

CRANFIELD UNIVERSITY

Cranfield Health

PhD Thesis

Academic Year 2006-2007

Judit Kumuthini

**Extraction of Genetic Network from
Microarray Data Using Bayesian
Framework**

Supervisors:

Conrad Bessant PhD

Steve Setford PhD

April, 2007

This thesis is submitted in partial fulfilment of the requirements for the
Degree of Doctor of Philosophy

ProQuest Number: 10820998

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10820998

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by Cranfield University.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

The aim of the work described in this thesis was to develop novel methods for the extraction of gene regulatory networks (GRN) from gene expression data, and use these methods to capture previously unknown relationships between genes in specific biological applications. This has been accomplished through the application of Bayesian Networks (BN) through minimum description length (MDL) and taboo search for parameter and structure learning respectively to three large scale microarray datasets from *Saccharomyces cerevisiae*, *Escherichia coli* and human stem cells.

The application of BNs for modelling the well characterised yeast cell cycle demonstrated the efficacy of the techniques employed. Using the cDNA microarray data from the yeast cell cycle project by Spellman *et al.* (1998), this study succeeded in extracting many biologically plausible genetic relationships, which were supported by evidence from publicly available genome and literature databases. Two novel knowledge extraction techniques were applied; Target Node (TN) analysis and learning through simulation. Further, it was demonstrated how the addition of prior knowledge to the extracted network can improve the network structure extracted purely from experimental data.

The second part of this thesis demonstrated how the BN approach could be adapted to a data set of very high dimensionality, specifically data from a 54,634 probe array used to monitor human adipose tissue. Genetic networks extracted included insulin receptor (IR) and Fatty acid binding proteins (FABP) families that play key roles in fatty acid uptake, transport, and metabolism

In the final part of this thesis, the genome-wide GRNs of a prokaryotic expression system were extracted from novel oligo cDNA microarray data from *E-coli* K12 to identify metabolic stress responsive genes during recombinant protein production. Also, detailed analysis of known metabolic stress related genes and the genes that are directly or indirectly associated in the GRN were used to establish possible markers for host system exhaustion.

In conclusion, the BN methods developed proved to be a powerful and effective means of extracting GRNs in a variety of applications.

*Once Again In memory
of my brother*

Acknowledgements

I would like to express my sincerest thanks to my supervisor Dr. Conrad Bessant for all his suggestions and especially for his kind and continued guidance through the problems encountered during the past four years.

I am grateful to Prof. Karl Bayer and his group at BOKU for their excellent hospitality they generously provided during my stay in Vienna which will never be forgotten.

I would also like to thank Theresa Scharl (TU, Vienna), Joanna Jankowicz (Seibersdorf International Atomic Energy Agency, Vienna) and Stephan Sykacek (TU, Vienna) whose help I could not have managed without.

I would also like to thank Daniel Crowther and Paul Wilson at GSK for their patience and support.

This thesis would be incomplete without a mention of the support given to me by my cherished friends and colleagues at Cranfield University, Silsoe, especially over the past two years. Without them lifting me up when this thesis seemed interminable, I doubt it should ever have been completed. I am grateful for being the surrogate family during the many years I stayed at Silsoe and for their continued moral support there after.

Finally, I am forever indebted to my family and Jeeva for their understanding, endless patience and encouragement when it was most required.

List of Contents

ABSTRACT	I
LIST OF CONTENTS	III
LIST OF FIGURES.....	VII
LIST OF TABLES.....	XIII
ABBREVIATIONS.....	XV
1 INTRODUCTION	1
1.1 GENE REGULATORY NETWORKS (GRNs).....	3
1.1.1 <i>Genomics and Gene Expression</i>	3
1.1.2 <i>Protein synthesis</i>	5
1.1.3 <i>Understanding Gene Expression</i>	7
1.1.4 <i>Functional Genomics and the Human Genome</i>	10
1.1.5 <i>Post Genomic Data</i>	11
1.2 GENE EXPRESSION MECHANISMS	13
1.2.1 <i>Gene Expression in response to external stimuli</i>	13
1.2.2 <i>Tissue specific gene expression</i>	13
1.2.3 <i>Transcriptional regulation of gene expression</i>	16
1.2.4 <i>Post-transcriptional regulation of gene expression</i>	17
1.2.5 <i>Epigenetic mechanisms and long-range control of gene expression</i>	17
1.2.6 <i>Regulatory Networks</i>	18
1.3 METHODS FOR MEASUREMENT OF GENE (mRNA) EXPRESSION.....	19
1.3.1 <i>Microarrays</i>	19
1.3.1.1 <i>Types of Microarray</i>	22
1.3.1.2 <i>Microarray Applications</i>	23
1.3.1.3 <i>Data Handling and Interpretation</i>	24
1.3.1.4 <i>Microarray Databases</i>	25
1.3.1.5 <i>Standards of microarray data</i>	27
1.4 CAPTURING GENE EXPRESSION DATA USING MICROARRAYS	29
1.5 MODELLING OF GENETIC NETWORK.....	30
1.5.1 <i>Processing Microarray Gene Expression Data</i>	30
1.5.1.1 <i>Normalisation</i>	31
1.5.2 <i>Modelling and Interpretation of Gene Expression Data</i>	33
1.5.2.1 <i>Pearson Correlation Coefficient</i>	34
1.5.2.2 <i>Clustering analysis</i>	34
1.5.2.3 <i>Limitations of Clustering Methods</i>	36
1.5.3 <i>The Case for Bayesian Approach</i>	36
1.6 BAYESIAN PROBABILISTIC NETWORKS	40
1.6.1 <i>Bayesian Network</i>	40
1.6.2 <i>Network Topology</i>	43
1.6.3 <i>Bayesian Learning</i>	45
1.6.3.1 <i>Complete Data</i>	46
1.6.3.2 <i>Incomplete Data</i>	57

1.6.3.3	Identifying Causal Relations	60
1.6.3.4	Dynamic Bayesian Networks	61
1.6.4	<i>Measures of Error: Kullback-Leibler Divergence</i>	63
1.6.5	<i>Bayesian Analysis of Gene Expression Data</i>	65
1.6.5.1	Application to Cell Cycle Expression Patterns.....	67
1.6.6	<i>Bayesian Analysis Software</i>	68
2	MODELLING YEAST CELL CYCLE REGULATION	69
2.1	INTRODUCTION	69
2.2	MATERIALS AND METHODS.....	70
2.2.1	<i>Gene Expression Data</i>	70
2.2.1.1	Data Acquisition and Data Pre-processing	71
2.2.1.2	Data Extraction	71
2.2.1.3	Discretisation	72
2.2.1.4	Replacement of Missing Values	72
2.2.2	<i>Learning Algorithms</i>	73
2.2.2.1	SopLEQ	73
2.2.2.2	Taboo search.....	74
2.2.2.3	Supervised Learning	75
2.2.3	<i>Computation</i>	75
2.2.4	<i>Extraction of Gene Regulatory Network</i>	76
2.2.5	<i>Validation of Extracted GRNs</i>	77
2.2.6	<i>Knowledge Extraction from the Learned Networks</i>	77
2.2.6.1	Target Node Analysis	77
2.2.6.2	Network Simulation.....	78
2.3	RESULTS AND DISCUSSION	78
2.3.1	<i>Performance of Learning Algorithms Employed</i>	78
2.3.2	<i>Analysis of Yeast Cell Cycle Phases</i>	79
2.3.2.1	GRN for Yeast Cell Cycle, Phase M/G1	79
2.3.2.2	Unsupervised Learning: GRN for Yeast Cell Cycle, Phase G1	85
2.3.2.3	Supervised Learning: GRN of Yeast Cell Cycle, Phase G1	86
2.3.2.4	GRN for Yeast Cell Cycle, Phase Late G1	88
2.3.2.5	GRN for Yeast Cell Cycle, Phase S	89
2.3.2.6	GRN for Yeast Cell Cycle, Phase S/G2	90
2.3.2.7	GRN for Yeast Cell Cycle, Phase G2/M	91
2.3.2.8	GRN for Yeast Cell Cycle, Phases M/G1 and G1	92
2.3.2.9	GRN for Yeast Cell Cycle, All Phases	93
2.3.2.10	GRN for Yeast Cell Cycle, Updated for KEGG.....	96
2.3.3	<i>Target Node Analysis</i>	99
2.3.4	<i>Validation of Gene Regulatory Networks</i>	101
2.3.5	<i>Incorporating prior knowledge</i>	106
2.4	DISCUSSIONS	114
2.4.1	<i>Issues and limitations</i>	116
2.4.1.1	Gene Regulatory Networks	118
2.4.2	<i>Yeast Cell Cycle Phases</i>	119
2.5	CONCLUSIONS	120
3	GENETIC NETWORK OF HUMAN ADIPOCYTE DIFFERENTIATION	123

3.1	INTRODUCTION	123
3.2	HUMAN ADIPOCYTE DIFFERENTIATION	124
3.3	REDUCTION OF DATA DIMENSIONALITY	127
3.3.1	<i>Gene Selection Methods</i>	130
3.3.2	<i>Dimension Reduction Methods</i>	133
3.3.3	<i>Clustering</i>	134
3.3.3.1	k-Means Clustering	135
3.3.3.2	Self-Organizing Map (SOM)	136
3.3.3.3	Gaussian clustering Algorithm	138
3.4	MATERIALS AND METHODS	141
3.4.1	<i>Data Structure</i>	141
3.4.2	<i>Data Pre-processing</i>	142
3.4.2.1	Artificial Experiments	142
3.4.2.2	Selection of Pre-processing Method	143
3.4.2.3	Gene Selection Methods	145
3.4.2.4	Clustering Genes	147
3.4.3	<i>Discretisation</i>	150
3.4.4	<i>Extraction of GRNs</i>	150
3.5	RESULTS AND DISCUSSIONS	150
3.5.1	<i>Gene Selection</i>	150
3.5.2	<i>Clustering Analysis</i>	154
3.5.3	<i>Extraction of Gene Regulatory Networks</i>	161
3.5.3.1	Extraction and Validation of Selected Genes of Interest	161
3.5.3.2	Analysis of Clusters	166
3.5.3.3	Analysis of Clusters with Outliers	170
3.5.3.4	GRN Extraction of Specific Clusters	173
3.6	CONCLUSIONS	176
4	GRN EXTRACTION OF METABOLIC STRESS OF <i>E. COLI</i>	180
4.1	INTRODUCTION	180
4.2	RECOMBINANT PROTEIN PRODUCTION	181
4.2.1	<i>Genome-wide GRN Extraction of <i>E. coli</i></i>	182
4.2.2	<i>The Role of Microarray Technology</i>	182
4.3	MATERIALS AND METHODS	183
4.3.1	<i>Data Acquisition and Structure</i>	183
4.3.2	<i>Data Pre-processing</i>	184
4.3.2.1	Missing Values and Data Noise	184
4.3.2.2	Filtering	185
4.3.2.3	Discretisation	186
4.3.2.4	Extraction of GRNs	189
4.4	RESULTS AND DISCUSSIONS	190
4.4.1	<i>Data Pre-Processing</i>	190
4.4.2	<i>Extracting Gene Regulatory Networks</i>	191
4.4.2.1	GRN Extraction for the CFP Protein Experimental Data	191
4.4.2.2	GRN Extraction for the SOD Protein Experimental Data	196
4.4.2.3	Relationship Analysis and Validation of extracted GRNs	199
4.5	CONCLUSIONS	203
5	CONCLUSIONS	205

5.1 SUMMARY.....	205
5.2 EXTRACTION OF GENE REGULATORY NETWORKS	207
5.2.1 Yeast Cell Cycle Regulation	207
5.2.2 Human Adipocyte Differentiation.....	209
5.2.3 Metabolic Stress System of <i>E. Coli</i>	212
5.3 CURRENT ISSUES LIMITING THE PERFORMANCE OF GRN EXTRACTION	213
5.3.1 Algorithm to extract Bayesian network.....	214
<i>The current algorithms available such as Taboo order search are relatively slow in terms of processing time when extracting larger networks. However, improving the efficiency of the search algorithm, whilst maintaining its ability to find the best solution, is a considerable challenge.</i>	214
5.3.2 Pre-Processing Issues	214
5.3.3 Incorporating prior knowledge from heterogeneous sources	215
5.3.4 Handling Continuous Data and Non-linear Dependencies.....	216
5.3.5 Dynamic Models.....	217
5.4 FUTURE WORK	219
REFERENCES	220
APPENDICES.....	230
APPENDIX I –	230
<i>Discussions of Yeast Cell Cycle Phases</i>	230
APPENDIX II	240
APPENDIX III	244
APPENDIX V	253
APPENDIX VI.....	262
Support Vector Machines (SVM).....	263
Modelling with differential equation.....	264
Fuzzy logic algorithm.....	264
<i>The fuzzified values get transformed into numerical value. Hence the predicted expression value for “C” is calculated for each time point. Predicted vales of target “C” are then compared with observed measurements. A residual, r^2, calculated between calculated and observed “C”, reflecting the goodness of strength between the expected and the predicted “C”.</i>	265
Smooth Response Surface Algorithm (SRS)	266
<i>The SRS algorithm (Xu, Wu et al., 2002) is aimed to improve the fuzzy logic approach. Response surface the methodology uses the input factors and response model. The underlying mechanism is widely employed by manufacturing and high-tech industries.</i>	266
APPENDIX VII.....	266

List of Figures

Figure 1: Prokaryotic and eukaryotic cell anatomy. The top part shows a typical human cell and a typical bacterial cell, drawn to scale. The bacterial DNA contained in the structure called the nucleoid. (All the figures from this chapter were obtained from www.bios.co.uk (Brown, 2002; Strachan and Read, 1999), unless otherwise stated).	4
Figure 2: Chromatin structure. There are two ways in which chromatin structure can influence gene expression. A region of unpacked chromatin in which genes are accessible is flanked by two more compact segments. The positioning of nucleosomes within the unpacked region influences the gene expression. On the left, it shows nucleosomes have regular spacing and on the right, the nucleosome positioning has changed. A short stretch of DNA, approximately 300 bp, is exposed.....	5
Figure 3: A) The central role of the proteome. B) The genome, transcriptome and proteome. Genome and gene expression is shown in orange colour, transcriptome in green and the proteome in blue. A detailed version of this figure is shown in Figure 7 below, with the same colour code.	6
Figure 4: RNA transcription as a single strand. It is complementary in base sequence to the template strand of a gene. Various transcription factors are required to bind to a promoter of the gene, in order to subsequently position and guide the RNA polymerase that will transcribe the gene.	7
Figure 5: Gene expression in an animal cell. A small proportion of nuclear RNA molecules can be converted naturally to cDNA by virally encoded and cellular reverse transcriptase. Thereafter, integrate into chromosomal DNA at diverse junction.	9
Figure 6: The organisation of the human genome.....	11
Figure 7: Two views of genome expression. A) Summary of simplistic belief of gene expression, ‘DNA makes RNA and RNA makes protein’. B) More accurate event of genome expression of protein coding genes.	15
Figure 8: Stages of gene expression and levels of regulation.....	16
Figure 9: Principle of the cDNA microarray analysis system. Basic procedures of microarray technique; microarray construction, probe preparation, hybridization, and measurements of signal intensities. Reproduced from: (Kelsuke Iida, 2002).	20
Figure 10: Nylon membrane cDNA microarray image. Each block contains 13 different cDNA spotted in duplicate. The first letter A, in ID represents the row, second number 12 the column and the third number locates the grid. (Reproduced from (http://udgenome.ags.udel.edu/chicest/SPL-GridMod1.JPEG)).	22
Figure 11: A) Scanned image of cDNA microarray. B) Hierarchical cluster image map (HCIM). Relating gene expression patterns C) Double clustering, using k-means and hierarchical methods. Adopted from (Bertone and Gerstein, 2001).	30

Figure 12: Example of a Bayesian network model. The nodes <i>A,B,C,D</i> in the above BN represent genes, whose expression levels are modelled as random variables. The direction of the pointer of the arc represents the association/dependency. For instance, the arc from <i>A</i> to <i>C</i> indicates a positive (upregulation) association whereas the arc from <i>B</i> to <i>C</i> shows a negative association (downregulation).	44
Figure 14: Predicted gene regulatory network for yeast cell cycle, phase M/G1. BN analysis using SopLEQ, Taboo Search and Taboo Order algorithms on genes identified by traditional methods as cell cycle regulated in phase M/G1. A) SopLEQ algorithm B) Taboo Search algorithm C) Taboo Order algorithm.	81
Figure 15: Marginal probability distributions of phase M/G1 BN. Each box represents the marginal probability distribution for each gene.	82
Figure 16: Use of a Bayesian network as a GRN simulation of M/G1 phase. The strength of the association between genes in the network structure is analysed by monitoring the impact of observing a variable (SIC1) on the marginal probabilities of other genes. The grey markers show the new probabilities.....	84
Figure 17: Unsupervised learning - predicted gene regulatory network for yeast cell cycle, phase G1. BN analysis using Taboo Order algorithm on genes identified by traditional methods as cell cycle regulated in phase G1.....	85
Figure 18: Supervised learning: predicted gene regulatory network for yeast cell cycle, phase G1. A) Naïve Bayes. B) Sons & Spouses. C) Markov Blanket. D) Augmented Markov Blanket.	86
Figure 19: Predicted gene regulatory network for yeast cell cycle, phase lateG1. BN analysis using Taboo Order algorithm on genes identified by traditional methods as cell cycle regulated in phase late G1.	88
Figure 20: Predicted gene regulatory network for yeast cell cycle, phase S. BN analysis using Taboo Order algorithm on genes identified by traditional methods as cell cycle regulated in phase S.	89
Figure 21: Predicted gene regulatory network for yeast cell cycle, phase S/G2. BN analysis using Taboo Order algorithm on genes identified by traditional methods as cell cycle regulated in phase S/G2.	90
Figure 22: Predicted gene regulatory network for yeast cell cycle, phase G2/M. BN analysis using Taboo Order algorithm on genes identified by traditional methods as cell cycle regulated in phase G2/M.....	91
Figure 23: Monitoring the extracted BN for changes in yeast cell cycle phase threshold. BN analysis for the combined phases of M/G1 and G1 using Taboo Order algorithm.....	92
Figure 25: Unsupervised learning, extracting information contribution of the arcs and the nodes to the overall BN of 161 genes. MCD1 is set as the target node (TN) represented by the wavy circular node. The shade of the squares in the centre of the node represents the amount of information brought to the target node. The lighter the shade of the square the more the information contributed to the (TN). The thickness of the arc represents the information contributed to the global structure. The strength of the arc is proportional to the strength of the probabilistic relation.	96
Figure 26: Extracting the information contribution of the nodes in cell cycle phases M/G1 and G1 to target node SIC1. SIC1 is set as the target node (TN) represented by the wavy circular node. The shade of the squares in the centre	

- of the node represents the amount of information brought to the target node. The lighter the shade of the square the more the information contributed to the (TN). The thickness of the arc represents the information contributed to the global structure. The strength of the arc is proportional to the strength of the probabilistic relation. 100
- Figure 27: Cell cycle control and DNA replication of S phase in *S. Cerevisae*. MIPS cYGD interactive physiological and genetic pathway of yeast. The black outer circle represents the phases of the cell cycle. The inner circle represents changes in the proteins bound to a typical replication origin during the cell cycle. Red colour indicates the origin is bound by a "pre-replicative complex" (pre-RC), consisting of ORC plus additional proteins. Yellow indicates the post-replicative state, when only ORC binding is detectable. 104
- Figure 28: Comparison of KEGG's Cell cycle genetic regulatory pathway of *S. Cerevisae* and genes identified by traditional methods. The pink colour boxes represents the gene identified traditional methods that is also present in KEGG cell cycle pathway database. 105
- Figure 29: GRN of common genes found in KEGG and yeast cell cycle project. Unsupervised learning of 31 genes that are common in both cell cycle pathway of KEGG and 104 genes traditionally identified as cell cycle regulated using Taboo Search algorithm. The presence of an arc between two genes represent that there is a direct probabilistic relation between them. The relation is identified as positive (solid arc) or negative (dotted arc) i.e. induce or suppress respectively. The Absence of an arc means that gene is independent of all other genes in the network. The cell cycle phases are colour coded according to findings from the traditional methods (<http://genome-www.stanford.edu/cellcycle/data/rawdata/KnownGenes.doc>). 107
- Figure 30: Learning of GRN of common genes of yeast cell cycle project and KEGG after adding biological information. Genes are colour coded according to the cell cycle phases as distinguished by traditional laboratory experiments and it is in yeast cell cycle project site. The lack of an arc means that gene is independent of all other genes in the network. The *direct* probabilistic relation is shown by the presence of an arc. Unless the relation is noted by dotted arc meaning the gene is suppressed i.e. negative and otherwise the relation is between the genes are induced in the direction of the arc i.e. positive. When two genes are associated only through a grandparent gene (no direct arcs), then they are regarded as *indirectly* associated. 108
- Figure 31: Adipogenesis. The development of mature adipocytes from a fertilized egg. Multipotent mesenchymal stem cells differentiate into cartilage, bone, muscle and adipose tissue. The Dark ovals and circles represent nuclei and pronuclei, respectively. Reproduced from (Avram, Avram *et al.*, 2007). 125
- Figure 32: Identifying differentially expressed genes using a t-test. Under this test the log-ratios of gene expression is assumed to follow a normal distribution with mean μ and standard deviation σ The genes whose log ratios fall under the shaded area ($\mu \pm z\sigma$, where z is the number of standard deviations) are classified as differentially expressed. 131
- Figure 33: Calculating the variable fold-change. The change in intensity and the fold-change necessary to provide 95% and 99% confidence, that intensity values are indeed significantly different. The base line intensity, calculated as

the Average Difference Value (ADV) of the signal intensities, was computed using Affymetrix MAS 4.0 software. Reproduced from (Mariani, Budhraj <i>et al.</i> , 2003).....	132
Figure 34: Self-organizing map. The initial, random mapping of nodes is shown on the left; the nodes after the SOM algorithm completed is shown on the right. Reproduced from: (Newton, 2002).....	137
Figure 35: The Gaussian Clustering in ArrayMiner2. Obtained from http://www.optimaldesign.com . As of 12/10/2006.....	140
Figure 36: Generation of artificial experiments. This shows, for day 2, 7 and 14, colour coded blue, purple and pink respectively, each sample data (denoted by R) was divided by its triplicate reference (denoted by W).....	143
Figure 37: Data pre-processing methods with and without artificial experiments. The method highlighted in red on the right shows the pre-processing method A and on the left the normal method that would have been followed if more experimental data was available.	144
Figure 38: Pre-processing methods, B and C. The parameters for probe filtering set the criteria followed for pre-processing methods B and C.	146
Figure 39: Measure of Fitness of a Gene within a Cluster. A) The closer the distance from the center of the own cluster, the better the fitness. B). The further the distance away from the center of closest cluster, the better the fitness.	149
Figure 40: Fold Change Filtration. Number of differentially expressed probes with at least one experiment value outside for the fold change class widths.....	151
Figure 41: Volcano plot of directional fold change and its significance p-value.	152
Figure 42: Directional fold change. The figures A and B show the up-regulated probes by 2 fold: 2729 probes (4.99%) and the down-regulated by 2 fold: 2266 probes (4.15%) respectively.....	152
Figure 43: Significant change in gene expression. A cut off point of 0.01 p value generated 1976 probes with significance change (3.62%) for up-regulated probes and 1540 probes with significance change (2.82%) for down-regulated probes as shown in A and B respectively.....	153
Figure 44: Cluster optimisation: SOM 10x10. Each grid shows if cluster is present or not. The empty grids means no cluster was found. Each cluster is shown in different colour with the profile it follows. In this instance, the grid shows the presence of 42 clusters. The 10 th column of the grid was omitted as it did not contain any clusters, and hence is not shown above.....	155
Figure 45: k-Means Clustering. This visualization shows the profiles in each of 60 calculated clusters. Number of iterations: 29, Total score: 6.718e+007.	156
Figure 46: Gaussian Clustering. The grids on top panel show the cluster ID, number of probes within that cluster and the overall profile of the cluster. On the left panel the heat map of individual clusters are shown.	157
Figure 47: PCA view of Gaussian clusters. Each clusters are shown in different colours with hulls. The red crosses in the middle represent the outliers.....	158
Figure 48: Average cluster profile. Profile of all 45 clusters over the 27 data points. Each colour represents different clusters. The vertical axis shows the average expression level in logarithms, and the horizontal axis (at the centre) shows the experiment time points 1-27.	159
Figure 49: Gene expression profiles of multiple probes from LPL and FABP4.....	160

