in a sense that we cover the gap in integration of knowledge that comes from different disciplinary domains such as

neuroscience, bioinformatics, genetics, computer and information sciences. We present some results in this direction and argue that our approach introduces an interesting new dimension to the problem that is likely to reveal novel knowledge in the future. The idea of BGO is to express and share knowledge that can be extended and adapted to support model-designed solutions and operations. The role of BGO is to catch up with the theme “how each gene affects the brain functions” (refer figure 8.1) and to produce a dynamic knowledge representation and evolving conceptual structure that can be used as knowledge of gene and protein roles in the brain and to facilitate new discoveries.

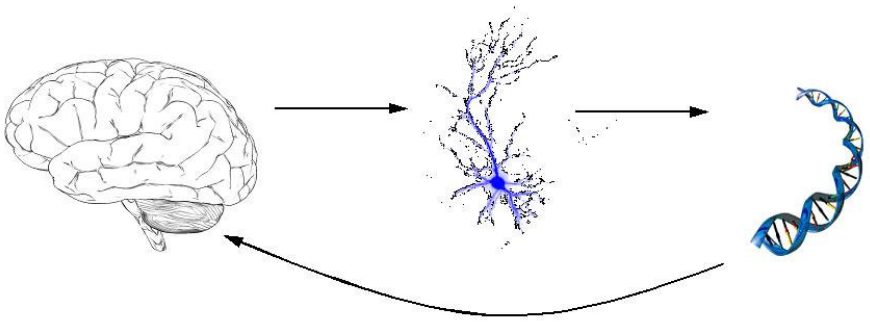


Figure 8.1: BGO is concerned with the accumulation and the use of data and knowledge for a better understanding and further discoveries of relationships between the brain, diseases and mammalian genes

In summary we can say the aims and goals of BGO (refer figure 8.2) are in the context of - (a) to facilitate education and science research (b) providing understanding on brain structure and hierarchical functioning system (c) teaching central dogma, bioinformatics databases and gene regulatory

networks within neurons (d) highlighting importance and use of ontologies in knowledge management and interpreting relationships among molecules.

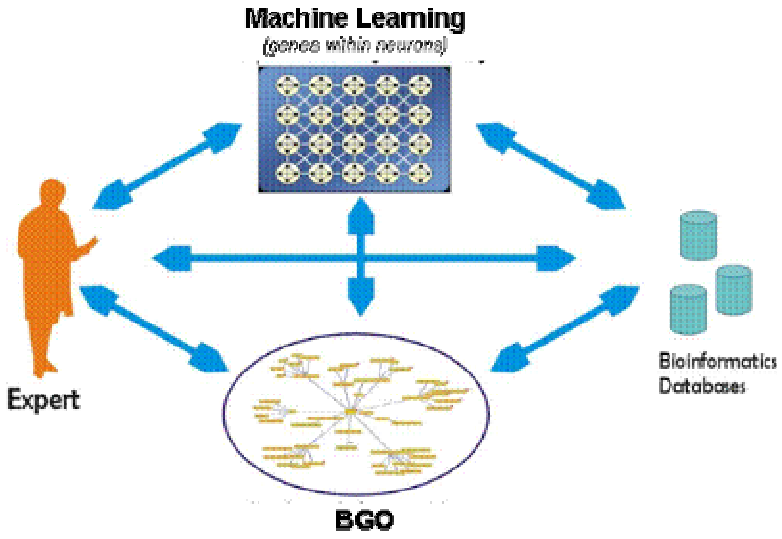


Figure 8.2: BGO – aims and goals; data or information can come from any source that can be analysed using machine learning tools and results may be interpreted by domain expert

The BGO is an evolving system that is changing and developing with the addition of new facts and knowledge in it by multiple users. Linking selected structured bodies of physiological, genetic and computational information provides a pathway for different types of users. Designing an interface that enables users with different levels of expertise, specialization and motivation to access the BGO – either through a familiar or specialist approach, or through a more general introduction is a critical issue that we solve with the help of specific Protégé engines and plugins. In this thesis chapter we describe how the information is organized in the BGO system, the environment in which it is implemented, and how we can use the system to aid novel discoveries by

mean of bioinformatics and computational intelligence methods. The role of the BGO to facilitate the education is discussed next and then we conclude this chapter with important discussion, summary and taking about BGO availability.

8.3 Implementation of brain-gene ontology system

In this section of this thesis chapter we suggest an integration of evolving and globally shared brain-gene ontology (BGO) and describe the implementation of BGO. The overall system is comprised of three main parts: (a) brain organization and functions; (b) genes and gene regulatory networks; and (c) a simulation module. Brain organization and function module contains information about neurons, their structure and the process of spike generation. It also describes processes in synapses and electroencephalogram (EEG) data for different brain states, in particular for the normal and epileptic state and also describes the processes in synapses. The genes and gene regulatory networks (GRN) part is divided into sections on neurogenetic processing, gene expression regulation, protein synthesis and abstract GRNs. The third large part, the simulation module, has sections on computational neurogenetic modelling (CNGM) [Benuskova and Kasabov 2007], evolutionary computation, evolving connectionist systems (ECOS) [Kasabov 2003 and 2007a] and other simulation tools. The user can navigate further into these sections and their subsections down to the genetic level and use the information for learning and research. Figure 8.3 shows the overall information structure of the BGO.

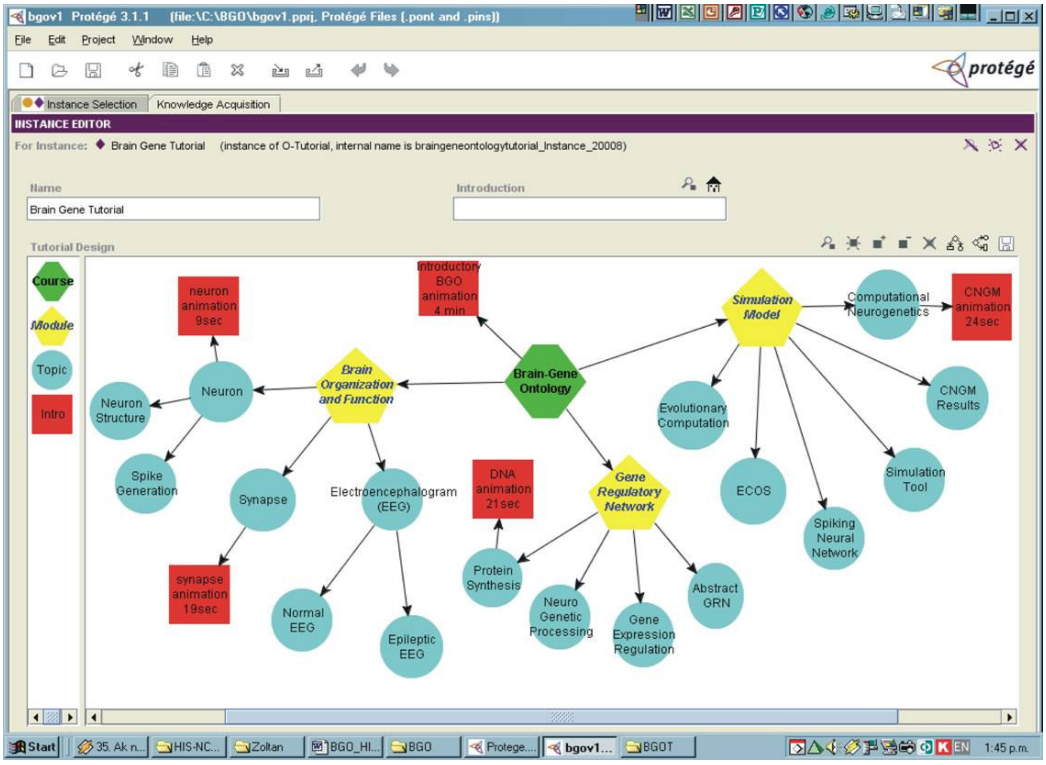


Figure 8.3: Snapshot of the BGO information structure with three main parts and their divisions. BGO is concerned with the accumulation and the use of data and knowledge for a better understanding and further discoveries of relationships between the brain, diseases and mammalian genes. In green: course, in yellow: modules, in blue: topics and in red: animations are shown

Evolving Implementation of BGO in Protégé

The first version of the BGO has been implemented in Protégé, which is open source ontology building environment developed by the Medical Informatics Department of the Stanford University (protege.stanford.edu). Protégé is an open-source ontology building environment developed by the

Medical Informatics Department of the Stanford University (http://protege.stanford.edu/index.html). At its core, Protégé implements a rich set of knowledge-modelling structures and actions that support the creation, visualization, and manipulation of ontologies in various representation formats. Protégé can be extended by way of a plug-in architecture and a Java-based Application Programming Interface (API) for building knowledge-based tools and applications. Our team has developed a set of plug-ins to enable to visualize, extract and import knowledge from/into different data sources and destinations. The Protégé platform supports two main ways of modelling ontologies:

- The Protégé-Frames editor enables users to build and populate ontologies that are frame-based, in accordance with the Open Knowledge Base Connectivity protocol (OKBC). In this model, ontology consists of a set of classes organized in a subsumption hierarchy to represent a domain's salient concepts, a set of slots associated to classes to describe their properties and relationships, and a set of instances of those classes - individual exemplars of the concepts that hold specific values for their properties.

- The Protégé-OWL editor enables users to build ontologies for the Semantic Web, in particular in the W3C's Web Ontology Language (OWL). OWL ontology may include descriptions of classes, properties and their

instances. Given such ontology the OWL formal semantics specifies how to derive its logical consequences, i.e. facts not literally present in the ontology, but entailed by the semantics. These entailments may be based on a single document or multiple distributed documents that have been combined using defined OWL mechanisms.

The information in the BGO is based on the two most used biological data sources, namely Gene Ontology, and Unified Medical Language System – UMLS, along with knowledge integrated from Entrez Gene, SwissProt, Interpro, Gene Ontology, Gene Expression Atlas, OPHID and others. It also incorporates knowledge acquired from biology domain experts and from different literature databases such as PubMed. Knowledge acquisition (KA) tab features the graphical presentation of relations by specific Protégé means (dynamic graphs, attached documents and pictures) and it describes concepts and their relations in a way that is both close to human language and formal. BGO can be viewed as a declarative model that defines and represents the concepts existing in the domain of brain and genes, their attributes and the relationships between them. The user can navigate into each instance to obtain the description, illustrations, and links to PubMed publications and other relevant web resources. BGO is represented as a knowledge base which is available to applications that need to use and/or share the knowledge of the domain. BGO utilizes a novel evolving conceptual metadata structure which

allows incorporation of new discoveries and adapt its structure. This evolving structure keeps track of change and provenance of source, date, among others [Gottgtroy, Kasabov and MacDonell 2006]. Thus, the ontology framework that we have developed enables hierarchical representation of relationships between genes, proteins, neurons and brain functions in a complex evolving structure [Jain et al. 2006, 2007].

Another feature of the BGO is the graphical presentation of relations by specific Protégé means (dynamic graphs, attached documents and pictures). There are many plugins that can be used to navigate, browse and visualize the information available within BGO. For example, as shown in figure 8.4, using OntoViz particular instances or appropriate classes can be selected and displayed in the form of a hierarchical graph. Few of the general domains are shown in this figure, out of many in the whole BGO.

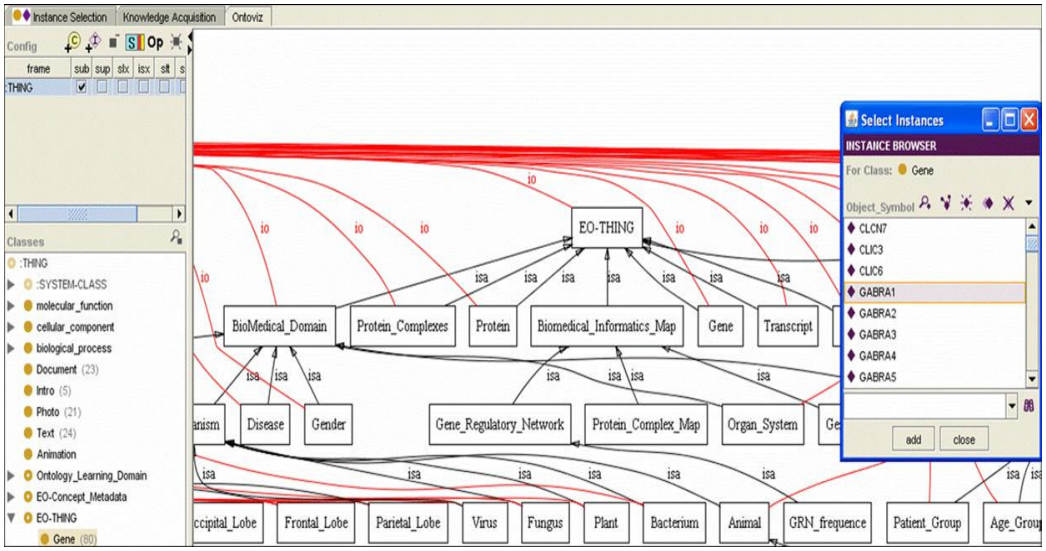


Figure 8.4: BGO domains visualization using OntoViz

As another example, using the plug-in called TGViz (touch graph visualization) we have explored the relationship of one gene, GABRA1, with several other molecules present in the BGO. The graph in figure 8.5 illustrates the detailed information available in the BGO about relations of GABRA1 with other genes, proteins, brain regions, molecular functions etc. The user can further navigate into each instance or class and their subsections down to the genetic level and use the relationship found for the learning and research.

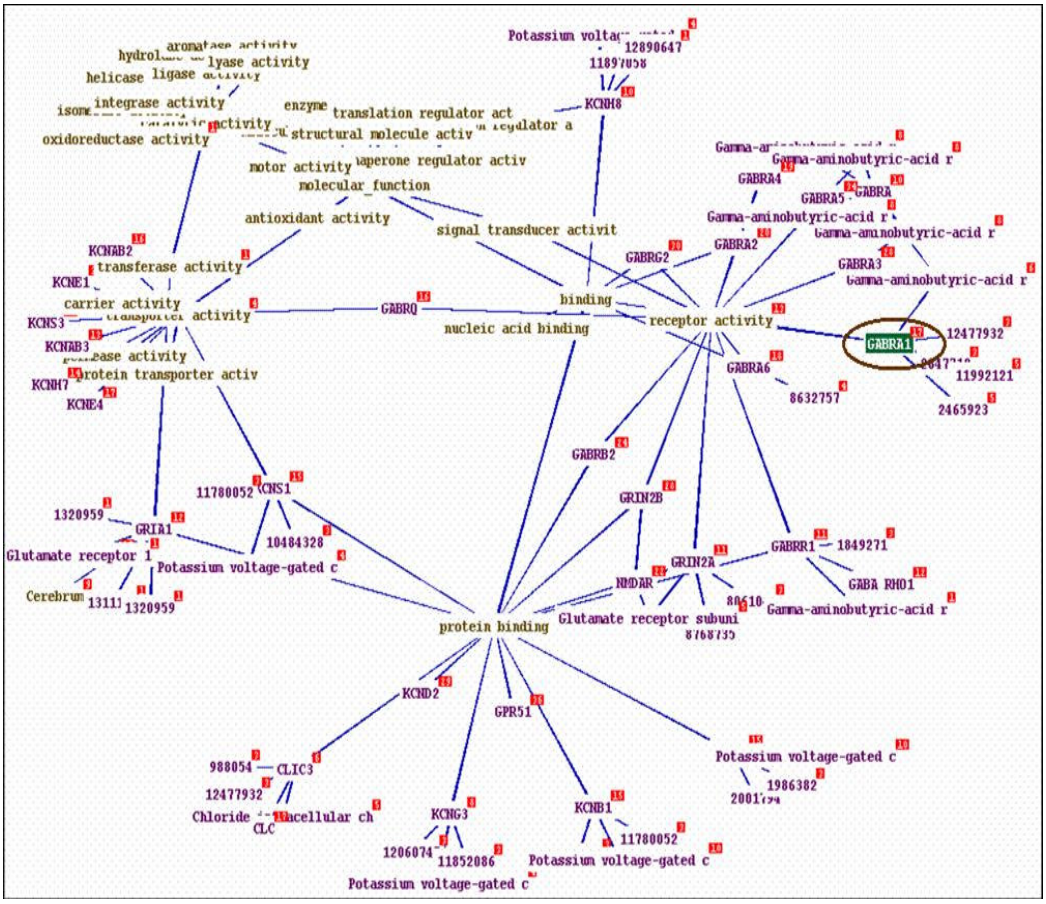


Figure 8.5: GABRA1 relationship visualization in BGO using TGViz

The data from the BGO can be used in a simulation system, such as computational neurogenetic simulation tool CNGM (www.kedri.info), NeuCom

(www.theneucom.com), Siftware (www.peblnz.com), Weka (http://www.cs.waikato.ac.nz/ml/weka/), CLUSTALW [Thompson 1994 and Chenna et al. 2003] and BLAST [Altschul et al. 1990]. Currently only the theory of the computational neurogenetic modelling (CNGM) and some illustrative results are included. These results include simulation of generation of local field potential (LFP) with the complete artificial genome within model neurons and then when one gene is knocked out that leads to an abnormal LFP. In the NeuroGenetic simulator, interaction of genes regulates the activity of neurons that consequently affects the dynamics of the whole spiking neural network (SNN). It can be shown that by tuning the interaction between genes and the initial gene/protein expression levels, different states of the neural network operation can be achieved. The behaviour of the SNN is evaluated by means of the spectral and bispectral analysis of field potential (LFP) and neural activity levels (spiking rate and neuronal synchronisations). The simulation of gene interactions is done through linear or non-linear gene regulatory network (GRN) models [Kasabov, Song et al. 2008]. NeuCom is a learning and reasoning computer environment based on connectionist models. It is designed to solve such problems as clustering, classification, prediction, adaptive control, data mining and pattern discovery from databases in a multidimensional, dynamic and changing data environment. Siftware is a software system for gene expression data analysis, modelling and profiling. Weka is a collection of machine learning algorithms for data mining tasks.

Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. CLUSTALW is widely accepted bioinformatics multiple sequence alignment program for various kinds of sequence investigation and to study the phylogenetic and clustering relationship between genes and proteins etc. BLAST is a program developed by National Centre for Biotechnology information (NCBI) and is widely used for revealing the homology patterns for genes/proteins of interest. Results using such kinds of methods and software can be added back to the BGO to update the BGO current knowledge base. Hence BGO evolves based on the knowledge input from outside and also based on creation of new knowledge by means of Computational Intelligence.

8.4 Knowledge reuse, elicitation and discoveries with BGO

The information processes at different levels like (quantum, molecular, single neuron, ensemble of neurons, cognitive and evolutionary) are very complex and difficult to understand as they evolve all the time, but much more difficult to understand is the interaction between the different levels [Kasabov 2007b]. All gene and protein related facts and other biological information is linked and stored within an ontological representation in order to enable scientists for further analysis and reusing this knowledge into their models through the high dimensional space of the BGO. Additional properties, such as uncertainty and importance, can be inferred and added by experts or

programmatically using the CNGM. This approach enables us to evolve the maps as new knowledge is discovered through the use of the simulator. Thus, different experts can work on the same project or share different experiments using the tools available in the ontology environment.

It may be that understanding the interaction through modelling would be a key to understanding each level of information processing in the brain and perhaps the brain as a whole. Using principles from different levels in one model and modelling their relationship can lead to a next generation of brain models as more powerful tools to understand the brain. A standardized ontology framework makes data easily available for advanced methods of analysis, including artificial intelligence algorithms, that can tackle the multitude of large and complex datasets for clustering, classification and rule inference for biomedical and bioinformatics applications. Results from the machine learning procedures can be entered back to the ontology thus enriching its knowledge base and facilitating new discoveries.

There have been several attempts to use ontology for cancer research. The National Cancer Institute Thesaurus (NCIT) has developed a biomedical ontology that provides consistent, unambiguous definitions for concepts and terminologies in cancer research domain [Ceusters et al. 2005]. It opens a way to integrate various types of information through semantic relationships, including cancer related disease, findings, gene data, drugs, etc. NCIT is also

linked to other internal or external information resources, such as caCore, caBIO and Gene Ontology (GO). In the study proposed by Dameron et al. 2006, ontology has been demonstrated that it is capable of automatically analyzing the grading of lung cancer. The cancer diagnosis and prognosis ontology will help the scientists in providing the relationships, either evidential or predicted, between genes; therefore, the scientists can target their research appropriately. The other benefit is to avoid repeatedly re-discovering any relationships that have been already been made by other researchers.

The above explained BGO system provides conceptual links between data on brain functions and diseases, their genetic basis, experimental publications, graphical illustrations and the relationships between the concepts. Each instance (information item) in BGO represents experimental research and these instances and their relationship also are traceable through a query plug-in that allows us, for example, to answer questions such as “Which genes are related to the occurrence of epilepsy?” by simply typing the key word epilepsy into the query plug-in and selecting the class gene. The system return the list of 198 genes and proteins potentially related to epilepsy. Query window may be also be used to investigate more specific discoveries, for example to find which genes are related to the occurrence of juvenile myoclonic epilepsy (JME)? – simply by typing the key word JME into the query window and selecting the class gene and the slot function comment (see figure 8.6). The

system returns the list of genes (here in this example: ten genes) potentially related to juvenile myoclonic epilepsy. By selecting any of them we can obtain detailed information about that particular gene, such as its GO function, chromosomal location, molecular weight, gene product, synonyms, function in neurons, mutations, brain expression profile, and literature etc. Here we have shown the navigation for the GABRA1 gene and the window shows the detailed information available within BGO (see figure 8.6). Next we can select gene(s) of interest to visualize their relationships to other concepts/instances in the BGO.

Own set of plug-ins that enables us to visualize, extract and import knowledge from/into different data sources and destinations were developed locally at KEDRI [Gottgtroy et al. 2004 and 2006]. BGO thus allows users to select and export the specific data of their interest like chromosomal location or molecular sequence length, or expression patterns, which can then be analyzed in a software machine learning environment, such as WEKA and NeuCom to train prediction or classification models and to visualize relationship information. Such exported gene/protein data can also be analysed in a different manner by standard bioinformatics software like BLAST and FASTA for revealing homology patterns for those genes/proteins of interest, etc.

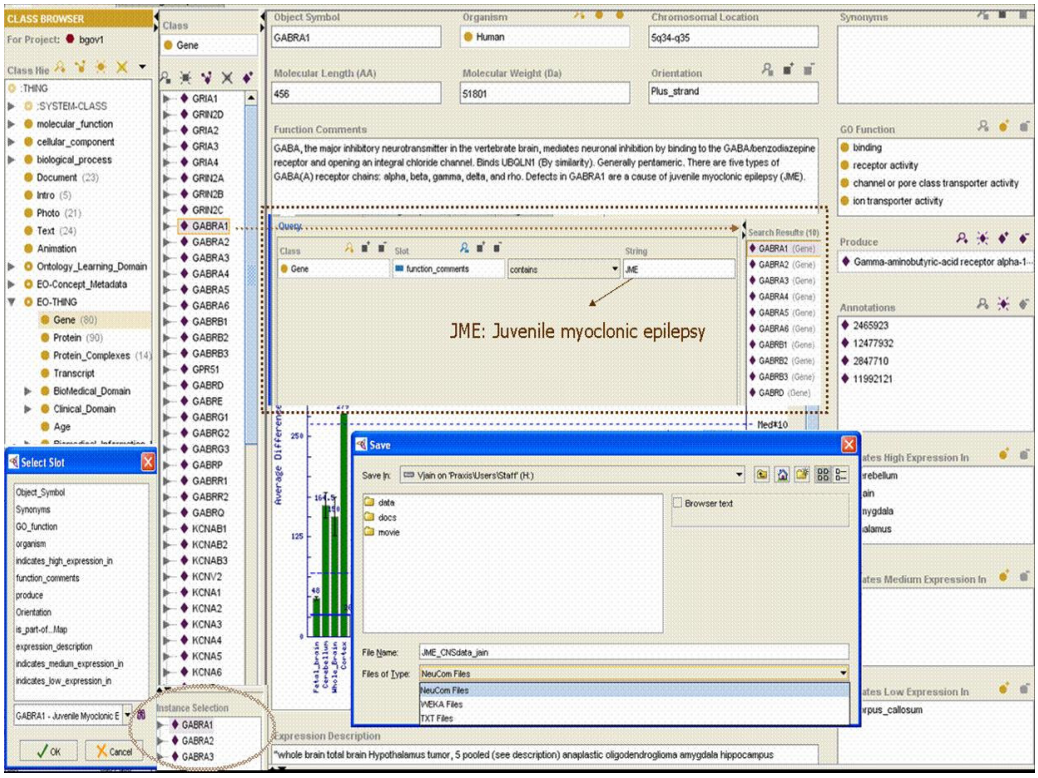


Figure 8.6: Query search system looking for JME related genes; navigation of GABRA1 in BGO; gene instance selection window to export and save the data in required format

One of the main applications of BGO is the integration between ontology and machine learning tools in relation to feature selection, classification and prognostic modelling with results incorporated back into the ontology. As an example, here we will take a BGO data, also publicly available, which is a gene expression data of 60 samples of CNS cancer (medulloblastoma) representing 39 children patients who survived the cancer after treatment, and 21 who did not respond to the treatment [Pomeroy, Tamayo, Gaasenbeek, Sturla et al. 2002]. We selected ten genes (out of 7129 genes) as top discriminating genes from the Central Nervous System (CNS) cancer data that discriminates two

classes - survivals and not responding to treatment. The Siftware system was used for the analysis and the method is called t-test. Below in the table 8.1 is the list of the ten selected genes (the first ID number is for reference of further analysis and the second ID number is the row number in the original data).

Table 8.1: Ten selected genes from CNS data using t-test of SIFTWARE

G1	G1352	High mobility group protein (HMG-I(Y)) gene exons 1-8, L17131, high mobility group AT-hook 1, HMGA1
G2	G327	D28124, NBL1 - neuroblastoma, suppression of tumorigenicity 1
G3	G348	Probable Ubiquitin Carboxyl-terminal Hydrolase, D29956 UBPY (ubiquitin specific peptidase 8, USP 8)
G4	G844	Dynein, Heavy Chain, Cytoplasmic, HG2417-HT2513
G5	G2196	Polyposis Locus Protein 1, M73547, adenomatosis polyposis coli, APC
G6	G2695	TAR (HIV-1) RNA binding protein 2, U08998, TARBP2
G7	G3645	Prostaglandin transporter hPGT mRNA, U70867
G8	G3320	Leukotriene C4 synthase (LTC4S) gene, U50136
G9	G2496	NTRK3 Neurotrophic tyrosine kinase, receptor, type 3 (TrkC), S76475 - (1 of 50 markers of survival from Pomeroy et al. 2002)
G10	G2996	Gps2 (GPS2, G protein pathway suppressor 2) mRNA, U2896

Evolving Connectionist System (ECOS) can be used for building adaptive classification or prognostic systems and for extracting rules (profiles) that characterize data in local clusters [Kasabov 2003 and 2007a]. This is illustrated on the ten CNS genes (as discussed above), where a classification system is evolved using the evolving classifier function method (ECF). Before

the final classifier is evolved in (see figure 8.7), a leave-one-cross validation method is applied to validate the ECOS model on the 60 samples, where 60 models are created – each one on 59 samples, after one example is taken out, and then the model is validated to classify the taken out example. The average accuracy over all 60 examples is 85% where 51 samples are classified accurately, out of 60 and 9 incorrectly. Then an ECF classifier is evolved on the ten CNS cancer genes. Aggregated (across all clusters) general profiles for each of the two classes are shown in (see figure 8.7). Class 1 is the non-responding group (21 samples, 71.43% accuracy) and class 2 is the group of survivals (39 samples, 92.31%). The results are better than the achieved in [Pomeroy, Tamayo, Gaasenbeek, Sturla et al. 2002] results of 78% (13 errors out of 60).

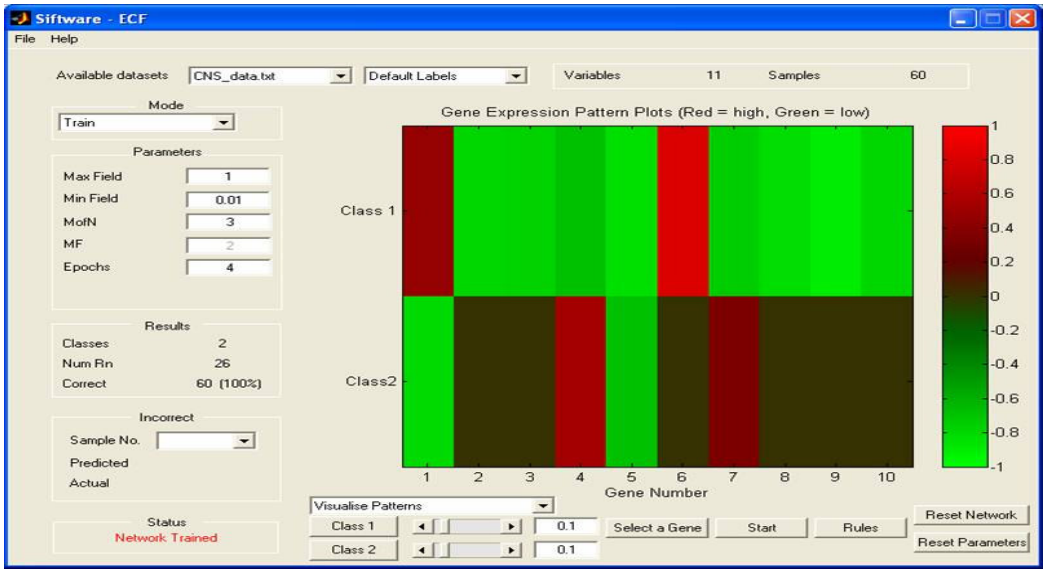


Figure 8.7: An ECOS classifier is evolved on the ten CNS cancer genes. Aggregated (across all clusters) general profiles for each of the two classes are shown. Class 1 is the non-responding group (21 samples) and class 2 is the group of survivals (39 samples). The analysis is performed with the use of a proprietary software system SIFTWARE (www.peblnz.com).

8.4.1 Biological validation and interpretation of results

The selected smaller number of genes, out of thousands, can be further analyzed in terms of their relation to cellular processes or other types of cancer or other diseases. The results then can be imported back to BGO and make conclusions about the genetic differences among the two groups of patients. For instance after entering the information about the ten selected genes from Entrez Gene Database and combining it with already present knowledge in BGO, we can discover that G1, the High mobility group protein (HMG-I(Y)), L17131, which is highly expressed in the treatment failures (see figure 8.7)

encodes a non-histone protein involved in many cellular processes, including regulation of inducible gene transcription, integration of retroviruses into chromosomes, and the metastatic progression of cancer cells. Our analysis has revealed that over-expression of this gene is associated with a bad prognosis for medulloblastoma in connection with the over-expression of G6, TAR (HIV-1) RNA binding protein 2. The protein encoded by this gene activates HIV-1 gene expression in synergy with the viral Tat protein. Thus, maybe as our analysis points to, over-expression of this latter gene it is related to a weaker immune response of an organism, which also makes sense from the point of view of failure to fight the disease. All other genes are under-expressed in the class of failures and relatively over-expressed in the class of survivors or at least not under-expressed. For instance, G2, NBL1 - neuroblastoma, D28124, which is involved in suppression of tumorigenicity, is not under-expressed in the class of survivors, but is under-expressed in failures, which again makes sense in terms of an outcome prognosis. G3, the probable Ubiquitin Carboxyl-terminal Hydrolase, D29956 UBPY, labeling proteins for proteasomal degradation, is not under-expressed in the class of survivors, but is under-expressed in the class of failures. G4, Dynein, Heavy Chain, Cytoplasmic, HG2417-HT2513, which mediates the perinuclear aggregation of phagocytosed melanosomes, participates in the formation of the supranuclear melanin cap and serves as a mechanism to help protect the nucleus from ultraviolet-induced DNA damage, is over-expressed in the class

of survivors meaning it might have a more general protective function not just against the UV light. G5, Polyposis Locus Protein 1, M73547, APC, adenomatosis polyposis coli, encodes a tumor suppressor protein that includes among its many intracellular functions one of nuclear export. Defects in this gene cause familial adenomatous polyposis (FAP), an autosomal dominant pre-malignant disease that usually progresses to malignancy. This gene is under-expressed in both classes reflecting the malignancy of medulloblastoma. G7, Prostaglandin transporter hPGT mRNA, U70867: so far only the role of PGT in the regulation of reproductive processes has been known. This study points to its role also in medulloblastoma, as one of the gene markers of survival, together with G4, Dynein, Heavy Chain, Cytoplasmic. The rest of the genes are under-expressed in failures and not under-expressed in survivals. G8, the Leukotriene C4 synthase (LTC4S) gene, U50136: This gene encodes an enzyme that catalyzes the first step in the biosynthesis of cysteinyl leukotrienes, potent biological compounds derived from arachidonic acid. Leukotrienes have been implicated as mediators of anaphylaxis and inflammatory conditions such as human bronchial asthma (under-expressed in failures). Mutations of G9 NTRK3/TrkC have been associated with secretory breast carcinomas and other cancers. Moreover, it plays a role in Long-Term Potentiation. It is under-expressed in failures. G10, Gps2 (GPS2, G protein pathway suppressor 2) mRNA, U2896, encodes a protein involved in G protein-mitogen-activated protein kinase (MAPK)

signalling cascades. When over expressed in mammalian cells, this gene could potently suppress a RAS- and MAPK-mediated signal and interfere with JNK (C-jun-amino-terminal kinase) activity, suggesting that the function of this gene may be the signal repression. Ras proteins transmit extracellular signals that promote the growth, proliferation, differentiation and survival of cells. This G protein pathway suppressor 2 is under-expressed in the class of failures. Thus, by means of BGO we can meaningfully interpret the results obtained by the CI analysis. In addition, for each of the genes, we can obtain the network of relations to other genes, gene functions, molecular processes and disease by means of Protégé TGVizTab as is illustrated in (see figure 8.8). We can navigate each node to obtain further information and links.

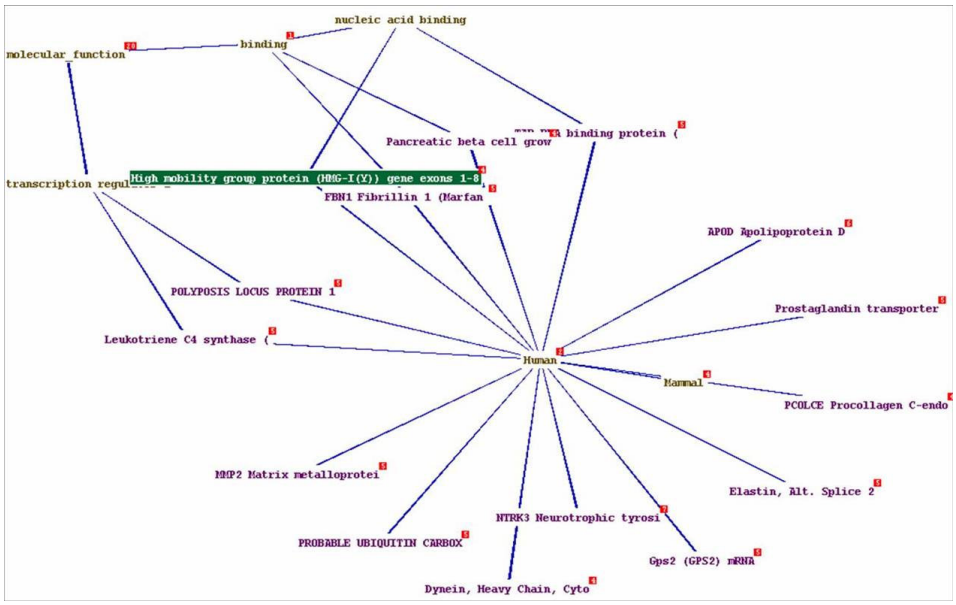


Figure 8.8: Relationship visualization in BGO using TGViz for the High mobility group protein (HMG-I(Y)), L17131, one of the genes, which is differentially expressed between the group of survivors and treatment failures in medulloblastoma data.

The BGO is an evolving ontology that evolves its structure and content so that new information can be added in the form of molecular properties, disease related information and so on. All of this information can be re-utilized to create further models of brain functions and diseases that include models of gene interactions. In future, one may hope that by linking and integrating simulation results from the CNGM simulations with genetic information in the BGO, researchers can facilitate better understanding of metabolic pathways and modelling of gene regulatory networks, and ultimately a more complete understanding of the pathogenesis of brain diseases.

8.5 Using BGO data for neuronal gene-protein sequence and clustering analysis

As discussed in the detailed review of this thesis chapter 2, we have seen that with the advancement of molecular research technologies huge amount of data and information are available about the genetic basis of neuronal functions and diseases. To understand human brain function we need to integrate knowledge from genomics, proteomics, neuroscience, psychology, and theoretical disciplines such as computer science and physics. As we have mentioned above, the data from BGO can be easily exported and may be used in various kinds of analysis. In this section this is what we have shown and

here as an example the clustering and sequence analysis was performed on brain gene data.

Complex interactions between genes and proteins in neurons affect the dynamics of the whole neural network. Gene and protein expression values may change due to internal dynamics of the gene regulatory (interaction) network, initial conditions of the genes and external conditions. It is observed that different initial gene conditions can lead to the same outcome in terms of neuronal activity. On the other hand, in the diseased brain, either altered initial conditions, mutated genes and/or altered interactions within gene network lead to abnormalities in network activity. As we know, for many Neurocomputing experiments, most of the valuable molecular information remains hidden in different databases and cannot be used to run simulations to judge the biological plausibility of computational models. But BGO brings all this information together and using several ontology modules, stored knowledge can be easily represented and reused for computational experiments. For example, besides the molecules analysed in our earlier CNGM experiments, later ontology helped us to identify other gene/proteins that are also believed to have a regulatory effect upon neuronal parameter's related genes and proteins, such as c-jun, mGluR3, Jerky, BDNF, FGF-2, IGF-1, GALR1, NOS and S100beta.

Even with the recent genetic discoveries, brain GRNs are still in an initial stage of development. Despite promising preliminary results, there are some issues like lack of information at the genetic level that must be addressed to perform more realistic simulations in computational neurogenetic modelling (CNGM). In this respect we rely on knowledge discovery from BGO and its integration with other genetic neuroscience analysis. A specific gene from the genome relates to the activity of a neuronal cell by means of a specific protein. Even in the presence of a mutated gene in the genome, which is known to cause a brain disease, the neurons can still function normally provided a certain pattern of interaction between genes is maintained. On the other hand we can assume, if there are no mutated genes, abnormalities in the brain functioning can be observed as a consequence of abnormal interaction between genes. Therefore to understand the function of genes, we also need to know how they are expressed, when they are expressed, conservation among their products and their response to therapeutic drugs. In general, we can say such kind of sequence analysis potentially may help to understand a broad range of fundamental questions about genetic influences on mental processes and diseases. The use of such methods and knowledge integration can facilitate new discoveries in the area of genetic neuroscience.

Genetic neuroscience uses principles from cellular and molecular biology to investigate questions about gene actions in neurons. Prominent approaches

in this area are genetic mutations, gene knockouts, protein purification, gene expression analysis, GAS chromatography, gel electrophoresis, high performance liquid chromatography (HPLC), western blotting, etc. Another multigene approach, high throughput genotyping technology, can be used to define single nucleotide polymorphisms that characterize diseases. However, all these techniques are useful but time consuming and are very expensive. Thus we use standard bioinformatics operating strategies like studying neuronal protein sequence conservation/variation. Vast comparative analysis of crucial protein sequences and structures can also reveal evolutionary relationships between specific brain regions of humans and other species. We have investigated through sequence analysis, the excitatory and fast inhibitory receptors (i.e., AMPAR, NMDAR and GABRA) gene groups, which participates in statistically significant gene interactions (based on CNGM of LFP/EEG generation, a working project at KEDRI). We mainly focused on them because they play a major role in most of the mental disorders through their direct or indirect interactions with several other genes/proteins and through specific parameter functions (e.g., excitation and inhibition). Each of these proteins is comprised of several subunits, and each subunit is coded by a separate gene (see Table 8.2).

Table 8.2: List of subunit proteins for AMPA, GABRA and NMDA receptors that define specific neuronal information-processing parameter functions. AMPAR = (amino-methylisoxazolepropionic acid) AMPA receptor, NMDAR = (N-methyl-D-aspartate acid) NMDA receptor, GABRA = (gammaaminobutyric acid) GABAA receptor, GABRB = GABAB receptor

gi|number|	Human Protein	Gene	Sequence length (aa)
gi|1169959|	Glutamate receptor ionotropic, AMPA 1; (GluR-1) (GluR-A) (GluR-K1)	GRIA1	906
gi|23831146|	Glutamate receptor ionotropic, AMPA 2; (GluR-2) (GluR-B) (GluR-K2)	GRIA2	883
gi|1169961|	Glutamate receptor ionotropic, AMPA 3; (GluR-3) (GluR-C) (GluR-K3)	GRIA3	894
gi|1346142|	Glutamate receptor ionotropic, AMPA 4; (GluR-4) (GluR4) (GluR-D)	GRIA4	902
gi|11496971|	NMDA receptor 1 isoform NR1-1 precursor	GRIN1	885
gi|14285603|	Glutamate [NMDA] receptor subunit epsilon 1 precursor; (NR2A) (NMDAR2A)	GRIN2	1464
gi|14548162|	Glutamate [NMDA] receptor subunit epsilon 2 precursor; (NR2B) (NMDAR2B)	GRIN3	1484
gi|2492629|	Glutamate [NMDA] receptor subunit epsilon 3 precursor; (NR2C) (NMDAR2C)	GRIN4	1233
gi|18201966|	Glutamate [NMDA] receptor subunit epsilon 4 precursor; (NR2D) (NMDAR2D)	GRIN5	1336
gi|38327554|	Gamma-aminobutyric acid (GABA) A receptor, alpha 1 precursor	GABRA1	456
gi|1346078|	Gamma-aminobutyric-acid receptor alpha-2 subunit precursor	GABRA2	451
gi|4557603|	Gamma-aminobutyric acid A receptor, alpha 3 precursor	GABRA3	492
gi|1346079|	Gamma-aminobutyric-acid receptor alpha-4 subunit precursor	GABRA4	554
gi|399519|	Gamma-aminobutyric-acid receptor alpha-5 subunit precursor	GABRA5	462
gi|23831128|	Gamma-aminobutyric-acid receptor beta-1 subunit precursor	GABRB1	474
gi|455946|	Gamma-aminobutyric acid A receptor beta 2 subunit	GABRB2	474
gi|120773|	Gamma-aminobutyric-acid receptor beta-3 subunit precursor	GABRB3	473
gi|27820121|	Gamma-aminobutyric-acid receptor gamma-1 subunit precursor	GABRG1	465
gi|38788155|	Gamma-aminobutyric acid A receptor, gamma 2 isoform 1 precursor	GABRG2	475
gi|13959689|	Gamma-aminobutyric-acid receptor gamma-3 subunit precursor	GABRG3	467

Initially, the information related to expression of these subunit genes, mutations, etc. was collected through relevant literature survey followed by sequences retrieval from NCBI database. We did preliminary analysis for searching the common motifs (patterns that occurs repeatedly in a group of related protein or DNA sequences) among these subunits. Through this initial analysis we observed that all detected motifs were belonging to similar protein

families, like ligand-gated ion channel, receptor family ligand binding region and neurotransmitter-gated ion channel ligand binding domain. This observation motivated us for performing the detailed residual inspection. For our investigation, we employed the standard operating bioinformatics procedure, i.e., comparative analysis from similarity measures that is widely accepted approach by most of neuroscientists. The method is familiarly known as multiple sequence alignments (MSA). Scope of this bioinformatics technique is twofold: (a) gains understanding to identify the shared regions of homology; (b) determines the consensus sequence of several aligned sequences. For such strategies, there are however many software available free to academic users, but here in our case we used the CLUSTALW package (http://align.genome.jp/) developed by Koichi Ohkubo, (Genome Net), for details see [Chenna et al. 2003]. A specific reason for picking this software was that it is mostly cited and secondly its output is a much readable representation than others in the field. Sequence similarity between all subunits was visualized, then we later used Box-Shade program V3.21 (http://www.ch.embnet.org/software/BOX form.html) in order to format our multiple alignment results for explanation purpose. Figure 8.9 is only a highlighted portion of most important observation and in general it indicates that comparison of the 20 sequences of subunit neuronal information-processing proteins has revealed a number of conserved residues. As a biological fact we know the number of structurally conserved residues

increases with the binding site contact size, thus we expect similar function for these amino acids in this case also. Here we have observed the extent to which these conserved residues are clustered, which is generally the case in most channel proteins and neuro-receptors. After carefully analysing the alignment, the most interesting investigation found was the consistent conservation of phenylalanine (F at position 269) and leucine (L at position 353) in all 20 proteins with no mutations. We expect these residues to play some role as a binding centre for interaction of these proteins with several other genes/proteins such as c-jun, mGluR3, Jerky, BDNF, FGF-2, IGF-1, GALR1, NOS and S100beta that are also believed to have a regulatory effect upon these receptors [Benuskova and Jain et al. 2006]. Therefore, we expect that the gene interaction observations may be justified up to certain extent because of such truly conserved residues. However we must say that all such hypotheses remain unproven and these predictions need to be tested through laboratory experimentation.