- Figure 50: GRN of genes of interest, colour coded according to the biological family they belong to. The respective Affymetrics probe ID annotation from TIGR database indicated by the sign “i”, is only available on the modelling view in BayesiaLab. 162
- Figure 51: Unsupervised learning, extracting information contribution of the arcs and the nodes to the overall BN. Target node (TN) represented by the wavy circular node. The shade of the squares in the centre of the node represents the amount of information brought to the target node. The lighter the shade of the square the more the information contributed to the (TN). 167
- Figure 52: Natural partitioning of GRN based on target node Cluster_1. The nodes in the GRN are grouped into sets of colour coded clusters based on their closeness as measured by the Kullback-Leibler divergence. 168
- Figure 53: Arc analysis of network of all 45 clusters. TN represented by the wavy circular node. The thickness of the arc represents the information contributed to the global structure. The strength and the length of the arc are proportional to the strength of the probabilistic relation and how close the node is to the TN respectively. 169
- Figure 54: GRN of all 45 clusters and outliers. GRN of all 45 clusters and outliers with natural partition. Cluster_1 is set as the TN as indicated by the wavy node. Independent probes and clusters were removed. 172
- Figure 55: GRN of all 45 clusters and 25 clusters of Outliers. The 25 clusters of outliers are named as Clust_out_1, Clust_out_2, etc., in order to differentiate them from the other 45 clusters. 173
- Figure 56: GRN of Cluster 37. Arc analysis of Cluster_37 showing the strength of the relationships between the probes; blue arc represents a positive relationship. 174
- Figure 57: GRN of Cluster 25. Arc analysis of Cluster_25 showing the strength of the relationships between the probes; blue arc represents a positive relationship, while a red arc represents a negative relationship. 175
- Figure 58: Discretisation of gene expression of GFP and SOD proteins under different thresholds. The number of genes classified as ‘Normal’ under discretisation thresholds (log) 0.1, 0.2, 0.4, and 0.5 for *E. coli* proteins GFP and SOD. 188
- Figure 59: Application of Fold Change Method for Gene Filtration of CFP. Vertical axis measures the fold change in logarithms; a 2-fold change is equivalent to $\log_2 0.5$. Horizontal axis shows the sample time points for the CFP experiment: taken 12, 32, 90, 150 and 180 minutes after induction. The red lines show the genes that are differentially regulated and the yellow lines at the centre shows the genes that are filtered out. 190
- Figure 60: Extraction of Genome-wide GRN for the GFP Experiment. The nodes represent the genes and the arcs represent the relationship between the genes. Genes that are conditionally independent of the GRN are shown at the bottom of the figure. Not all independent genes are shown, and gene names shown in small letters due to space restrictions. 192
- Figure 61: Inset view of the Circular Cluster of Genes in the GRN for GFP. The genes that are directly associated with the gene pflB; pflB is the ‘father’ node of this cluster. The colour coded sets of genes are generated using the Taboo algorithm, which creates groups of variables that are close in terms of their

mutual information as measured by Kullback-Leibler divergence (Eguchi and Copas, 2006). This technique is known as variable clustering.	193
Figure 62: Arc analysis of the Circular Cluster of Genes in the GRN for GFP. The strength and the length of the arcs are proportional to the strength of the probabilistic relation and closeness of the relationship respectively. Blue arc represents a positive relationship, while a red arc represents a negative relationship.....	195
Figure 63: Extraction of Genome-wide GRN for the SOD Experiment. Genes that are conditionally independent of the GRN are shown at the bottom of the figure. Not all independent genes are shown due to space restrictions.	196
Figure 64: Inset view of the Small Circular Cluster of Genes in the GRN for SOD. The genes that are directly associated with the gene <i>cstA</i> ; <i>cstA</i> is the ‘father’ node of this cluster. Blue arc represents a positive relationship, while a red arc represents a negative relationship.....	197
Figure 65: <i>E. coli</i> Oxidative Phosphorylation Pathway in KEGG. The pink cells represent the genes identified by the extracted GRN for GFP experiment that are also present in the KEGG oxidative phosphorylation pathway database. ...	202
Figure 66: Cell cycle genetic regulatory pathway of <i>S. Cerevisae</i> . Each box represents a gene and rounded rectangle represent an interconnecting pathway.	244
Figure 67: MAPK signalling pathway in yeast. The pink box represents the gene that is identified as cell cycle regulated by traditional methods. Appendix IV	245
Figure 69: Support vector machine: mapping non-separable data from input space to a higher-dimensional feature space, where a separating hyperplane can be constructed. Reproduced from: Newton J., Analysis of Microarray Gene Expression Data Using Machine Learning Techniques. Available on the World Wide Web: http://www.cs.ualberta.ca/~newton/academic/ml.pdf	263
Figure 70: Fuzzy membership as a function of a normalised expression level.....	265
Figure 71: Decision matrix describing an activator (A) and repressor (B) acting on a target	265

List of Tables

Table 1: Summary of examples of genome sequencing projects.....	12
Table 2: Public databases of microarray data. Adapted from (Gardiner-Garden and Littlejohn, 2001; Kohane, Kho <i>et al.</i> , 2003).	26
Table 3: Algorithms/methods employed in the analysis of microarray gene expression data.....	33
Table 4: Example of a gene expression matrix	35
Table 5: Network topology employed in BayesiaLab and in manually extracted GRNs.....	43
Table 6: List of available software packages for Bayesian Networks.....	68
Table 7: Data description. Data sets used for Bayesian learning and validation. Yeast cell cycle mRNA expression. Data subset late G1 is SCB regulated and late G1 is (MCB regulated). The data subset classification was obtained from Spellman <i>et al.</i> (1998).	73
Table 8: Computational specification. The hardware and software used for extraction of GRN from yeast cell cycle data.....	76
Table 9: BayesiaLab processing time for M/G1 phase and all 104 genes. Showing all 3 structure learning algorithms, SopLEQ, Taboo Search and Taboo Order Search employed.	79
Table 10: Node relative significance with respect to the information gain brought by the node to the knowledge of SIC1.	99
Table 11: Comparison of extracted genetic associations with published information.....	109
Table 12: Impact of Discretisation Thresholds. Discretisation of gene expression values using thresholds of $\log_2$ 0.1 and $\log_2$ 0.2. Under threshold of $\log_2$ 0.1, expression level between ≤ 0.1 and ≥ -0.1 is classified as ‘normal’, > 0.1 as ‘up’, and < -0.1 as ‘down’; under threshold of $\log_2$ 0.2, expression level between ≤ 0.2 and ≥ -0.2 is classified as ‘normal’, > 0.2 as ‘up’, and < -0.2 as ‘down’.	187
Table 13: Relationship Analysis Report for <i>cstA</i> cluster from the GRN for SOD experiments. The nature and strength of the relationships between genes from relationship analysis report. The higher the value of Kullback-Leibler (K-L) divergence and Pearson correlation coefficient (PCC), the stronger is the association.	198
Table 14: Relationship analysis report for GFP and SOD GRNs. The nature and strength of the relationships between genes; genes considered include operons <i>atp</i> , <i>nuo</i> and <i>gat</i> family and heat shock response genes and stress induced genes.....	199
Table 15: Gene regulatory reference databases.....	240
Table 16: Cross-reference databases of biological data. Adapted from (Gardiner-Garden and Littlejohn, 2001).	242
Table 17: Gene classification. All 104 traditionally determined to be yeast cell cycle regulated genes names, ORFs and their biological processes. (based on data from SGD)	246

Table 18: The pathways and biological functions for the E. coli K12 genes obtained from the KEGG 253

Abbreviations

Abbreviation	Expansion
AC	Accession number
ASO	Allele-Specific Oligonucleotide
API	Application Programming Interface
BIC	Bayesian Information Criterion
BN	Bayesian Network
BNJ	Bayesian Network tools in Java
BNT	Bayes Net Toolbox
CADLIVE	Computer-Aided Design of LIVing systEms
CDs	Coding sequence
CDK	Cyclin-Dependent Kinase
CLICK	CLuster Identification via Connectivity Kernels
CLIFF	Clustering Iterative Feature Filtering
CPU	Central Processing Unit
CPDs	Conditional Probability Distributions
CPTs	Conditional Probability Tables
CYGD	Comprehensive Yeast Genome Database
DB	Database
dBEST	Database of Expressed Sequence Tags
DDBj	DNA Data Bank of Japan
EU	Elutriation
DNA	Deoxyribo Nucleic Acid

DoTS	Database of Transcribed Sequences
DS	Double Stranded
E	E-value
EBI	European Bioinformatics Institute
EcoCyc	Encyclopedia of <i>Escherichia coli</i> K-12 Genes and Metabolism
EST	Expressed Sequence Tags
FTP	File Transfer Protocol
GATO	Graph Animation Toolbox
GN	Genetic Network
GRN	Gene Regulatory Network
GUS	Genomics Unified Schema
HCS	High Connectivity Subgraph
HMM	Hidden Markov Model
IMM	Interpolated Markov Model
HTML	Hyper Text Mark up Language
KEGG	Kyoto Encyclopaedia of Genes and Genomes
MATLab	Matrix Laboratory
MCM	Mini Chromosome Maintenance proteins
MCB	MluI Cell cycle Box
MGD	Mouse Genome Database
MIPS	Munich Information Centre for Protein Sequences
mRNA	messenger RNA
MV	Missing Value

NCBI	National Center for Biotechnology Information
ORC	Origin Recognition Complex
ORF	Open Reading Frame
Pre-RC	Pre-Replicative Complexes
RNA	Ribonucleic Acid
rRNA	ribosomal RNA
tRNA	transfer RNA
RTF	Rich Text Format
SCB	Swi4, 6 dependent Cell cycle Box
SGD	Saccharomyces Genome Database
SNP	Single Nucleotide Polymorphism
SS	Single Stranded
TN	Target Node
TIGR	The Institute for Genomic Research
URLs	Uniform Resource Locators
UTR	UnTranslated Regions
WWW	World Wide Web
YPD	Yeast Proteome Database

The best model of a cat is another cat or, better, the cat itself

N.Wiener

1 Introduction

In the context of this doctoral thesis, it is important to understand not only the biological structure and mechanisms of the genome, but also the tools and technology available for data capture and analysis, so that the complex relationships between the genes embodied in such networks can be inferred, interpreted, and more pertinently, better understood.

The first section of this chapter describes the general structure and workings of the gene regulatory network, while the second section discusses the various gene expression mechanisms, an understanding of which is essential for modelling the gene regulatory networks. Sections 3 and 4 explore the ways in which gene expression data can be captured, in particular, the development of microarray technology. Section 5 considers the range of methods and algorithms available for the analysis of gene expression data, their appropriateness and limitations, and the arguments for why the Bayesian probabilistic approach is well suited to modelling gene expression data. The final section is devoted to the Bayesian approach: it's theoretical underpinning and model structure, learning scenarios and issues, alternative learning strategies, algorithms, tools & techniques, and the software available, and finally its applications, in particular for the extraction of gene regulatory networks.

The chapters that follow present the application of the network structure learning based on Bayesian inference to modelling gene regulatory networks for three distinct datasets from yeast, human and *E. coli* genomes respectively.

Chapter 2 exhibits, how the Bayesian methods can be employed in temporal microarray data from the yeast cell cycle regulation to extract biologically plausible relationships between the genes that are supported by evidence from the published literature and the genome databases.

It also demonstrates two novel knowledge extraction techniques, target node analysis and network simulation, which show how the information contained within the extracted network can be used to simulate the interaction between the genes. The final section of the chapter shows how the addition of prior knowledge to the learnt network can improve the network structure.

Chapter 3 highlights and addresses the issues faced in handling data from large scale microarray experiments, the so called '*curse of dimensionality*'. It demonstrates the application of various dimensionality reduction techniques to a novel dataset containing 54,634 gene probes isolated from human adipose tissue, but with only 3 sample time points, to yield a data matrix that is reduced in size but statistically sound and one that still achieves good predictive performance.

Chapter 4 presents the extraction of genome-wide genetic networks for the metabolic stress system of the recombinant protein production using *E. coli* as a host/vector system. It further demonstrates how multi-discretisation method can be applied to tackle the noise present within microarray data.

Chapter 5 summaries the findings of this doctoral thesis, and highlights the areas where the future work could be undertaken to improve the modelling of the GRN.

1.1 Gene Regulatory Networks (GRNs)

1.1.1 Genomics and Gene Expression

The genome is the complete genomic sequence of an organism, or more specifically the catalogue of genetic material in a living organism. For most organisms, the genome is made of DNA, and for some viruses it is made of RNA, for example Human Immunodeficiency Virus (HIV) has only RNA (Tuma, 2002).

Genomics is the study of genomes, and is operationally defined as investigations into the structure, genome map, annotation and function of very large numbers of genes undertaken in a simultaneous fashion (Brown, 2002). Genomics is further divided into two main sub groups; functional genomics and structural genomics.

As the name suggests, functional genomics is the study of identifying the functions of genes and other parts of the genome whereas structural genomics is the molecular study of structure of proteins and genomes in general.

All living organisms are divided into two types: eukaryotes, whose cells contain compartments or organelles, such as ribosome, mitochondria, chloroplast and the nucleus; prokaryotes, whose cell do not possess these organelles and have smaller genomes (see **Figure 1 A**). Most prokaryotes have a more simple cellular structure and biochemical and signalling pathways compared to eukaryotes. Animals, plants, and the yeast are examples of eukaryotes; bacteria and viruses are prokaryotes.

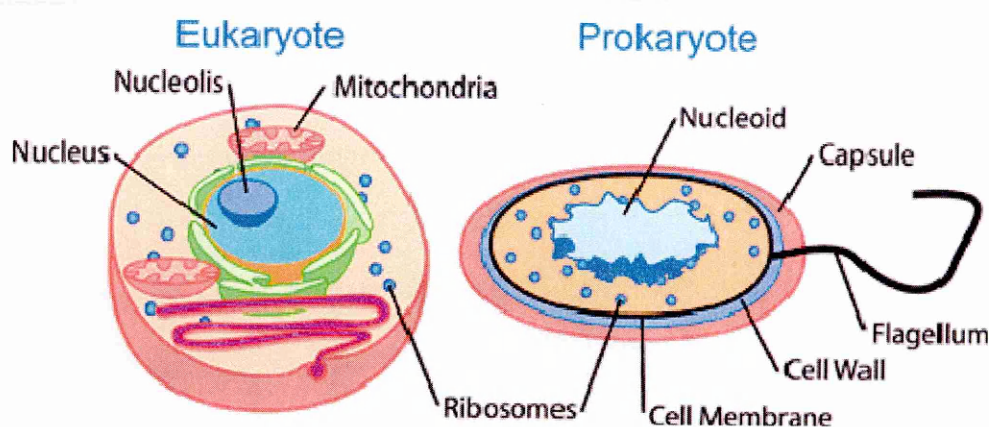


Figure 1: Prokaryotic and eukaryotic cell anatomy. The top part shows a typical human cell and a typical bacterial cell, drawn to scale. The bacterial DNA contained in the structure called the nucleoid. (All the figures from this chapter were obtained from www.bios.co.uk (Brown, 2002; Strachan and Read, 1999), unless otherwise stated).

Selected regions of coding cellular DNA molecules serve as a template for synthesising RNA molecules. The majority of RNA molecules that are synthesised in this way are then involved in the synthesis of specific amino acids that makes the polymer molecules of polypeptides, the only constituent of proteins. These processes of synthesis of basic units of life are known as transcription and translation respectively. Figure 2 collectively the process of flow of genetic information from DNA → RNA → protein polypeptide has been described as the central dogma of molecular biology, (Kohane, Kho *et al.*, 2003; Strachan and Read, 1999).

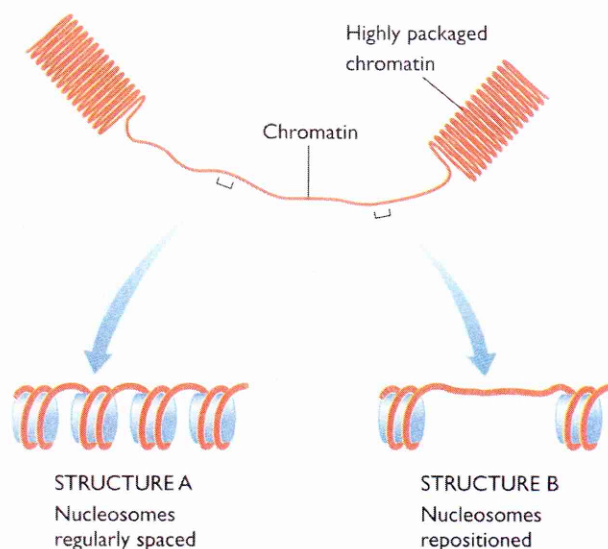


Figure 2: Chromatin structure. There are two ways in which chromatin structure can influence gene expression. A region of unpacked chromatin in which genes are accessible is flanked by two more compact segments. The positioning of nucleosomes within the unpacked region influences the gene expression. On the left, it shows nucleosomes have regular spacing and on the right, the nucleosome positioning has changed. A short stretch of DNA, approximately 300 bp, is exposed.

1.1.2 Protein synthesis

Translation is the formation of protein from mRNA. The mRNA code, which is the result of the translation of DNA, is in the form of codons¹. Each codon is a 3-nucleotide sequence. Since there are 4 nucleotides, there could be $4 \times 4 \times 4$ (64) different codes. But since there are actually only 20 amino acids, there is a lot of redundancy, with many amino acids having multiple codons.

Generally, it is the protein content of a cell that is responsible for the chemical activities in that particular cell, and to a certain extent the external factors such as stimuli from neighbouring cells play a part as well.

¹ Codon is a nucleotide triplet, in mRNA that specifies an amino acid or translation stop signal.

It should be remembered that the protein content of cells differ, it is controlled by gene expression, and the entirety of synthesis of gene expression controlled protein is called the proteome, this is shown in **Figure 3**.

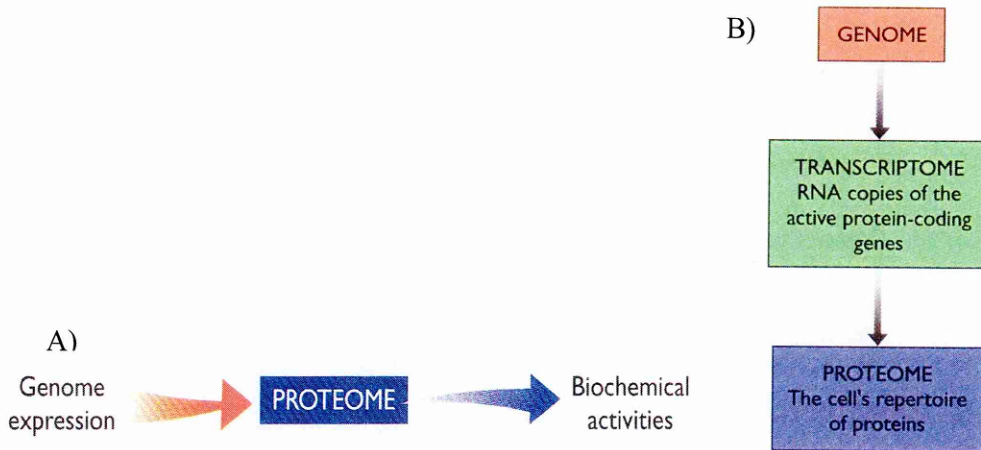


Figure 3: A) The central role of the proteome. B) The genome, transcriptome and proteome. Genome and gene expression is shown in orange colour, transcriptome in green and the proteome in blue. A detailed version of this figure is shown in Figure 7 below, with the same colour code.

The first step in protein synthesis is essentially RNA synthesis; using a DNA-dependant RNA polymerase, is described as transcription and occurs in the nucleus of eukaryotic cells and to a certain extent in organelles with genetic capacity; that is mitochondria and chloroplasts (Strachan and Read, 1999; Brown, 2002). This is illustrated in **Figure 4**.

Translation is a little different in all most every organism. The current model organism is *E. coli*. Translation is well understood in this organism and serves as a model for the other systems.

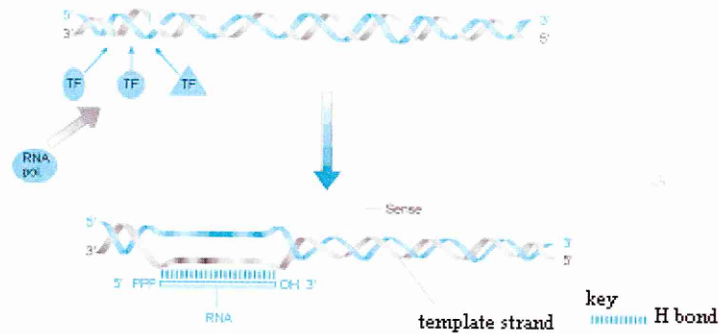


Figure 4: RNA transcription as a single strand. It is complementary in base sequence to the template strand of a gene. Various transcription factors are required to bind to a promoter of the gene, in order to subsequently position and guide the RNA polymerase that will transcribe the gene.

This project aims to model the gene regulatory network described above using mRNA expression data and not the final protein product. Nevertheless, the biological knowledge from sources other than the gene expression data, for example, protein-protein interactions, protein-DNA interactions, and binding site information can also be exploited in learning the GRNs, provided these can be incorporated into the learning process in a structured manner. This is discussed in detail in Chapter 5, section 5.3.3.

1.1.3 Understanding Gene Expression

The term gene expression commonly refers to the entire process whereby the information encoded in a particular gene is translated and transcribed into a particular protein. As mentioned earlier, the actions and properties of each cell type are determined by the proteins it contains (Brazma and Vilo, 2000; Vilo and Kivinen, 2001). The amount and types of proteins are in turn determined by the concentration of their corresponding mRNAs. This mRNA concentration depends on the level of its corresponding gene expression. So studying gene expression helps us gain a better understanding of intracellular mechanisms and a more complete picture of the biological system of interest (DeRisi, Iyer *et al.*, 1997)

Out of the entire DNA of a cell, only a fraction of it gets transcribed in order to synthesise proteins. This means a great majority of the cellular DNA is never transcribed in any cell. Different cells transcribe different segments of the DNA called transcription units, which are discretely spaced irregularly along the DNA sequence.

Moreover, only a part of the RNA synthesised by transcription is ever translated into polypeptide polymer. This is due to three reasons, as explained by Strachan and Read, 1999.

- 1) Transcription units are expressed to synthesise an RNA molecule other than mRNA and so the polypeptides are not directly specified. These RNA molecules are mature molecules that serve as ribosomal RNA (rRNA) or transfer RNA (tRNA) and diverse small nuclear (sn) and cytoplasmic RNA.
- 2) The initial transcription product (primary transcript) of such transcription units, which codes for the polypeptides are subject to RNA processing events. This results in much of the initial RNA sequence being discarded to give a much smaller mRNA.
- 3) Only a central part of the mature mRNA is translated. Hence resulting in further reduction of translated mRNAs.

The gene expression control mechanisms of human are very similar to those found in other mammals, and generally the fundamental gene expression mechanisms are the same in all eukaryotes. Many basic principles apply as in other eukaryotes. A major level at which gene expression is controlled is the initiation of transcription, although, these mechanisms are much more complex than equivalent mechanisms in organisms with smaller genomes (Brown, 2002).

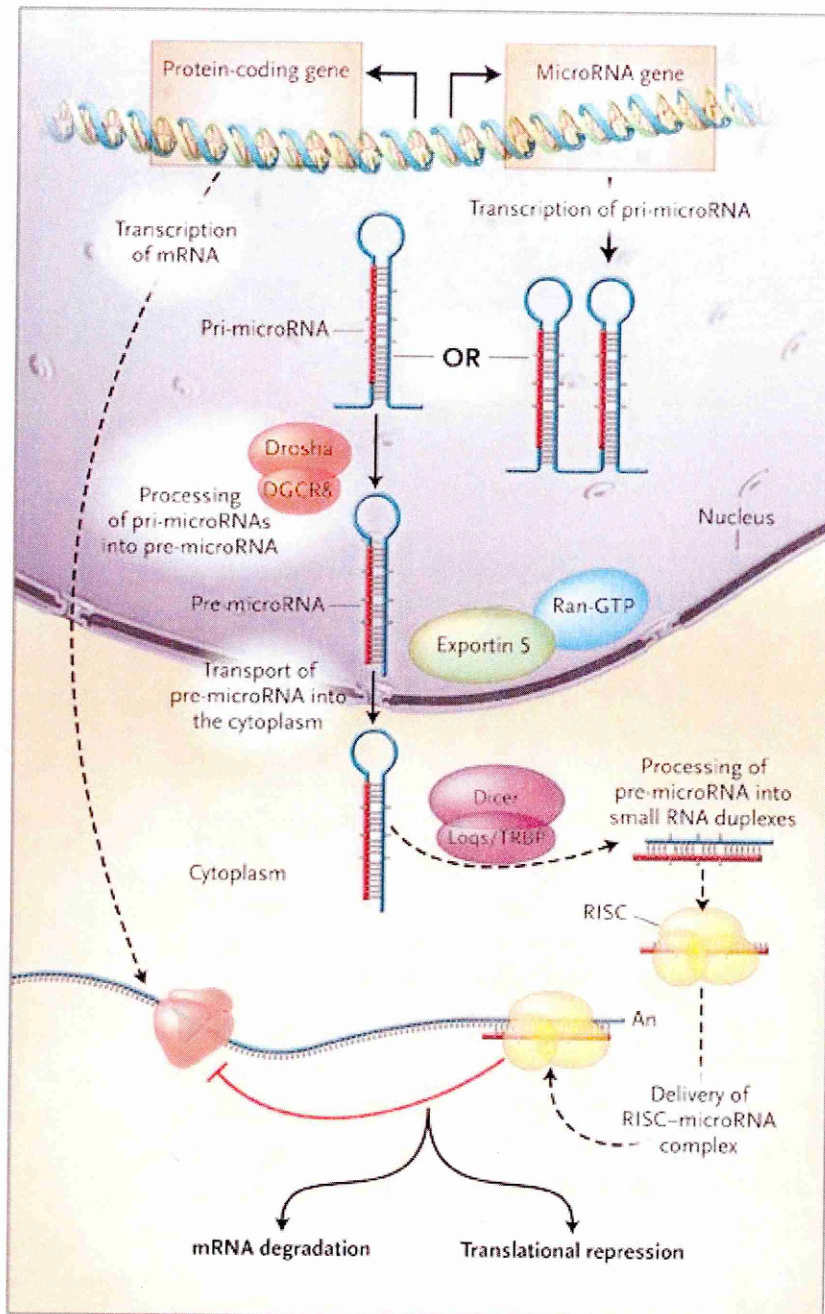


Figure 5: Gene expression in an animal cell. A small proportion of nuclear RNA molecules can be converted naturally to cDNA by virally encoded and cellular reverse transcriptase. Thereafter, integrate into chromosomal DNA at diverse junction.

Various mechanisms are required to maintain all the facets of gene expression, the entire process of gene expression in an animal cell is shown in **Figure 5**. The gene expression controls the production of proteins and hence maintains all the cell's functions and activities. There is a unique cascade of gene expression that belongs to each cell type and yet the level of expression itself is different both at spatial and temporal levels. The gene expression level itself is strictly controlled by many regulatory mechanisms. It is also regulated by other factors such as external stimuli; heat, radiation, *etc.* Normally, the changes in gene expression in response to environmental changes are beneficial to the cell, protecting it from starvation or injury. When this regulation is uncontrolled, i.e. when there are errors in the proteins that react to these environmental changes, it can lead to serious conditions such as cancer and heart diseases (Xu, Wu *et al.*, 2002).

1.1.4 Functional Genomics and the Human Genome

The genomic and proteomic studies of other species provide important insights into gene function in human and other species. The understanding of how genes function and its expression level in normal and disease states is a challenge that functional genomics brought upon to this era. Proteomics and genomics in general involve the study of functional genomics: identification of genes involved in disease and complex traits, novel gene discovery, higher order genome organisation, chromosome structure, functional genetic bases for human diversity, and the structure and functional analysis of proteins.

The capacity to sequence the entire genomes of living organisms, and to analyse such genomes in their entirety, and carry out comparative genomics has considerable implication for understanding human evolution, biology, embryology and medicine. The tools of modern genomics and proteomics are now used to specifically guide the discovery of drugs for the prevention, diagnosis, and treatment of human disease. It is hoped that the knowledge fostered by understanding the genome is hoped to resolve which human characteristics are innate or acquired and how the interplay between

heredity and environment factors contributes to defining susceptibility to illness. Such an understanding will make it possible to study how genomic DNA and gene expression varies between individuals, and especially the role of such variation in the causation of diseases and responses to drugs.

Figure 6 shows the organisation of the entire human genome, published together by Celera Genomics and the publicly funded human genome mapping project; the blue blocks represent the mitochondrial DNA and the grey colour blocks represent the nuclear genome.

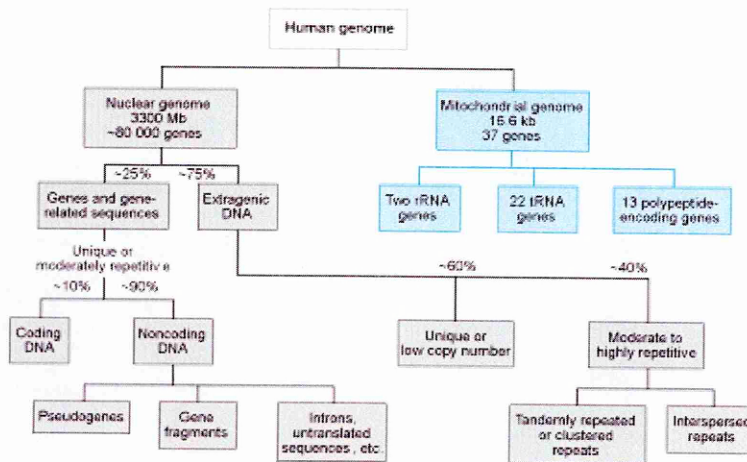


Figure 6: The organisation of the human genome

1.1.5 Post Genomic Data

Genetic information is being uncovered quickly and in vast amounts through the largely automated sequencing of genomes from all kinds of organisms. As this information becomes available, enormous challenges are emerging on three levels:

- First, functions will have to be assigned to individual genes and their products.
- Second, factors that influence the expression level of these gene products will have to be identified.

- Third, allelic variants that act alone or in combination to give rise to complex traits will have to be characterised. Because of the large size of genomes, methods, or to be more specific, bioinformatics programs, that can streamline or automate these processes are highly desirable (Strachan and Read, 1999; Brown, 2002).

Table 1: Summary of examples of genome sequencing projects

Organism	Number of genes	Percent of genes with an inferred function	Completion date for genome sequencing
<i>E. coli</i>	4,288	60	1997
<i>Yeast</i>	6,600	40	1996
<i>C. elegans</i>	19,000	40	1998
<i>Drosophila</i>	12,000-14,000	25	1999
<i>Arabidopsis</i>	25,000	40	2000
Mouse	60,000-100,000	10	2002
Human	25,000-30,000	10	2001

Obtained from <http://genomics.ucdavis.edu/what.html> and updated.

Now that more and more genomes are being sequenced, (see **Table 1**), the wealth of data generated is enormous, not to mention the results of microarray technology (discussed in section 1.3.1) that are used to measure the gene expression. Thus, analysing and interpreting such vast quantities of data when modelling gene regulatory networks presents a formidable challenge. Understanding gene expression mechanisms is, subsequently, of paramount importance in tackling this challenge.

1.2 Gene Expression Mechanisms

1.2.1 Gene Expression in response to external stimuli

Translational control of gene expression can allow more rapid response to alteration of external stimuli. Iron metabolism is a good example that best describes this response system. Often proteins are modified after their synthesis. Therefore, a cell-wide analysis of the population of mature proteins is expected to correlate much more closely with cellular processes than the expressed genes that are measured today (Strachan and Read, 1999; Brown, 2002).

1.2.2 Tissue specific gene expression

The expression of non-housekeeping genes² that may be largely restricted to a specific cell type is known as tissue specific gene expression. Tissue specific gene expression involves selective activation of specific genes and transcriptionally active regions of chromatin. The DNA that is present in any cell of a specific tissue type is virtually identical, e.g. cheek cells, hepatocytes, monocytes or any other nucleated cell from the same organism. What makes the different cell types different is gene expression i.e. they are tissue specific, in that the morphology, growth, differentiation, apoptosis (natural cell death) and function of different cell types vary considerably.

Clearly, the genes that are expressed are the ones fully responsible for the functions of the cell. Some of these functions are common in most or all cell types, which are essential for general cell functions, e.g. energy production etc, and are specified by group of genes called the housekeeping genes. Only certain genes in each cell type are expressed - these cell specific genes can be identified by microarray, see section 1.3.3.

² While housekeeping genes are constitutively expressed, Non-housekeeping genes are expressed in a particular type of cell, under specific environmental conditions, or at a certain stage of development.

Three broad levels at which gene regulation can operate in eukaryotes (Strachan and Read, 1999; Brown, 2002) and there are many gene expression control points in genome expression as shown in **Figure 7** with an asterisk. The mRNA expression level is used to quantify gene expression in microarrays such as the one that is used in Spellman *et al.* (1998), it should be noted that mRNA expression level could only be an estimate of gene expression, as there many more control points after translation.

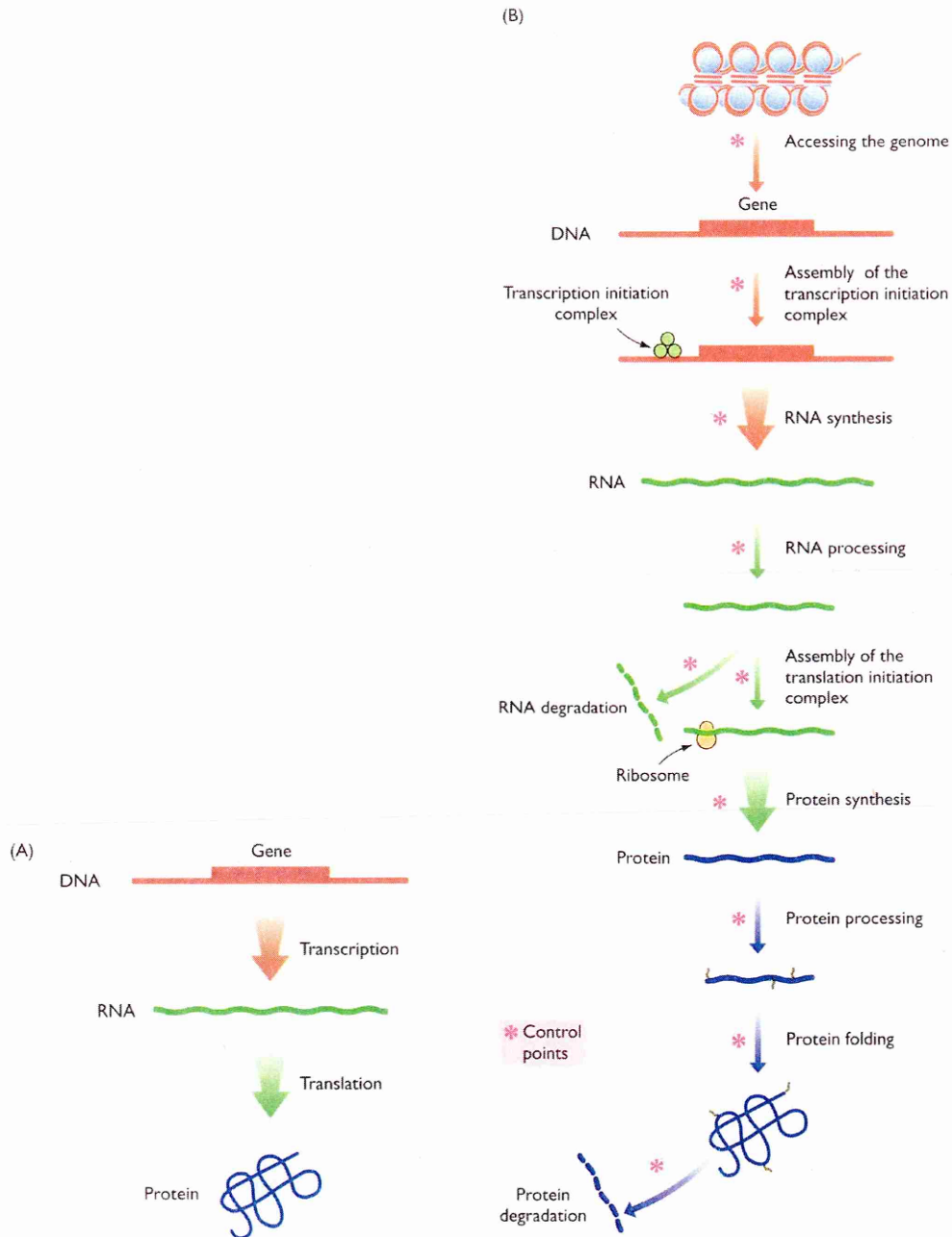


Figure 7: Two views of genome expression. A) Summary of simplistic belief of gene expression, 'DNA makes RNA and RNA makes protein'. B) More accurate event of genome expression of protein coding genes.

1.2.3 Transcriptional regulation of gene expression

The primary control of gene regulation in eukaryotes occurs at the level of initiation of transcription is a well known fact. However, in eukaryotes, the regulation of gene expression can also occur through the core promoter of a gene and the expression level depends on the availability and activity of relevant RNA polymerase. Expression of genes is initiated by the binding of transcription factors to the promoter. The initiation of transcription can be modulated by the proteins factors binding to other regulatory regions. These regulatory regions particularly occur in sequences flanking the gene or sometimes within the introns³ of the gene (see Figure 8).

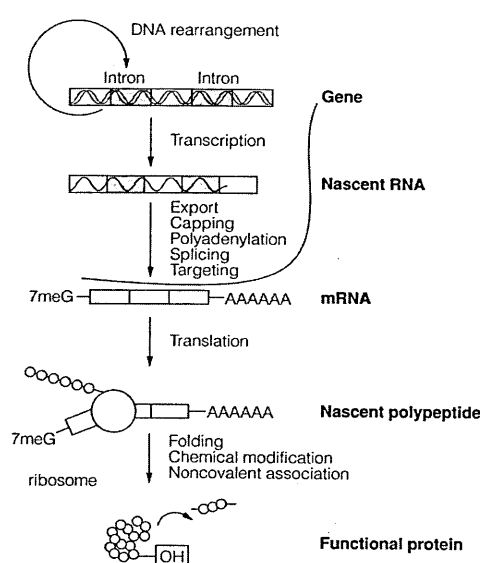


Figure 8: Stages of gene expression and levels of regulation.

³ A non-coding gene within a discontinuous gene

1.2.4 Post-transcriptional regulation of gene expression

The mechanisms involved in transcriptional regulation and post-transcriptional gene regulation slightly overlap, in that, RNA splicing which may more accurately be considered as co-transcriptional rather than post transcriptional. The control of gene expression is also exerted at various levels, including mRNA transport, translation, mRNA stability, protein processing, protein targeting, protein stability, etc. Several mechanisms are involved in regulating gene expression at the level of translation and it has been reported that an increasing number of regulatory units have been identified in the 5' and 3' of untranslated regions of mRNA. In this section, the control of gene expression is extended to the level of protein processing, targeting and stability.

For example, some protein peptides hormones activation requires post-translational cleavage from precursor forms. The entire network system of mechanisms that govern the gene expression is very complex and this system gets even more complicated in multicellular organisms such as higher eukaryotes.

1.2.5 Epigenetic mechanisms and long-range control of gene expression

The term epigenetic means, heritable, from mother cell to daughter cell but not produced by a change in DNA sequence. DNA methylation is an epigenetic mechanism that plays a part in mammalian gene control by maintaining repression of transcription. Many other mechanisms are also known to affect the chromatin environment of a gene and hence affect the gene expression in mammalian cells. In some cases, the mechanisms ensure only one of the two epigenetic alleles is normally expressed, even when the genetic code for the non-expressed allele is the same as the expressed one. It is not yet exactly known why one particular allele is expressed in the above case.

1.2.6 Regulatory Networks

In regulatory networks, signals are integrated through a cascade of gene products until they exert an effect on the biological process; this activity is known as the metabolic process. Metabolic processes can be represented by metabolic maps, which provide a framework for studying the consequences of genotype changes and the relationships between genotypes and phenotypes.

In eukaryotes, unraveling cell networks i.e. how biochemical and signaling pathways are put together, is considerably more complicated compared to a prokaryotic cell network. This is because of the enormity of the genome size, the complexity of intron and exon splicing, and the vast number of perturbations possible (Kurella, Hsiao *et al.*, 2001).

The ultimate goal of a genetic regulatory study is to be able to extract models of transcriptional regulatory networks of living systems that are an exact model of the biological system being studied. Such models should be accurate and should have a standard genetic topology. The data from such studies should be able to be shared within the scientific research community by means of public regulatory databases; given in Appendix II.

One of the emerging problems in bioinformatics is identifying the relationships between various regulatory components of a system, as to what and how one component impacts on the production of another. Increasingly gene networks and especially their graphical representation are being used to represent genetic interactions. These genetic network models have major advantages and several applications. Gene regulatory networks not only give a representation of cell function, state, differentiation and biochemical activity at a system level, they also provide a large-scale, coarse-grained view of the physiological state of an organism at the mRNA level. Thus GRN extracted from microarray investigations are able to give us a concise and precise way of observing gene interaction under the experiments studied and also their dynamic properties of the underlying interactions (Dewey, 2002; Brazhnik, de la Fuente *et al.*, 2002) and provides

an overall view of the system at a genetic level. The knowledge within gene regulatory networks might provide valuable clues and lead to new ideas for treating complex diseases. Such knowledge can aid pharmaceutical research to fulfil what drug and gene therapy now lack, providing possibilities for tailoring target drugs for individual patients to unravelling the cause factor of some serious genetic diseases.

1.3 Methods for Measurement of Gene (mRNA) Expression

As described earlier the mRNA composition of a particular cell represents the gene expression level in that particular cell, see **Figure 7** and **8** The mRNA composition can be measured using nucleic acid hybridization kinetics i.e. hybridization of a complementary DNA (cDNA) copy of mRNA to cDNA generated from same type of tissue. Hybridization based assays include Northern blotting, ribonuclease protection assays, quantitative reverse transcription polymerase chain reaction (RT-PCR). However, the number of genes that can be analysed simultaneously using the above two methods is very limited (Pennington, 2001). Many methods have been developed to identify differentially expressed genes. These include serial analysis of gene expression (SAGE), differential display PCR (DD-PCR) and of course the cDNA microarray technology, which allows gene expression measurement on a genome wide scale (Pennington and Dunn, 2001).

1.3.1 Microarrays

Like the techniques mentioned above, microarray technology is based on the principle that complementary nucleic acids will hybridise. However, unlike these techniques, microarrays allow the simultaneous observation of expression levels for thousands of genes. The underlying principle of cDNA microarray technology is shown in **Figure 9**.

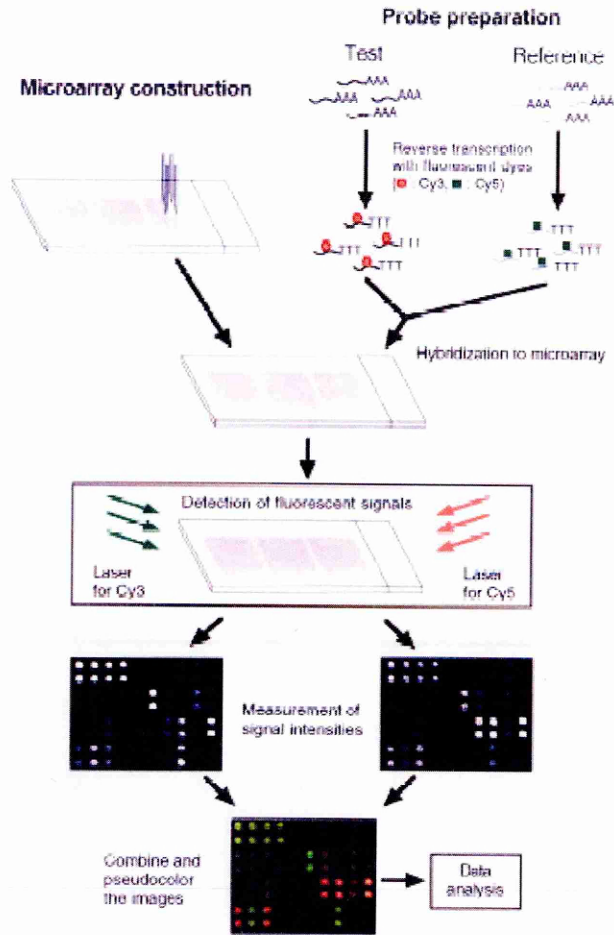


Figure 9: Principle of the cDNA microarray analysis system. Basic procedures of microarray technique; microarray construction, probe preparation, hybridization, and measurements of signal intensities. Reproduced from: (Kelsuke Iida, 2002).

The target cDNAs are cloned, and amplified by PCR. Purified PCR products are printed onto glass microscopic slides with a robotic microarrayer. cDNA probes (test or control) labelled with different fluorescent dyes (Cy3 and Cy5) are synthesised from total RNA or mRNA derived from test and control samples. Pooled probes are hybridized to the microarray. Hybridised fluorescent signals are detected with a dual-wavelength laser scanner. Separately scanned images are combined and pseudo coloured by means of specialised software e.g. GenePix[®], (see **Figure 9**).

Typically, microarrays are constructed by depositing the DNA of the genes of interest (target) in particular locations (spots) printed on a glass slide in grids of rows and columns. mRNAs (probes) are extracted from both the experimental and control samples.

The samples are labelled with fluorescent dyes, Cy3 (Green, Simons *et al.*, 2001) for control and Cy5 (red) for experimental use. They are then allowed to hybridize with the immobilised DNAs on the glass slide. The degree to which the experimental sample and reference sample bind to each of the spots (targets) on the microarray reflects the activity level for the gene located in that spot. The hybridization signals are captured by scanning the microarray slide with a dual wavelength laser and two monochrome images are generated. In the pseudo-coloured merged image of the original images, a gene whose expression is high when compared to the control appears as a red spot; that which is low appears as a green spot and whose activity level remains unchanged appears yellow, (see **Figure 9**).

Microarrays consist of large numbers of DNA, RNA or cDNA strands spotted onto a solid substrate made of glass (microscope slide), nylon or silicon. Depending on the size of the spot, the arrays can be macroarrays (spot diameter $>300\mu\text{m}$) or microarrays (spot diameter $<250\mu\text{m}$) (Schena, 1996). If the substrate itself is small, they are referred to as chips. There may be hundreds or thousands of spots per array. Microarrays can be manufactured in a number of different ways. One technique used for DNA microarray manufacture is called photolithography.

1.3.1.1 Types of Microarray

(a) cDNA Microarrays

cDNA microarrays were first described by Schena in 1996, ;(Schena, 1996; Schena, Shalon *et al.*, 1996), as small devices containing immobilised gene specific cDNA fragments on a solid substrate. The matrix is then probed with fluorescently tagged cDNA, reverse transcribed from total RNA of cells of interest. An example of nylon cDNA microarray is shown in **Figure 10**.

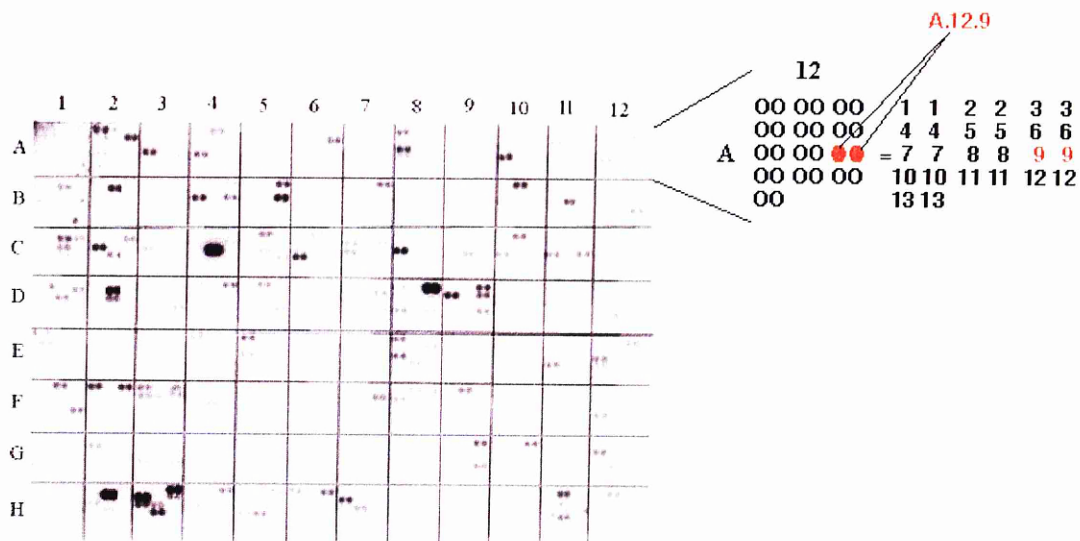


Figure 10: Nylon membrane cDNA microarray image. Each block contains 13 different cDNA spotted in duplicate. The first letter A, in ID represents the row, second number 12 the column and the third number locates the grid. (Reproduced from <http://udgenome.ags.udel.edu/chickest/SPL-GridMod1.JPEG>).