		Position			
Gene	**...**	**260 270**	**...**	**350 360**	**...**
GABRA3	...	IGEYVVMIIHFHLKRKIGYFV	...	AVCYAFVFSALIEFAIVNYFI	...
GABRA5	...	IGEYIIMIAHFHLKRKIGYFV	...	AVCYAFWFSALIEFAIVNYFI	...
GABRA1	...	IGEYVVMIIHFHLKRKIGYFV	...	AVCYAFVFSALIEFAIVNYFI	...
GABRA2	...	IGEYIVMIAHFHLKRKIGYFV	...	AVCYAFVFSALIEFAIVNYFI	...
GABRA4	...	IGEYIVMIVYFHLRRKMGYFM	...	AVCFAFVFSALIEFAAVNYFI	...
GABRG1	...	SGDYVIMIIFFDLSRRMGYFI	...	SVCFIFVFAALMEYGILHYFI	...
GABRG2	...	SGDYVVMSVYFDLSRRMGYFI	...	SVCFIFVFSALVEYGILHYFV	...
GABRG3	...	AGDYVVMIIYFELSRRMGYFI	...	IVCFLFVFAALMEYAILNYYS	...
GABRB2	...	IGSYPRLSLSFKLKRNIGYFI	...	MGCFVFVFMALLEYALVNYIF	...
GABRB3	...	IGAYPRLSLSFRLKRNIGYFI	...	MGCFVFVFLALLEYAFVNYIF	...
GABRB1	...	IGAYPRLSLSFRLKRNIGYFI	...	MGCFVFVFLALLEYAFVNYIF	...
GRIA2	...	PQKSKPGVFSFLDPLAYEIWM	...	LIIISSYIANLAAFLIVERMV	...
GRIA3	...	PQKSKPGVFSFLDPLAYEIWM	...	LIIISSYIANLAAFLIVERMV	...
GRIA4	...	PQKSKPGVFSFLDPLAYEIWM	...	LIIISSYIANLAAFLIVERMV	...
GRIA1	...	PQKSKPGVFSFLDPLAYEIWM	...	LIIISSYIANLAAFLIVERMV	...
GRIN2	...	.SNGIVSPSAFLEPFSASVWV	...	VIFLASYIANLAAFMIQEEFV	...
GRIN3	...	.SNGIVSPSAFLEPFSADVWV	...	VIFLASYIANLAAFMIQEEYV	...
GRIN4	...	.SNGIVSPSAFLEPYSPAVWV	...	VIFLASYIANLAAFMIQEQYI	...
GRIN5	...	.SNGIVSPSAFLEPYSPAVWV	...	VIFLASYIANLAAFMIQEEYV	...
GRIN1	...	EIPRS.ILDSFMQPFQSILWL	...	MIIVASYIANLAAFLVLDRPE	...

Figure 8.9: Multiple alignments of all 20 subunits of three neuronal information-processing proteins showing consistent conservation of phenylalanine (F) at position 269 and leucine (L) at 353.

For these neuronal proteins we also obtained a dendogram (see figure 8.10) that represents the related closeness between these subunit proteins. More closely related pairs of neuronal protein sequences have aligned most readily to each other than more divergent pairs.

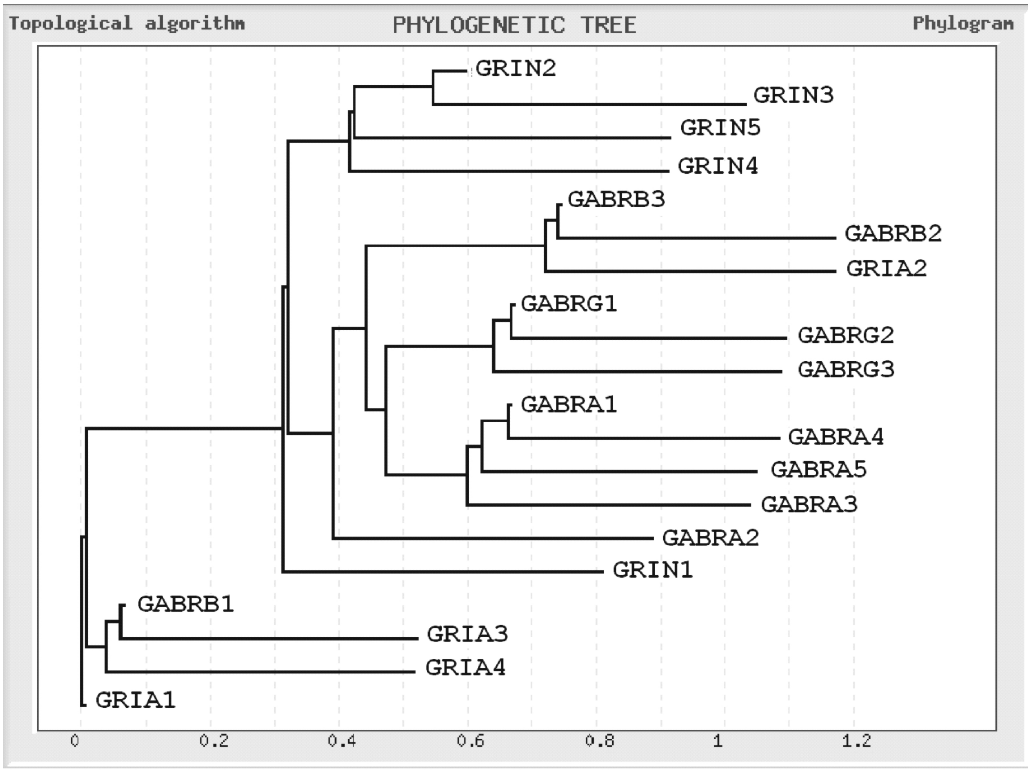


Figure 8.10: Clustering tree representing 20 subunit proteins of AMPA, GABAA and NMDA receptors

Based on the result of analysis that clearly showed us the extent of conservation of many amino-acid residues among all investigated receptor subunits, we assume that the expression of these individual subunits should be coordinated within one gene group. In addition, these regions can be the basis for mutual interactions. Mutual interaction between subunits of different receptors has been recently confirmed experimentally. Furthermore, this biological insight about our findings is also supported by most of similar biological statements available in neuroscience area, especially in case of channel proteins and receptors.

8.6 Facilitating education with BGO

The BGO is an electronic learning environment that can be used as a teaching tool for undergraduate and postgraduate students as well as researchers in bioinformatics, neuroinformatics, computer and information sciences and related areas. It exemplifies the importance of use of ontologies in current knowledge management and interpreting relationships between molecules and brain functions. It enables teaching the basics of molecular biology and gene regulatory networks as well as introducing the area of computational neurogenetic modelling [Benuskova and Kasabov 2007]. The BGO can be used to better understand and explain various topics related to brain, genes and their modelling, for example: the structure of the brain; main functions of the brain; the importance of gene mutation on brain functions and behaviour; importance of gene regulatory networks in neurons; mental/neurological disorders and main receptor/ion channel genes/proteins involved; understanding neural signal propagation and the role of synapses; analysis of LFP/EEG data and its relevance to brain functions; neurogenetic modelling and the role of its parameters for the outcome.

Animated visualisations of information are becoming more prevalent as technology advances. In today's environment, businesses and institutions need to present an increasingly complex range of information to their market or

audience who want relevant information and a clear way to differentiate it from competitive sources. The pressure to meet these demands raises the issues of efficiency, communicability, accuracy and training. The BGO interface is built using textual, graphical, audio and visual media. The inclusion of 3D animation, gives both a dynamic narrative introduction and overview to the BGO (refer to appendix G to view a snapshots of animations). Animations are used to represent complex scientific information, by visualising it in a more comprehensive, integrated form. The animation navigates and provides a sophisticated visual method of integration across the different domains of the BGO. This approach, drawn from the aesthetics and immersive experience of computer games and special effects technologies is introduced as a way of engaging a younger or novice audience in this complex, emergent, cross disciplinary field of linking genes to brain functions.

In point-wise summary, BGO can be used for research informed teaching as follows: (1) Teaching students about brain structure and hierarchical functioning of the brain from genes to behaviour (2) Highlighting importance and use of ontologies in knowledge management and interpreting relationships among concepts (3) Teaching central dogma of molecular biology and gene regulatory networks within neurons and Bioinformatics methods (4) Animations with narrations and explanative texts accompanying the BGO contents.

8.7 Conclusion and System availability

In this thesis chapter we have presented brain-gene ontology as a software system for knowledge integration and information fusion. Our system includes conceptual and factual information about brain and gene functions and their relationships. BGO can be viewed as a declarative model that defines and represents the concepts existing in the domain of brain and genes, their attributes and the relationships between them. It is represented as a knowledge base which is available to applications that need to use and/or share the knowledge of the domain. BGO is a tool for research and teaching across areas of bioinformatics, neuroinformatics, computer and information sciences at different levels of education and expertise. Different parts of it can be used by different users, from a school level to postgraduate and PhD student level.

BGO allows users to navigate through the rich information space of brain functions and brain diseases, brain related genes and their activities in certain parts of the brain and their relation to brain diseases; to run simulations; to select and download data that can be used in a software machine learning environment, such as SIFTWARE and CLUSTALW to train prediction or classification models and to perform molecular sequence analysis; to visualize relationship information; and to add new information as the BGO has an

evolving structure. The BGO contains also a description of a computational model and a simulation tool for modelling complex relationships between genes and neural oscillations. The BGO is an evolving ontology that evolves its structure and content so that new information can be added in the form of molecular properties, disease related information and so on. All of this information can be re-utilized to create further models of brain functions and diseases that include models of gene interactions. We hope that by linking and integrating simulation results from the CNGM simulations with genetic information in the BGO, we can facilitate better understanding of metabolic pathways and modelling of gene regulatory networks, and ultimately a more complete understanding of the pathogenesis of brain diseases. In future (which is beyond the aim of this doctoral research) more data and information can be added, that will include both higher level information on cognitive functions and consciousness, and lower level quantum information. In addition novel plugins may also be developed to derive hidden knowledge. Also, the system may be automated to bridge the loop between computational neurogenetic modelling (CNGM) and brain gene ontology (BGO) so that information exchange might occur without the need of domain experts.

In summary we can say that BGO is novel approach in the direction of information fusion and has multi functionality ranging from knowledge discovery for research to facilitating education. Some of the capabilities and

usefulness of the system can be listed as follows: (1) Protégé query interface to find all the genes related to a particular brain disease (Alzheimer, Parkinson, epilepsy, mental retardation, Rett syndrome, and schizophrenia) (2) Visualisation of new relationships via Protégé TGVizTab (3) Export of data (sequence and molecular info) and perform analysis using standard bioinformatics methods (BLAST, FASTA, CLUSTALW etc.) (4) Export and analysis of data using methods of computational intelligence (NeuCom, WEKA, etc) (5) Import new data from SwissProt, UniH, Entrez, etc. or any relevant experimental data (6) Aid in obtaining computational neurogenetic models.

Access to the current "BGO" System can be gained in two ways: (a) BGO 2007 version 1 (BGOv1) is released on a KEDRI CD and it is freely available to academic users and is for non-commercial use only (b) Our system is also available through world wide web (WWW) and may be downloaded from online KEDRI Computational Intelligence Repository (KCIR) at "http://kcir.kedri.info" or from KEDRI website at "www.kedri.info" under the neuroinformatics centre. Please note, due to size limits the WWW version does not contain the animations and demonstration movie created for explaining the system. We will conclude the thesis with the next chapter by discussing some implications, potential applications, ethics and future directions. As this research is a "never ending topic", we have also put a section for an open discussion.

9. Implications and Future Directions

In this last chapter, we discuss some of the implications and future directions of the undertaken research. We also talk about some of the potential applications of our developed system and applied methods. This discussion is followed by our declaration on the ethics before we conclude the thesis and discuss some open questions in this research area.

9.1 Introduction

In this doctoral thesis we have presented our work on understanding the molecular interactions in a cell and knowledge integration from a different viewpoint which is our novel idea and has not been discussed so far. We have proposed an integrated framework (based on computational intelligence methods like Genetic algorithm, Kalman filter, Least Angle Regression, Expectation Maximization, Evolving Fuzzy Neural Network, Quantum Inspired algorithm, Gabor filtering, BLAST and CLUSTALW etc.) to analyse range of information types like time series gene expression, gene and protein sequences, promoters, microRNAs and literature data etc. for studying the gene regulation area. Finally, we focus on knowledge integration and information fusion (using ontology) and discuss the utility of the machine learning tools in integrated framework. However we argue that our approach on each domain of the problem i.e. gene regulatory networks, functional

classification of microRNAs and information fusion and knowledge integration in our system called brain gene ontology (BGO) is unique but we do not claim that this is the end of the story to tackle such complicated problem to understand the complete central dogma of molecular biology. We can only say and expert may agree that we have shown the world a first step in this research direction and each of our study domains has future directions to go further that can be the work of post doctoral research or can lead to another doctoral thesis. Also, we are aware that the scientific community may argue that our approaches/methods are not without limitations and implications and not above criticism. But as mentioned earlier, understanding molecular interactions is a recent research area and the issue cannot be addressed by a single expert in few years of time. Considering this discussion above, we have talked below about some of the implications and future directions of this research topic. Next, we talk about some of the potential applications of the developed system and applied methods, ethics consideration and open questions.

9.2 Implications and future directions of the suggested approach:

In this section of this thesis, we discuss some of the implications of our developed methods and systems. We have also discussed future directions where the vision is clear for the further research. We discuss this topic

separately into three interlinked domains, i.e. gene regulatory networks, microRNA regulations and brain gene ontology.

9.2.1 Gene regulatory networks

There are some implications of the methodology that we have explained and applied for the problem of understanding molecular interactions of the cell, i.e. inferring gene regulatory networks. First obvious point is microarray data are noisy and any minor experimental errors might occur during the production of such data. Slight negligence while using the microarray chips can in turn ultimately leads to crucial problems and results can be erroneous.

Inference of GRN using the computational intelligence approach has few implications, for example the use of differential equations has drawbacks such as it requires the estimation of n^2 parameters for the transition matrix **F** and n (n-1)/2 parameters for the noise covariance **E**. In addition, while genetic algorithms (GA) are very effective approach but when using GA as more candidates are identified in the future, the search space grows exponentially in size and exhaustive search can soon become infeasible. So such methods are usually computationally very expensive. And last but not least, the results we have obtained from our novel computational intelligence methods demands the experimental verification in laboratories. In the respective chapters of this thesis we have discussed the way we have dealt with such problems. Just to remind that in many cases we have also verified our results based on

published experimental evidences and in other cases where appropriate, we have explained the result validation choices that we have adopted in this research.

In future apart from LARS and QiEA that we have used for inferring GRNs, researchers may explore the integration of several other novel computational intelligence methods. It should also be noted that one method may work well on the given dataset for extracting knowledge while it may not perform similarly well on the other dataset, therefore future research may also be done with the viewpoint of further generalizing the methodology that we have suggested in this research work. It should be noted that domain experts are always needed to validate and interpret the meaning of results.

9.2.2 MicroRNA regulations

The novel methodology that we have described in this thesis chapter 7 may have some implications like, first; the single structure obtained from RNAfold may not be biologically occurring one, as there can be other alternative conformations having approximately similar deltaG binding energy, i.e. similar thermodynamic stability. The folded RNA may also have further tertiary structure and pseudoknots, or associated proteins in the cell may force a different form within natural conditions in the cell. Second, if there are several alternative conformations in the cell for the same nucleotide sequence of the

microRNA precursor (called Bolzmann ensembles), a better (but much more complicated) approach would be to compute the complete Bolzmann ensemble of likely conformations and compare then sets of bitmap images from each RNA sequence with each other to find conserved structural features between the Bolzmann ensembles. Software for computing such ensembles with probabilities/lifetimes of each conformation has been developed recently [Ding and Lawrence 2003], as well as software to evaluate existence of alternative riboswitch type RNA conformations [Björn Voss et al. 2004]. Third, some other bitmap image analysis algorithm may extract other, more relevant features from the 2D topology of the molecular structure represented by the bitmap image. Fourth, alternative methods of drawing the hairpin molecules might be more informative, e.g. using different geometrical shapes or colours for the four nucleotides A, C, G and U, and representing the AU, CG and GU molecular bonds differently. These representations would give more structural information available from a (colour) bitmap image for the image classifier.

Therefore, whether the Gabor Filter features are most suitable for this task if beside the point; we could have used several other general methods for extracting potentially more informative features from our bitmap images. For example, one could use the classical approach of hierarchical neural networks imitating human visual system [Amari and Kasabov 1998], like the [Fukushima

1988]. Another approach would be to use various shape, contour and curvature feature extraction methods, chain codes or boundary descriptors.

A limited set of features extracted from bitmap images can limit the usefulness of the retrieved information for discovery and classification, so several different methods should be tried concurrently. This is beyond the scope of the topic of this doctoral research. A further important point is to use the bitmap derived information in conjunction with any other kind of information for multimodal data integration by artificial intelligence methods. For example, for the human microRNAs studied here, we envisage combined use of Gabor filter features, sequence similarity measures, nucleotide composition features, genomic location information (e.g. is the gene in an intron, exon, intergenic region, near a promoter etc) and so on. Such a rich feature set would then be analysed as multimodal data input for more informative clustering of known microRNAs and learning of pertinent feature combinations to discover novel microRNAs and even other types of related non-coding RNAs having different functions. Thus the methodology we have adopted here and described is merely a prelude to further work for integrated multimodal data analysis for all kinds of nanoscale 2D representations of such macromolecules. It could be useful even for classification of the huge and complicated organic molecule databases, giving in effect a simple short-cut via bitmap images to discovery of novel classification and clustering methods of such databases.

If one is interested in further research on this topic, then especially interesting are the prospects for high-throughput simulated folding of non-coding RNA molecule candidates along the genomes of various organisms. Such work seems well warranted, because the RNA World is still largely unknown and unexplored, and many interesting discoveries will thus be forthcoming in this hot area of molecular biology, where the nanoscale interactions of RNA molecules with other molecules control the life and death of all eukaryotic cells in minute detail.

9.2.3 Brain gene ontology

"BGO" system has been shown as a potential contribution in the area of bioinformatics research and teaching; however there are some implications on developed system. Also, along with the implications here we have suggested the future directions (beyond the aim of this doctoral research) of "BGO" for further development in six phases:

(1) WWW Open Source: One may easily criticize that for developing such an integrated system why only the protégé was used? In this direction we can only say that there are some other tools also available like chimaera but Protégé ontology editor was freely available to the academic users and is quite prominent tool in this area for developing ontologies. In various conferences where we have presented the "BGO" it has been suggested that in future it can

be given the shape of Web-based, multiple users, shared, open source environment entity. This can help in a way that scientists and researchers from different countries/locations will be able to access our system without the need of installing anything and also can feed the data of their interest into our system.

(2) Information integration to keep BGO evolving: As suggested in phase 1, more data and information from different source ontologies and databases (like GO, SwissProt etc.) can be kept added in future also to maintain the BGO as fully updated rich source of information with the similar pace as the science is evolving. This phase will also depend on the development and/or usage of specialized engines and interfaces to merge information from different platforms.

(3) Inference and Knowledge Discovery (KD): Knowledge discovery (KD) has always been a critical aspect in the ontology usage. The growth of BGO will by far exceed the human capacity to analyse the ontology in order to find implicit regularities, relations or clusters hidden in the facts. Therefore, knowledge discovery becomes more and more important in the ontology usage. Typical tasks for KD are the identification of classes (clustering), the prediction of new unknown objects belonging to these classes (classification), and the discovery or inference of associations and relations between facts. In future, researchers

may use machine learning approaches for ontologies and suggested means are CLIPS, fuzzy CLIPS, Algernon and Jess, etc, for this purpose.