(b) Oligonucleotide Microarrays

Oligonucleotide microarrays, also known as ‘oligo microarrays’, are the second-most popular class of microarrays after cDNA microarrays, also known as oligonucleotide chips or oligo chips. These are currently being developed by a few companies, most notably Affymetrix, who produce a product called a GeneChip. Currently each chip accommodates over 1.5×10^5 oligonucleotides of 25-mers each. The main advantage of oligo microarrays over cDNA array is its higher density of the probe pairs. This allows more genes to be arrayed on a single chip. A disadvantage of this type of array is that it allows only one experiment to be run on each chip at any one time.

1.3.1.2 Microarray Applications

Microarrays are used as capturing tools to find out more about the presence of a DNA or RNA sequence in an unknown sample from an organism. They are incubated with cellular extracts and labelled complementary molecules will hybridise to produce fluorescence. The presence and intensity of this fluorescence is measured to indicate the level of hybridisation.

In genomic studies, DNA extract can be incubated with an oligonucleotide microarray. This approach can be used to capture novel mutations, deletions or polymorphisms in the extract. Also genomic microarrays can be used to detect the presence or absence of certain genes. Knowledge of how genes that causes disease differs from the normal (healthy) genes, or which genes indicate a predisposition to a particular disease can aid in the accurate prediction or diagnosis of diseases.

For gene expression analysis, cDNA microarrays are incubated with an RNA extracts to produce results from what is considered one of the most biologically informative techniques. The RNA hybridises with the cDNA, capturing gene transcripts in the cellular extract. This can be used to compare extracts from healthy and diseased tissues to identify the genes that are likely to cause the disease. This has implications for drug

design and can also be used for medical diagnosis. Also gene expression under certain metabolic or environmental conditions can be examined.

DNA microarray data produces a genomic map, which is quite different to a genetic map in that there is no logical framework on which to build it. Therefore certain approaches must be taken to ensure a functional map. The data must be organised to find orderly features and clustering algorithms, parameterisations or profiling methods, can do this. Then the ordered data can be displayed graphically to represent the huge volumes of quantitative data. This allows biologists to look at the patterns of gene expression on a broad and fine scale. It reflects information about gene function and even physical associations of gene products (Schena, 1996).

1.3.1.3 Data Handling and Interpretation

Transforming the data obtained in a microarray experiment into useful information is a huge challenge. The informatics tools such as software that operates arraying devices and perform image analysis of data from readers are vital. The objective of the microarray image analysis is to extract sample intensities or ratios, at each printed cDNA location in a given microarray scan, and then cross-link printed clone information so that biologists can easily interpret the outcomes and perform further data integration and analysis.

A common analysis method is the cluster analysis method (Brazma and Vilo, 2000), which groups together genes with similar patterns of expression using a mathematical description of similarity. There are two classes of clustering methods, supervised, where vectors are classified with respect to known reference vectors, and unsupervised where no predefined reference vectors are available. Unsupervised is mainly used. Graphical representations of the data, typically in the form of dendrograms, are produced which allows data exploration in a natural intuitive manner.

Once the data has been deciphered, databases are required to hold it. Some of the laboratories that produce large amounts of microarray data have their own web-based databases to share information with other research communities. The integrity of the data would be screened first to ensure that it is accurate and of high quality. Two well known groups have set up such databases (Nature Editorial, 2001). These are the European Bioinformatics Institute (EBI) who have created ArrayExpress and the National Centre for Biotechnology Information (NCBI) who have created GEO (Gene Expression Omnibus). GEO provides a public repository for gene expression data where it is catalogued and stored. ArrayExpress is both an archive for the data as well as a database of gene expression results, which can be searched directly. The features of these databases are discussed in detail in the next section.

1.3.1.4 Microarray Databases

DNA microarray analysis has become the most widely used genome-wide technique in functional genomics, after genome sequencing. This has resulted in a proliferation of microarray databases.

As it can be difficult for researchers to know which best meet their requirements, it is imperative to obtain a better understanding of the databases currently available (Gardiner-Garden and Littlejohn, 2001); (Sherlock, Hernandez-Boussard *et al.*, 2001). For example, the data required for this project was a time series data of genes that are involved in cell cycle regulation of bakers yeast, and this was obtained from the yeast cell cycle project database from Stanford University. The reason for the choice of this dataset is that it is widely used and well documented.

Detailed analysis of the main features of the above microarray databases, for example, types of gene expression data that can be stored (e.g. spotted arrays, Affymetrix, SAGE), analytical packages provided or supported, etc., are given in a review by Gardiner-Garden, (2001) and summarised in Table 2.

Table 2: Public databases of microarray data. Adapted from (Gardiner-Garden and Littlejohn, 2001; Kohane, Kho *et al.*, 2003).

Microarray database	URL
Affymetrix Analysis Data Model (previously Genetic Analysis Technology Consortium)	www.affymetrix.com/support/aadm/aadm.html
Another Microarray Database (AMAD)	www.microarrays.org/software.html
ArrayDB	genome.nhgri.nih.gov/arraydb
ArrayExpress	www.ebi.ac.uk/arrayexpress/Design/design.html
ChipDB	www.young39.wi.mit.edu/chipdb_public/
ExpressDB	twod.med.harvard.edu/ExpressDB
GeneX	www.ncgr.org/research/genex
GEO	www.ncbi.nlm.nih.gov/geo/
GXD	www.informatics.jax.org/
mADb	www.madb.nci.nih.gov/
MicroArray Database	pompos.swmed.edu/exptbio/microarrays/mad
RAD (RNA Abundance Database)	www.cbil.upenn.edu/rad2/servlet
Stanford Microarray Database	www.genome-www4.stanford.edu/MicroArray/SMD
PharmGKB (cancer microarray database)	http://www.pharmgkb.org/index.jsp
YMD (Yale Microarray Database)	info.med.yale.edu/microarray

(a) Cross-reference Databases

Often microarray users wish to relate their microarray results to published non-array information on their genes of interest: for example, information on DNA sequence, function, genetic and physical map locations or homologous genes in other species. Therefore, it is important that the microarray databases support contextual links to external databases containing such information. Examples of such external databases are given in Appendix II. (Gardiner-Garden and Littlejohn, 2001) provides a detailed list on microarray databases that support contextual links to a host of external databases, including those listed in Appendix II.

(b) Integrated Databases

An integrated database is a novel database management system that integrates large-scale studies of gene expression with other molecular biology databases. An example of such database is KEGG/EXPRESSION (Goto, Kawashima *et al.*, 2000). For example, one can map the genes listed as a result of a search on the database to the KEGG pathway database or KEGG genes database.

1.3.1.5 Standards of microarray data

The adoption of common standards and ontologies for the management and sharing of microarray data is essential and it provides immediate benefit to scientists (Stoeckert, Causton *et al.*, 2002). Before a researcher can deposit the microarray experiment results in a microarray database, there are certain guidelines and standards that they must meet. Minimum information about a microarray experiment (MIAME) (Brazma and Vilo, 2001), is a set of guidelines that these databases adhere to.

(a) Minimum Information About a Microarray Experiment (MIAME)

MIAME aims to outline the minimum information required to unambiguously interpret microarray data and to subsequently allow independent verification of this data at a later stage if required. Although details for particular experiments may be different, MIAME aims to define the core that is common to most experiments. MIAME is not a formal specification, but a set of guidelines (Taylor, 2002). This set of guidelines will then assist with the development of microarray repositories and data analysis tools. MIAME is being developed continuously as microarray technology and its applications develop.

A major objective of MIAME is to guide the development of microarray databases and data management software (Brazma and Vilo, 2001). A standard microarray data model and exchange format Microarray Gene Expression (MAGE), which is able to capture information specified by MIAME. This has been submitted by EBI and Rosetta Biosoftware and recently became an Adopted Specification of the OMG standards group. Many organizations, including Agilent, Affymetrix, and Iobion, have contributed ideas to MAGE. Although MIAME concentrates on the content of the information, it also tries to provide a conceptual structure for microarray experiment descriptions.

(b) MAGE-ML

MAGE-ML, Microarray Gene Expression Markup Language (Taylor), developed by the “MAML” working group of MGED, is a first attempt to provide a standard platform for submitting and analysing the enormous amounts of microarray expression data generated by different laboratories around the world.

MAGE-ML proposes a framework for describing information about a DNA-array experiment and a data format for communicating this information, including details about:

- Experimental design: the set of the hybridisation experiments as a whole
- Array design: each array used and each spot on the array
- Samples: samples used, the extract preparation, and labelling
- Hybridisations: procedures and parameters
- Measurements: images, quantification, and specifications
- Controls: types, values, and specifications.

MAGE-ML is independent of the particular experimental platform and provides a framework for describing experiments done on all types of DNA-arrays, including spotted and synthesized arrays, as well as oligo and cDNA arrays. It also provides the opportunity to represent microarray data in a flexible way, which allows analysis of data obtained from not only any existing microarray platforms, but also many of the possible future variants, including protein arrays.

1.4 Capturing Gene Expression Data Using Microarrays

As discussed earlier, microarrays enable monitoring of the expression levels of large number of genes in a highly parallel process. An example of this is the *GeneChip expression analysis system*. Quantitative and comparative data on gene expression are obtained by measuring the abundance of the various mRNAs present in a sample. During analysis, the software calculates the average intensity of a specific area on the probe array that contains one unique oligonucleotide species called the probe cell using the intensities of the pixels contained in the cell.

Pixels on the edges of each cell are not included, as it prevents neighbouring cell data from affecting a cell's calculated average (mean) intensity. The calculated average intensity is assigned a XY-coordinate position, which corresponds to the cell's position on the array. The software then applies the selected probe array algorithm to determine expression levels of each gene using information contained in library files for each probe array (Kohane, Kho *et al.*, 2003; Baldi and Hatfield, 2002).

This microarray based sequence system can identify variations if any in nucleotide sequences such as point mutations and polymorphisms in specific nucleic acid sequences. During analysis the software automatically calls mutations in the target sequences and displays them thereby enabling the user to make and edit composite sequences on combining sense and antisense nucleotide base call files, for example those for sense and antisense target strands.

Nucleotide analysis can then be performed through processing the scan images to calculate all the signal intensities on the probe array. Using the signal intensity data, statistical algorithms in the software will generate genotyping results as well as tabular and graphical reports. Specific software tools have been designed to manage the huge datasets generated with oligonucleotide arrays. It is possible to query, sort, cluster and visualize different experimental parameters such as behaviour, function, chromosomal position, pathway or phenotype.

Moreover, the results can be linked to other information databases e.g. Gene Spring[®], from Agilent Technologies, USA and Spotfire[®], USA.

1.5 Modelling of Genetic Network

1.5.1 Processing Microarray Gene Expression Data

Image analysis (Brazma and Vilo, 2000) is the first and very important step in processing microarray gene expression data as it can have a profound influence on further analysis and interpretation of data. The processing of scanned microarray images can generally be separated into three tasks, (Figure 11).

- **Addressing** or **gridding** is the process of assigning coordinates to each of the spots.
- **Segmentation** allows the classification of pixels either as foreground – that is, within printed DNA spot – or as background.
- **Intensity extraction step** includes calculating, for each spot on the array, red and green foreground fluorescence intensity pairs along with background intensities.

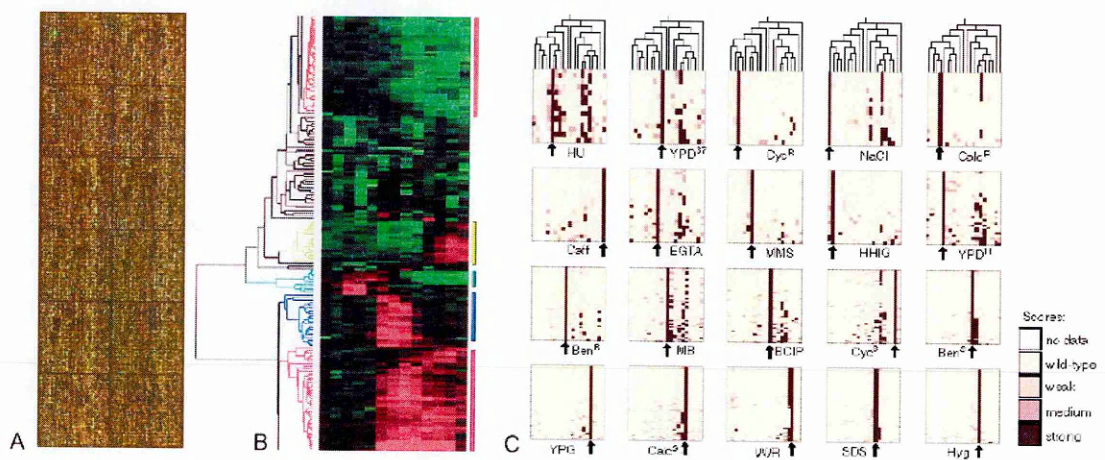


Figure 11: A) Scanned image of cDNA microarray. B) Hierarchical cluster image map (HCIM). Relating gene expression patterns C) Double clustering, using k-means and hierarchical methods. Adopted from (Bertone and Gerstein, 2001).

Image analysis results in numerical values of the red and green fluorescence intensities of individual spots reflecting the level of hybridization or expression of the gene in that particular spot. Knowledge discovery aims at extracting meaningful information from this data. Before any analysis of this data can be performed, it is essential to normalise this data.

1.5.1.1 Normalisation

In a microarray experiment, there are many sources of systematic variation that affect the measured levels of gene expression. Sources of systematic variation include:

- Differences in labelling efficiency between the two dyes
- Differences in the power of the two lasers
- Differing amounts of RNA labelled between the two channels
- Spatial bias in ratios across the surface of the microarray.

Normalisation is the term used to describe the process of accounting for such variation (Lodish, Berk *et al.*, 2000). The commonly used normalisation methods are:

(a) House keeping genes

This method is based on the assumption that some genes have relatively constant expression levels and do not change under the tested condition. A normalisation factor is then calculated so that, this set of genes have a mean ratio of 1 and is applied to all the data. This method is appropriate when measuring large changes in expression levels but unreliable when changes on a relatively small scale need to be measured. Furthermore, all genes show some change under some conditions.

(b) Globalization method

In this method, each gene's expression level is divided by the total expression level for the array. This is based on the assumption that the total amount of mRNA per cell is constant and is widely used. The major weakness of this method is that the sum of all expression levels is influenced by genes having the strongest intensities. This results in widely varying total expression levels.

(c) Total intensity normalisation

This method relies on the assumption that the quantity of initial mRNA is the same for both labelled samples. In a microarray experiment, it is assumed that some genes are upregulated in the experimental sample compared to the control and some are down regulated. These changes should balance out so that the total quantity of RNA hybridizing to the array from each sample is the same. So, the total intensity should be the same in both the red and green channels. A normalisation factor is then calculated and used to rescale the intensity for each gene in the array, (Lodish, Berk *et al.*, 2000). After normalisation, the data for each gene is usually reported as an 'expression ratio' or as a logarithm of the expression ratio. This ratio is computed by dividing the normalised value of the expression level for a particular gene in the experimental sample by the normalised value of that gene in the control sample.

1.5.2 Modelling and Interpretation of Gene Expression Data

Once gene expression data has been processed, there is a whole range of methods that can be applied for inferring meaningful relationships amongst the genes of interest. These are summarised in **Table 3**.

Table 3: Algorithms/methods employed in the analysis of microarray gene expression data

Clustering Analysis	Bayesian Probabilistic Networks	Other Methods
Hierarchical k-means Self Organising Maps (SOM) CAST VizCluster ClusterRasch CLIFF HCS CLICK Gaussian	Maximum Likelihood Estimate - (MLE) Maximum A Priori Estimate - (MAP) Score based approach Naïve Bayes Markov Blanket Expectation-Maximization (E-M) Structural E-M algorithm Sparse Candidate algorithm Dynamic Bayesian Networks - (DBNs)	Differential equations Plaid models Fuzzy Logic approach Smooth response surface (SRS) Golub-Slonim (G-S) Principal component analysis - (PCA) Similarity Pearson Correlation - (SPC) ANOVA (Analysis Of Variance) Partial Least Squares - (PLS) Sliced Inverse Regression (SIR) Support Vector Machines - (SVM)

Some of the methods/algorithms have been around for many years and more widely used than others e.g. *Pearson correlation coefficient*, *Principal Component Analysis (PCA)*. The following section describes the *Pearson correlation coefficient*, which is often employed for correlating the expression profiles of genes (e.g. Spellman *et al*, 1998). This is followed by discussion of clustering techniques, and Bayesian probabilistic approach. The non-Bayesian methods *PCA*, *Partial Least Squares (PLS)*, and *Sliced Inverse Regression (SIR)* are discussed in **Chapter 4** where they are more directly relevant. Other non-Bayesian methods such as *Support Vector Mechanisms (SVM)*, *Differential equations*, *Fuzzy Logic algorithm*, and *Smooth Response Surface Algorithm (SRS)* are described in Appendix VI.

1.5.2.1 Pearson Correlation Coefficient

The *Pearson product-moment correlation coefficient* (PMCC), r , is a measure of the tendency of two variables X and Y , measured on the same object or organism to increase or decrease together (see Appendix VI).

The square of r is conventionally used as a measure of the association between X and Y . For example, if the coefficient is 0.90, then 81% of the variance of Y can be "accounted for" by changes in X and the linear relationship between X and Y .

Spellman *et al.* (1998) applied Pearson correlation measurement to correlate the expression profile of each gene to five different profiles (phases) representing genes known to be expressed in cell cycle phases M/G1, G1, S, G2, and M.

1.5.2.2 Clustering analysis

Cluster analysis is used to group the genes according to the similarity of their expression profiles over a number of experiments. The calculation of a 'distance' between any two objects is fundamental to placing them into groups. The distance calculation is done by using various distance metrics.

To determine similarity (Vilo and Kivinen, 2001), an 'expression vector' for each gene is defined that represents its location in 'expression space'. In a series of microarray experiments, each experiment represents a separate, distinct axis in space and the $\log_2$ ratio of that gene in each experiment represents its geometric co-ordinate. Assuming there are three experiments, the $\log_2$ ratio for a gene in experiment 1 is its x coordinates, the $\log_2$ ratio in experiment 2 is its y coordinate and $\log_2$ ratio in experiment 3 is its z coordinate, **Table 4**. So, all the information about the expression level of a gene can be represented by a point in x - y - z expression space. Expression vectors for all the genes can be determined in this way and a *gene expression matrix* is created.

	Exp.1	Exp.2	Exp.3
Gene 1	x	y	Z
Gene 2	x	y	Z
Gene 3	x	y	Z

Table 4: Example of a gene expression matrix

In this matrix, each column represents a single experiment and each row represents the expression vector for a particular gene. Cluster analysis is then used to group together genes based on their separation or distances in expression space. Genes with a similar expression profile are near each other spatially and those that are different are far away in expression space.

The various clustering techniques such as *k-Means*, *Self Organising Maps (SOM)*, and *Gaussian clustering* are discussed in detail in **Chapter 4**, where the application of such techniques is more directly relevant.

Clustering of expression profiles was successfully used in many cases. In a time course study of diauxic shift in yeast (DeRisi, 1997) identified genes that are functionally related and containing relevant transcription factor binding sites upstream of their open reading frames. (Tavazoie, 1999) clustered expression profiles of 3000 most variable yeast genes during the cell cycle into 30 clusters by using *k-means* clustering.

1.5.2.3 Limitations of Clustering Methods

Cluster analysis methods have some drawbacks despite their success in microarray data analysis: The selection of different clustering algorithms is subjective and this affects the results. Clustering unrelated data also will produce clusters, although this is biologically irrelevant (Lodish, Berk *et al.*, 2000). Even if a gene plays more than one role, it is still placed only in one cluster. It is also not possible to map the interactions between genes and regulatory pathways with clustering methods. Whilst clustering methods reveal similarities that exist over all of the measurements, they obscure *context-specific* relationships, relationships that exist over only a subset of the dataset.

In addition, clustering methods cannot readily incorporate additional information such as experimental conditions, temporal indicators (the time/stage that the sample was taken from), background variables (e.g. clinical procedure employed in obtaining the sample), and exogenous cellular conditions.

1.5.3 *The Case for Bayesian Approach*

Bayesian networks are a probabilistic tool that presents a way to circumvent the shortcomings of the aforementioned techniques, allowing a model to be *learned* from the data, and to discover the patterns in the data and the dependencies between gene expression levels and additional attributes.

DNA microarray data is characterised by a high degree of measurement noise and variability. Biological systems also have very high *dimensionality*: typically, what is measured is only a very small subset of the relevant variables, even in a large or controlled array experiment. Therefore, it is crucial that the model or learning algorithm is capable of handling noisy data and hidden variables.

For example, suppose the underlying structure really is that of a known type, such as a *boolean network*. However, due to the noisy observation of some of the variables, there might not be any boolean⁴ network which can model it. Rather than giving up, one should instead look for the most *probable* model given the dataset; this requires a well-defined probabilistic framework for modelling the data (Murphy, 2001).

BNs can be used to provide a *graphical framework* for Bayesian reasoning, a probabilistic approach to inference based on combining *prior* knowledge with *observed* data using *Bayes' rule*:

$$\text{Posterior belief} = \frac{\text{Conditional Likelihood} * \text{Prior belief}}{\text{Likelihood}}$$

$$P(H | D) = \frac{P(D | H)P(H)}{P(D)}$$

where $P(H)$ is the *prior* probability of *hypothesis* H , $P(D)$ is the probability (*likelihood*) of observing data D , $P(D|H)$ denotes the probability (or *conditional likelihood*) of observing D if *hypothesis* H holds, and $P(H|D)$ is the *posterior* probability of H after observing data D .

Consider the following example⁵. Human males have one X -chromosome and one Y -chromosome, whereas females have two X - chromosomes, each chromosome being inherited from one parent. Haemophilia is a disease that exhibits X - chromosome-linked recessive inheritance, meaning that a male who inherits the gene that causes the disease on the X - chromosome is affected, whereas a female carrying the gene on only one of her two X - chromosomes is not affected. Consider a woman who has a 50-50 chance of having the Haemophilia gene. The unknown quantity of interest, the state of the women, has just two values: the woman is either a carrier ($\theta = 1$) or not ($\theta = 0$). So the prior distribution is $Pr(\theta = 1) = Pr(\theta = 0) = 0.5$.

⁴ particular case of discrete dynamical networks, where time and states are discrete, i.e. they take only integer values.

⁵ Adapted from Gelman *et al*, 2003.

The data used to update this prior information is the affection status of the woman's sons. Suppose she has two sons (not twins), neither of whom is affected. Let $Y_i = 1$ denote affected and $Y_i = 0$ unaffected son. Then the conditional likelihood function on θ is

$$\Pr(Y_1 = 0, Y_2 = 0 \mid \theta = 1) = (0.5)(0.5) = 0.25$$

$$\Pr(Y_1 = 0, Y_2 = 0 \mid \theta = 0) = (1)(1) = 1$$

This follows from the fact if the woman is a carrier, then each of her sons will have a 50% chance of inheriting the gene, whereas if she is not, then there is a probability very close to 1 that of the sons not being affected⁶.

Now Bayes' rule can be used to combine the information in the data with the prior probability to derive the posterior probability that the woman is a carrier:

$$\begin{aligned}\Pr(\theta = 1 \mid Y) &= \frac{\Pr(Y \mid \theta = 1) \Pr(\theta = 1)}{\Pr(Y \mid \theta = 1) \Pr(\theta = 1) + \Pr(Y \mid \theta = 0) \Pr(\theta = 0)} \\ &= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = 0.20.\end{aligned}$$

The fact that the woman has unaffected children has made it less probable that she is a carrier, compared to the prior probability of 0.5.