(4) Knowledge Visualisation (KV): Currently the BGO uses some well known visualization techniques like TGViz and OntoViz etc. But we feel that there is a better scope in the area of visual KD. It focuses on integrating the user in the KD process in terms of effective and efficient visualization techniques, interaction capabilities and knowledge transfer especially when only uncertain information is available. Effective visualisation actually maps the data to some kind of valid, novel, potentially useful and understandable knowledge through clusters, trees, graphs, etc. Obviously, just the user can determine whether the resulting knowledge satisfies these requirements. Moreover, the usefulness of some kind of knowledge varies from user to user.

(5) Education: As we have learnt in this thesis chapter 8, the BGO system facilitates education and we have used it as a tool to teach master's students of AUT in the bioinformatics paper. On this layer, ontology schema describes the semantic relationships of stored education knowledge. Broader goal is to facilitate teachers and students to exploit the myriad possibilities of using ontology resources to meet individual's learning or teaching needs. It is not unreasonable to assume that in the long run, the BGO will facilitate the development of methods for helping students to understand and to recreate in

new contexts the content and knowledge produced by experts in several disciplines like genetics, neuroscience, information science and bioinformatics.

6) Simulation of CNGM and development of BGO: Currently we use the knowledge obtained from BGO in our CNGM models. This knowledge can either be the data integration from diverse source of databases or novel information/hypothesis derived based on computational intelligence modules used in conjunction with the BGO. We then manually enter back the findings of CNGM within BGO. Such process may be automated in some way so that the results obtained by CNGM can become new facts to be integrated in the BGO automatically, which will close the loop of new knowledge discovery and representation in the ontology knowledge base.

9.3 Ethical considerations

We understand that community-based research usually raises ethical issues which are not normally encountered in research conducted in academic settings. In this section we declare the ethical implications on our work. This doctoral research does not involve any clinical trials and we were not using any such datasets, use of which requires ethics approval. All of our datasets were either publicly available or obtained through confidential agreement as part of collaboration. Any of our results or publications does not point out the

individual patient information or reveal similar facts. This had been more of a systems biology study and had nothing to do with the clinical domain.

9.4 Potential applications of the developed methods and systems

In this section, we will discuss the importance of this doctoral study and talk about some of the applications of our research. The study has a wide variety of applications and the methods can easily be generalized to suit the appropriate datasets for investigations of interest. Below we have talked about potential applications of this research:

(1) GRN study - cancer prognosis and aid in more complete understanding pathogenesis of diseases

In the chapter 4 of this thesis, the discovery of gene regulatory networks (GRN) from time series of gene expression observations using KF and GA is discussed. The integrated method is designed to deal effectively with irregular and scarce data collected from a large number of variables (genes) and it can be easily generalized to extract GRN from other time series gene expression data. We believe that such generalized methods can be used to identify important genes in relation to any disease or a biological function. For example, we took the study on cancer (leukaemia) and reported that one outstanding gene “TCF-1” potentially co-regulates the telomerase activity (crucial for leukaemia study). In addition, several GRNs were found in relation

to the telomerase gene and a global GRN was also predicted. Such discoveries that we have reported using our novel methods have been verified in National Cancer Institute (using experimental biology techniques) and are very critical for the cancer prognosis. Other discoveries can also help researchers to carry out wet experimentation on selected number of molecules in cancer study.

Similarly we have addressed some of the main questions in the chapter 5, like: (a) what knowledge can be derived from different models? (b) Would an integrated approach be more suitable to reveal the controls of gene regulation? Each method (LARS, EM with GA and EFuNN) revealed some new aspects of the problem and it is agreed that to infer the GRN and to understand the processes behind gene regulation it is more suitable to adopt such integrative approach as ours through which some new knowledge is discovered. It seems that the proposed integrated approach for inferring gene regulatory network makes it suitable to be applied on proteomics and metabolomics time-series to derive possible regulatory networks between genes, proteins and various metabolites. Ultimately, information fusion methods could be used to link all these interaction networks for a global integrated analysis. Such approach can be easily extended by incorporating other novel computational intelligence methods (based on their computing efficiency and parameter-free operation like LARS) and comparative analysis

may also be accounted. Such findings using integrative methods of computational intelligence opens new doors in studying the genetic basis of various kinds of diseases and this can constitute a proper step towards understanding gene regulation to model the behaviour of biological cells.

In the chapter 6 we used the integrated power of clustering approach and a novel optimization method called quantum inspired evolutionary algorithm (QIEA) to infer an abstract matrix of interactions between temporal clusters of genes, which we have validated using documented interactions and also run virtual transgenic and gene knock-in mice experiments. A kind of analysis and model development scenario presented in this chapter has a general application to any time-course microarray data accompanying any cellular process to obtain further insights into the underlying genetic and molecular mechanisms. In addition, the abstract interaction network can serve to simulate virtual gene knock-out and knock-in experiments to predict the effect of mutations upon the rest of interacting genes, which is very crucial for studying any disease like cancer.

Using BGO and machine learning tools, the sort of analysis on brain cancer tumour response (Central Nervous System data) that we have performed in chapter 8 of this thesis is the first step that opens new door in cancer diagnosis and prognosis in providing the relationships, either evidential or predicted, between genes. The other benefit of using such integrated

systems is to avoid repeatedly re-discovering any relationships that have already been made by other researchers.

(2) MicroRNA study - step towards exhaustive genome-wide surveys for discovering microRNA, aid in cancer and brain disease study

Over recent years, miRNAs have emerged as major players in the complex networks of gene regulation and have been implicated in various aspects of human disease. Only five years after the first study reported a direct involvement of miRNAs in cancer, these small RNAs have already significantly improved our understanding of carcinogenesis [Stefanie Sassen et al. 2007]. MicroRNAs play a key role in diverse biological processes, including development, cell proliferation, differentiation, apoptosis and brain development [Miska 2004 and 2005]. Accordingly, altered miRNA expression is likely to contribute to human disease, including cancer and brain diseases.

The Myc oncogene encodes the transcription factor c-Myc (it was one of the top 32 selected genes for our Leukaemia case study in the chapter 4 of this thesis) that regulates cell proliferation, growth, and apoptosis, and overexpression of c-Myc is common in cancer [Kasabov and Dimitrov 2002]. He et al. (2005) demonstrated that additional expression of the microRNA "mir-17-92" cluster accelerated c-Myc-induced tumorigenesis in mice. Another research from Rockefeller University shows that neurons that cannot produce

microRNAs, tiny single strands of RNA that regulate the expression of genes, slowly die in a manner similar to what is seen in such human neurodegenerative disorders as Alzheimer's and Parkinson's diseases (Source: Science Daily 26 July 2007). Gerhard et al. (2006) showed that a brain-specific microRNA, miR-134, is localized to the synapto-dendritic compartment of rat hippocampal neurons and negatively regulates the size of dendritic spines - postsynaptic sites of excitatory synaptic transmission. This effect is mediated by miR-134 inhibition of the translation of an mRNA encoding a protein kinase, Limk1, that controls spine development. Exposure of neurons to extracellular stimuli such as brain-derived neurotrophic factor relieves miR-134 inhibition of Limk1 translation and in this way may contribute to synaptic development, maturation and/or plasticity.

Regulatory RNAs may also have therapeutic applications by which disease-causing miRNAs could be antagonized or functional miRNAs restored. The most intuitive choice of molecules to correct altered miRNA–messenger RNA interactions is RNA oligonucleotides. The miRNA profiles may become useful biomarkers for brain tumour diagnostics, and miRNA therapy could be a powerful tool for brain tumour prevention and therapeutics. Therefore, in addition to protein-coding oncogenes and tumour suppressor genes, we will have to take into account miRNAs and their regulatory networks if we aim to understand the complex processes underlying

malignant transformation. The methodology that we have described for addressing the classification problem of microRNAs may help researchers in connecting the knowledge between human miRNA biology and different aspects of carcinogenesis and brain diseases. A novel method of classifying 2D shapes of simulated, thermodynamically optimal folded RNA structures was introduced in the chapter 7 of this thesis. Our approach represents a simplified hypothesis-less ab-initio discovery tool for relevant novel structural features from 2D representations of RNA molecular structures using various image analysis algorithms. We have showed that using visual information from bitmap images of 2D structures of microRNA precursors one can extract potentially novel and useful information that can be used for discovery and classification of related molecules. It bypasses the need for complex 3D structural data comparisons similar to protein threading methods (e.g. sequence similarity guided 3D folding to a known structure). Our promising results on 222 human microRNA precursors suggested that the method can reveal new and useful discriminatory features of RNA molecules.

This can be useful additional information in multimodal data integration for improved microRNA gene discovery and classification. There are plenty of prospects for expanding the methodology for other types of molecules and utilization of other image analysis methods, as well as possibilities for exhaustive genome-wide surveys for discovery and classification of novel

non-coding RNA molecules from various organisms. It could be useful even for classification of the huge and complicated organic molecule databases, giving in effect a simple short-cut via bitmap images to discovery of novel classification and clustering methods of such databases. Especially interesting are the prospects for high-throughput simulated folding of non-coding RNA molecule candidates along the genomes of various organisms. Drug discovery has traditionally started with a biochemical pathway implicated in a pathophysiological process. We know that microRNAs controls the gene regulation by playing interactive roles therefore such information may be used to design novel drugs in order to potentially block the target of microRNAs. Once the binding of two molecules may be prevented the gene mutations leading to abnormalities or diseases can be controlled. Such work seems well warranted, because the RNA World is still largely unknown and unexplored, and many interesting discoveries will thus be forthcoming in this hot area of molecular biology, where the nanoscale interactions of RNA molecules with other molecules control the life and death of all eukaryotic cells in minute detail. Therefore, predicting novel microRNAs within the genome using the innovative computational intelligence approach as ours should help scientists to think in the direction of analysing their potential functions in various diseases like cancer and mental disorders.

(3) Knowledge integration - BGO as another step in brain disease modelling and cure

We have designed the BGO (for more details see chapter 8 of this thesis) to facilitate active learning and research in the areas of bioinformatics, neuroinformatics, information engineering, and knowledge management and we claim that different parts of it can be used by different users, from a school level to postgraduate and PhD student level.

It may be that understanding the interaction through modelling would be a key to understanding each level of information processing in the brain and perhaps the brain as a whole. Using principles from different levels in one model and modelling their relationship can lead to a next generation of brain models as more powerful tools to understand the brain. We believe that integrated framework of our BGO system makes data easily available for advanced methods of analysis, including artificial intelligence algorithms, that can tackle the multitude of large and complex datasets for clustering, classification and rule inference for biomedical and bioinformatics applications. Results from the machine learning procedures can be entered back to the ontology thus enriching its knowledge base and facilitating new discoveries. Such methods and knowledge integration vision can facilitate new discoveries in the area of genetic neuroscience. We hope that by linking and integrating simulation results from the various experiments with genetic information in the

BGO, one can facilitate better understanding of metabolic pathways and modelling of gene regulatory networks, and ultimately a more complete understanding of the pathogenesis of brain diseases for cure.

9.5 Conclusion and Open discussion

Understanding molecular interactions (gene regulatory networks) has "never been an easy task". Such knowledge in molecular biology may be very useful in discovering the new metabolic pathways, gaining knowledge on gene expression levels in future time, predicting effect of drugs over time and ultimately the pathogenesis of disease. MicroRNAs have recently been highlighted and their potential role in gene regulation is now being accepted broadly by the scientific community. Experimental methods to detect the binding or interacting events between two molecules are slow and very costly in this direction. Same is indeed true for classifying and predicting microRNAs as well. Therefore, there has always been the demand of computational intelligence bioinformatics methods to study these areas in system biology. Research groups have applied computer science methods in this direction but it is well understood that each of these methods have several limitations (refer chapter 2). In addition, studying each domain of this problem is one aspect of the research but another perspective which has not been taken into account by this scientific community is the integration of vast amount of knowledge and reusing this knowledge to enable further discoveries in this era of science.

This thesis has addressed some of the above aspects of research in which we have suggested and used some novel bioinformatics methods. We have proposed an integrated framework (refer chapter 3) to study the gene regulation problem and fulfil the desire of scientific community that molecular interactions should be studied at different checkpoints of the central dogma and range of data should actually be taken into account for analysis using novel computational intelligence methods. Within the framework we have suggested the possible use of multiple data types, for example, gene expression time series, sequence and literature etc. and included the area of microRNAs prediction. Finally, after examining each regulatory stage of a cell – we have proposed that integration is required to make the possible reuse of the obtained knowledge. In chapters 4, 5 and 6, three separate gene expression time series datasets were analysed using different integrative computational intelligence methods like Genetic algorithm (GA) with Kalman filter (KF), Least Angle Regression (LARS), Expectation Maximization (EM) with Kalman Filter (KF), Evolving Fuzzy Neural Network (EFuNN) and Quantum Inspired Algorithm (QiEA) with the aim of inferring the meaningful GRNs. To rely on any discoveries made we have potentially used the valuable inputs from diverse sources of information, like gene and protein sequence data analysis, promoter prediction and analysis, functional analysis, clustering, and literature data and so on. Next, in chapter 7 we have done the miRNA structure analysis (folding) based on our novel approach of the Gabor filtering.

Again, while doing the miRNA analysis we have used multiple data information, for example, miRNA sequence analysis (BLAST) and clustering (CLUSTALW). Later in chapter 8, we focus on knowledge integration and information fusion (using ontology) and discuss the utility of the machine learning tools in integrated framework.

We have learnt that multiple sources of data and hybrid methods reveals some new aspects of the problem and we suggest that to infer the GRNs it is more suitable to adopt such integrative approach as ours through which some new knowledge is discovered. We are aware that our suggested methods are not the end of this research and each method has certain advantages and limitations. There is also a scope of further development in each domain of this research study and possibly one may integrate the some other computational approaches with our methods to obtain even better results. The undertaken research has wide variety of applications in the science area. This thesis is a novel research work that we have successfully published in journals, books and conferences etc. and we hope this study will definitely help scientific community to reconsider the way to tackle this problem in molecular biology.

References

Abraham, W. C., Mason, S. E., Demmer, J., Williams, J. M., Richardson, C. L., Tate, W. P., Lawlor, P. A., & Dragunow, M. (1993). Correlation between immediate early gene induction and the persistence of long-term potentiation. Neuroscience, vol. 56(3), pp. 717-727

Abraham, W. C., & Williams, J. M. (2003). Properties and mechanisms of LTP maintenance. The Neuroscientist, vol. 9(6), pp. 463-474

Aleman, A., Hijman, R., Haan, E. H. F. d., & Kahn, R. S. (1999). Memory impairment in schizophrenia: a meta-analysis. Am. J. Psychiatry, vol. 156(9), pp. 1358-1366

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic Local Alignment Search Tool. J Mol Biol. vol. 5; 215(3), pp. 403-10

Allen E, Xie Z, Gustafson AM, Sung GH, Spatafora JW, Carrington JC. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat Genet. vol. 36(12), pp.1245-6

Amy E Pasquinelli, Shaun Hunter and John Bracht (2005). MicroRNAs: a developing story. Current Opinion in Genetics & Development, vol. 15, pp. 200–205

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., & Cherry, J. M. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics, vol. 25, pp. 25–29

Arevalo-Rodriguez M. and J. Heitman (2005). "Cyclophilin A is localized to the nucleus and controls meiosis in Saccharomyces cerevisiae," Eukaryot Cell, vol. 4, pp. 17-29

Arnone, M. I., and Davidson, E.H. (1997). "The hardwiring of development: Organization and function of genomic regulatory systems." Development, vol. 124, pp. 1851-1864

Baeck, T. (1995). Evolutionary algorithm in theory and practice: evolution strategies, evolutionary programming, and genetic algorithms. New York, Oxford University Press

Baeck, T., D. B. Fogel, et al. (2000). Evolutionary Computation I and II. Advanced algorithm and operators. Bristol, Institute of Physics Pub

Balakrishnan S., and D. Madigan (2006). Algorithms for Sparse Linear Classifiers in the Massive Data Setting," J. Machine Learning Research, vol. 1

Baldi, P. and S. Brunak (2001). Bioinformatics - a Machine Learning Approach. Cambridge, MA, MIT Press

Bardoni, B., Schenck, A., & Mandel, J. L. (1999). A novel RNA-binding nuclear protein that interacts with the fragile X mental retardation (FMR1) protein. Human Molecular Genetics, vol. 8(13), pp. 2557-2566

Bar-Joseph Z. (2004). "Analyzing time series gene expression data," Bioinformatics, vol. 20(16), pp. 2493-2503

Bay, J. S. (ed.) (1999). Fundamentals of Linear State Space Systems, WCB/McGraw-Hill

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). "Clustering gene expression patterns." J. Comp. Biol. vol. 6(3-4), pp. 281-297

Benuskova L. (2000). The intra-spine electric force can drive vesicles for fusion: a theoretical model for long-term potentiation. Neurosci. Lett., vol. 280(1), pp. 17-20

Benuskova L., Jain V., Wysoski S. G. and Kasabov N. (2006). Computational Neurogenetic Modelling: A pathway to new discoveries in Genetic Neuroscience. Intl. Journal of Neural Systems, ISSN 0129-0657, vol. 16 (3), pp 215-226

Benuskova L. and Kasabov N. (2007). Computational Neurogenetic Modeling. Springer, New York. ISBN:978-0-387-48353-5

Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, Cuppen E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. Cell. vol. 14;120(1), pp. 21-24

Berriz G. F., O. D. King, B. Bryant, C. Sander, and F. P. Roth (2003). "Characterizing gene sets with FuncAssociate," Bioinformatics, vol. 19, pp. 2502-2504

Bertram, L., & Tanzi, R. E. (2005). The genetic epidemiology of neurodegenerative disease. J. Clin. Invest., vol. 115(6), pp. 1449-1457

Björn Voss, Carsten Meyer and Robert Giegerich (2004). Evaluating the predictability of conformational switching in RNA. Bioinformatics vol. 20(10), pp. 1573-1582

Bliss, T. V., & Lomo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbitt following stimulation of perforant path. J. Physiol., vol. 232(2), pp. 331-356

Bolouri, J. M. B. a. H. (2001). Computational modelling of Genetic and Biochemical Networks. London, The MIT Press

Bompfuenewerer, Athanasius F.; Flamm, Christoph; Fried, Claudia; Fritzsch, Guido; Hofacker, Ivo L.; Lehmann, Joerg; Missal, Kristin; Mosig, Axel; Mueller, Bettina; Prohaska, Sonja J.; Stadler, Baerbel M. R.; Stadler, Peter F.; Tanzer, Andrea; Washietl, Stefan; Witwer, Christina (2005). Evolutionary patterns of non-coding RNAs. Theory in Biosciences vol. 123(4), pp. 301-369

Bozon, B., Kelly, Á., Josselyn, S. A., Silva, A. J., Davis, S., & Laroche, S. (2003). MAPK, CREB and zif268 are all required for the consolidation of recognition memory. Philos. Trans. R. Soc. Lond. B Biol. Sci., 358(1432), 805-814

Bramham, C. R., Southard, T., Sarvey, J. M., Herkenham, M., & Brady, L. S. (1996). Unilateral LTP triggers bilateral increases in hippocampal Neutrophin and trk receptor mRNA expression in behaving rats: evidence for interhemispheric communication. J. Comparative Neurology, vol. 368, pp. 371-382

Brown, R. G. (1983). Introduction to Random Signal Analysis and Kalman Filtering, John Wiley & Son

Brown M. P. S., G. W. N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares Jr M. and Haussler D. (2000). "Knowledge-based analysis of microarray gene expression data by using support vector machines." Proc. Natl. Acad. Sci. USA, vol. 97(1), pp. 262-267

Brownstein, M. J., Trent, J.M., and Boguski, M.S. (1998). "Functional genomics In M. Patterson and M. Handel, eds." Trends Guide to Bioinformatics, pp. 27-29

Carter, Richard J., Dubchak, Inna and Holbrook, Stephen R. (2001), A computational approach to identify genes for functional RNAs in genomic sequences, Nucleic Acid Research, vol. 29(19), pp. 3928-3938

Ceusters W, Smith B, Coldberg L (2005). A Terminological and Ontological Analysis of the NCI Thesaurus. Methods Inf Med, vol. 44(4), pp. 498-507

Chan, S. H., Havukkala, I., Jain, V., Hu, Y. and Kasabov, N. (2008). Soft Computing Methods to predict Gene Regulatory Networks: An Integrative approach on Time-Series Gene Expression Data. Applied Soft Computing Journal, vol. 8(3), pp 1189-1199

Chan, S. H., Kasabov N. and Collins L. (2005). A hybrid genetic algorithm and expectation maximization method for global gene trajectory clustering, Journal of Bioinformatics and Computational Biology, Imperial College Press, vol. 3 (5), pp. 1227-1242

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them? Intelligent Systems and Their Applications, vol. 14(1), pp. 20-26

Chang, E. H., Savage, M. J., Flood, D. G., Thomas, J. M., Levy, R. B., Mahadomrongkul, V., Shirao, T., Aoki, C., & Huerta, P. T. (2006). AMPA receptor downscaling at the onset of Alzheimer's disease pathology in double knock in mice. Proc. Natl. Acad. Sci. USA, vol. 103(9), pp. 3410-3415

Chen, A., Muzzio, I. A., Malleret, G., Bartsch, D., Verbitsky, M., Pavlidis, P., Yonan, A. L., Vronskaya, S., Grody, M. B., Cepeda, I., Gilliam, T. C., & Kandel, E. R. (2003). Inducible enhancement of memory storage and synaptic plasticity in transgenic mice expressing an inhibitor of ATF4 (CREB-2) and C/EBP proteins. Neuron, vol. 39(4), pp. 655-669

Chen D., W. Toone, J. Mata, R. Lyne, G. Burns, K. Kivinen, A. Brazma, N. Jones, and J. Bahler (2003). "Global transcriptional responses of fission yeast to environmental stress," Molec. Biol. Cell, vol. 14, pp. 214-229

Chen, J., Park, C. S., & Tang, S.-J. (2006). Activity-dependent synaptic Wnt release regulates hippocampal long-term potentiation. J. Biol. Chem., vol. 281(17), pp. 11910–11916

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. vol. 31(13), pp. 3497-500

Cho, R. J., Campbell, M.J., Winzeler, E.A. Steinmetz, L. Conway, A. Wodicka, L. Wolsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W.