In a Bayesian analysis, in other words, a set of observations should be seen as something that changes a belief or an opinion, rather than as a means of determining the ultimate truth. This is far more subtle than the *traditional (classical)* way of presenting results, in which an outcome is deemed statistically significant only if there is a better than 95% chance that it could not have occurred by chance.

⁶ There is a very small, but non-zero probability of being affected even if the mother is not a carrier due to the mutation rate.

For example, in the case of a drug trial, it is possible to evaluate and compare the degree to which a sceptic and an enthusiast would be convinced by a particular set of results. Only if the sceptic can be convinced should a drug be licensed for use ((Murphy, 2001).

The general Bayesian statistical framework codifies how to proceed with data analysis and inference in a rational way. *"The essence of the Bayesian approach is to provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence. In other words, it allows scientists to combine new data with their existing knowledge or expertise"* (Murphy, 2001).

Bayesian networks are well suited to analysing gene expression patterns. The expression levels of some genes are regulated by a handful of other genes whose regulation may depend upon many other factors as discussed in section 1.2. If the expression level of those genes is known, then it is possible to determine the expression level of the gene of interest. Bayesian networks help to find this interaction among various genes. Although Bayesian networks are mathematically defined strictly in terms of probabilities and conditional independence statements, a connection can be made between this characterisation, and the notion of *direct causal influence* (Heckerman, Meek *et al.*, 1999).

“Bayesian networks also provide a model for causal influence, allowing us to encode statements such as ‘*gene A being upregulated causes gene B to be upregulated*’ ” (Kerr, Martin *et al.*, 2000).

Thus, the advantages of Bayesian networks can be summarised by their ability to model stochasticity⁷, to incorporate prior knowledge, to infer causal relations, and to handle hidden variables and missing data in a principled way.

The next section describes the main features of the Bayesian probabilistic approach and how this can be applied for modelling gene regulatory networks.

⁷ High degree of measurement noise and/or variability.

1.6 Bayesian Probabilistic Networks

1.6.1 Bayesian Network

Bayesian Networks (BNs), also known as Belief Networks, despite their name, do not necessarily imply a dedication to Bayesian methods. Rather, they are so called because they use Bayes' rule for making inferences. They are a special case of a more general class called graphical models⁸ in which nodes represent random variables, and the arcs represent conditional independence assumptions (see Section 1.6.2).

A Bayesian network is a representation of a joint probability distribution. This representation consists of two components, (G, Θ) . The first component, G , is a *directed acyclic graph* (DAG), whose vertices or nodes correspond to the random variables of interest $X_1, \dots, X_n$, and whose edges represent direct dependencies between variables.

The second component of the pair, Θ , denotes the set of parameters that quantifies the network, the *conditional probability distributions* for each variable X_i , given its parents in G , i.e. it contains a parameter $\theta_{x_i, pa_i} = P(X_i = x_i \mid Pa_i = pa_i)$ for each possible value x_i of X_i , and $pa(x_i)$ of $pa(X_i)$, where $pa(X_i)$, denotes the set of parents of X_i in G . Together, these two components specify a *unique* distribution on $X_1, \dots, X_n$.

The simplest conditional independence assumption encoded in BNs is the *Markov Assumption*: each variable X_i is independent of its ancestors given its parents G . Conditional independence assumptions allow the joint distributions to be decomposed, economizing on the number of parameters.

⁸ Undirected graphical models, also known as *Markov Random Fields*, have a simple definition of independence; Two (sets) of nodes A and B are conditionally independent given all other nodes if they are separated in the graph.

The distributions $P(X_i|Pa_i)$, associated with each node X_i , are called *local probability distributions*. By applying the chain rule of probabilities and properties of conditional independencies, the Bayesian network can be decomposed as a product of local distributions:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i = x_i | Pa_i = pa_i) \quad (1)$$

Under Bayesian approach, each gene's expression level is modelled as a random variable. For n number of genes, the variables are $X_1, \dots, X_n$ and the dependencies between them are calculated.

In specifying the conditional probability distributions (CPDs), one can choose from several representations. Suppose the parents of a variable X are $(Y_1, \dots, Y_n)$. The choice of representation depends on the type of variables under consideration:

- **Discrete variables:** X and $Y_1, \dots, Y_n$ take discrete values from a finite set, then the CPDs are represented by *conditional probability tables (CPTs)*, where $P(X|Y_1, \dots, Y_n)$ specifies the probability of values for X given each joint assignments of $Y_1, \dots, Y_n$.

For example, if all the variables are binary valued (two states) the CPT will specify 2^K distributions. This is a general representation (known as *multinomial models*), which can describe any discrete conditional distribution.

However, the major drawback with the above general representation is that the number of free parameters is exponential in the number of parents. Another potential problem is, when discretizing data for application of the model, there is a risk of losing information. For example, when discretizing the gene expression levels into three categories, *under-expressed(-1)*, *normal(0)*, and *over-expressed(1)*, the control expression thresholds against which the values are compared needs to be determined. In such cases, one must test for the *robustness* of results to various assignments of thresholds.

- **Continuous variables:** When the variables are real valued, there is no representation that can represent all possible densities. A natural choice is the use of *Multivariate Gaussian* distributions. These can be represented in BNs using *linear Gaussian* conditional densities:

$$P(X | Y_1, \dots, Y_n) \sim N(a_0 + \sum_i a_i \cdot u_i, \sigma^2) \quad (2)$$

That is, X is normally distributed around a mean that depends *linearly* on the values of its parents, Y_i , but the variance of this *normal* distribution is independent of the parents' values.

However, the linear Gaussian model can only detect dependencies that are close to linear. It is not likely to discover combinatorial effects, for example, it portrays that a gene is over expressed only if several genes are jointly over expressed, but not when just one of them is not over expressed (Friedman, 2000).

- **Hybrid networks:** when the BN contains both discrete and continuous variables, one need to consider representations for instances of a continuous variable with discrete parents and vice versa. For example, when a continuous variable X has discrete parents, conditional Gaussian distributions can be used in which for each joint assignment to the discrete parents of X , a linear Gaussian distribution of X given its continuous parents is represented.

1.6.2 Network Topology

In recent years, a standard way of representing Bayesian network graph topology has emerged ((Murphy, 2001); (Rung, Schlitt *et al.*, 2002); (Herrero, Valencia *et al.*, 2001) and (Kurata, Matoba *et al.*, 2003)). Table 5 summarises the network components used to build such networks.

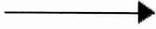
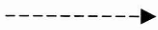
Network Component	Name	Representation and Function
	Node	Variable
	Target node	Target node is the specific variable under consideration; all relationships are measured relative against the target.
	Target node analysis	Lighter the square in the node, the more information is brought to the target node by that node in a given network structure.
	Arc (Positive relation)	Connector. Links nodes based on the parametric value (of association). The direction of the pointer represents positive probabilistic dependency.
	Arc (Negative relation)	The direction of the pointer represents negative probabilistic dependency

Table 5: Network topology employed in BayesiaLab and in manually extracted GRNs.

In the example (**Figure 12**), the upregulation of *gene A* causes *gene C* to be upregulated and in turn *gene D*. The upregulation of *gene B* causes the down regulation of *gene C* and upregulation of *gene D*. Bayesian networks are used to exactly capture such dependencies among genes.

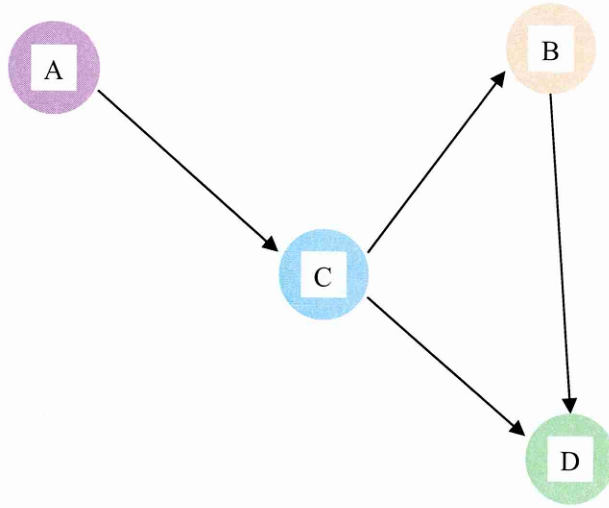


Figure 12: Example of a Bayesian network model. The nodes A, B, C, D in the above BN represent genes, whose expression levels are modelled as random variables. The direction of the pointer of the arc represents the association/dependency. For instance, the arc from A to C indicates a positive (upregulation) association whereas the arc from B to C shows a negative association (downregulation).

Since any subset of the nodes can be observed, Bayesian Networks can be used for both *causal reasoning* (from known causes to unknown effects) and *diagnostic reasoning* (from known effects to unknown causes), or any combination of the two.

In addition to causal and diagnostic reasoning, Bayesian Networks support the powerful notion of “*explaining away*” (known as *selection bias* in statistics): if two causes (parents) compete to explain the observed data, they become conditionally dependent given that their common child is observed, even though they are marginally independent. For example, suppose the grass is wet, but that we also know it is raining and hence the probability that the wet grass was caused by the use of a sprinkler goes down (Murphy, 2001).

1.6.3 Bayesian Learning

A Bayesian network is described jointly by its graph topology (structure) and the parameters of each conditional probability distribution (CPD). It is possible to learn both of these from the data. However, learning the structure is harder than learning the parameters of the model and learning when some of the nodes are hidden, or when there is missing data, is much harder than when fully observed data is available. For example, it would be relatively easier to estimate the conditional probabilities that establish the strength and nature of the associations between genes (parameters) once the underlying genetic network, which identifies the genes and their topological order (structure), is extracted.

As discussed earlier, when modelling genome-wide data, one is faced with very high dimensionality. In addition, when measuring gene expression levels some missing data values can be expected due to experimental errors, as in the case of the cell cycle regulation data used in this study (6% of total data points). So it is vital that the model and learning algorithm are capable of handling noisy data and hidden variables.

Figure 13 summarises the Bayesian learning scenario and each condition is discussed in detail in sections 1.6.3.1 and 1.6.3.2.

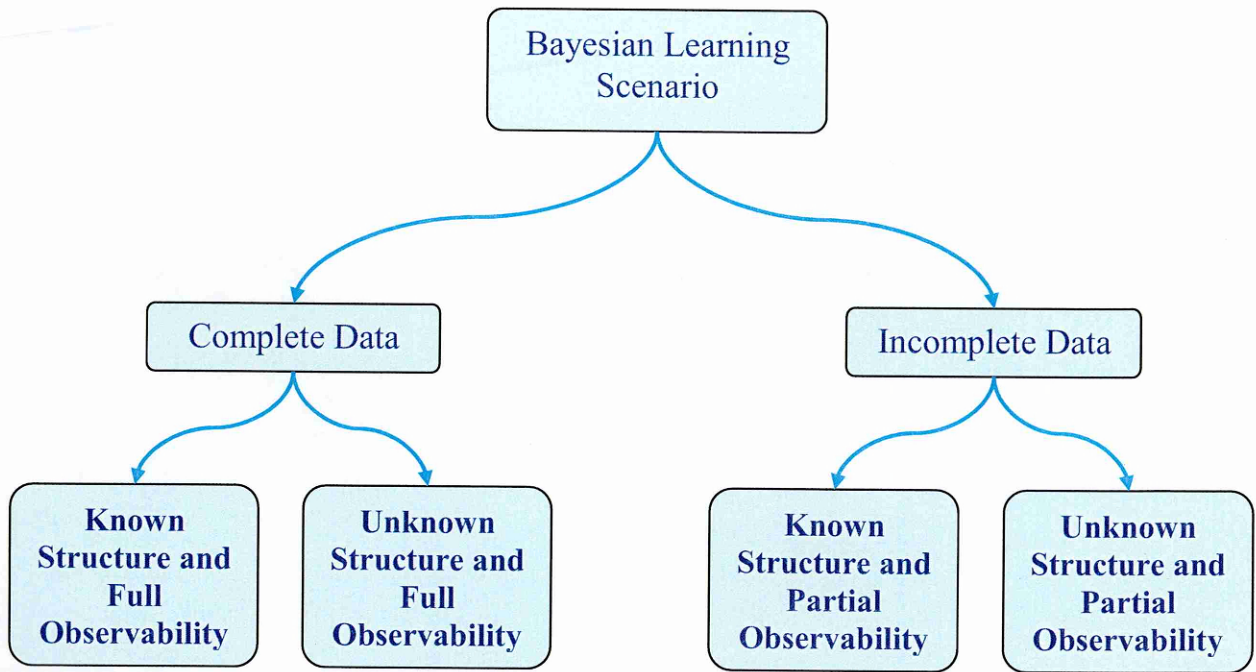


Figure 13: The Bayesian Learning (BL) using Bayesian network. It is described by its graph topology (DAG) and the parameters of each CPD, and both can be learned from given data set. However, the goal and feasibility of learning the model to best fit the data is highly dependent on data set itself and the available information in the data set. This is clearly shown in the flowchart above.

1.6.3.1 Complete Data

A complete data set is a dataset in which each instance contains the values of all the variables in the network, i.e. full observability.

Known Structure, Full Observability

The goal of learning in this case is to find the values of parameters of each CPD, which maximises the likelihood of the data set, $D = \{X^1, \dots, X^N\}$ which is assumed to be independent. This amounts to finding a Maximum Likelihood Estimate (MLE) of the conditional probability table (CPT) that decomposes according to the structure of the graph, and maximises the contribution to the log-likelihood,

$$L = \frac{1}{N} \int \log P(X_i | Pa(X_i), D) dX_i \quad (3)$$

of each node independently⁹.

For discrete nodes with CPTs, this amounts to simple counting. For Gaussian nodes, one can compute the sample mean and variance, and use linear regression to estimate the weight matrix. For other kinds of distributions, more complex procedures are necessary.

In cases where N is small compared to the number of parameters, known as *sparse data problem*, one can use numerical prior, e.g. using (mixtures of) Dirichlet priors (pseudo counts), to solve this problem (Murphy, 2001). In this case, the estimates are called *Maximum A Posteriori (MAP)* estimates, as opposed to *Maximum Likelihood (ML)* estimates.

Unknown Structure, Full Observability

The problem of learning a Bayesian network structure in this scenario can be stated as follows (Friedman, Goldszmidt *et al.*):

Given a dataset $D = \{X^1, \dots, X^N\}$ of independent instances of X , find a model $B = (G, \Theta)$ that best fits or matches D . The common approach to this problem is to introduce a *scoring function* that evaluates the ‘fitness’ of each model with respect to the dataset, and then search for the optimal network according to this score.

One method for deriving a score is based on Bayesian considerations.

⁹ Equation (3) is the logarithm of equation (1).

Bayesian Information Criterion (BIC)

In this score, the posterior probability of a structure (graph) is evaluated given the data:

$$S(G : D) = \log P(G | D) \quad (4)$$

Using Bayes's rule, and making a standard assumption that the different structures are equally likely *a priori*,

$$= \log P(D | G) + \log P(G) + C, \quad (5)$$

where C is a constant independent of G , and

$$P(D | G) = \int P(D | G, \Theta) P(\Theta | G) d\Theta, \quad (6)$$

is the marginal likelihood which averages the probability of the data over all possible parameterizations of the model G . This averaging regularises the score and avoids over fitting the data with complex models.

In other words, the problem again is to find the model with maximum likelihood. However, the maximum likelihood will be a complete graph containing all possible explanations, since this has the largest number of parameters, and hence can fit the data best. A well-principled way to avoid this kind of over-fitting is to put a prior on models, which amounts to specifying that sparse models are preferred. This implies use of the *Maximum A Posteriori (MAP)* estimate defined above would be appropriate. A *MAP model* is the one that maximises Equation (5).

For example, since there are $n \approx 6400$ genes for yeast (*S. cerevisiae*) all of which can be simultaneously measured, some pre-processing must be done to exclude the “irrelevant” genes, and just try to learn the connections between the rest. After all, many of them do not change their expression level during a given experiment. “In a dataset considered by (DeRisi *et al*, 1997), which consisted of 7 times steps, only 416 genes showed a significant change, the rest were completely uninformative” (Murphy and Mian). However, even $n = 416$ is too big for the techniques that were discussed, so one must process further, perhaps by clustering genes with similar properties.

The particular choice of priors $P(G)$ and $P(\Theta|G)$ for each G determines the exact Bayesian score. The effect of structure prior $P(G)$ is equivalent to penalising over complex models.

However, as seen above, this is not strictly necessary as the *marginal likelihood* term has a similar effect of penalising models with too many parameters, this is known as *Occam's razor*.

An exact Bayesian approach to model selection is usually unfeasible, since it involves computing the marginal likelihood $P(D) = \sum_G P(D, G)$, which is a sum over an exponential number of models.

However, one can use an asymptotic approximation to the posterior called Bayesian Information Criterion (BIC),

$$\text{Log } P(G | D) \approx \text{Log } P(D | G, \hat{\Theta}_G) - \text{Log } \frac{N}{2} * \#G \quad (7)$$

Where N is the number of samples, $\#G$ is the dimension of the model¹⁰, and $\hat{\Theta}_G$ is the ML estimate of the parameters. The first term is computed using the Equation (1); the second term is a penalty for model complexity. Since the number of parameters in the model is the sum of the number of parameters in each node, the BIC score decomposes just like the log-likelihood of Equation (1), and hence the same techniques apply (Murphy and Mian).

This scoring is *asymptotically consistent* under mild assumptions on the prior probabilities: given a large number of samples, graph structures that exactly capture all dependencies in the distribution, will receive, with high probability, a higher score than all other graphs, and therefore, given a large number of instances, learning procedures can pinpoint the exact network structure up to the correct *equivalence class*. Two Bayesian network structures G and G' are equivalent if they imply the same set of

¹⁰ In the Fully observable case, the dimension of the model is the number of free parameters. In a model with hidden variables, it might be less than this.

independence statements, e.g. the graphs $X \rightarrow Y$ and $X \leftarrow Y$ both imply the same set of independencies and therefore are equivalent.

When learning is from *complete data*, the following properties hold. The priors are *structure equivalent* in the sense that equivalent networks receive the same score. The score is *decomposable*, i.e. it can be written as the sum

$$S(G : D) = \sum_i \text{ScoreContribution}(X_i, \text{Pa}(X_i) : D), \quad (8)$$

where the contribution of every variable X_i to the total network score depends only on its own value and the values of its parents in G .

Finally, these local contributions for each variable can be computed using a closed form equation. (Friedman, Linial *et al.*, 2000).

Once the prior is specified and data is given, learning amounts to finding the structure G that maximises the score, which is usually an intractable problem. This problem is known to be *NP-hard*, thus *heuristic search* is usually used to avoid examining all possible subsets. The decomposition of the score is crucial for this optimisation problem. A *local* search procedure that changes one arc at each move can efficiently evaluate the gains made by adding, removing or reversing a single arc.

An example of such a procedure is a *greedy hill-climbing algorithm* that at each step performs the local change that results in the maximal gain, until it reaches a local maximum. Although this procedure does not necessarily find a global maximum, it does perform well in practice (Friedman *et al.*, 2000). Other algorithms used for such search include *stochastic hill climbing*, *beam-search*, and *simulated annealing*.

For example, the method of *simulated annealing* avoids the local maxima associated with the greedy search procedure by taking a random step with some probability, which decays exponentially as the search progresses, rather than always taking the highest scoring move in each search step.

An alternative approach that uses the *BIC* is the *Extended Dependency Analysis (EDA)*, which evaluates all subsets of size up to two, keep all the ones with significant (in the χ^2 test sense) *mutual information (MI)*¹¹ with the target node, and take the union of the resulting set as the set of parents.

Away from *BIC* methods, one could use a χ^2 test to decide whether an increase in the *MI* score is statistically significant. This also gives us some kind of confidence measure on the connections that we learn (Murphy and Mian).

For deterministic relationships, a useful algorithm is the *REVEAL algorithm*, which starts at the bottom of the lattice, and evaluate the score at all points in each successive level, until a point is found with a score of 1.0.

Another more popular alternative to *BIC* in evaluating probabilistic networks is the *Minimum Description Length (MDL)* method.

Minimum Description Length (MDL)

The *Minimum Description Length Principle* is a relatively recent method for inductive inference that provides a generic solution to the model selection problem (Grunwald, 2000). *MDL* is based on the following insight: any regularity in the data can be used to compress the data. The more regularities there are, the more the data can be compressed. In other words, *MDL* equates 'learning' with 'finding regularity': the more we are able to compress the data, the more we have learned about the data. Thus, the *MDL* principle dictates that given a data set D of instances, the optimal model is the one that minimizes the total description length: sum of the length of the compressed version of D and the length of the description of the model used in the compression, (Grunwald, I.J. *et al.*, 2004).

¹¹ Between each node and its parents

In the context of learning Bayesian Networks, the model under consideration is a network, B , which describes a probability distribution, P_B , over the instances appearing in the data. Using this distribution an encoding scheme that assigns shorter codes to more probable instances can be built. Examples of such codes are *Shannon encoding* and *Huffman Code* (Cover and Thomas, 1991).

The MDL Principle, we should choose a network, B , such that the combined length of the network description and the encoded data with respect to P_B is minimized. The implication of this method is that the learning procedure balances the complexity of the deduced network with the degree of accuracy with which the network represents the frequencies in D (Friedman *et al*, 1996).

As discussed above, the problem of learning a Bayesian network can be stated as follows: given a dataset $D = \{X^1, \dots, X^N\}$ of independent instances of X , find a model $B = (G, \Theta)$ that best fits or matches D .

To store a network $B = (G, \Theta)$, we need to describe X , G , and Θ ¹².

To describe X , we need to store the number of variables, N , and the cardinality of each variable X_i . Since X is the same for all candidate networks, the description length of X can be ignored in the comparisons between networks.

To describe the structure (DAG) G , it would be sufficient to store for each variable X_i a description of its parents in G , Pa_i . This description consists of the number of parents, K , and the index of the set Pa_i in some agreed upon enumeration of all $\binom{N}{K}$ sets of this cardinality. Since the number K can be encoded using $\log N$ bits, and the index using $\log \left(\frac{N}{K} \right)$ bits, the description length of the graph structure is:

¹² Adapted from (Friedman *et al*, 1996).

$$DL_{graph}(G) = \sum_i \left(\log N + \log \binom{N}{|Pa_i|} \right), \quad (9)$$

To describe the CPDs in Θ , the parameters, $\|Pa_i\|(\|X_i\| - 1)$, in each conditional probability table must be stored. The encoding length of X_i 's in CPD is¹³:

$$DL_{tab}(X_i, Pa_i) = \frac{1}{2} \|Pa_i\|(\|X_i\| - 1) \log N. \quad (10)$$

Finally, to encode the training data, the probability measure defined by the network B to construct a Huffman code for the instances in D is used. Here an approximation, $-\log P_B(x)$ is used as the optimal encoding length of each distance of x in line with (Cover *et al*, 1991). Thus, the description length of the data is approximated by

$$DL_{data}(D | B) = -\sum_{i=1}^N \log PB(X_i) \quad (11)$$

It can be proven that $DL_{data}(D | B)$ is equivalent to $N \sum_i H(X_i | Pa_i)$ where, $H(X | Y) = -\sum_{x,y} P_{(x,y)} \log P(x | y)$ is the conditional entropy of X , given Y . This formula provides an information-theoretic interpretation to the representation of the data: it measures how many bits are necessary to encode the value of X_i , once we know the value of Pa_i (Friedman *et al*, 1996). This encoding of the data can be decomposed as sum of terms that are 'local' to each CPD.

Thus the MDL score of a candidate network structure G is defined as the total description length

$$DL(G, D) = DL_{graph}(G) + \sum_i DL_{tab}(X_i, Pa_i) + N \sum_i H(X_i | Pa_i) \quad (12)$$

The aim would be to find a network structure which minimizes this length. In practice this is done by searching over the space of possible networks.

¹³ Usual choice for number of bits per parameter is $\frac{1}{2} \log N$.

The MDL and BIC scores are asymptotically equivalent (i.e. for large samples, the two methods tend to yield same results).