(1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." Mol Cell vol. 2, pp. 65-73

Clare A. and R. King (2002). "How well do we understand the clusters found in microarray data?," In Silico Biology, vol. 2, pp. 511-522

Cloninger, C. R. (2002). The discovery of susceptibility genes for mental disorders. Proc. Natl. Acad. Sci. USA, vol. 99(21), pp. 13365-13367

Collado-Vides, J. and R. Hofestadt, Eds. (2002). Gene Regulation and Metabolism. Post-Genomic Computational Approaches. Cambridge, MA, MIT Press

Collado-Vides, J. (1989). "A transformational-grammar approach to study the regulation of gene expression." J. Theor. Biol. vol. 136, pp. 403-425

Dameron I, Roques E, Rubin D, Marquet G, Burgun A (2006). Grading lung tumors using OWL-DL based reasoning. Paper presented at the 9th International Protege Conference, Stanford, USA

Davidson E. H., D. R. McClay, and L. Hood (2003). "Regulatory gene networks and the properties of the developmental process," Proc.Natl.Acad.Sci. USA, vol. 100, pp. 1475-1480

Davis, S., Vanhoutte, P., Pagès, C., Caboche, J., & Laro, S. (2000). The MAPK/ERK cascade targets both Elk-1 and cAMP response element-binding protein to control long-term potentiation-dependent gene expression in the dentate gyrus in vivo. Journal of Neuroscience, vol. 20(12), pp. 4563-4572

Defoin-Platel M., Schliebs S., & Kasabov N. (2007). A versatile quantum-inspired evolutionary algorithm. In IEEE Congress on Evolutionary Computation (CEC'07), Singapore, pp. 423–430

De Jong, H. (2002). Modelling and simulation of genetic regulatory systems: a literature review, Journal of Computational Biology vol. 9 (1), pp. 67-103

Dickey, C. A., Loring, J. F., Montgomery, J., Gordon, M. N., Eastman, P. S., & Morgan, D. (2003). Selectively reduced expression of synaptic plasticity-related genes in amyloid precursor protein + presenilin-1 transgenic mice. J. Neurosci., vol. 23(12), pp 5219-5226

Ding Y, Lawrence CE. (2003). A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. 2003 Dec 15; vol. 31(24), pp. 7280-301

Dorf, R. and R. H. Bishop (1998). Modern Control Systems, Prentice Hall

Dow, L. a. M. (1996). Biochemistry: Molecules, cells and the body, Addison-Wesley

Edelman, D. B., Meech, R., & Jones, F. S. (2000). The homeodomain protein Barx2 contains activator and repressor domains and interacts with members of the CREB family. J. Biol. Chem., vol. 275(28), pp. 21737-21745

Egan, M. F., Goldberg, T. E., Kolachana, B. S., Callicott, J. H., Mazzanti, C. M., Straub, R. E., Goldman, D., & Weinberger, D. R. (2001). Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. Proc. Natl. Acad. Sci. USA, vol. 98(12), pp. 6917-6922

Efron B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). "Least Angle Regression," Ann. Statist. vol. 32(2), pp. 407–499

Eisen, M. B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc. Natl. Acad. Sci. USA vol. 95, pp 14863-14868

Encyclopaedia Britannica, http://www.encyclopedia.com/

Endy, D., and Brent, R. (2001). Modelling cellular behaviour, Nature vol. 409, pp. 391-395

Fensel, D. (2004). Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce (2 ed.). Heidelberg: Springer

Fields, S., Kohara, Y. and Lockhart, D. J. (1999). "Functional genomics." Proc Natl. Acad. Sci USA vol. 96, pp. 8825-8826

Fiser J, King I. (1997). Gabor-wavelet decomposition based filtering of gray-level images for object and scene recognition experiments. Spat Vis. vol. 11(1), pp. 117-119

Friedman, L., Nachman, Pe'er (2000). Using Bayesian networks to analyze expression data. Journal of Computational Biology, vol. 7, pp. 601-620

Friedman N., M. Linial, I. Nachman, and D. Pe'er (2000). "Using Bayesian networks to analyze expression data," Journal of Computational Biology, vol. 7, pp. 601-620

Fukushima, K (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. Neural Networks, Vol. 1(2), pp. 119-130

Galor, O. (2004). Introduction to Stability Analysis of Discrete Dynamical Systems. Macroeconomics 0409011, EconWPA, http://ideas.repec.org/p/wpa/wuwpma/0409011.html.

Galperin Michael Y (2008). The Molecular Biology Database Collection: 2008 update. Nucleic Acids Research, Vol. 36 (Database issue):D2-D4

Gerhard M. Schratt, Fabian Tuebing, Elizabeth A. Nigh, Christina G. Kane, Mary E. Sabatini, Michael Kiebler and Michael E. Greenberg (2006). A brain-specific microRNA regulates dendritic spine development, Nature, vol. 439, pp. 283-89

Gilks W.R., B.D.M. Tom, and A. Brazma (2005). “Fusing microarray experiments with multivariate regression,” Bioinformatics, vol. 21(2), pp. ii137-ii143

Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization and machine Learning. Reading, MA, Addison-Wesley

Gottgtroy P, Kasabov N, MacDonell S (2006) Evolving ontologies for intelligent decision support. In: Sanchez E (ed) Fuzzy Logic and the Semantic Web (Capturing Intelligence), Elsevier, Amsterdam, Chapter 21, pp 415-439

Gottgtroy P., Kasabov N., & MacDonell S. (2004). An ontology driven approach for knowledge discovery in Biomedicine. In C. Zhang, H. W. Guesgen & W. K. Yeap

(Eds.), Berlin: Springer-Verlag, PRICAI: Trends in Artificial Intelligence. Proc. VIII Pacific Rim Intl. Conf. AI, Lecture Notes in Artificial Intelligence, Vol. 3157, pp. 53-67

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. (2008). miRBase: tools for microRNA genomics. Nucleic Acids Res., vol. 36, (Database issue):D154-158

Gruber TR (1993). A translation approach to portable ontologies. Knowledge Acquisition, vol. 5, pp. 199-220

Gu, Y., McIlwain, K. L., Weeber, E. J., Yamagata, T., Xu, B., Antalffy, B. A., Reyes, C., Yuva-Paylor, L., Armstrong, D., Zoghbi, H., Sweatt, J. D., Paylor, R., & Nelson, D. L. (2002). Impaired conditioned fear and enhanced long-term potentiation in Fmr2 knock-out mice. Journal of Neuroscience, vol. 22(7), pp. 2753-2763

Guldener U., M. Munsterkotter, G. Kastenmuller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. Garcia-Martinez, J. E. Perez-Ortin, H. Michael, A. Kaps, E. Talla, B. Dujon, B. Andre, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H. W. Mewes (2005). "CYGD: the Comprehensive Yeast Genome Database," Nucleic Acids Res, vol. 33, pp. D364-368

Gustafsson M., Hörnquist M., and A. Lombardi (2005). “Constructing and analyzing a large-scale gene-to-gene regulatory network — lasso-constrained inference and biological validation,” IEEE/ACM Transact. Comp. Biol. Bioinformatics, vol. 2(3), pp. 254-261

Håkansson P, Dahl L, Chilkova O, Domkin V, Thelander L. (2006). The Schizosaccharomyces pombe replication inhibitor Spd1 regulates ribonucleotide reductase activity and dNTPs by binding to the large Cdc22 subunit. J Biol Chem. vol. 281(3), pp1778-83

Halees, A. S., Leyfer, D., & Weng, Z. (2003). PromoSer: a large-scale mammalian promoter and transcription start site identification service Nucleic Acids Research, vol. 31(13), pp. 3554-3559

Han J-D.J., N. Bertin, T. Hao, D. S. Goldberg, G.l F. Berriz, L. V. Zhang, D. Dupuy, A.J.M. Walhout, M.E. Cusick, F.P. Roth, and M. Vidal (2004). “Evidence for dynamically organized modularity in the yeast protein–protein interaction network,” Nature, vol. 430, pp. 88-93

Han K. H., & Kim, J. H. (2003). On setting the parameters of quantum-inspired evolutionary algorithm for practical application. Paper presented at the IEEE Congress on Evolutionary Computation (CEC'03)

Han K., B. Ju, and H. Jung (2004). "WebInterViewer: visualizing and analyzing molecular interaction networks," Nucleic Acids Res., vol. 32, pp. W89-95

Harlan Robins, Ying Li, and Richard W. Padgett (2005). Incorporating structure to predict microRNA targets. PNAS, vol. 102 (11), pp. 4006–4009

Hastie T., R. Tibshirani, and J. Friedman (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction: Springer-Verlag

Havukkala I., Pang S., Jain V., and Kasabov N. (2005). Novel Method for Classifying MicroRNAs by Gabor Filter Features from 2D Structure Bitmap Images. Journal of Theoretical and Computational Nanoscience, vol. 2 (4), pp. 506-513

Havukkala I., Benuskova L., Pang S., Jain V., Kroon R., and Kasabov N. (2006). Image and Fractal Information Processing for Large-Scale Chemoinformatics, Genomics Analyses and Pattern Discovery. Pattern Recognition in Bioinformatics, Lecture Notes in Computer Science (LNCS), ISBN 3-540-37446-9, vol. 4146, pp. 163-173

He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, Powers S, Cordon-Cardo C, Lowe SW, Hannon GJ, Hammond SM (2005). A microRNA polycistron as a potential human oncogene. Nature, vol. 435, pp. 828–833

Hey T. (1999). Quantum computing: an introduction. Computing & Control Engineering, vol. 10, pp. 105-112

Hicks A., Davis, S., Rodgera, J., Helme-Guizona, A., Laroche, S., & Malleta, J. (1997). Synapsin I and syntaxin 1B: key elements in the control of neurotransmitter release are regulated by neuronal activation and long-term potentiation in vivo. Neuroscience, vol. 79(2), pp. 329-340

Hofestadt R. a. M., F. (1995). "Interactive modelling and simulation of biochemical networks." Comput. Biol Med., vol.25, pp. 321-334

Holter, N. S., Maritan A., Cieplak, M. Fedoroff, N.V. and Banavar, J.R. (2001). "Dynamic modelling of gene expression data." Proc Natl. Acad. Sci USA, vol. 98(4), pp. 1693-1698

Huang X., W. Pan, S. Grindle, X. Han, Y. Chen, S.J Park, L.W. Miller, and J. Hall (2005). “A comparative study of discriminating human heart failure etiology using gene expression profiles,” BMC Bioinformatics, vol. 6(205), pp. 1-15

Impey, S., Obrietan, K., Wong, S. T., Poser, S., Yano, S., Wayman, G., Deloume, J. C., Chan, G., & Storm, D. R. (1998). Cross talk between ERK and PKA is required for Ca2+ stimulation of CREB-dependent transcription and ERK nuclear translocation. Neuron, vol. 21, pp. 869-883

Ivo L. Hofacker (2003). RNA secondary structure analysis using the Vienna RNA Package. In A.D. Baxevanis and D.B. Davison, editors, Current Protocols in Bioinformatics, volume 1. John Wiley & Sons

Ivo L. Hofacker (2003). The Vienna RNA secondary structure server. Nucl. Acids Res., vol. 31, pp. 3429–3431

Joseph J. D., J. Heitman, A.R.Means (1999). “Molecular cloning and characterization of Aspergillus nidulans cyclophilin B,” Fungal Genet Biol.,vol. 27(1), pp. 55-66

Jain V., Benuskova L., Gottgtroy P., and Kasabov N. (2006). Brain Gene Ontology, Proceedings of International Australasian Winter Research Conference on Brain Research, ISSN 1176-3183, vol. 24, pp 30

Jain V., Kasabov N., Gottgtroy P., Benuskova L., Joseph F. (2007). Brain Gene Ontology (BGO): Tool to facilitate Education and Research in Neuroinformatics area. Proceedings of the Fifth New Zealand Computer Science Research Student Conference (NZCSRSC), April 10–13th at University of Waikato, Hamilton, vol. 5, pp 262-265

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., & Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. Trends Biochem Sci., vol. 23, pp. 403-405

Jee, S. W., Cho, J. S., Kim, C. K., Hwang, D. Y., Shim, S. B., Lee, S. H., Sin, J. S., Kim, Y. S., Park, J. H., Lee, S. H., Choi, S. Y., & Kim, Y. K. (2007). Analysis of differentially expressed genes in early- and late-stage APPsw-transgenic and normal mice using cDNA microarray. Int J Mol Med., vol. 19(3), pp. 461-468

Jia-Fu Wang, Hui Zhou, Yue-Qin Chen, Qing-Jun Luo and Liang-Hu Qu (2004). Identification of 20 microRNAs from Oryza sativa. Nucleic Acids Research, Vol. 32(5), pp. 1688-1695

John G. Doench and Phillip A. Sharp (2004). Specificity of microRNA target selection in translational repression. GENES & DEVELOPMENT, vol. 18, pp. 504–511

Kahn P. (1995). "From genome to proteome: Looking at cell's proteins." Science, vol. 270, pp. 369-370

Kasabov N. (2007a). Evolving Connectionist Systems. The Knowledge Engineering Approach. 2nd edition, Springer, New York. ISBN-10: 1-84628-345-0

Kasabov N. (2007b). Global, local and personalised modelling and profile discovery in Bioinformatics: An integrated approach, Pattern Recognition Letters, vol. 28(6), pp 673-685

Kasabov N. (2006). Adaptation and Interaction in Dynamical Systems: Modelling and Rule Discovery through Evolving Connectionist Systems, Applied Soft Computing, vol. 6(3), pp 307-322

Kasabov N. (2004). Knowledge based neural networks for gene expression data analysis, modelling and profile discovery, Drug Discovery Today: BIOSILICO, vol. 2(6), pp. 253-261

Kasabov N. (2003). Evolving Connectionist Systems. Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines. London, Springer-Verlag

Kasabov N. (2001a). Evolving Fuzzy Neural Networks for Supervised/Unsupervised On-Line, Knowledge-Based Learning, IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, vol. 31(6), pp. 902-918

Kasabov N. (2001b). Artificial Neural Networks for Intelligent Information Processing, Transactions of Chemical Engineering, London, pp. 27-28

Kasabov N. (2001c). On-line learning, reasoning, rule extraction and aggregation in locally optimised evolving fuzzy neural networks, *Neurocomputing,* vol. 41, pp 25-41

Kasabov N. (2001d). Adaptive learning system and method, Patent USA, PCT WO 01/78003

Kasabov N. (1996). Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering. Cambridge, Massachussets, MIT Press, pp. 570 [ISBN 0 -262-11212-4]

Kasabov N. and Benuskova L. (2004). *Computational Neurogenetics,* International Journal of Theoretical and Computational Nanoscience, vol. 1(1) American Scientific Publisher, pp. 47-61

Kasabov N. and D. Dimitrov (2002). A method for gene regulatory network modelling with the use of evolving connectionist systems. ICONIP - International Conference on Neuro-Information Processing, Singapore, IEEE Press, pp. 596-601

Kasabov N., Jain V., Benuskova L., (2008). Integrating Evolving Brain-Gene Ontology and Connectionist-based System for Modeling and Knowledge Discovery, Neural Networks, vol. 21(2-3), pp 266-275

Kasabov N., Song Q., Benuskova L., Gottgtroy P., Jain V., Verma A., Havukkala I., Rush E., Pears R., Tjahjana A., Hu Y., MacDonell S. (2008). Integrating Local and Personalised Modelling with Global Ontology Knowledge Bases for Biomedical and Bioinformatics Decision Support, in: Smolin et al (Eds.), Computational Intelligence in Bioinformatics, Springer

Kasabov N., Chan S. H., Jain V., Sidirov I., and Dimitrov S. D. (2004). Gene Regulatory Network Discovery from Time-Series Gene Expression Data – A Computational Intelligence Approach, Lecture notes in computer science (LNCS), Springer-Verlag, vol. 3316, pp. 1344-1353

Kasabov N., Chan S. H., Jain V., Sidirov I., and Dimitrov S. D. (2005). Computational modeling of gene regulatory networks, In Vladimir B Bajic and Tan Tin Wee (Eds.), Information Processing and Living Systems. Imperial college press (ICP), World scientific Publishers, vol. 2, pp. 673-686

Kasabov N., Sidorov I., D S Dimitrov (2005). Computational Intelligence, Bioinformatics and Computational Biology: A Brief Overview of Methods, Problems and Perspectives, Journal of Computational and Theoretical Nanoscience, vol. 2(4), pp 473-491

Kasabov N., Jain V., Gottgtroy P., Benuskova L., Wysoski S., Joseph F. (2007). Evolving Brain-Gene Ontology System (EBGOS): Towards Integrating Bioinformatics and Neuroinformatics Data to Facilitate Discoveries, Proceedings of International Joint Conference on Neural Networks (IJCNN), Orlando, Florida, USA, ISBN: 1-4244-1380-X, pp. 1054

Kasabov N., Jain V., Gottgtroy P, Benuskova L, Joseph F (2007). Brain gene ontology and simulation system (BGOS) for a better understanding of the brain, Cybernetic and Systems: An international Journal, vol. 38 (5), pp. 495-508

Kasabov N., Jain, V., Gottgtroy, P., Benuskova, L. and Joseph, F. (2006). Brain-Gene Ontology: Integrating Bioinformatics and Neuroinformatics Data, Information and Knowledge to Enable Discoveries, IEEE, ISBN: 0-7695-2662-4, pp. 13

Kasabov N., Jain V., Benuskova L., Gottgtroy P., Joseph F. (2008). Integration of Brain-Gene Ontology and Simulation Systems for Learning, Modelling and Discovery, Computational Intelligence in Medical Informatics, Springer, ISBN: 978-3-540-75766-5, vol. 85, chapter 11, pp. 221-234

Kasabov N., and Song, Q. (2002). DENFIS: Dynamic Evolving Neural-Fuzzy Inference System and its Application for Time Series Prediction, IEEE Transactions on Fuzzy Systems, vol. 10 (2), pp. 144-154

Kolchanov, N. A. (2003). GeneExpress system, version 2.2.11. Russia, http://wwwmgs.bionet.nsc.ru/mgs/systems/geneexpress/

Lai EC, Tomancak P, Williams RW, Rubin GM. (2003). Computational identification of Drosophila microRNA genes. Genome Biol., vol. 4(7), R42

Lander, E. S. (1996). "The new genomics: Global views of biology." Science vol. 274, pp. 536-539

Levenson, J. M., Choi, S., Lee, S.-Y., Cao, Y. A., Ahn, H. J., Worley, K. C., Pizzi, M., Liou, H.-C., & Sweatt, J. D. (2004). A bioinformatics analysis of memory consolidation reveals involvement of the transcription factor c-Rel. J. Neurosci., vol. 24(16), pp. 3933–3943

Lewin, B. (1999). Genes VII. Oxford, Oxford University Press

Levine M, Tjian R (2003). Transcription regulation and animal diversity. Nature., vol.424(6945), pp. 147-151

Li L., Carter, J., Gao, X., Whitehead, J., & Tourtellotte, W. G. (2005). The neuroplasticity-associated arc gene is a direct transcriptional target of early growth response (Egr) transcription factors. Mol. Cell Biol., vol. 25(23), pp. 10286-10300

Likhoshvai V. A., Matushkin, Yu G., Vatolin, Yu N. and Bazan, S. I (2000). A generalized chemical kinetic method for simulating complex biological systems. A computer model of lambda phage ontogenesis." computational technol., vol. 5(2), pp. 87-89

Lledo, P.-M., Zhang, X., Sudhof, T. C., Malenka, R. C., & Nicoll, R. A. (1998). Postsynaptic membrane fusion and long-term potentiation. Science, vol. 279, pp. 399-403

Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. (2003). Vertebrate microRNA genes. Science, vol. 299(5612), pp.1540

Lin He and Gregory J.Hannon (2004). MicroRNAs: Small RNAs with a big tole in gene regulation. Nature genetics, vol. 5, pp. 522

Liu C, Powell KA, Mundt K, Wu L, Carr AM, Caspari T. (2003). Cop9/signalosome subunits and Pcu4 regulate ribonucleotide reductase by both checkpoint-dependent and -independent mechanisms. Genes Dev, vol. 17(9), pp.1130-40

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000). Molecular Cell Biology (4th ed.). New York: W.H. Freeman & Co.

Loomis, W. F., and Sternberg, P.W. (1995). "Genetic networks." Science, vol. 269, pp. 649

Lynch, M. A., Voss, K. L., Rodriguez, J., & Bliss, T. V. P. (1994). Increase in synaptic vesicle proteins accompanies long-term potentiation in the dentate gyrus. Neuroscience, vol. 60(1), pp. 1-5

Mann, M. (1999). "Quantitative proteomics." Nat. Biotechnol., vol. 17, pp. 954-955

Mattick John S. and Makunin Igor V. (2005). Small regulatory RNAs in mammals. Human Molecular Genetics, vol. 14, Review Issue 1, R121–R132

Marguerat S., T.S. Jensen, U. de Lichtenberg, B.T. Wilhelm, L.J. Jensen, and J. Bähler (2006). “The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast,” Yeast, vol. 23, pp. 261-277

Marianthi Kiriakidou, Peter T. Nelson, Andrei Kouranov, Petko Fitziev, Costas Bouyioukos, Zissimos Mourelatos, and Artemis Hatzigeorgiou (2004). A combined computational-experimental approach predicts human microRNA targets. Genes & Development, vol. 18, pp. 1165–1178

Martın-Cuadrado A. B., E, Duẽnas, M. Sipiczki, C.R. V´azquez de Aldana, and F. Del Rey (2003). “The endo-β-1,3-glucanase eng1p is required for dissolution of the primary septum during cell separation in Schizosaccharomyces pombe,” J. Cell Sci., vol. 116, pp. 1689–1698

Mayford, M., & Kandel, E. R. (1999). Genetic approaches to memory storage. Trends in Genetics, vol. 15(11), pp. 463-470

Mc Adams, H. H. a. A. A. (1997). Stochastic mechanism in gene expression. Proc. Natl. Acad. Sci. USA, vol. 94, pp. 814-819

Meberg, P. J., Valcourt, E. G., & Routtenberg, A. (1995). Protein F1/GAP-43 and PKC gene expression patterns in hippocampus are latered 1-2 h after LTP. Mol. Brain Res., vol. 34, pp. 343-346

Michel J. Weber (2005). New human and mouse microRNA genes found by homology search. FEBS Journal, vol. 272, pp. 59–73

Milhavet, O., Martindale, J. L., Camandola, S., Chan, S. L., Gary, D. S., Cheng, A., Holbrook, N. J., & Mattson, M. P. (2002). Involvement of Gadd153 in the pathogenic action of presenilin-1 mutations. J. Neurochemistry, vol. 83, pp. 673-681.

Miller, S., Yasuda, M., Coats, J. K., Jones, Y., Martone, M. E., & Mayford, M. (2002). Disruption of dendritic translation of CaMKIIalpha impairs stabilization of synaptic plasticity and memory consolidation. Neuron, vol. 36(3), pp. 507-519

Miska EA (2005). How microRNAs control cell division, differentiation and death. Curr Opin Genet Dev, vol. 15, pp. 563–568

Miska EA, Alvarez-Saavedra E, Townsend M, Yoshii A, Sestan N, Rakic P, Constantine-Paton M, Horvitz HR (2004). Microarray analysis of microRNA expression in the developing mammalian brain. Genome Biol, vol. 5, R68

Moretti, P., Levenson, J. M., Battaglia, F., Atkinson, R., Teague, R., Antalffy, B., Armstrong, D., Arancio, O., Sweatt, J. D., & Zoghbi, H. Y. (2006). Learning and memory and synaptic plasticity are impaired in a mouse model of Rett syndrome. The Journal of Neuroscience, vol. 26(1), pp. 319-327

Muhlenbein, H. (1992). How genetic algorithms really work: I. mutation and hillclimbing. Parallel Problem Solving from Nature 2. B. Manderick. Amsterdam, Elsevier

Murphy K. and S. Mian (1999). "Modelling Gene Expression Data using Dynamic Bayesian Networks," Technical report, Computer Science Division, UC Berkeley, CA

Nagl S.B. (2002). “Computational function assignment for potential drug targets: from single genes to cellular systems,” Current Drug Targets, vol. 3, pp. 387-399

O., P. B. (1997). "What lies beyond Bioinformatics?" Nat. Biotechnology, vol. 15, pp. 3-4

Okazaki Y et al (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature, vol. 420(6915), pp. 563-73

Pandey, A., and Mann M. (2000). "Proteomics to study genes and genomes." Nature, vol. 405, pp. 837-846

Pang P. T., & Lu B. (2004). Regulation of late-phase LTP and long-term memory in normal and aging hippocampus: role of secreted proteins tPA and BDNF. Ageing Research Reviews, vol. 3, 407-430

Pardossi-Piquard, R., Petit, A., Kawarai, T., Sunyach, C., Costa, C. A. d., Vincent, B., Ring, S., D’Adamio, L., & J. Shen, U. M. (2005). Presenilin-dependent transcriptional control of the Abeta-degrading enzyme neprilysin by intracellular domains of betaAPP and APLP Neuron, vol. 46(4), 541-554

Park C. S. Gong, R. & Tang S.-J. (2006). Molecular network and chromosomal clustering of genes involved in synaptic plasticity in the hippocampus. Journal of Biological Chemistry, vol. 281(40), pp. 30195-30211

Pasquinelli AE (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature, vol. 408(6808), pp. 86-89

Paul W.C. Hsu, H.-D. H., Sheng-Da Hsu, Li-Zen Lin, Ann-Ping Tsou, Ching-Ping Tseng, Peter F. Stadler, Stefan Washietl and Ivo L. Hofacker (2006). "miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes." Nucleic Acids Research, vol. 34, pp.135-139

Pe'er D., A. Regev, G. Elidan, and N. Friedman (2001). "Inferring subnetworks from perturbed expression profiles," Bioinformatics, vol. 17, pp. 215-224

Perrière, G., & Gouy, M. (1996). WWW-Query: An on-line retrieval system for biological sequence banks. Biochimie, vol. 78, pp. 364-369

Peng X., R. Karuturi, L. Miller, K. Lin, Y. Jia, P. Kondu, L. Wang, L. Wong, E. Liu, M. Balasubramanian, and J. Liu (2005). "Identification of Cell Cycle-regulated Genes in Fission Yeast," Molec. Biol. Cell, vol. 16, pp. 1026-1042

Penkett C.J., J.A. Morris, V. Wood, and J. Bähler (2006). “YOGY: a web-based integrated database to retrieve protein orthologs and gene ontologies,” Nucleic Acids Research, vol. 34, pp. W330-334

Pemberton T.J., and J.E. Kay (2005). “The cyclophilin repertoire of the fission yeast Schizosaccharomyces pombe,” Yeast, vol. 22, pp. 927–945

Pisanelli, D. M. (Ed.). (2004). Ontologies in Medicine. Amsterdam: IOS Press.

Pomeroy S. L., Tamayo P., Gaasenbeek M., Sturla L. M., & et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, vol. 415(6870), pp. 426

Poser, S., & Storm, D. R. (2001). Role of Ca2+-stimulated adenylyl cyclase in LTP and memory formation. Int. J. Devl. Neurosci., vol. 19, pp. 387-394

Puzzo, D., Vitolo, O., Trinchese, F., Jacob, J. P., Palmeri, A., & Arancio, O. (2005). Amyloid-beta peptide inhibits activation of the nitric oxide/cGMP/cAMP-responsive element-binding protein pathway during hippocampal synaptic plasticity. Journal of Neuroscience, vol. 25(29), pp. 6887-6897

Ropers, H.-H., Hoeltzenbein, M., Kalscheuer, V., Yntema, H., Hamel, B., Fryns, J.-P., Chelly, J., Partington, M., Gecz, J., & Moraine, C. (2003). Nonsyndromic X-linked mental retardation: where are the missing mutations? Trends in Genetics, vol. 19(6), pp. 316-320

Rustici G., J. Mata, K. Kivinen, P. Lió, C. J. Penkett, G. Burns, J. Hayles, A. Brazma, P. Nurse, and J. Bähler (2004). "Periodic gene expression program of the fission yeast cell cycle," Nature Genetics, vol. 36, pp. 809-817

Sam Grifths-Jones (2004). The microRNA Registry. Nucleic Acids Research, vol. 32, D109-D111

Sam Griffiths-Jones, R. J. G., Stijn van Dongen, Alex Bateman and Anton J. Enright (2006). "miRBase: microRNA sequences, targets and gene nomenclature." Nucleic Acids Research, vol. 34, pp. 140-144

Sanchez, L., van Helden, J. and thieffry, D. (1997). Establishment of the dorso-ventral pattern during embryonic development of Drosophila melanogaster. A logical analysis. J. Theor. Biol., vol. 189, pp. 377-389

Segal M.R., K.D. Dahlquist, B.R. Conklin (2003). “Regression approaches for microarray data analysis,” J. Comput Biol.; vol. 10(6), pp. 961-80

Sehgal M S, Gondal I and Dooley Laurence (2008). Computational Modelling Strategies for Gene Regulatory Network Reconstruction, Computational Intelligence in Medical Informatics, Springer, ISBN: 978-3-540-75766-5, Vol. 85, chapter 10, pp 207-220

Shapiro H. H. M. a. L. (1995). "Circuit simulation of Genetic Networks." Science, vol. 269(4), pp. 650-656

Shi, S. H., Hayashi, Y., Petralia, R. S., Zaman, S. H., Wenthold, R. J., Svoboda, K., & Malinow, R. (1999). Rapid spine delivery and redistribution of AMPA receptors after synaptic NMDA receptor activation. Science, vol. 284, pp. 1811-1816

Shimazu, K., Zhao, M., Sakata, K., Akbarian, S., Bates, B., Jaenisch, R., & Lu, B. (2006). NT-3 facilitates hippocampal plasticity and learning and memory by regulating neurogenesis Learning and Memory, vol. 13(3), pp. 307-315

Shun-Ichi Amari and Nikola K. Kasabov, editors (1998). Brain-Like Computing and Intelligent Information Systems, ISBN: 9813083581, Springer Verlag

Simon I., J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, and R. A. Young (2001). "Serial regulation of transcriptional regulators in the yeast cell cycle," Cell, vol. 106, pp. 697-708

Someren E, L. Wessels, and M. Reinders (2003). "Multi-criterion optimization for genetic network modelling," Signal Processing, vol. 83, pp. 763-775

Spellman P., Sherlock G., Zhang M., Iyer V., Anders K., Eisen M., Brown P., Botstein D., and Futcher B. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol. Biol. Cell, vol. 9, pp. 3273-3297

Stefanie Sassen, Eric A. Miska & Carlos Caldas (2007). MicroRNA—implications for cancer. Springer-Verlag, DOI 10.1007/s00428-007-0532-2

Stefan Washietl, I. L. H., Melanie Lukasser, Alexander Hüttenhofer and Peter F. Stadler (2006). "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome." Nature Biotechnology, vol. 23, pp. 1383-90

Stoenica, L., Senkov, O., Gerardy-Schahn, R., Weinhold, B., Schachner, M., & Dityatev, A. (2006). In vivo synaptic plasticity in the dentate gyrus of mice deficient in the neural cell adhesion molecule NCAM or its polysialic acid. Eur. J. Neurosci., vol. 23(9), pp. 2255-2264

Sudhof, T. C. (1995). The synaptic vesicle cycle: a cascade of protein-protein interactions. Nature, vol. 375, pp. 645-653

Sugai, T., Kawamura, M., Iritani, S., Araki, K., Makifuchi, T., Imai, C., Nakamura, R., Kakita, A., Takahashi, H., & Nawa, H. (2004). Prefrontal abnormality of schizophrenia revealed by DNA microarray: impact on glial and neurotrophic gene expression. Ann. N. Y. Acad. Sci., vol. 1025, pp. 84-91

Tchuraev, R. N. (1991) A new method for the analysis of the dynamics of the molecular genetic control systems. I. Description of the method of generalized threshold models. J. Theor. Biol., vol. 151, pp. 71-87

Thieffry, D. (1999). "From global expression data to gene networks." BioEssays, vol. 21(11), pp. 895-899

Thieffry, D. a. T. R. (1995). Dynamical behaviour of biological regulatory networks-II. Immunity control in bacteriophage lambda. Bull. Math. Biol, vol. 57, pp. 277-297

Thompson J., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," Nucleic Acids Research, vol. 22, pp. 4673-4680

Tibshirani R. (1996). "Regression Shrinkage and Selection via the Lasso," J. Royal Statistical Society B, vol. 58, pp. 267-288

Tikka J., and J. Hollmén (2004). “Learning linear dependency trees from multivariate time-series data,” In: Proceedings of the Workshop on Temporal Data Mining:

Algorithms, Theory and Applications (in conjunction with The Fourth IEEE International Conference on Data Mining), Brighton, UK

Vitolo, O., Sant'Angelo, A., Constanzo, V., Battaglia, F., Arancio, O., & Shelanski, M. (2002). Amyloid b-peptide inhibition of the PKA/CREB pathway and long-term potentiation: reversibility by drugs that enhance cAMP signaling. Proc. Natl. Acad. Sci. USA, vol. 99(20), pp. 13217-13221

Wang P. and J. Heitman (2005). "The cyclophilins," Genome Biology, vol. 6, pp. 226

Washietl S., H. I. L., Stadler P.F. (2005). "Fast and reliable prediction of noncoding RNAs." Proc. Natl. Acad. Sci., vol. 102, pp. 2454-2459

Watson JD, Crick FH (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid". Nature 171, vol. 4356, pp. 737–8

Wille A., P. Zimmermann, E. Vranova, and A. Furholz (2004). "Sparse graphical Gaussian modelling of the isprenoid gene network in Arabidopsis thaliana," Genome Biology, vol. 5(11), R92

Williams, J., Dragunow, M., Lawlor, P., Mason, S., Abraham, W. C., Leah, J., Bravo, R., Demmer, J., & Tate, W. (1995). Krox20 may play a key role in the stabilization of long-term potentiation. Molecular Brain Research, vol. 28, pp. 87-93

Wood V. (2006). “Schizosaccharomyces pombe comparative genomics; from sequence to systems,” In: Comparative Genomics using fungi as models . (P. Sunnerhagen, J. Piskur, eds.) Topics in Current Genetics, vol. 15, pp. 233-285

Wood, V., et al. (2002). The genome sequence of Schizosaccharomyces pombe. Nature, vol. 415, pp. 871-880

Wu, G.-Y., Deisseroth, K., & Tsien, R. W. (2001). Activity-dependent CREB phosphorylation: convergence of a fast, sensitive calmodulin kinase pathway and a slow, less sensitive mitogen-activated protein kinase activity. Proc. Natl. Acad. Sci. USA, vol. 98(5), pp. 2808-2813

Wu, Z., Ciallella, J., Flood, D., O’Kane, T., Bozyczko-Coyne, D., & Savage, M. (2006). Comparative analysis of cortical gene expression in mouse models of Alzheimer's disease. Neurobiol. Aging, vol. 27(3), pp. 377-386

Xiu-JieWang, José L Reyes, Nam-Hai Chua and Terry Gaasterland (2004). Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. Genome Biology, vol. 5, R65

Yamazaki, M., Matsuo, R., Fukazawa, Y., Ozawa, F., & Inokuchi, K. (2001). Regulated expression of an actin-associated protein, synaptopodin, during long-term potentiation. J. Neurochem., vol. 79, pp. 192-199

Yan, S. D., Fu, J., Soto, C., Chen, X., Zhu, H., Al-Mohanna, F., Collison, K., Zhu, A., Stern, E., Saido, T., Tohyama, M., Ogawa, S., Roher, A., & Stern, D. (1997). An intracellular protein that binds amyloid-b peptide and mediates neurotoxicity in Alzheimer's disease. Nature, vol. 389, pp. 689-695

Yeung M., J. Tegner, and J. Collins (2002). "Reverse engineering gene networks using singular value decomposition and robust regression," Proc. Natl. Acad. Sci. USA, vol. 99, pp. 6163-6168

Ying, S.-W., Futter, M., Rosenblum, K., Webber, M. J., Hunt, S. P., Bliss, T. V. P., & Bramham, C. R. (2002). Brain-derived neurotrophic factor induces long-term potentiation in intact adult hippocampus: requirement for ERK activation coupled to CREB and upregulation of Arc synthesis. The Journal of Neuroscience, vol. 22(5), pp. 1532-1540

Zerbinatti, C. V., Wahrle, S. E., Kim, H., Cam, J. A., Bales, K., Paul, S. M., Holtzman, D. M., & Bu, G. (2006). Apolipoprotein E and low density lipoprotein receptor-related protein facilitate intraneuronal Abeta42 accumulation in amyloid model mice. J. Biol. Chem., vol. 281(47), pp. 36180-36186

Zhang, J., Vinkemeier, U., Gu, W., Chakravarti, D., Horvath, C. M., & Darnell, J. E. (1996). Two contact regions between Stat1 and CBP/p300 in interferon gamma signaling. Proc. Natl. Acad. Sci. USA, vol. 93(26), pp. 15092-15096

Zhuo, M., Holtzman, D. M., Li, Y., Osaka, H., DeMaro, J., Jacquin, M., & Bu, G. (2000). Role of tissue plasminogen activator receptor LRP in hippocampal long-term potentiation Journal of Neuroscience, vol. 20(2), pp. 542-549

Zhou, Z., Hong, E. J., Cohen, S., Zhao, W.-n., Ho, H.-y. H., Schmidt, L., Chen, W. G., Lin, Y., Savner, E., Griffith, E. C., Hu, L., Steen, J. A. J., Weitz, C. J., & Greenberg, M. E. (2006). Brain-specific phosphorylation of MeCP2 regulates activity-dependent Bdnf transcription, dendritic growth, and spine maturation. Neuron, vol. 52(2), pp. 255-269

Zou H. and T. Hastie (2003). "Regularization and Variable Selection via the Elastic Net," J. Royal Society of Statistics B, vol. 67, pp. 301-320

Appendices:

A. Kalman filter (KF)

Kalman filter (KF), originally developed by Rudolf Kalman, is a set of recursive equations capable of computing optimal estimates (in the least-square sense) of the past, present and future states of the state-space model based on the observed data or in the other words, this efficient recursive filter estimates the state of a dynamic system from a series of incomplete and noisy measurements. The Kalman filter may be regarded as analogous to the hidden Markov model, with the key difference that the hidden state variables take values in a continuous space (as opposed to a discrete state space as in the hidden Markov model). Additionally, the hidden Markov model can represent an arbitrary distribution for the next value of the state variables, in contrast to the Gaussian noise model that is used for the Kalman filter.

The Kalman filter is a recursive estimator; this means that only the estimated state from the previous time step and the current measurement are needed to compute the estimate for the current state. In contrast to batch estimation techniques, no history of observations and/or estimates is required. The Kalman filter has two distinct phases: Predict and Update. The predict phase uses the state estimate from the previous timestep to produce an estimate of the state at the current timestep. In the update phase, measurement information at the current timestep is used to refine this prediction to arrive at a new, (hopefully) more accurate state estimate, again for the current timestep. More details about Kalman filter method and various equations may be referred in [Brown 1983].

In the chapter 4 of this thesis, we have used it to estimate gene expression trajectories given irregularly sampled data. To specify the operation of Kalman filter, we define the conditional mean value of the state x_t^s and its covariance P_{tu}^s as:

$$\mathbf{x}_t^s = E(\mathbf{x}_t \mid \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_s)$$

$$\mathbf{P}_{tu}^s = E\left[(\mathbf{x}_t - \mathbf{x}_t^s)(\mathbf{x}_u - \mathbf{x}_u^s)' \mid \mathbf{y}_1, ..., \mathbf{y}_s\right]$$

For prediction, we use the KF forward recursions to compute the state estimates for (s<t). For likelihood evaluation and parameter estimation, we use the KF backward recursions to compute the estimates called the smoothed estimates based on the entire data, i.e. (s=T; T>t is the index of the last observation), which in turn are used to compute the required statistics. Later the search method like GA for the selection of a gene subset for a GRN was used, the background on such evolutionary computation method may be found in the appendix B.

B. Evolutionary Computation and Genetic Algorithm (GA)

The evolution of nature inspired computational methods called evolutionary computation (EC). EC are stochastic search methods that mimic the behaviour of natural biological evolution. They differ from traditional optimization techniques in that they involve a search from a population of solutions, not from a single point, and carry this search over generations. So, EC methods are concerned with population-based search and optimisation of individual systems through generations of populations [Goldberg 1989; Koza 1992; Holland 1992, 1998]. Several different types of evolutionary methods have been developed independently. These include Genetic Programming (GP) which evolve programs, Evolutionary Programming (EP), which focuses on optimizing continuous functions without recombination, Evolutionary Strategies (ES), which focuses on optimizing continuous functions with recombination, and Genetic Algorithms (GAs), which focuses on optimizing general combinatorial problems, the latter being the most popular technique. To solve some of the GRN problems we have used GAs (refer chapter 4 of this thesis), here more details about the method is provided.

Genetic algorithms (GA) were introduced for the first time in the work of John Holland in 1975. They were further developed by him and other researchers [Holland, 1992, 1998; Goldberg 1989; Koza 1992]. The most important terms used in GA are analogous to the terms used to explain the evolution processes. They are:

- *Gene* – a basic unit that defines a certain characteristic (property) of an individual;

• *Chromosome* – a string of genes; used to represent an individual or a possible solution to a problem in the solution space population – a collection of individuals;

• *Crossover (mating) operation* – substrings of different individuals are taken and new strings (offspring) are produced – random change of a gene in a chromosome;

• *Fitness (goodness) function* – a criterion which evaluates how good each individual is;

• *Selection* – a procedure of choosing a part of the population which will continue the process of searching for the best solution, while the other individuals "die".

A simple genetic algorithm consists of steps shown in figure B1 given below. The process over time has been "stretched" in space. While figure B1 shows graphically how a GA searches for the best solution in the solution space, figure B2 gives an outline of the GA.