The MDL method has several attractive properties (Grunwald *et al.*, 2000):

- **Occam's Razor:** MDL chooses a model that trades-off goodness-of-fit on the observed data with 'complexity' or 'richness' of the model. As such, MDL embodies a form of *Occam's Razor*, a principle that is both intuitively appealing and informally applied throughout all the sciences.
- **Avoid Over-fitting:** MDL procedures automatically and inherently protect against overfitting and can be used to estimate both the parameters and the structure (e.g. number of parameters) of a model. In contrast, to avoid overfitting when estimating the structure of a model, traditional methods such as maximum likelihood must be modified and extended with additional, typically adhoc principles.
- **Bayesian interpretation:** MDL is closely related to Bayesian inference, but avoids some of the interpretation difficulties of the Bayesian approach, in particular, the use of prior information which may be subjective.
- **No need for 'underlying truth':** In contrast to other statistical methods, MDL procedures have a clear interpretation independent of whether or not there exists some underlying 'true' model.

However, if none of the distributions under consideration represents the data generating machinery very well, then both MDL and Bayesian inference may failed to find the 'best' approximation. This has been observed in practice in some cases, even though it generally works quite well (Grunwald *et al.*, 2000 ; 2004)

Taboo Search Algorithm

The learning algorithm used in this project to extract GRNs is based on the MDL score described above. The search strategy employed to find a network that optimizes that score is the ***Taboo Search***, also known as *Tabu search*, was first presented in its present form by (Glover, 1986), and has since been refined, and extensively used in many practical applications (Glover, 1997; Fox, 1993; Hertz *et al*, 1990). The name taboo comes from the method of the search, where it avoids ‘revisiting’ already explored states, or ‘taboos’, by the use of adaptive memory.

Taboo Search is a dynamic neighbourhood search technique, which is more elaborate than the descent methods such as *greedy hill-climbing algorithm* described above, which suffer from often getting trapped in a local minimum, and hence fail to find a global minimum.

In order to improve the efficiency of the search process, one needs to keep track not only of local information such as the current value of the objective function, but also of information related to the search process. This systematic use of memory is an essential feature of *Taboo Search*. While most exploration methods keep in memory essentially the value $f(i^*)$ of the best solution i^* visited so far, *Taboo Search* will also keep the information on the last solutions visited.

To avoid being trapped in a local minimum, which may be far from a global minimum, any iterative exploration process should in some instances accept also non-improving moves from i to j (i.e. with $f(j) > f(i)$) if one would like to escape from a local minimum. *Simulated annealing* does this also, but it does not guide the choice of j , *Taboo Search* in contrast chooses a best j in V^* , a subset of solution in $N(i)$ (Hertz *et al*, 1990). The choice of V^* will be crucial; in order to define it at each step *Taboo Search* will use systematic memory to exploit knowledge extending beyond the function f and the neighbourhood $N(i)$. As soon as non-improving moves are possible, the risk of visiting again a solution and more generally of cycling is present. This is the point where the use of memory, known as ‘short term memory’, is helpful to forbid moves which might lead to recently visited solutions.

Since *Taboo Search* will in fact include in its own operating rules, some heuristic techniques, it would be more appropriate to characterize *Taboo Search* as *meta-heuristic*. Its role will most often be to guide and to orient the search of another (more local) search procedure.

Memory is also present at a deeper level, known as ‘long term memory’. In the search process it is sometimes fruitful to intensify the search in some region of the set S because it may contain some acceptable solutions. Such *intensification* can be carried out by giving a high priority to the solutions which have common features with the current solution. On the other hand, *diversification* tends to spread the exploration effort over different regions of S ; this term will penalise at some stage solutions which are close to the present one. The simplest way to do it is to perform several random restarts, known as *responsive exploration*.

Both the intensification and diversification terms have weights which are modified in the process in such a way that intensification and diversification phases alternate during the search. This long term memory may be viewed as a kind of learning process.

Many computational experiments have shown that *Taboo Search* has now become an established optimization technique which can compete with almost all known techniques and which, by its flexibility, can beat many classical procedures such as maximum likelihood (Hertz *et al*, 1990). Current applications of *Taboo Search* span the fields of molecular engineering, biomedical analysis, artificial intelligence, resource planning, telecommunications, financial analysis, energy distribution, mineral exploration, logistics, waste management, environmental conservation and scores of others (Glover *et al*, 1997).

Although basically simple, the *Taboo Search* technique owes its efficiency to a rather fine-tuning of an apparently large collection of parameters (CPDs). In fact experiments have shown that the degrees of freedom in most of these choices are not so restricted (Hertz *et al*, 1990).

Theoretical considerations based on a probabilistic model of *Taboo Search* seem to confirm partially this statement (Faigle & Kern, 1992): when moves from a solution i to a solution j are performed according to a probability distribution p_{ij} , then under rather mild assumptions one can prove that convergence is obtained to an optimal solution for probabilities p_{ij} which are allowed to be chosen in a large subinterval of $[0,1]$.

Taboo Order Algorithm

Taboo Order is a novel technique, applied here for the first time. It uses the Taboo Search algorithm for finding the best order for the nodes. This ensures local maximisation at each node is achieved when choosing the best Bayesian network. Thus, it aims to achieve both global and local optima. Hence this search is more complete and takes much longer than Taboo.

All results in this project were obtained using this algorithm. However, the efficiency and accuracy of other two methods were also tested in the initial stages.

1.6.3.2 Incomplete Data

Known structure, Partial Observability

So far it was assumed that the dataset D is complete in that it specifies the values of all the attributes. This may not be so in practicality. In many situations, we find that we are dealing with models with *latent* (hidden) variables, that are never observed in the training data (Segal, Taskar *et al.*, 2001). Learning from partially observed data is substantially more difficult than when the data is fully observable. One is faced with the likelihood function that has multiple local maxima, and no general method exists for finding the global maximum.

The *Expectation Maximisation (EM) algorithm* is an approach for parameter estimation with incomplete data. The EM is an iterative method: starting from an initial guess for the parameters, it repeatedly performs two steps. The basic idea behind that is, if one knew the values of all the nodes, learning (the M step) would be easy as discussed above.

So in the E step, the expected values of all the nodes are computed using an inference algorithm, and these expected values are treated as though they were observed (distributions). Given the expected counts, we maximise the parameters, and then recompute the expected counts, etc. This iterative procedure is guaranteed to find a local maximum of the likelihood function.

It is also possible to do *gradient ascent* on the likelihood surface (the gradient expression also involves the expected counts), but EM is usually faster (since it uses the natural gradient) and simpler (since it has no step size parameter and takes care of parameter constraints (e.g., the "rows" of the CPT having to sum to one) automatically).

In any case, we see that when nodes are hidden, inference becomes a subroutine, which is called by the learning procedure; hence fast inference algorithms are crucial. (Murphy and Mian).

Inference Algorithms

A Bayesian network specifies a complete joint probability distribution (JPD) over all the variables. Given this JPD, inference amounts to summing out over irrelevant variables known as *marginalisation*. However, as the number of variables (nodes n) is exponential in the number of parents, summing over the JPD takes exponential time. For example, the number of fan-in can be quite large for certain genes. Therefore, methods that increase the efficiency of the search should be used.

One method is a factored representation of the JPD known as *Variable Elimination Algorithm*. The idea is to 'push sums in' as far as possible when summing out irrelevant

variables, i.e. by distributing sums over products. This forms the basis of many common algorithms, such as *Viterbi decoding* and *Fast Fourier Transform* (Murphy, 2001).

When computing several marginals at the same time, one can use **Dynamic Programming (DP)** to avoid the redundant computation involved in repeated use of variable elimination. The most commonly used algorithm is the **junction tree algorithm**. The idea is that if the underlying undirected graph of the Bayesian Network is acyclic (i.e. a tree), one can use a **local message passing algorithm**¹⁴.

However, if the BN has undirected cycles, this runs the risk of double counting. For example in **Figure 12**, the information flowing from *C* and *D* are not independent, because they come from a common cause *B*. This is avoided by converting the BN into a tree, by clustering the nodes together, to form what is called a junction tree. The **forwards-backwards algorithm** for *Hidden Markov Models (HMM)* is a special case of the junction tree. An alternative to junction tree is the **peeling algorithm**; however, junction tree algorithm is much more efficient for learning (Murphy and Mian, 1999).

The running time of the DP algorithm is exponential in the size of the largest cluster¹⁵. This is called the *induced width* of the graph. Minimizing this is NP-hard (Murphy, 2001).

When there is large induced width, e.g. as in multivariate time-series or image analysis, exact inference becomes very slow and often computationally intractable. Therefore approximation methods should be used. There are many approaches, including *sampling (Monte Carlo) methods* and *variational methods*.

The simplest kind of sampling (*importance sampling*) involves, drawing random samples X from $P(X)$, and then weighting the samples by their likelihood $P(D|X)$. In high dimensions, a more efficient approach would be **Monte Carlo Markov Chain (MCMC)**; **Gibbs sampler** and the **Metropolis-Hastings Algorithm** are special cases of MCMC.

¹⁴ Due to Pearl.

¹⁵ These clusters correspond to the intermediate terms created by variable elimination.

The simplest example of variational method is the *mean-field approximation*, which exploits the law of large numbers to approximate large sums of random variables by their means. This produces a lower bound on the likelihood.

Unknown structure, Partial Observability

Learning an unknown structure with incomplete data, in the form of hidden variables and/or missing data, presents the hardest cases of all. One difficulty is that the score does not *decompose*. However, as seen above, the *expected score* does decompose, and for that *Structural EM Algorithm* can be used.

In this situation, missing attributes are replaced with their most probable value (expected score) given the current model, and structure learning is run on the completed data. When structure learning converges, the hypothesised values for the unobserved attributes are removed, EM is run to fit the parameters of learned structure, and then a hard assignment is selected for the missing attributes. This process is iterated until convergence.

1.6.3.3 Identifying Causal Relations

Having learned the model from observations, next obvious objective would be to understand the mechanism that generates those dependencies between multiple interacting quantities (e.g. gene expression levels) in a BN. Thus, it would be logical to model the flow of causality in a system of interest, say gene transcription. A *causal network* is a model of such causal processes. It is represented similarly to a Bayesian network, with similarly defined (G, Θ) . However, causal networks have a stricter interpretation of the meaning of edges: the parents of a variable are its immediate causes (Friedman *et al*, 2000).

Having a causal interpretation helps determining the effect of an *intervention*, setting the values of some variables at domain using forces outside the causal model, as opposed to an *observation*, which is a passive measurement of the domain (e.g. a

sample). For example, if X causes Y , then manipulating the value of X affects the value of Y . On the other hand, if Y is a cause of X , then manipulating X will not affect Y . Thus, even though $X \rightarrow Y$ and $X \leftarrow Y$ are equivalent Bayesian networks, they are not so as causal networks.

Whilst interventions remain an important tool for inferring causality, some causal relations can be learned from observations alone. This is facilitated by the use of *Partially directed acyclic graph* (PDAG), which given a DAG, further identifies the members of equivalence classes along the lines of a directed edge $X \rightarrow Y$ (implies all members of the equivalence class contain the arc $X \rightarrow Y$), and an undirected edge $X \cdots Y$ (implies some members contain the arc $X \rightarrow Y$ and others contain the arc $X \leftarrow Y$).

For example, if a directed arc $X \rightarrow Y$ is in the PDAG, then all the networks in the equivalence class agree that X is an immediate cause of Y .

Thus, by employing the *Causal Markov Assumption* that given the values of a variable's immediate causes, it is independent of its earlier causes, along with PDAG, one can learn some causal directions from observations.

1.6.3.4 Dynamic Bayesian Networks

Dynamic Bayesian networks (DBNs) refer to the application of Bayesian networks to time series data. They are directed graphical models of stochastic processes as opposed to the HMM or *Linear Dynamical Systems* (LDSs)¹⁶ which are undirected graphical models. Here, what one is interested in is how the random variable X evolves over time, i.e. the *inter-slice* connections.

Again using the first-order Markov assumption, which states that the future is independent of the past given the present, X_{t+1} is independent of X_{t-1} given X_t . However, in addition to connections *between* time slices, one can also allow connections *within* a

¹⁶ This has the same topology as an HMM, but all the nodes are assumed to have linear-Gaussian distributions, compared to the HMM which has one discrete hidden node and one discrete or continuous observed node per slice. Examples of LDSs include *Auto Regressive model (AR)* and the *Kalman Filter*.

time slice (intra-slice relations). For example, one may want to measure the expression levels of genes simultaneously, as they change over time and react to external stimuli, as well as any co-occurrence effects. Therefore, the techniques for learning DBNs are similar to those developed above for BNs.

The learning involve maximising the log-likelihood of the training dataset, which contains S independent sequences, each of which has the observed values of all n nodes per slice for each of T time slices:

$$L = \frac{1}{N} \int \int_{i=1}^N \log P(X_i | Pa(X_i), D_j) dD_j dX_i \quad (13)$$

$D = \{D_1, \dots, D_S\}$, is a sum of terms, for each node.

Despite its name, a DBN assumes that the model structure does not change and that the parameters do not change, i.e. the model is time-invariant. However, one can add extra hidden nodes to represent the ‘current regime’, thereby creating mixtures of variables to capture periodic non-stationarities; also the size of the state space, i.e. structure, can be allowed to change over time. One can also allow for correlations between non-adjacent time slices by adding *lag variables* to switch to a higher-order Markov chain.

Inferring in *temporal model* such as DBN, is even more computationally intractable in the sense that, even if the connections between two slices are sparse, correlations can arise over several time steps. Nevertheless, one can use variants of approximation algorithms to handle this. Examples of DBNs include boolean network models (deterministic), stochastic boolean networks (e.g. *noisy-OR model*), linear models of mRNA expression levels by D’haeseleer *et al.*, and the nonlinear model of Weaver *et al* (Murphy and Mian).

1.6.4 Measures of Error: Kullback-Leibler Divergence

The main measurement of error in testing the significance of the structures generated in our analysis is the *Kullback-Leibler divergence* (also known as *Entropy Distance*), which measures the loss of power of the likelihood ratio test when the wrong distribution is used for one of the hypotheses. The entropy distance from a distribution P to an approximation Q is defined as:

$$D(P|Q) = \sum P(x) \log \frac{P(x)}{Q(x)} \quad (14)$$

Equation 14 is a measure of the inefficiency incurred by assuming that the distribution is Q when the real distribution is P .

Two fundamental properties of D are:

- non-negativity : $D(P, Q) \geq 0$, where equality holds if and only if $P = Q$.
- asymmetry : $D(P, Q) \neq D(Q, P)$.

Note: $\log P(x)|Q(x)$ is the density function of P relative to Q . Another way of writing D is $D(P, Q) = L(P, P) - L(P, Q)$ where $L(P, Q) = EP \{\log q(x)\}$, and the symbol EP means taking the expectation over the distribution P . Hence K-L divergence D is the expected log-likelihood ratio.

Just as likelihood measures how well a model explains the data, so one can think of D as measuring the lack of fit between model and data relative to a perfect fit. The larger is the likelihood ratio, the more evidence is there for the alternative hypothesis. “In this interpretation, D is reminiscent of the power function (1 minus the Type II Error) in hypothesis testing, measuring the degree to which the data will reveal that the null hypothesis is false when the alternative is in fact true” (Eguchi and Copas, 2006). There are several possible justifications for using the entropy distance, one is employment of the concept in data compression method for finding the best structure such as MDL, where the entropy distance measures the loss incurred by using Q , instead of the true

distribution P . “On the axiomatic side, Shore and Johnson (1980) suggest several desirable properties of approximation measures, and show that entropy distance is the only function that satisfies all of them” (Friedman, Linial *et al.*, 2000). An alternative way of viewing the inherent error (relative to the empirical distribution P^*) of a network structure G , i.e. $D(P|G)$, is as a measure of how ‘reasonable’ are the independence assumptions encoded in G . This can be measured by estimating to what degree each of these independencies is violated in P^* .

One way of measuring the strength of the dependency between variables is the measure of *conditional mutual information*.

Let X, Y, Z be three sets of variables; the conditional mutual information between X and Y , given Z , is defined as

$$I_p(X; Y | Z) = H_p(X | Z) - H_p(X | Y, Z). \quad (15)$$

In other words, this term measures how much the knowledge of Y helps us compress X when Z is already known. It should be noted that $I_p(X; Y | Z) \geq 0$, and $I_p(X; Y | Z) = 0$, if and only if X is independent of Y , given Z .

Using the mutual information as a quantitative measure of strength of dependencies, the extent to which the independence assumptions represented in G are violated in the real distribution P can be measured.

(Friedman and Goldszmidt, 1996) provides more technical details on the incorporation of entropy in model evaluation in practice.

The Taboo Search algorithm employed in this project for finding the best structure takes into consideration the principles of entropy outlined above.

1.6.5 Bayesian Analysis of Gene Expression Data

Bayesian analysis of gene expression data considers probability distributions over all possible states of the system in question, namely a cell or an organism and its environment. It describes the state of the system using random variables. These random variables denote the expression level of individual genes. For example, at any given time or sample, the expression level of a gene could take one of three possible states, “over expressed”, “normal”, and “under expressed” and hence can be regarded as a random variable. Other attributes that affect the system, such as experimental conditions, the time/stage that the sample is taken from, clinical procedure employed, etc. can be captured by including random variables to denote these attributes (Friedman, Linial *et al.*, 2000). Thus, the aim would be to build a model which is a joint distribution over a collection of random variables.

Having such a model can answer a wide range of questions about the system, for example, how or to what extent expression levels of a particular gene is influenced by other genes or the time/stage the sample was taken. In learning such a model from gene expression data several issues need to be considered, such as choice of local probability models, algorithmic complexity, and statistical aspects. The main feature that distinguishes expression data is that it involves transcript levels of thousands of genes, which raises problems in computational complexity and the statistical significance of the resulting networks. However, one can use the techniques discussed above, such as imposing priors, clustering genes together that have similar expression levels, and efficient learning techniques, to deal with this.

The use of these techniques is justified by the fact that genetic regulation networks are sparse, i.e. given a gene, it is assumed that no more than a few dozen genes directly affect its transcription. Bayesian networks are specially suited for learning in such sparse domains (Friedman, Linial *et al.*, 2000).

In the next sub-sections, two examples of application of Bayesian probabilistic approach to modelling gene expression data are discussed.

Context-Specific Bayesian Clustering

(Barash and Friedman, 2001) developed a class of clustering models that cluster genes based on the random variables of two types, expression level of a gene (or more precisely its mRNA transcript), and the occurrences of putative binding sites in the promoter region. Each gene is described by the values of the random variables. Thus, learning involves a combined probability model of binding sites and expression patterns which enable improving of clustering of the genes based on the discovery of putative binding sites and to detect which binding sites and experiments best characterise a cluster. In addition, the learned model provides insight on the regulation of genes within each cluster.

The key features of the approach are: (1) detect the number of clusters; (2) detect variables irrelevant to clustering; (3) robust clustering; (4) context-depended representation that describes the attributes (random variables) that characterise each cluster. Learning involves the use of a Bayesian approach that uses the Structural EM: a special subclass of BNs known as *Selective Naïve Bayesian Models*.

The *Naïve Bayesian Clustering Model* assumes that there is an unobserved random variable C that takes values $1, \dots, K$, and describes which cluster the example belongs to. It is then assumed that if one know the value of C , all the observed variables become independent of each other. Thus, the form of the distribution in Equation (1) we saw earlier becomes:

$$P(X_1, \dots, X_N) = \sum_k P(C = k) P(X_1 | C = k) \dots P(X_N | C = k)$$

The local probability distributions $P(X_i | C = k)$ assumed to have a Gaussian distribution. The learning procedure is no different to what has been considered so far. For example, the decision between any two clusters is centred on the sum of terms that represent the contribution of each attribute to this decision.

1.6.5.1 Application to Cell Cycle Expression Patterns

Friedman *et al.*, (2000) considered a dataset (Spellman *et al.*, 1998), which contained 76 gene expression measurements of the mRNA levels of 6177 *S. cerevisiae* ORFs. These experiments measure six time series under different cell cycle synchronisation methods. Spellman *et al.*, identified 800 genes whose expression varied over the different cell-cycle stages. In learning from this data, Friedman's team treated each measurement as an independent sample, and did not take into account the temporal aspects. Introducing an additional variable to denote the cell cycle phase compensated for this.

Friedman *et al.*, considered two forms of local probability distributions: one with discrete multinomial distributions, the other with the linear Gaussian distribution. The search procedure they used focussed on the relevant regions of the search space, known as the *Sparse Candidate algorithm* for the efficient learning. The main idea of this technique is that one can identify a relatively small number of *candidate* parents for each gene, based on simple local statistics such as correlation; and then restrict the search to networks in which only the candidate parents of a variable can be its parents, resulting in a much smaller search space. To avoid this leading to an overly restricted search space, Friedman's team used an iterative algorithm that adapted the candidate sets during search. Using this methodology, Friedman *et al.*, were able to extract a GRN, which contained many biologically plausible relationships.

1.6.6 Bayesian Analysis Software

The Table 6 shows a list of commercial and freely available software packages for applying Bayesian probabilistic approach to data modelling.

Table 6: List of available software packages for Bayesian Networks

Commercial Name	Analysis	Platform	WWW Link
BayesiaLab	Supervised and unsupervised learning, clustering, analysis toolbox, adaptive questionnaires, dynamic models	BayesiaLab requires JRE 1.3.1. Download it freely at java.sun.com .	www.bayesia.com/GB/products/bLab/BlabPresentation.php
Freeware			
Bassist	hierarchical Bayesian models in complex analysis tasks.	Unix	www.cs.helsinki.fi/research/fdk/bassist/
BNT	Employs many algorithms, also handles dynamic models, like HMMs and Kalman filters.	Windows, Unix, Mac. Matlab/C.	www.ai.mit.edu/~murphyk/Software/BNT/bnt.html
BNJ v1.051 beta	graphical models of probability, Dynamic Bayesian networks, Probabilistic Relational Models (PRMs)	Windows	http://bndev.sourceforge.net/
PMT (Probabilistic Modeling Toolkit for Matlab)	basic static & dynamic probabilistic models (based on maximum likelihood estimation (EM))	Windows, Matlab/C	http://people.bu.edu/vladimir/pmt/index.html
JavaBayes – version 0.346	Probabilistic Bayesian network, maximum a posteriori probability	Unix	www-2.cs.cmu.edu/~javabayes/Home/

2 Modelling Yeast Cell Cycle Regulation

2.1 Introduction

This chapter describes the application of structure learning algorithms based on Bayesian networks (BNs) to temporal microarray gene expression data from the yeast, *Saccharomyces cerevisiae*. Yeast is an especially favourable organism for our purposes of learning genetic network structure and dependencies from microarray data: it was the first genome to be sequenced, and since then its genomic data has frequently been used to study its biological processes including the transcriptional regulatory mechanisms, and hence it is well documented. The cell cycle gene regulatory network (GRN) of the yeast has been well characterised and thus provides a rich validation set for comparison.

BNs were extracted that model the yeast cell cycle GRN. Two novel knowledge extraction techniques are demonstrated: target node analysis and network simulation, which showed how quantitative and qualitative information captured within the BBN can be used to simulate the nature of interaction between genes, for example to test if a gene is influencing or suppressing another and the strength of these relations in the network. The nature and strength of these relationships is also represented visually.

The GRNs extracted are compared and validated against genome databases, *Saccharomyces* Genome Database (SGD), Kyoto Encyclopaedia of Genes and Genomes (KEGG), Munich Information Centre for Protein Sequences (MIPS) and other published literature, which confirms the efficacy of the techniques employed. Further, it is demonstrated how the addition of prior knowledge to the learnt network can improve the network structure, in terms of agreement with the genomics and pathway databases mentioned above.

2.2 Materials and Methods

2.2.1 Gene Expression Data

The yeast cell cycle expression data from Spellman *et al.* (1998) is used as the training data set in this chapter. This set of data was obtained from a cell cycle analysis project by Spellman's team, whose aim was to identify all genes regulated by cell cycle. This time series dataset has been used by Spellman *et al.* (1998), Friedman *et al.* (2000), Yihui *et al.* (2002), and many more for genome wide gene expression and cell cycle regulation studies, and hence provides a rich set of literature for validation of learned relationships. The data is freely available from the yeast cell cycle project site¹⁷.

The dataset of Spellman *et al.* (1998) contained 76 gene expression measurements of the mRNA levels of 6177 *S. Cerevisae* ORFs (Open Reading Frame). The data was collected by using DNA microarrays to analyse mRNA levels in yeast cell cultures that had been synchronised by three independent methods. These experiments comprise six time series collected over the different cell cycle synchronisation methods. These data were analysed by Spellman *et al.* (1998), deriving a numerical score based on a Fourier algorithm (testing periodicity) and by a correlation function that identified genes whose RNA levels were known to be regulated by the cell cycle. This protocol allowed Spellman's team to include data from previously published studies such as Cho *et al.* (1998). As a result 800 were identified that met an objective minimum criterion for cell cycle regulation.