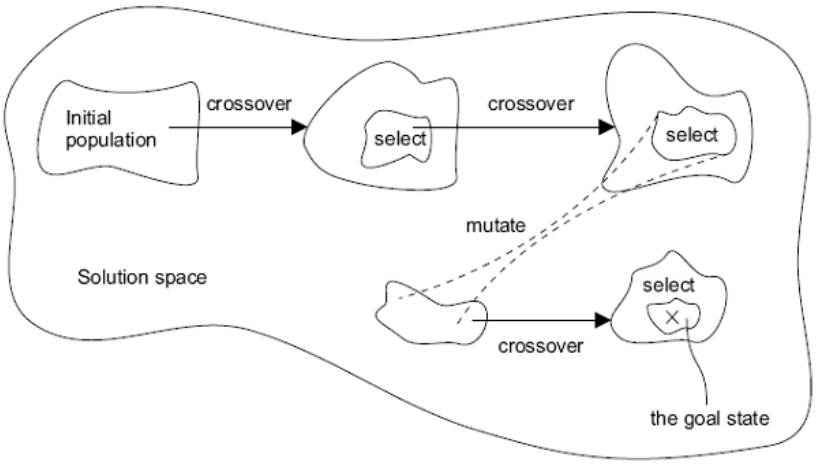


Figure B1: A schematic diagram of how a gene algorithm (GA) works in time (figure is taken from Kasabov 1996, MIT press)

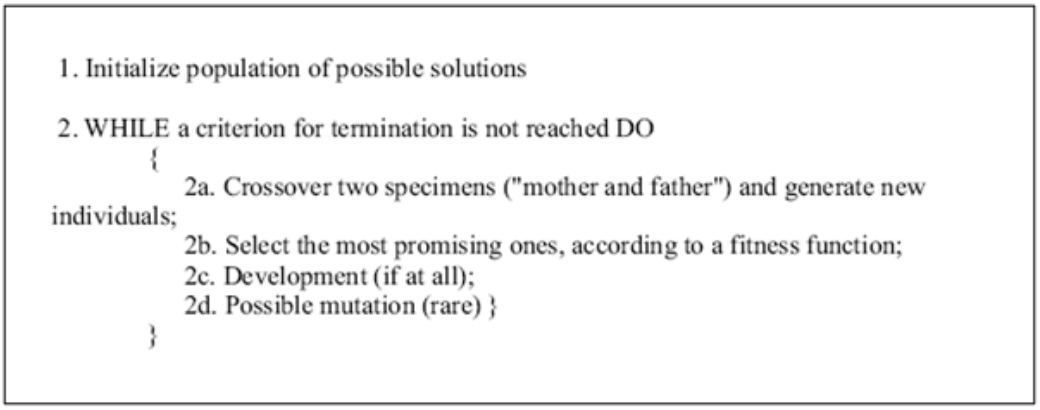


Figure B2: A general representation of the GA (figure is taken from Kasabov 1996, MIT press)

In short, the major characteristics of a GA are the following. They are heuristic methods for search and optimisation. In contrast to the exhaustive search algorithms, GA does not evaluate all variants in order to select the best one. Gene repair function can be used in the algorithm that replaces the missing information using the corrective template. Therefore they may not lead to the perfect solution, but to one which is closest to it taking into account the time limits. But nature itself is imperfect too (partly due to the fact that the criteria for perfection keeps changing), and what seems to be close to perfection according to one “goodness” criterion may be far from it according to another. For more details on EC and GA, readers are advised to refer to the second edition of the book on Evolving Connectionist Systems – the knowledge engineering approach, by Kasabov (2007a) published with Springer.

C. EFuNN and ECF

This appendix presents some background knowledge on well known connectionist methods for supervised learning, such as Evolving Classifier Function (ECF) and evolving fuzzy neural networks (EFuNN). For more details and examples on such methods, readers are advised to refer to the second edition of book on Evolving Connectionist Systems – the knowledge engineering approach, by Kasabov (2007a) published with Springer.

A simple evolving connectionist method (ECOS) for classification is Evolving Classifier Function, ECF (see figure C1). The learning and the recall algorithms of ECF are shown in Box C1 (a) and (b). Internal nodes in the ECF structure capture clusters of input data that belong to the same class. We have applied the ECF as an artificial intelligence method in conjunction with the BGO to analyse the central nervous system (CNS) cancer dataset [Pomeroy 2002]. Further details of our approach can be obtained in chapter 8 of this thesis and see figure C2 in which a leave-one-cross validation method was applied to validate an ECF ECOS model on the 60 CNS cancer samples.

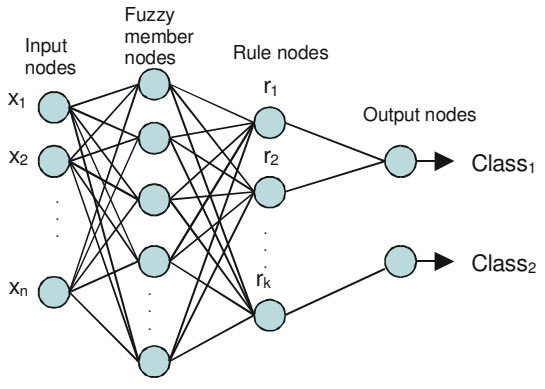


Figure C1: A simplified structure of an evolving classifier function ECF (figure is taken from Kasabov 2007a)

Box C1 (a): Learning algorithm of ECF:

1. Enter the current input vector from the data set (stream) and calculate the distances between this vector and all rule nodes already created using Euclidean distance (by default). If there is no node created, create the first one that has the coordinates of the first input vector attached as input connection weights.

2. If all calculated distances between the new input vector and the existing rule nodes are greater than a max-radius parameter *R*max, a new rule node is created. The position of the new rule node is the same as the current vector in the input data space and the radius of its receptive field is set to the min-radius parameter *R*min; the algorithm goes to step 1; otherwise it goes to the next step.

3. If there is a rule node with a distance to the current input vector less then or equal to its radius and its class is the same as the class of the new vector, nothing will be changed; go to step 1; otherwise:

4. If there is a rule node with a distance to the input vector less then or equal to its radius and its class is different from those of the input vector, its influence field should be reduced. The radius of the new field is set to the larger value from the two numbers: distance minus the min-radius; min-radius. New node is created as in 2 to represent the new data vector.

5. If there is a rule node with a distance to the input vector less than or equal to the max-radius, and its class is the same as of the input vector's, enlarge the influence field by taking the distance as a new radius if only such enlarged field does not cover any other rule nodes which belong to a different class; otherwise, create a new rule node in the same way as in step 2, and go to step 1.

Box C1 (b): Recall procedure (classification of a new input vector) in a trained ECF :

1. Enter the new vector in the ECF trained system; If the new input vector lies within the field of one or more rule nodes associated with one class, the vector is classified in this class;

2. If the input vector lies within the fields of two or more rule nodes associated with different classes, the vector will belong to the class corresponding to the closest rule node.

3. If the input vector does not lie within any field, then take m highest activated by the new vector rule nodes, and calculate the average distances from the vector to the nodes with the same class; the vector will belong to the class corresponding to the smallest average distance.

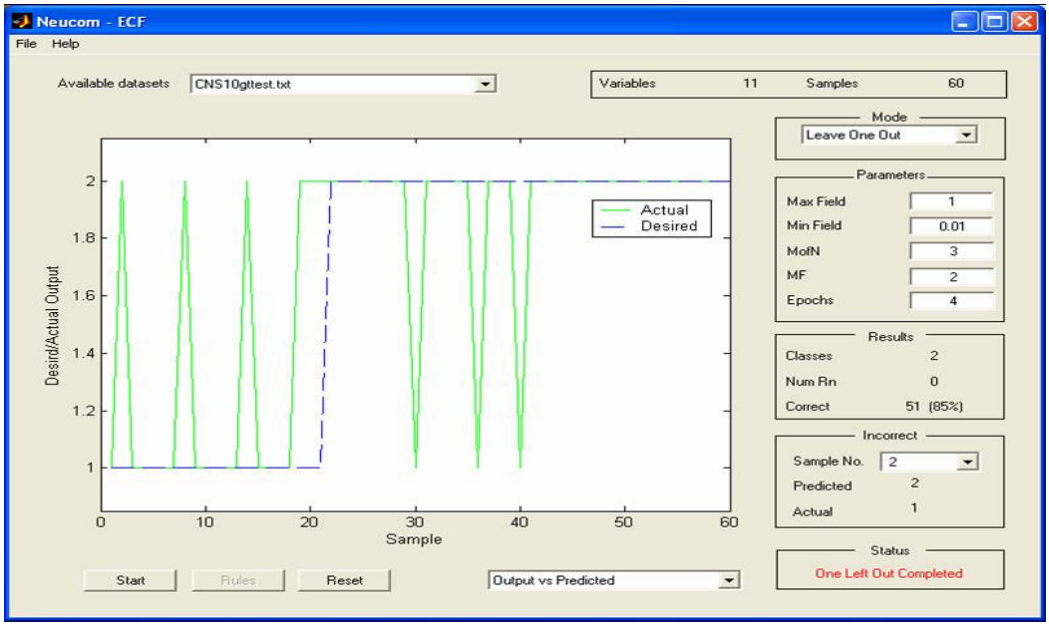


Figure C2: A leave-one-cross validation method was applied to validate an ECF ECOS model on the 60 CNS cancer samples [Pomeroy et al. 2002], where 60 models were created – each one on 59 samples, after one example was taken out, and then the model was validated to classify the taken out example. The average accuracy over all 60 examples was 85%, where 51 samples were classified accurately and 9 incorrectly. Class 1 is the non-responding group (21 samples, 71.43% accuracy) and class 2 is the group of survivals (39 samples, 92.31%). The results were better than the achieved in [Pomeroy et al. 2002] results of 78% (13 errors out of 60).

Fuzzy neural networks are connectionist structures that can be interpreted in terms of fuzzy rules [Yamakawa et al. 1992; Furuhashi et al., 1993; Lin and Lee 1996 and Kasabov 1996]. Evolving Fuzzy Neural Networks (EFuNNs) have a five-layer structure (figure C3). Here nodes and connections are created/connected as data examples are presented. An optional short-term memory layer can be used through a feedback connection from the rule (also called, case) node layer (see figure C4). The layer of feedback connections could be used if temporal relationships of input data are to be memorized structurally.

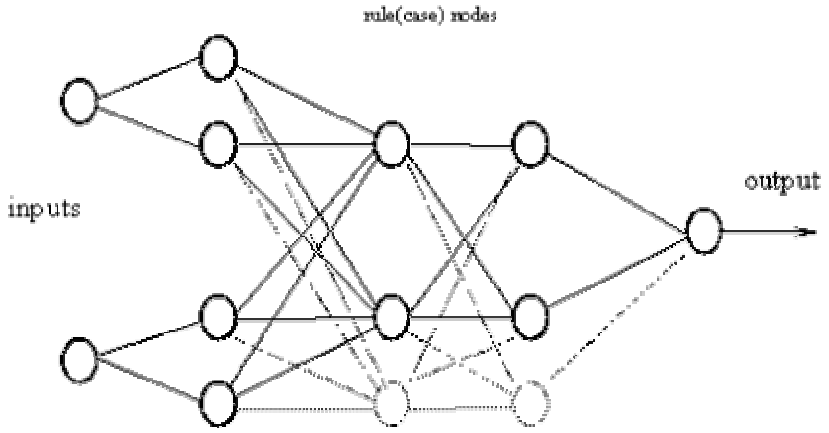


Figure C3: Evolving Fuzzy Neural Network (EFuNN): an example of a simplified standard feedforward EFuNN system (figure is taken from Kasabov 2007a)

The input layer represents input variables. The second layer of nodes (fuzzy input neurons, or fuzzy inputs) represents fuzzy quantisation of each input variable space. For example, two fuzzy input neurons can be used to represent "small" and "large" fuzzy values. Different membership functions (MF) can be attached to these neurons. The number and the type of MF can be dynamically modified. The task of the fuzzy input nodes is to transfer the input values into membership degrees to which they belong to the corresponding MF.

The third layer contains rule (case) nodes that evolve through supervised and/or unsupervised learning. The rule nodes represent prototypes (exemplars, clusters) of input-output data associations that can be graphically represented as associations of hyper-spheres from the fuzzy input and the fuzzy output spaces. Each rule node r is defined by two vectors of connection weights – $W1(r)$ and $W2(r)$, the latter being adjusted through supervised learning based on the output error, and the former being adjusted

through unsupervised learning based on similarity measure within a local area of the problem space. A linear activation function, or a Gaussian function, is used for the neurons of this layer.

The fourth layer of neurons represents fuzzy quantization of the output variables, similar to the input fuzzy neuron representation. Here, a weighted sum input function and a saturated linear activation function is used for the neurons to calculate the membership degrees to which the output vector associated with the presented input vector belongs to each of the output MFs. The fifth layer represents the values of the output variables. Here a linear activation function is used to calculate the defuzzified values for the output variables.

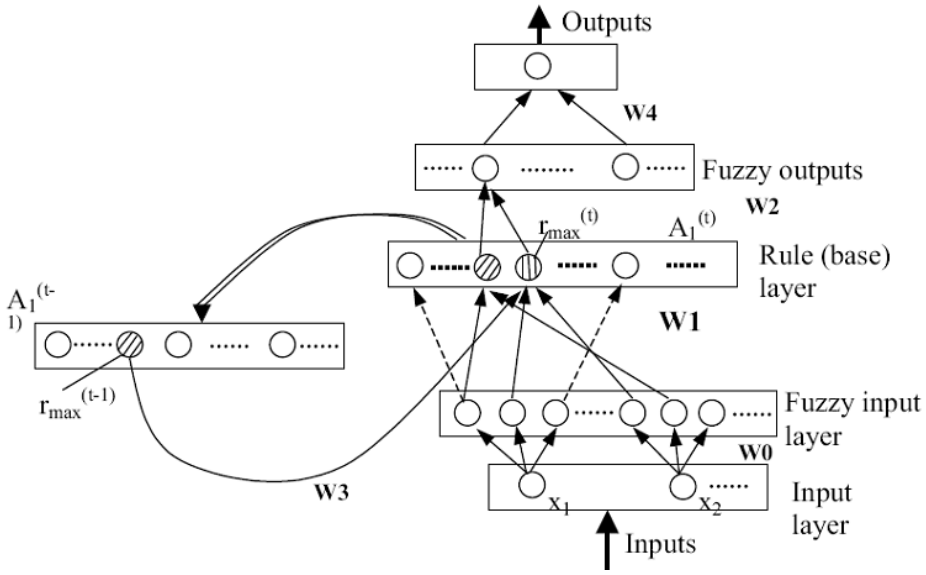


Figure C4: An example of EFuNN with a short term memory realised as a feedback connection (figure is taken from Kasabov 2007a)

EFuNNs were used by us as a part of integrative approach to infer GRN, so further details on the kind of analyses done on a yeast dataset are provided in chapter 5. At KEDRI, some of these well known ECOS methods have been implemented in the machine learning tool NEUCOM. The tool SIFTWARE also uses a computational intelligence approach. A brief background on NEUCOM and SIFTWARE is provided in the appendix D.

D. Neucom and Siftware

Neucom

NeuCom is a self-programmable, learning and reasoning computer environment based on connectionist (neurocomputing) modules (figure D1).

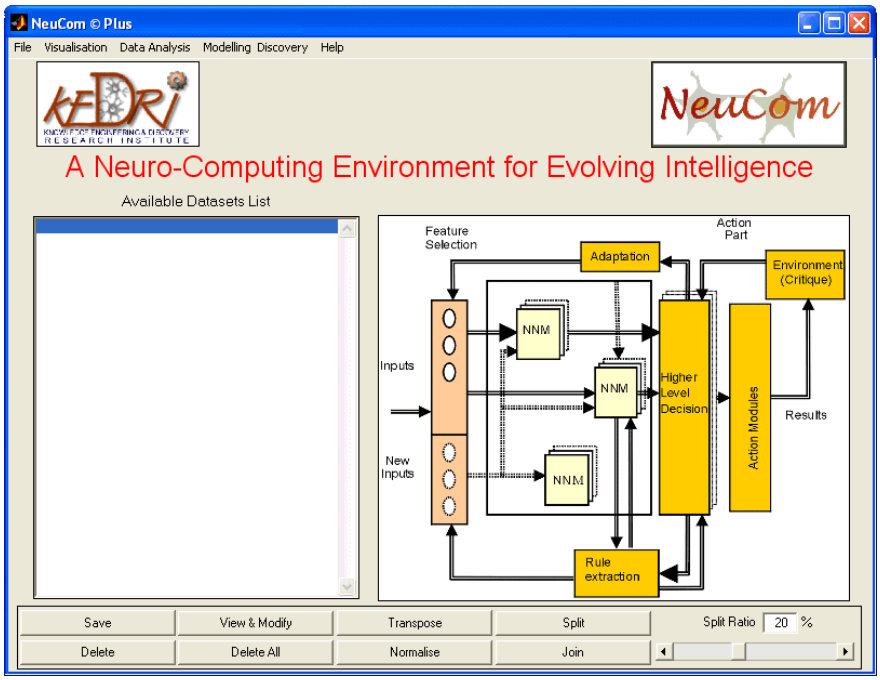


Figure D1: Screenshot of main GUI of NEUCOM

NeuCom is a complete software package for generic data processing that learns from data, thus uses evolving new connectionist modules. It contains a set of tools for data visualisation, normalisation, analysis, feature extraction, clustering and modelling with cross validation and genetic algorithm with an consistent and easy to use interface. The modules can adapt to new incoming data in an on-line incremental, life-long learning mode, and can extract meaningful rules that would help people to discover new knowledge in their respective fields. NeuCom is based on the theory of Evolving Connectionist Systems (ECOS) [Kasabov 2003].

NeuCom can be used either as a decisions support system (DSS), where users specify their task and define data to be used, in order to obtain a solution, or as a DSS development environment for building sophisticated problem oriented intelligent DSS. The end users in the former case are people who have never programmed computers, but have databases available and need a decision to be made based on existing data and/or human knowledge. In the latter case users are professional system developers who can develop DSS for various applications in collaboration with experts in the field.

NeuCom can be used to solve complex problems like clustering (see figure D2), classification, prediction, adaptive control, data mining and pattern discovery from databases in a multidimensional, dynamic and possibly changing data environment. Applications span all areas of Science, Engineering, Medicine, Bio-informatics, Business, Arts and Design, Education. NeuCom is currently being used by 35 universities from all over the world for teaching and research. To provide the detailed background on the NeuCom software and description of the example leukemia dataset etc. is beyond the scope of this thesis. Readers are however advised to visit www.theneucom.com for further information.

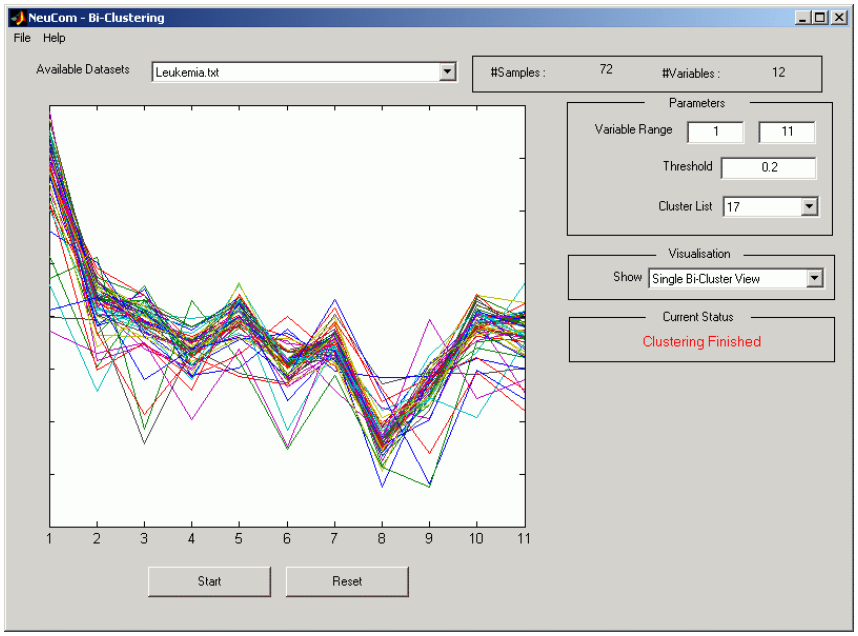


Figure D2: An example of clustering of Leukemia dataset using NEUCOM

Siftware

Gene Expression Profiling software – Siftware (see figure D3) was also developed in house at KEDRI especially for bioinformatics data analysis.

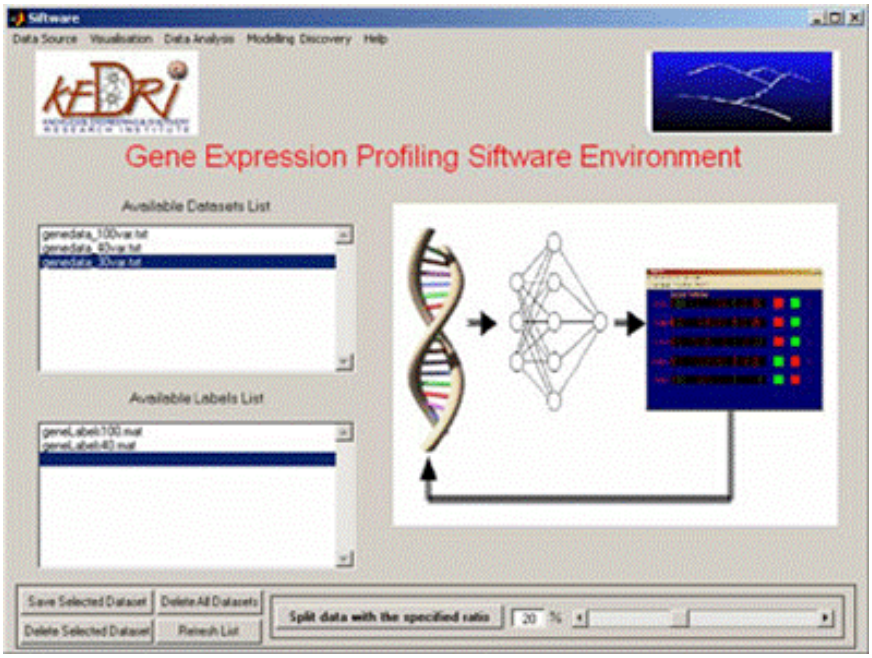


Figure D3: Screenshot of main GUI of a software system “SIFTWARE”

It implements the modules of (1) Visualisation which includes 3D visualisation and Principal Component Analysis Data analysis. This module includes several dimension reduction mechanisms like Signal to Noise Ratio (SNR) and correlation coefficient analysis. It also contains several clustering algorithms, K-means clustering and Hierarchical clustering. Next module (2) is for Modelling and discovery that includes cross validation for modelling and feature selection in an unbiased way including the essential cross validation module from the NeuCom project. The available methods are: SNR, t-Test (see example in figure D4, and for more details on this analysis refer to the chapter 8 of this thesis), Multiple Linear Regression, Support Vector Machine, K Nearest Neighbour, Weighted K Nearest Neighbour, Multi-Layer Perceptron, Radial Basis Function, Evolving Classification Function (ECF) and Evolving Clustering Method (ECM) for Classification.