In contrast, traditional wet lab methods identified 104 genes as cell cycle regulated, while the study by Cho *et al.* (1998), using a manual decision process identified 421 genes. In fact, the study by Spellman *et al.* included 304 of the genes identified by Cho *et al.*, perhaps indicating that the technical differences in the way the two studies were carried out may have contributed to the differences between the results.

¹⁷ <http://genome-www.stanford.edu/cellcycle/info/>

Spellman *et al.* (1998) in particular, employed a statistical rather than a manual approach to identification, using larger and more diverse experiments.

2.2.1.1 Data Acquisition and Data Pre-processing

The raw, normalised microarray intensity values were in tab delimited text format, required pre-processing before being analysed in BayesiaLab (Bayesia®, Paris). The BayesiaLab requirement for input data is .txt, .csv or .data. The raw data was downloaded from the project site <http://genome-www.stanford.edu/cellcycle/>.

2.2.1.2 Data Extraction

The genome wide time course microarray data for mRNA expression, was downloaded from the yeast cell cycle project (combined.txt). These data sets were generated by synchronizing with respect to alpha, CDC15, CDC28 and eu. The number of data points for each synchronization are 18 (alpha), 24 (CDC15), 17 (CDC28) and 14 (eu) and it also includes data from induction experiments on CLB2 and CLN3. (Spellman, Sherlock *et al.*, 1998).

Normalised microarray log ratio data for all 800 genes identified by Spellman. *et al.* (1998) as cell cycle regulated were extracted from combined.txt, which contained all 6300 yeast genes. Separately, 104 genes identified previously by traditional methods as cell cycle regulated (Spellman *et al.*, 1998) were extracted from combined.txt according to the cell cycle phase they belonged to.

The extracted data files were filtered so that they were in an acceptable data format for BayesiaLab version 1.3. The dataset was transposed in order for BayesiaLab to recognise the genes as the target variables.

The data from experiments to study the effect of inducing G1 cyclin CLN3p and B-type cyclin CLB2 were completely removed, as it is a separate experiment and the microarray slide for these showed substantially different expression levels.

2.2.1.3 Discretisation

As discussed in Chapter 1, the discretisation process for discovering gene to gene associations is vital as the learning is based on the probabilistic conditional dependency between the outcomes of gene expression. These data are captured continuously as real numbers¹⁸, and hence require discretisation. The alternative is to directly use the continuous datasets, but this means assuming a distribution for the datasets, for example, linear Gaussian, Gamma etc., which is rather restrictive.

The log ratio of the expression levels were discretised into three classes: the expression level more than zero i.e. all positives are over expressed; less than zero i.e. all the negative values and log ratios equal to zero are classified as 'Normal'(sometimes referred to as 'Off') and the other two 'Up' and 'Down' respectively. The different class width selection of expression values for possible ORF states due to experimental noise are discussed in detail in **Chapter 3**.

2.2.1.4 Replacement of Missing Values

All the datasets used had missing values as shown in **Table 7**. The Structural Expectation Maximisation (EM) algorithm employed in BayesiaLab was used for parameter estimation with incomplete data. The EM is an iterative method: starting from an initial guess for the parameters, it estimates and replaces the missing data until it finds the best Bayesian network structure in terms of maximisation of the likelihood function.

¹⁸ Non-integers.

Data Set: Microarray mRNA Gene Expression	Data Subset	No Of Variables	Total No Of Data Points	Data Type	Percentag e Missing Data
Yeast Cell Cycle	6: M/G1, G1, Late G1, S, S/G2, G2/M.	104 ORFs	104 x 73 = 7592	Continuous Log2 ratio	5.97%

Table 7: Data description. Data sets used for Bayesian learning and validation. Yeast cell cycle mRNA expression. Data subset late G1 is SCB regulated and late G1 is (MCB regulated). The data subset classification was obtained from Spellman *et al.*(1998).¹⁹

2.2.2 Learning Algorithms

Unsupervised learning was used to discover all probabilistic relationships in the data. This process is known as association discovery. Three different algorithms are employed for association discovering, SopLEQ, Taboo Search and Taboo Order Search. The Taboo search and Taboo order search are described in detail under section 1.6.3.1.

2.2.2.1 SopLEQ

This search is based on a total characterisation of the data and the exploitation of the equivalence properties of Bayesian networks. The networks that generate the same score are considered equivalent Bayesian networks (Murphy & Mian, 1999). This method is the faster learning method, and it extracts BNs close to those that are generated by more accurate, but more time consuming algorithms, Taboo and Taboo Order Search as described below. Local optimization however may not be achieved.

¹⁹ Can be found <http://genome-www.stanford.edu/cellcycle/data/rawdata/KnownGenes.doc>. As of 07/08/03.

2.2.2.2 Taboo search

This method is particularly useful for refining a network built by human experts or for updating a network learned on a different data set. Indeed, whereas the two other learning methods begin with a fully unconnected network (no arc), this one begins the search with the network in its current state, i.e. with existing arcs.

The Taboo Search algorithm is based on the Minimum Description Length (MDL) score (Grunwald, 2000). This score, based on information theory, is two fold. One part is used to score how well the network fits data; the second part represents the complexity of the network. The structural part prevents over-fitting and can be viewed as a probabilistic significance threshold that is set automatically (see Friedman, 1996).

The search strategy employed to find a network that optimizes that score is the so called Taboo Search (Glover, 1997). As discussed in **Chapter 1**, this algorithm works like the greedy search algorithm, but, whereas the greedy strategy consists in always choosing the operation that leads to the best score improvement (here adding, deleting and inverting arcs), the Taboo Search always selects the best operation, even if this operation degrades the score. By accepting such score deterioration, it expects to escape from local optima.

For variables $X_1, \dots, X_n$, where n is the number of genes, one uses the data that describes the experiments to find the probabilistic dependencies between these variables. In specifying the conditional probability distributions (CPDs), one can choose from several representations. Suppose the parents of a variable X (i.e. the variables that have a direct probabilistic relation with X) are $(Y_1, \dots, Y_n)$. The choice of representation depends on the type of variables being dealt with. As described in section 2.2.1.3, we chose in this case to discretize the expression levels of the genes. This data pre-processing allows us to work with discrete random variables. As X and $Y_1, \dots, Y_n$ take discrete values from a finite set, the CPDs can be represented with conditional probability tables (CPTs), where $P(X/Y_1, \dots, Y_n)$ specifies the probability of values for X given each joint assignments of $Y_1, \dots, Y_n$.

There are two main approaches to learn the structure, i.e. finding the arcs, of Bayesian networks from data. The first one, the constraint-based approach, consists of constructing the network according to the conditional independence relations found in the data. The other approach is to define an evaluation function (or score) that accounts for the candidate network quality with respect to the data, and to use a search algorithm to find a network optimizing that score. Score-based algorithms are less sensitive to the quality of the available data than the constraint-based algorithms, which have to use independence tests that suffer from a lack of reliability for conditional parts of high-dimensionality.

2.2.2.3 Supervised Learning

This learning in BayesiaLab is dedicated to characterising a particular variable, in that induction of the Bayesian network is completely guided throughout the characterisation of a particular node. The tools for this learning are: Naïve Bayes, Sons and Spouses, Markov Blanket and Augmented Markov Blanket. A node is targeted as the father of the entire network in the structure and the value of the target node (weight) makes each node independent of others. The application of these tools is demonstrated in section 2.3.2.3.

2.2.3 Computation

The Bayesian structure learning process is very computationally demanding, in terms of the processing power and memory needed to execute the algorithm. When the variables (nodes/genes) are binary valued (two states), the conditional probability table, CPT will specify 2^K distributions, where K is the number of nodes in the network. This is a general representation (known as *multinomial models*), which can describe any discrete conditional distribution. The CPT size and time taken to learn the network structure grow super exponentially as the number of nodes in the network increases. The hardware and software used and the times taken for SopLEQ, Taboo and Taboo Order Search algorithms to learn the GRN are summarised in **Table 8** and **Table 9** respectively.

Hardware	1 GHz Intel Pentium II,	512 MB of RAM
Software specification	BayesiaLab 1.3	JRE 1.4

Table 8: Computational specification. The hardware and software used for extraction of GRN from yeast cell cycle data.

2.2.4 Extraction of Gene Regulatory Network

The extraction of the gene regulatory network was carried out on a subset (see section 2.2.1) of the 800 genes identified by Spellman *et al* as cell cycle regulated using the BayesiaLab software. The gene names were used as the node names. The gene names and their ORFs are given in **Table 17** in Appendix IV.

GRNs were extracted for each of the cell cycle phases identified by the traditional methods. These were then compared to the extracted GRN from the dataset of all 104 genes. The nodes in extracted GRNs were tagged according to their cell cycle phase presented in Spellman *et al.* (1998).

Traditional laboratory analysis indicates that the yeast cell cycle consists of six separable phases. These are M/G1, late G1 (SCB²⁰ regulated), late G1 (MCB²¹ regulated), S, S/G2, and G2/M. In this chapter, late G1 (SCB regulated) is denoted as G1, and late G1 (MCB regulated) is denoted as Late G1. At the time of writing, the KEGG database for the yeast cell cycle pathway only contained 88 genes, and these were grouped according to cell cycle phases, G1, S, G2 and M.

²⁰ The cell cycle regulation of these genes is modulated through an upstream regulatory element termed the SCB (SWI4,6-dependent cell cycle box, CACGAAA), which is dependent on both SW14 and SW16. The SCB binding factor is known as SBF, (Harrington *et al*, 1996)

²¹ The cell cycle regulation of these genes is modulated through an upstream regulatory element known as the MluI cell cycle box (MCB), ACGCGTNA, which is present in the promoters of these genes. The *Mlu* I cell cycle box binding factor (MBF, also known as DSC1) is composed of Swi6 and the SWI4 homolog Mbp1, (Harrington *et al*, 1996).

The relation between two genes (the arc connecting their two nodes) is identified as positive or negative based on the analysis of the CPT. If the conditional probability of the parent and child nodes being of the same state ('Over', 'Under' or 'Normal') is over 50%, the relationship is designated as positive, otherwise it is designated as negative.

2.2.5 Validation of Extracted GRNs

Each parent to child node association was validated using queries against KEGG and MIPS databases. The associations that were not found in these databases were then queried separately against SGD and published literature databases, PubMed and Science Direct.

2.2.6 Knowledge Extraction from the Learned Networks

2.2.6.1 Target Node Analysis

Target node analysis was used to measure the influence of conditional dependencies of the network variables on the knowledge of the target node. It also allows the visualization of the quantity of information contributed by each node to the knowledge of the target node. The brightness of the squares appearing inside the nodes is proportional to this quantity of information. Application of this technique is demonstrated in section 2.3.3.

2.2.6.2 Network Simulation

Once the probabilistic relationships between the genes have been determined, network simulation allows one to observe changes in marginal probabilities of genes in response to user defined interventions. Such a causal interpretation helps determine the effect of an intervention, simulated by manually setting the state of a node outside the causal model, as opposed to an observation, which is a passive measurement of the domain (e.g. a sample). For example, if X causes Y , then manipulating the value of X affects the value of Y . This is equivalent to a laboratory experiment where we study the effect of one particular gene on other genes through an intervention, setting its value in such a way that the manipulation itself does not affect the other variables, except via established regulatory pathways.

2.3 Results and Discussion

2.3.1 Performance of Learning Algorithms Employed

Initially, to evaluate the efficiency and effectiveness of unsupervised learning algorithms in terms of their processing time and accuracy, all three algorithms SopLEQ, Taboo Search and Taboo Order Search were used for discovering probabilistic relations in the data.

In many cases, e.g. phase M/G1 of cell cycle, it was found that the accuracy of the Bayesian network was improved by employing Taboo Order Search, however the processing times were longer than the other two methods employed. This is shown in **Figure 14**. However, the benefits were not clear-cut in relation to the processing times for larger networks as given in **Table 9**. The differences in association between genes in phase M/G1 using all three algorithms show variance in the overall network structure, yet nevertheless preserve on average 90% of overall individual gene to gene associations in each structure.

Data Set	SopLEQ	Taboo Search	Taboo Order Search
M/G1 phase (19 variables)	1 min	3 mins	10 mins
All 104 Genes (104 variables)	50 mins	5 .15 hrs	28 hrs

Table 9: BayesiaLab processing time for M/G1 phase and all 104 genes. Showing all 3 structure learning algorithms, SopLEQ, Taboo Search and Taboo Order Search employed.

2.3.2 Analysis of Yeast Cell Cycle Phases

In this section all six cell cycle specific genes were analysed separately and a combination of two phases M/G1 and G1 and finally all 104 cell cycle specific genes (all phases) analysed using the Taboo Order Search algorithm. The Taboo Order algorithm was preferred for Bayesian network identification, as it is a more complete method which achieves local (best order of nodes given the data) as well as global maximum (best overall score).

Results were then compared with those from traditional methods, Spellman *et al.* (1998), and also gene database KEGG (Kyoto Encyclopedia of Genes and Genomes).

2.3.2.1 GRN for Yeast Cell Cycle, Phase M/G1

The derived network structure for phase M/G1, shown in **Figure 14**, shows that gene SIC1 is the most prominent member of the network architecture in all three cases. In that, there are direct probabilistic relations between SIC1 and at least seven other genes. The following seven genes, EGT2, PCL9, CDC6, CHS1, TEC1, ASH1, and RME1 are identified as directly associated by SopLEQ, and Taboo Order Search, while the Taboo Search algorithm shows CTS1 instead of RME1 among the seven genes directly associated with SIC1. This is shown by arcs from these genes directly associating with (pointing at) SIC1.

Figure 14 also shows the existence of mini clusters within the network structure. The Taboo Order Search algorithm identified a small cluster of 3 genes, SST2, FUS1 and AGA1 among the M/G1 associated genes. However SST2 and AGA1 were not found in the KEGG cell cycle pathway even though these two genes are known to be cell cycle regulated (Oehlen, McKinney *et al.*, 1996). This suggests that genes in this cluster are possibly associated with other phases in cell cycle regulation or other pathways associated with cell cycle regulation. Another possible reason could be that their role is only marginally influenced for a short time frame that microarray sampling time points didn't capture it. Hence the marginal probability doesn't reflect this.

The marginal probability distributions calculated directly from pre-processed microarray data of all 19 genes (nodes) in phase M/G1 are shown in **Figure 15**. The expression level for each gene as 'Up', 'Down' and 'Off' are shown on the horizontal axis. The systematic gene name, ORF represents each node and graph box.

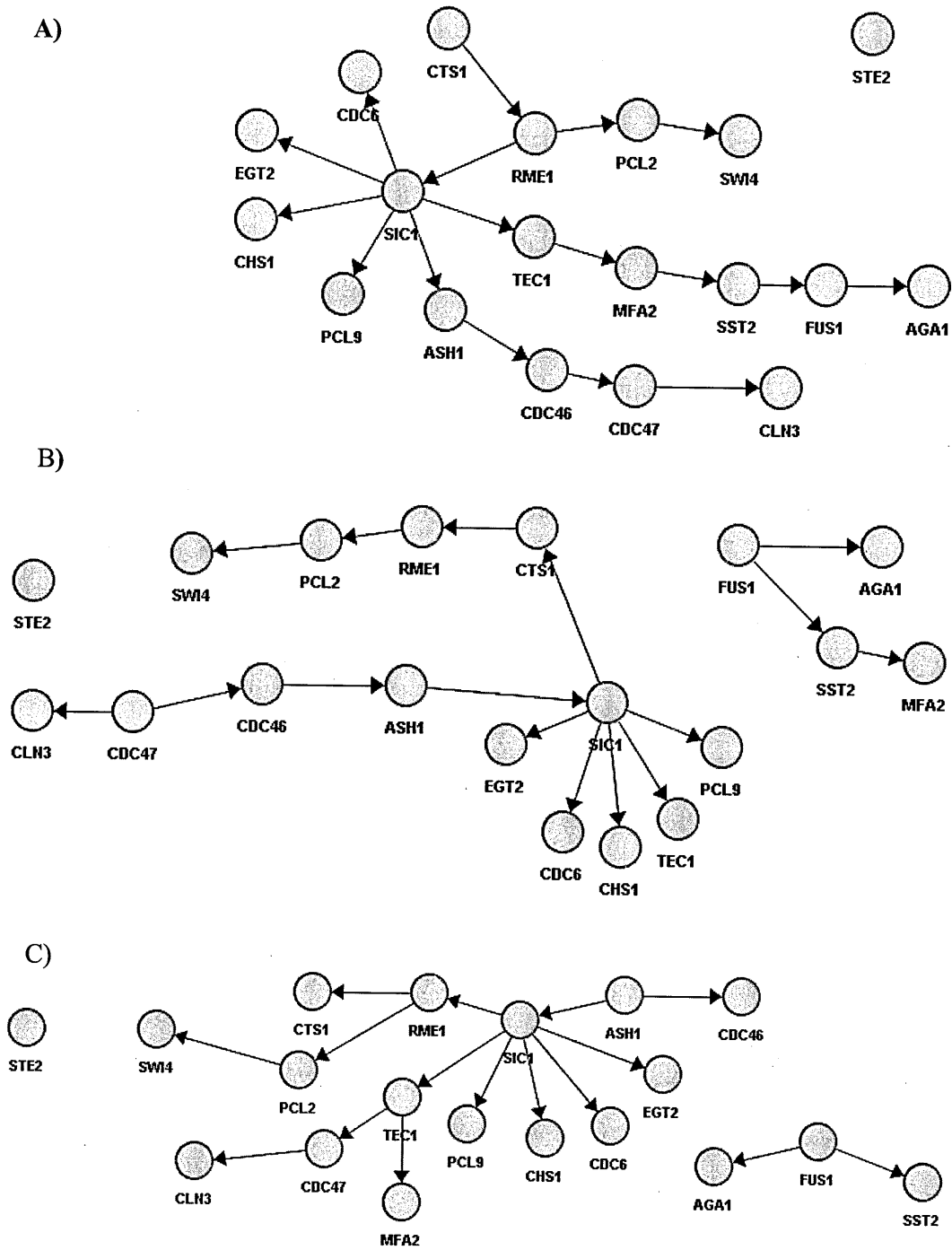


Figure 14: Predicted gene regulatory network for yeast cell cycle, phase M/G1. BN analysis using SopLEQ, Taboo Search and Taboo Order algorithms on genes identified by traditional methods as cell cycle regulated in phase M/G1. A) SopLEQ algorithm B) Taboo Search algorithm C) Taboo Order algorithm.

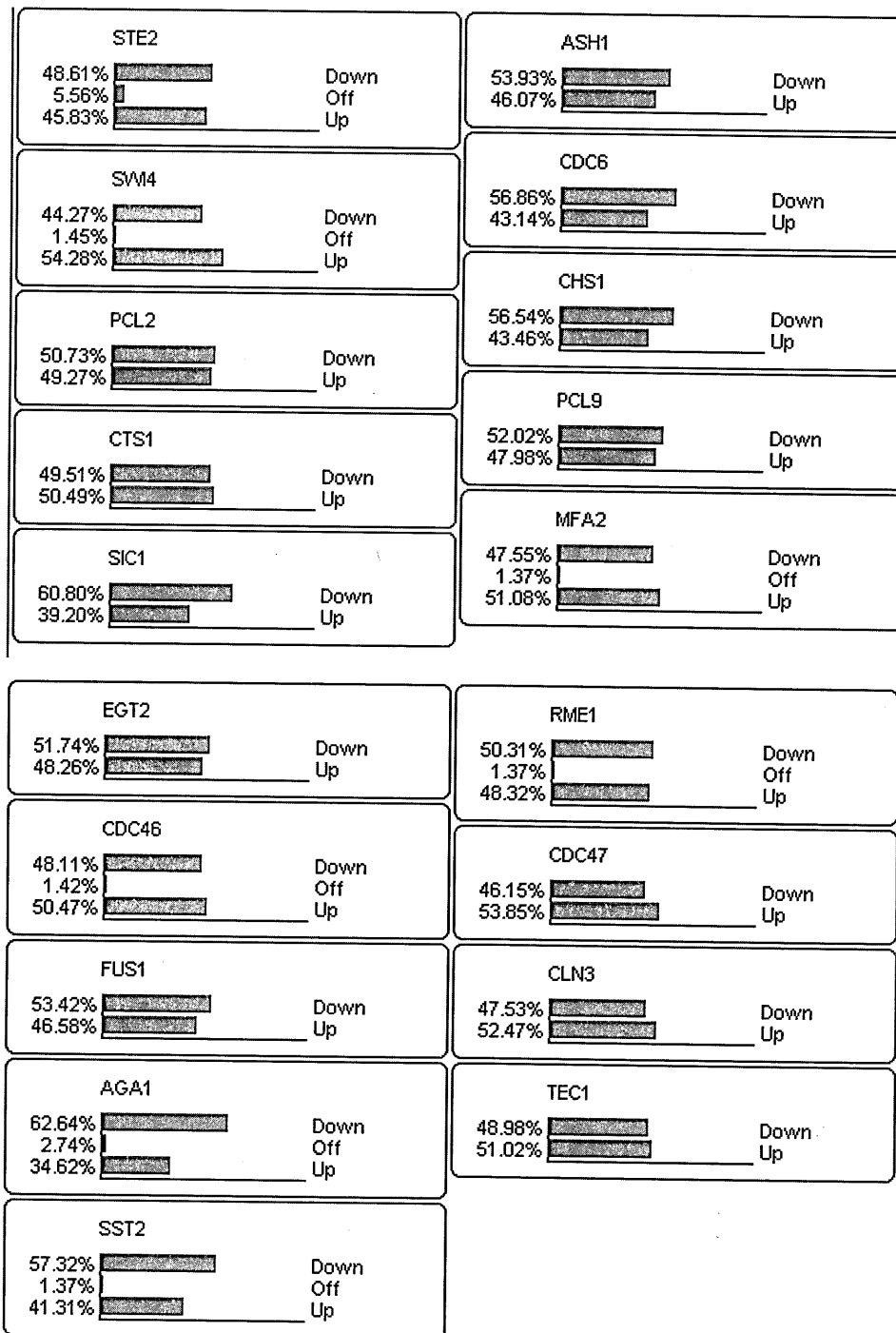


Figure 15: Marginal probability distributions of phase M/G1 BN. Each box represents the marginal probability distribution for each gene.

The marginal probability for the father node²², SIC1 was set to 100% to monitor the impact of gene expression on all the other nodes in M/G1 phase. This allows monitoring of ‘what if’ scenarios based on the marginal probabilities as the ‘resting’ position under the assumption the network structure holds. The marginal probability value can be modified between 0-100% for any node. This can be based on prior knowledge from wet lab findings.

Figure 16 shows that the probability of a gene being ‘under-expressed’ (“Down”) has increased for 11 of the nodes, 5 nodes show no difference and a slight increase for 2 of the nodes. The decrease is shown by the grey arrow pointing to the left of the graph and the length of arrow represents the change in the probability. The overall length of the grey stalk represents the percentage probability change in that particular gene, given the overall network structure and the new input value and class for target node: 100% and “Down” respectively.