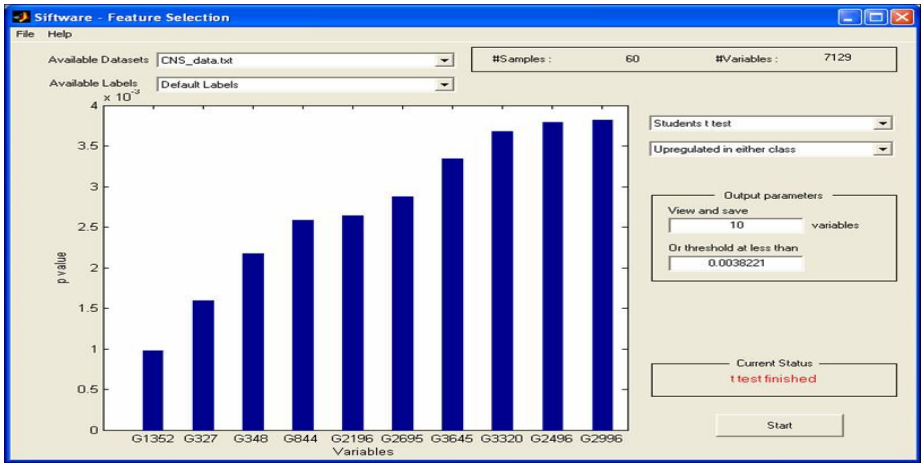


Figure D4: Ten genes selected as top discriminating genes from the Central Nervous System (CNS) cancer data [Pomeroy et al. 2002] that discriminates two classes - survivals and not responding to treatment. The Siftware system was used for the analysis and the method is called t-test.

To learn more about NEUCOM and SIFTWARE, please visit: "www.kedri.info"

E. GnetXP – description and user manual

GNetXP (also called hybridClust) is a software system (see figure E1) for gene time course data clustering and gene interaction network discovery. While developing this system our objective was to combine the strength of Genetic Algorithm and Expectation Maximization algorithm to produce a global yet efficient clustering algorithm.

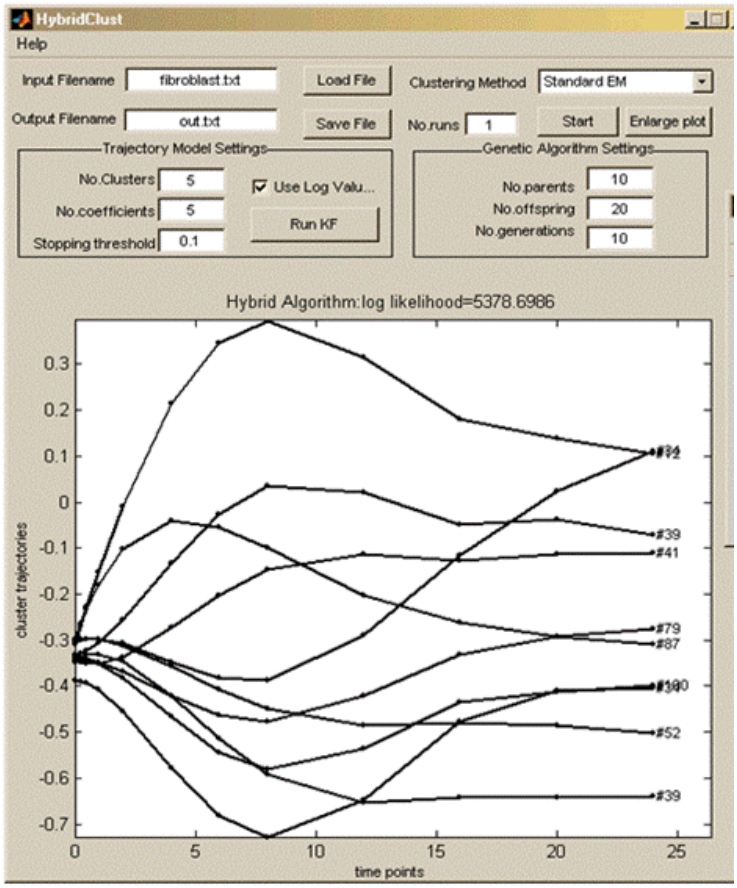


Figure E1: Screenshot of main GUI of GnetXP

At the higher level, GA searches for the optimal subset of genes that act as initial cluster centres and at the lower level, the local learning method, Expectation Maximization algorithm (EM), performs local clustering from these

initial centres. These two learning algorithms are implemented with mixture of Multiple Linear Regression models (MLRs) (A) Standard Expectation Maximization (EM) algorithm that uses random initialized cluster centres and (B) Hybrid Genetic Algorithm (GA) and EM that uses GA for initializing the cluster centres.

Using GA, the hybrid algorithm searches the clustering solution space more thoroughly, offering more consistent and more optimal solutions than the standard EM algorithm by far. Our software extracts GRN using two stage process (1) Hybrid Genetic Algorithm and Expectation Maximization algorithm is applied on clustering the large number of gene trajectories using the mixture of multiple linear regression models for fitting the trajectory data (2) Kalman Filter (parameter estimation) is applied to identify a set of first-order differential equations that describe the dynamics of the representative trajectories, and use these equations for discovering important gene interactions and predicting gene expression values at future time points. Demonstration results from the clustering and Kalman filter modules are shown in figure E2.

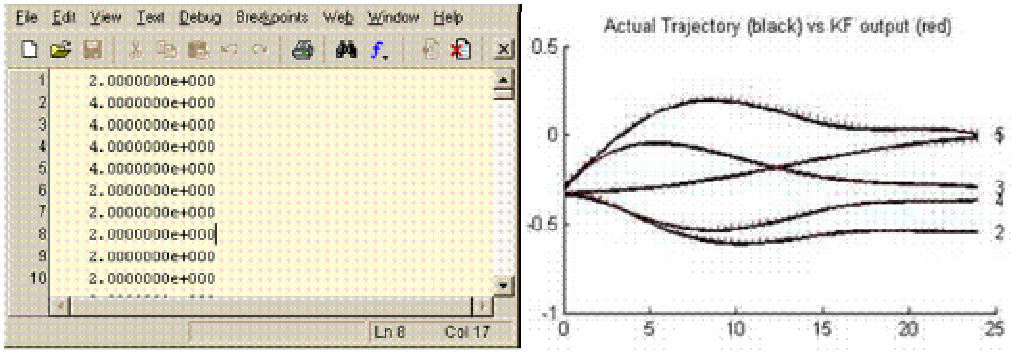


Figure E2: An example of clustering and Kalman Filter results

The software is a generalized one and may be used on any time series data. To use HybridClust, one must save the time series gene expression file into an appropriate structure (figure E3). The time points must be expressed as a row vector and stacked on top of the gene expression data. This matrix should then be saved as an ascii file. HybridClust will accept most ascii file format, including delimited and fixed width partitioning.

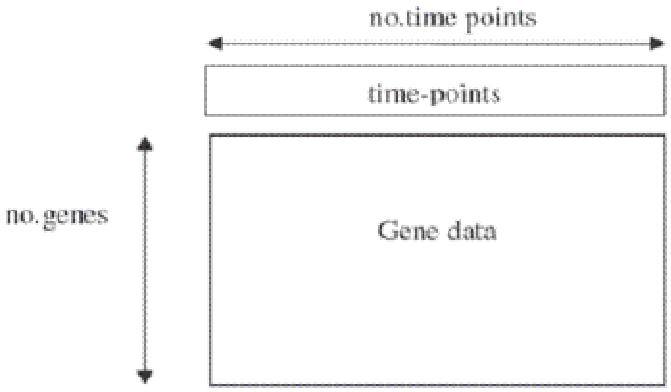


Figure E3: Suggested dataset structure for using GnextXP

The GnetXP system was used to derive GRN in the chapter 5 of this thesis where the more details on actual results can be obtained. The system is freely available to academic users and is for “non-commercial” use only. Potential researchers may download this software from KEDRI website, “www.kedri.info” under the centre for bioinformatics. Alternatively, one may visit the KEDRI’s computational intelligence repository at “kcir.kedri.info” and look for the project “gene regulatory network modelling”. Gnext XP software, readme file, related publications, user manual and demonstration movie etc. are available to download from this repository.

F. Supplementary information for chapter 6

This appendix contains some additional information for this thesis chapter 6. Therefore readers are requested to refer to the respective sections of original chapter. Figure F1 correspond to 14 temporal clusters and table F1 list the genes according to the time they were over-expressed and/or under-expressed.

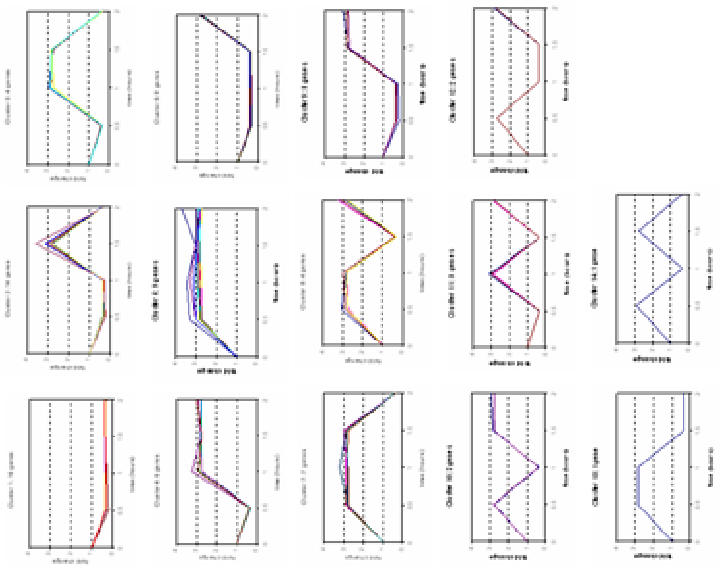


Figure F1: 14 temporal clusters of gene expression of the 79 selected genes based on gene expression time series. Each box represent one cluster and temporal profiles reflect the fold change (y axis) in gene expression over time in hours (x axis).

Table F1: Genes according to the time they are over-expressed and/or under-expressed. Information is developed based on the gene expression data that was used for experiments in the chapter 6

Function	30 min	60 min	90 min	120 min	Under- expressed
Transcription factors / DNA binding	CREB, ICER/Crem, Barx2, Camk4, Gadd153/Chop.	CREB, Atf2/Creb2, beta-catenin, c-fos, Egr2/Krox20, Camk4, ICER/Crem, Barx2, Stat1	CREB, ICER/Crem, Atf2/Creb2, Atf4/CREB-2, Cebpa, c-Rel, beta-catenin, c-fos, Egr2/Krox20, Camk4, Eg1/Zif268, Chop	beta-catenin, c-fos, Egr2/Krox20, Camk4, Elk1, Fosb, Junc, Lef1, Eg1/Zif268, Chop, Stat1	Abeta, Aplp2, Arx, Fmr2
Translation regulating factors / RNA binding	Fmr1/FMRP and Nufip (isoform AA681274), Akap1 (isoform U95145)	Akap1 (isoforms U95145 and U95146), Qk, Fmr1/FMRP, Nufip (isoform AA681274)	Akap1 (isoforms U95145 and U95146), Qk, Fmr1/FMRP, Nufip (isoform AA681274)		
Immediate early gene	Homer1_AB019479	c-fos, Egr2/Krox20, Homer1_AB019479	c-fos, Egr2/Krox20, Homer1_AB019479,	c-fos, Egr2/Krox20, Homer1_AB019479, Elk1, Fosb, Junc,	Junb

			Egr1/Zif268	Egr1/Zif268	
Cell cycle regulation	Tgm3	Tgm3, Mecp2	Cebpa, c-Rel, Ncam1 (isoform X15050), Tgm3	Tgm3, Mecp2	Arx, Junb
Cell matrix and/or adhesion	Aplp1	Aplp1	Ncam1 (isoform X15050)	Reln, Aplp1	Abeta, Aplp2, Ptpns1
Growth factors	Bdnf, Ntf3, tPA	Bdnf, Ntf3, Stat1	Bdnf, Ntf3, Ncam1 (isoform X15050)	Bdnf, Ntf3, Ncam1 (isoform X15052), Stat1, tPA	
Synaptic structure and plasticity	Bdnf, Fmr1/FMRP, Nufip (isoform AA681274), Synaptopodin	Bdnf, Fmr1/FMRP, Nufip (isoform AA681274), Pak3, Synaptopodin	Bdnf, Fmr1/FMRP, Nufip (isoform AA681274), Egr1/Zif268	Bdnf, Mapk1, Ncam1_X15052, Synaptophysin, Synaptopodin Pak3, Egr1/Zif268	Abeta, PSD95
Cytoskeleton, structure related activity	Synapsin 1, Synaptopodin	Beta-catenin, Synapsin 1, Synaptopodin	Mapt, beta-catenin, Synapsin 1, Egr1/Zif268	beta-catenin, Synapsin 1, Egr1/Zif268, Synaptopodin	PSD95, Ptpns1

Membrane fusion	Synapsin 1, Synaptotagmin	Synapsin 1, Synaptotagmin	Camk2a_X14836, Cplx2, syntaxin, Synapsin 1, Synaptotagmin	Synapsin 1	PSD95
Calcium-regulated	Camk4, Tgm3, PKC, Lrp1, Synaptotagmin	Camk4, Tgm3, PKC, Stat1, Synaptotagmin	Calm3, Camk2a, Camk4, Grin1, Grin2a, PKC, Tgm3, Synaptotagmin	Camk4, Tgm3, Synaptophysin, Stat1, Lrp1	Grin2b
Intracellular signalling	Wnt3a, Agtr2	Wnt3a, Agtr2, Stat1	Calm3, Agtr2	Wnt3a, Agtr2, Stat1	Gucy1b3, Ptpns1, Homer1-pending (isoform AF093257)
Retrograde signalling		nNOS/Nos1	nNOS/Nos1		
Receptor-associated binding	Ntf3, Synaptotagmin, Wnt3a	Ntf3, Synaptotagmin, Wnt3a	Ntf3, Synaptotagmin,	Ntf3, Wnt3a	Homer1-pending (isoform AF093257), PSD95
Ion-channel activity	Chrna7		Grin1, Grin2a, Chrna7		Gria1, Grin2b
Receptor activity	Agtr2, Lrp1, Chrna7	Agtr2	Grin1, Grin2a,	Agtr2, Lrp1, Trkb,	Gria1, Grin2b,

			Chrna7, Ncam1 (isoform X15050), Agtr2, Trkb, Trkc	Trkc	Ptpns1
Protein kinase activity	Camk4, MEK2, PKC, cGK/Prkg2, Akap (isoform U95145)	Camk4, Akap1 (isoforms U95145 and U95146), MEK2, PKC, cGK, Pak3	Camk2a, Akap1 (isoforms U95145 and U95146), Camk4, MEK2, PKC, Trkb, Trkc	Camk4, MEK2, Mapk1, Mapk14, Trkb, Trkc, cGK, Pka3	Camk2a_X87142, Gucy1b3, JNK/Mapk8, PKA
Transferase activity (not related to kinase activity)	Comt, Tgm3	Comt, Tgm3	Comt, Tgm3	Comt, Tgm3	
Hydrolase activity	tPA, Psen1	Psen2	Psen1, Psen2	Psen1, Psen2, Reln, tPA	Cdc25b (phosphatase)
Beta-amyloid or tau related	Psen1	COX-2	Psen1, COX-2	COX-2	ApoE

In continuation, figure F3, F4 and F5 correspond to the obtained clustering trees for proteins, genes and genes promoter sequences respectively.

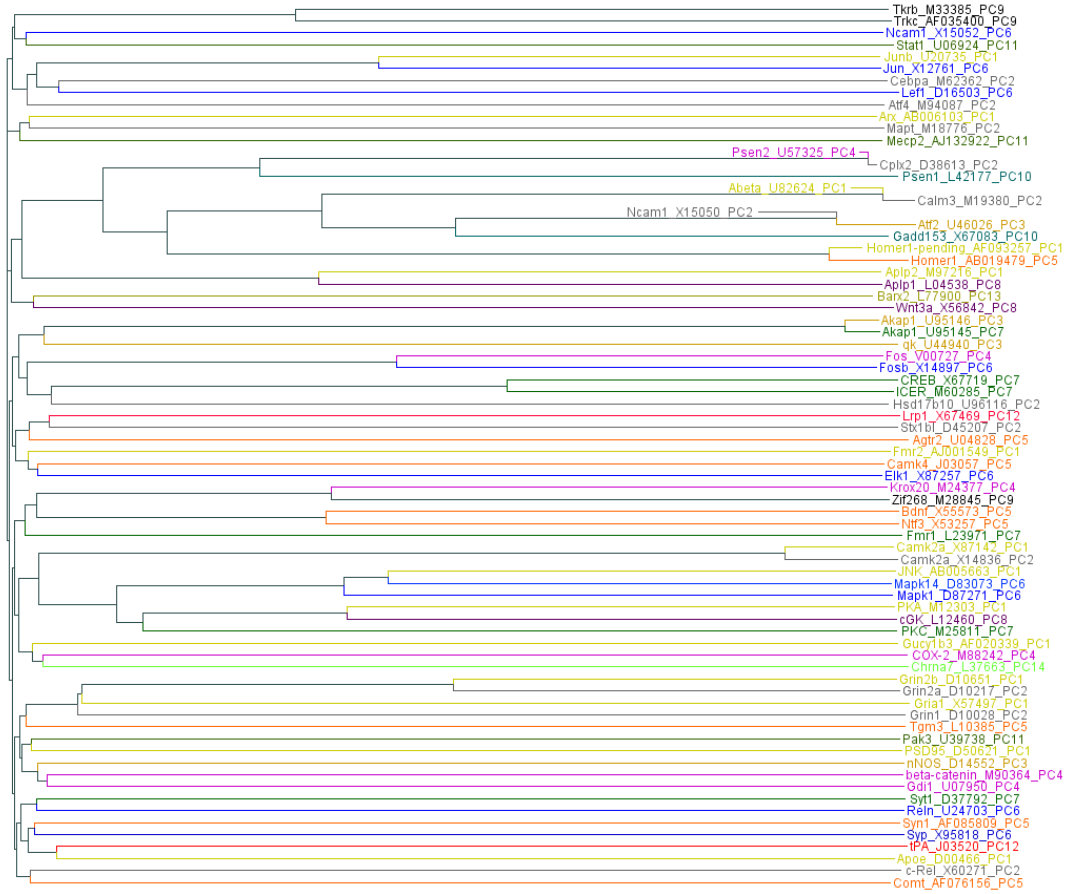


Figure F3: Phylogenetic tree of proteins coded for by 79 selected genes. Each temporal cluster is denoted by a different color and by the number at the end of the protein name, i.e. PC1 means protein cluster 1, etc., referring to clusters based on corresponding gene expression profiles from table 6.1 (chapter 6)

Figure F4: Phylogenetic tree of gene sequences of 79 selected genes. Each temporal cluster is denoted by a different color and by the number at the end of the gene name, i.e. GC1 means gene cluster 1, etc., referring to clusters from table 6.1 (chapter 6)

Figure F5: Phylogenetic tree of gene promoter sequences for the subset of 79 selected genes. Each temporal cluster is denoted by a different color and by the number at the end of the gene name, i.e. PrC1 means promoter cluster 1, etc., referring to clusters from table 6.1 (chapter 6). Some accessions could not be mapped very well to the mouse genomic loci or they were below the specified quality criteria that we have mentioned in chapter 6 for promoter selection, therefore, in our clustering analysis we were unable to include the promoters from each of the 79 genes

G. Snapshots from animations of BGO

Brain gene ontology (BGO) system has several voice enabled illustrative movies. We have captured few snapshots from the animations and they are presented below in the figure G1.

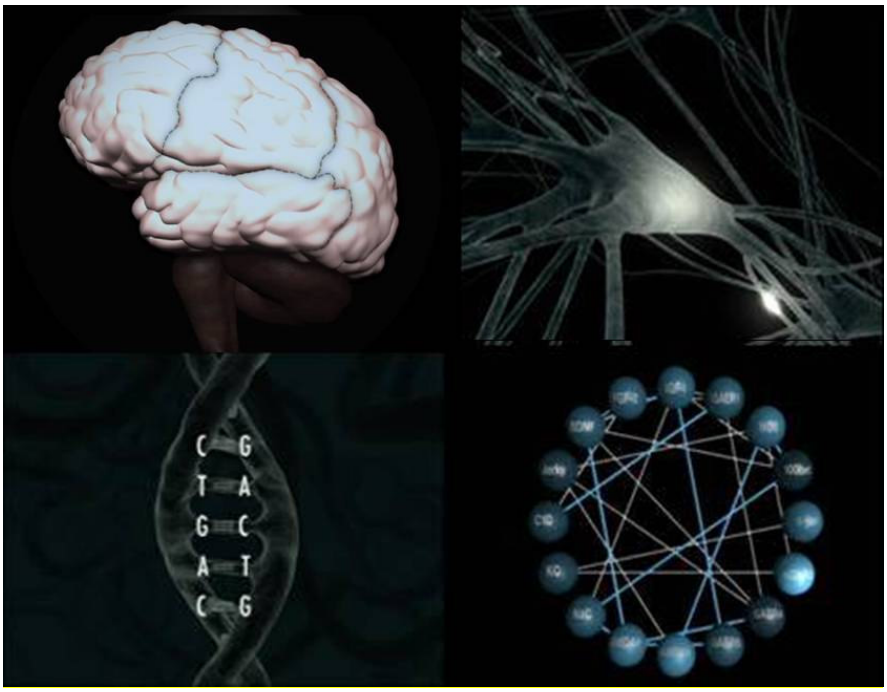


Figure G1: Snapshots from animations embedded in BGO system (upper left) the brain (upper right) signal propagation in and within neurons (lower left) gene sequence (lower right) building an abstract GRN

As mentioned in the chapter 8 of this thesis, the BGO system is freely available to academic users and is for “non-commercial” use only. Potential researchers may download this system from KEDRI website, “www.kedri.info” under the centre for neuroinformatics and brain study.