²² This structure thus stipulates that the target node is the cause of all, i.e. father of all other nodes in the network

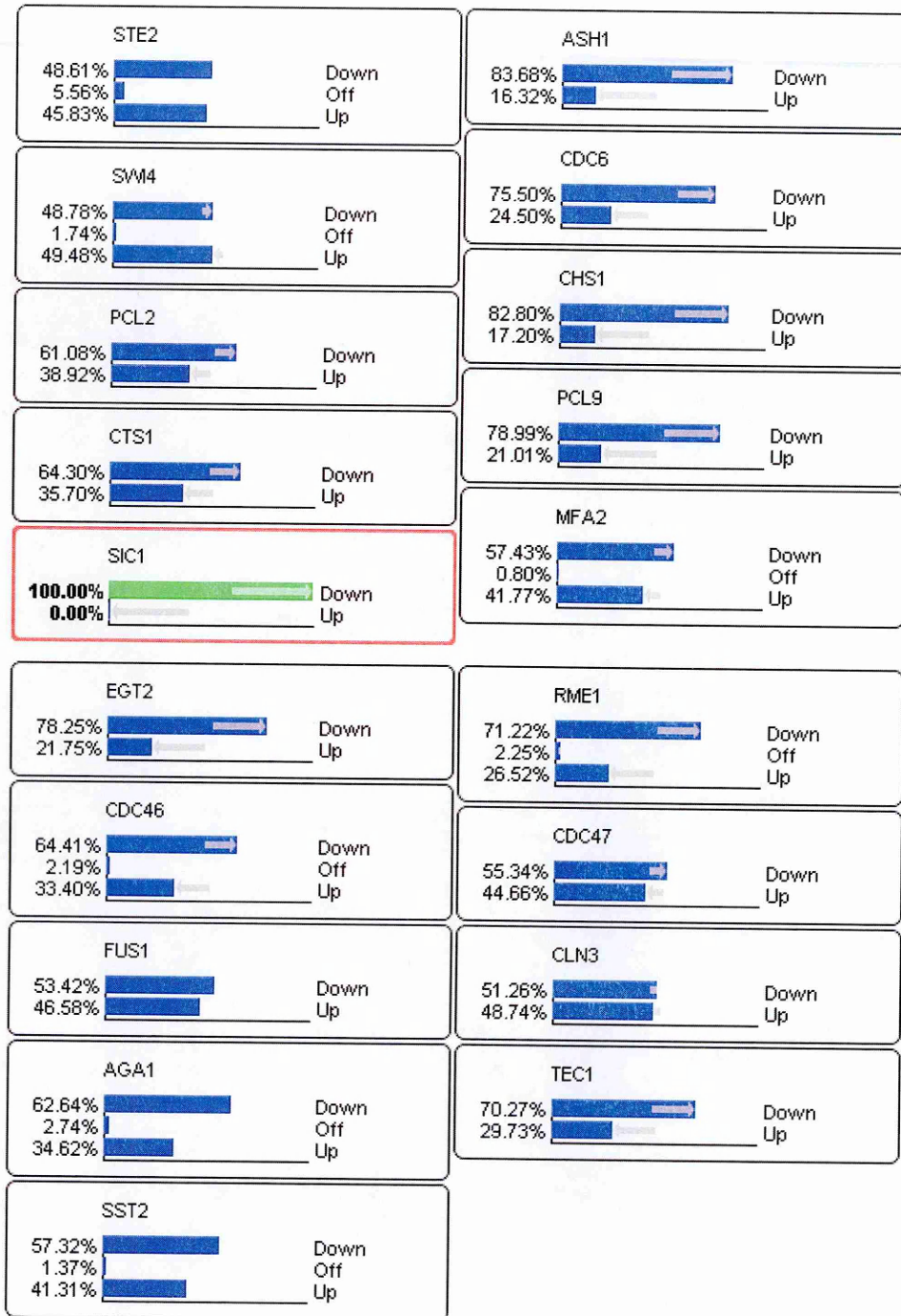


Figure 16: Use of a Bayesian network as a GRN simulation of M/G1 phase. The strength of the association between genes in the network structure is analysed by monitoring the impact of observing a variable (SIC1) on the marginal probabilities of other genes. The grey markers show the new probabilities.

2.3.2.2 Unsupervised Learning: GRN for Yeast Cell Cycle, Phase G1

The extracted BN using the Taboo Order Search on genes identified by traditional methods as cell cycle regulated in phase G1, shows CLN1 as a ‘father’ node with strong association identified with FKS1, MNN1, SWE1, and GAS1. KEGG however, did not identify GAS1 as a known regulated gene (see Figure 17).

It also shows the other associations as indirect, with their main roles in activities associated with the cell cycle process itself; FKS1 (MAPK signalling pathway), MNN1 (N-Glycos Biosynthesis), and SWE1 (Inositol Phosphate Metabolism).

The Taboo Search did not find any correlation with CHS3/CSD2, PCL1, TIP1, and KAR4. This contradicts the results from traditional laboratory techniques and KEGG. Only in the case of KAR4 are the results similar. KEGG identifies that PCL1 is cell cycle regulated. The associations with TIP1 (Butanoate Metabolism), and CHS3/CSD2 are identified as indirect.

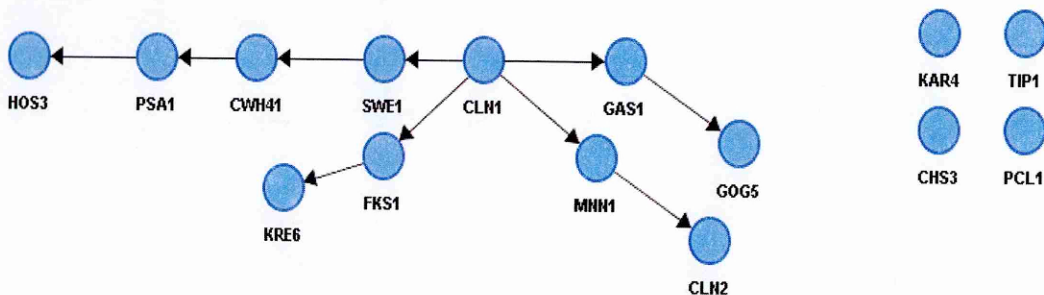


Figure 17: Unsupervised learning - predicted gene regulatory network for yeast cell cycle, phase G1. BN analysis using Taboo Order algorithm on genes identified by traditional methods as cell cycle regulated in phase G1.

2.3.2.3 Supervised Learning: GRN of Yeast Cell Cycle, Phase G1

BN analysis using supervised algorithms on genes identified by traditional methods as cell cycle regulated in phase G1 is shown in **Figure 18**. In supervised learning, the induction of the BN is completely guided by the characterisation of a particular node, in this case, CLN1. This gene was chosen as the target node on the basis of the unsupervised learning results presented in **Figure 17**.

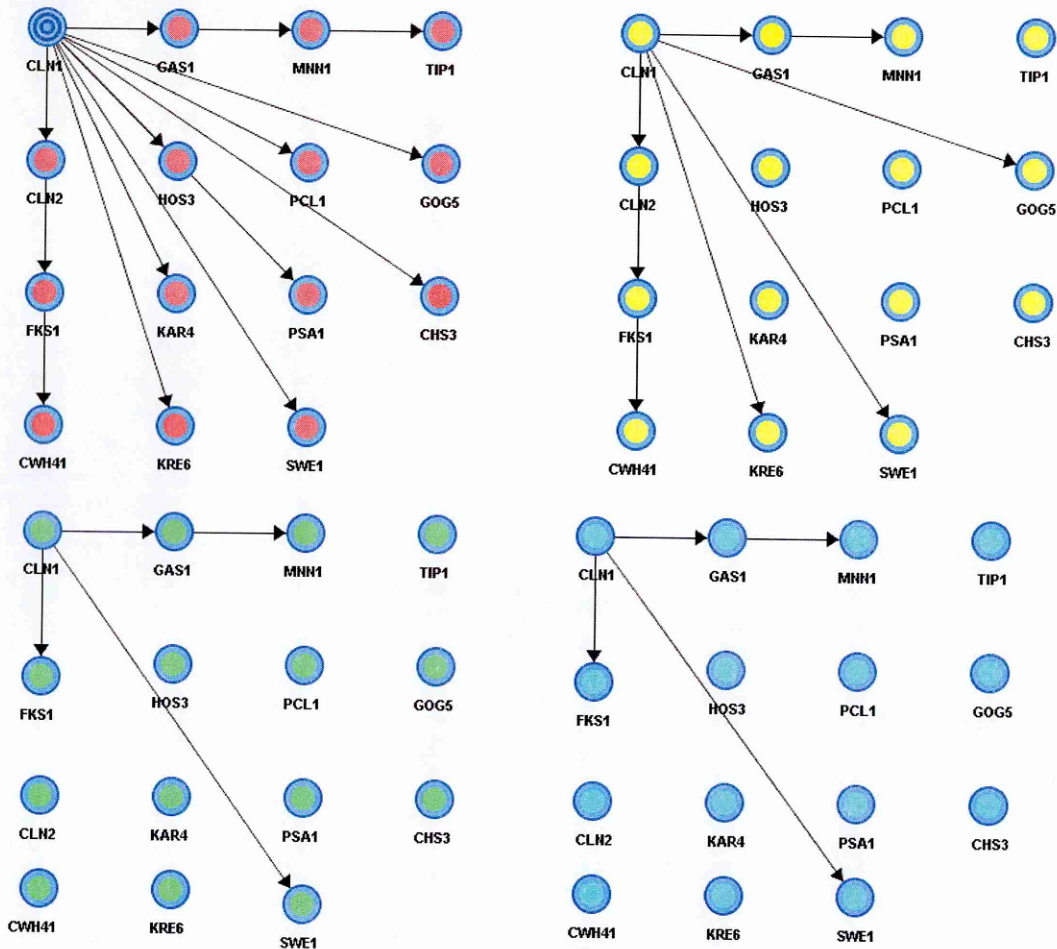


Figure 18: Supervised learning: predicted gene regulatory network for yeast cell cycle, phase G1. A) Naïve Bayes. B) Sons & Spouses. C) Markov Blanket. D) Augmented Markov Blanket.

In the *Naïve BN*, the target node CLN1 is the father or cause of all other nodes, and knowing the value of the target makes each node independent of others. In spite of this strong assumption, which is unlikely in the majority of the cases, the low number of probabilities to estimate makes this structure very robust, with a very short time of learning. However, due to the supervised method of learning, the results are generally not precise, with relations between the other nodes and the information content in them being ignored.

The *Augmented Naïve BN* allows for relations between the child nodes, knowing the value of the target node, by relaxing the conditional independence assumptions. This generates more precise results, however, the learning of child relations is unsupervised, and hence takes longer than Naïve BN. The result for the Augmented Naïve BN is not shown above as they are the same as the Naïve BN in this particular instance.

The *Sons & Spouses BN* has the advantage of highlighting the nodes that are not correlated to the target, by allowing for other parents (spouses) to the son nodes. The cost in time is comparable with the augmented naive architecture one.

With the *Markov Blanket algorithm*, the search, based on Markov relations, is entirely focussed on the target node, and only the sub-set of the nodes that are relevant is identified. The processing time is hence largely lower than the learning times of the two preceding architectures.

The *Augmented Markov Blanket* involves the Markov Blanket learning, with an addition of an unsupervised search of the probabilistic relations between each variable belonging to this Markov Blanket by relaxing the independence assumptions. This search involves an additional cost but reaches a better precision compared to the preceding version.

If no additional relationships are identified, the above algorithms may yield the same results as its preceding version, as in the instance above.

2.3.2.4 GRN for Yeast Cell Cycle, Phase Late G1

The Taboo Order Search on the late G1 subset identified probabilistic association between only 28 of the 38 genes in the set as shown in Figure 19. Out of these 28, only 18 genes were found in the KEGG pathway database and 6 of those have been identified as cell cycle regulated. They are MCD1, mitotic chromosome determinant; DBF4, regulatory sub-unit for CDC7p protein kinase; PDS1, cell cycle regulator primary relationship; CLB6, cyclin, B-type; CDC45, CDC45p assembles into a complex with CDC46p/MCM5p; RAD53, ser/thr/tyr protein kinase. PDS1 may be an anaphase inhibitor that plays a critical role in control of anaphase by both the anaphase promoting complex (APC) and DNA-damage checkpoints.

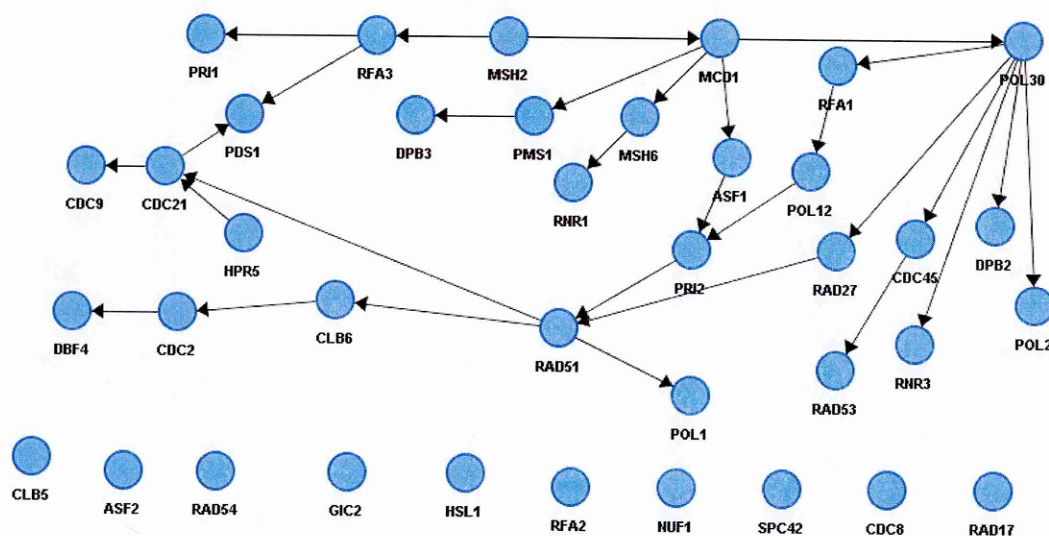


Figure 19: Predicted gene regulatory network for yeast cell cycle, phase late G1. BN analysis using Taboo Order algorithm on genes identified by traditional methods as cell cycle regulated in phase late G1.

2.3.2.5 GRN for Yeast Cell Cycle, Phase S

The S phase of the yeast cell life cycle identified by traditional methods consists of 8 genes (Histones) and the above results show that there is an association among all genes. The gene HHF2, is acting as a father to all other nodes. This is shown by direct association with genes, HTB1, HTA1 and HHT1, and all other genes are associated indirectly. When all the 104 genes identified by traditional methods as cell cycle regulated were considered, the extracted BN also showed the same associations as above, except for the association between HTA1 (Histone H2A) and HHF2 (Histone H4). However, while Figure 20 indicates that the direction of causation is from HHF2 to HTA1, the extracted BN of all 104 genes (see Figure 25) shows the relation as that from HTA1 to HHF2. As generally addition of more variables improve the accuracy of the model, the latter is more likely to be the true relationship. However, one may need to refer to biological knowledge to confirm this.

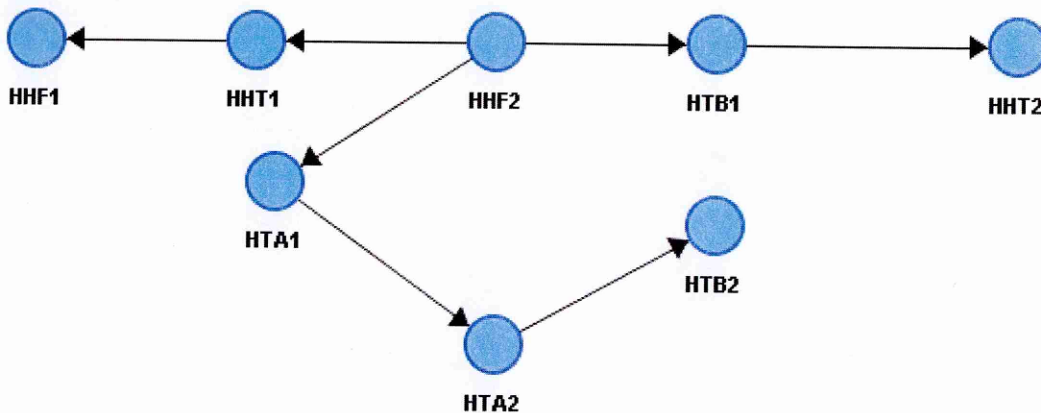


Figure 20: Predicted gene regulatory network for yeast cell cycle, phase S. BN analysis using Taboo Order algorithm on genes identified by traditional methods as cell cycle regulated in phase S.

2.3.2.6 GRN for Yeast Cell Cycle, Phase S/G2

The S/G2 phase dataset contains nine genes and two of those (CDC14 and CLB3) are identified as independent of the association structure extracted using the Taboo Order algorithm, as shown in Figure 21. This suggests that CDC14 and CLB3 might be involved in earlier in the S phase or at a later stage in G2 phase or that their regulation activity is acyclic in the cell cycle. This implies that the behaviour of other 7 genes in the structure are not suppressing nor inhibiting their activities.

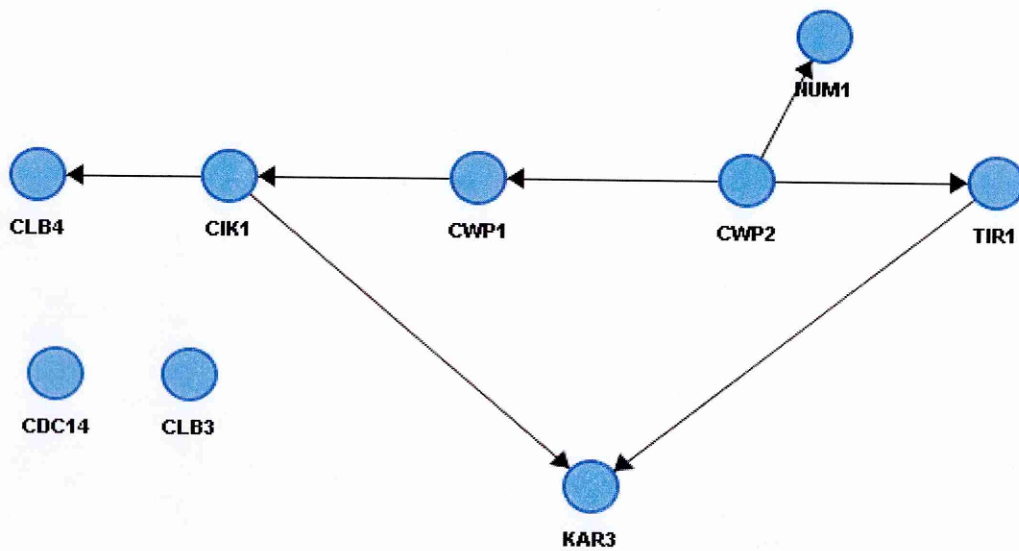


Figure 21: Predicted gene regulatory network for yeast cell cycle, phase S/G2. BN analysis using Taboo Order algorithm on genes identified by traditional methods as cell cycle regulated in phase S/G2.

2.3.2.7 GRN for Yeast Cell Cycle, Phase G2/M

Figure 22 consists of a set of 16 genes in total and 2 of them, MST1 and ASE1 are shown to be independent genes of the overall structure, i.e. given the Bayesian network, there is no Bayesian probabilistic association between these 2 genes and other genes in the network structure.

Figure 22 also shows gene KIN3 is acting as a “father” in the network, with direct relations with CDC20, YRO2, FAR1, DBF2 and SPO12. All other 8 genes are related to KIN3 indirectly. This suggests that these 13 genes are influenced by KIN3 in this network.

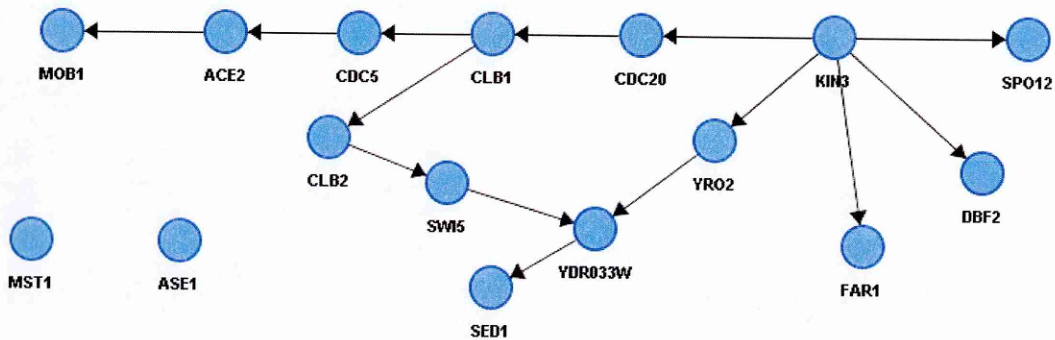


Figure 22: Predicted gene regulatory network for yeast cell cycle, phase G2/M. BN analysis using Taboo Order algorithm on genes identified by traditional methods as cell cycle regulated in phase G2/M.

However, as seen in the case of Phase S, incorporation of more variables, i.e. genes, might provide different information, with the possibility of a different network structure. To demonstrate this all the genes involved in the two phases, M/G1 and G1 have been added together and the Bayesian structure extracted using the Taboo Order Search algorithm. This is shown in **Figure 23**.

2.3.2.8 GRN for Yeast Cell Cycle, Phases M/G1 and G1

The results for unsupervised Bayesian network analysis for phases M/G1 and G1 together show 2 main clusters within the overall structure. Most nodes of both clusters are separated according to the phases.

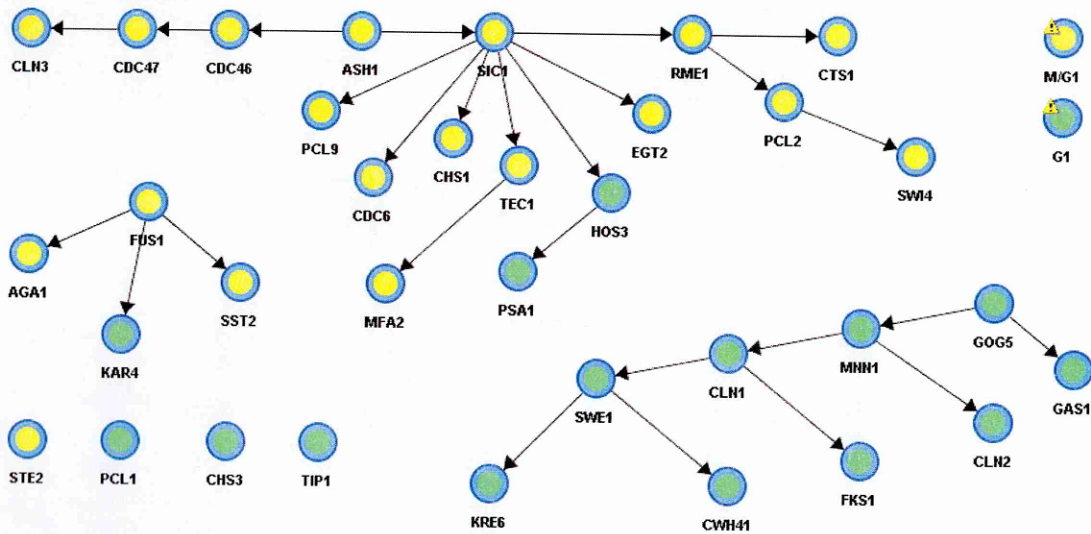


Figure 23: Monitoring the extracted BN for changes in yeast cell cycle phase threshold. BN analysis for the combined phases of M/G1 and G1 using Taboo Order algorithm.

However, in the combined network, PSA1 and HOS3 from phase G1 are shown as being associated with phase M/G1, while genes FUS1, AGA1, and SST2 from phase M/G1 and gene KAR4 from phase G1 form a third, smaller cluster. Nevertheless, one should take into account the fact that the separation of the different cell cycle phases, (both traditional methods and Spellman *et al*, 1998) are based on rather crude cut-off points. Therefore, this can lead to some overlapping of phases when extracting the underlying networks.

2.3.2.9 GRN for Yeast Cell Cycle, All Phases

Finally all 104 genes were analysed together to extract a genetic regulatory network using the Taboo Order Search algorithm. The resulting network shows strong similarity to the actual classification of the phases of yeast cell cycle. Generally there are 4 phases in the yeast cell cycle, namely, growth phase (G1), stationary phase (S), growth phase 2 (G2), and mitosis (M). However, the molecular activities, cellular processes vary considerably in each phase. Therefore sometimes, the cell cycle phase is divided into 6 phases, namely M/G1, G1, Late G1, S, S/G2 and G2/M in the order of cell cycle process.

The cell cycle data used here were originally classified into 6 phases. The genes involved in these phases are clearly distinguishable in GN extracted, according to the cluster they belong to. The nodes are tagged with different colour codes. Each colour codes represent a different phase. For example, phase M/G1 is represented by a yellow tag and the S phase is shown in red.

Phase M/G1, showed two clusters and an independent gene within this phase on individual phase analysis as shown in **Figure 14 C**. These characteristics of network architecture were preserved even when genetic association was extracted with all phases together.

This shows that the associations of genes in phase M/G1 in cell cycle are probabilistically dependant despite the influence of the other 85 genes said to be cell cycle associated. From phase M/G1 there are 3 independent genes (MFA2, CLN3 and STE2) out of 19 in the overall structure. This means that their role could be one off, i.e. to initiate or stimulate mediate pathway or could be that they have to be present at all times in order for other genes involved in the cell cycle pathway to function properly.

Nodes from a particular phase that link nodes from different phases reveal a consistency of associations between consecutive phases in the cell cycle. For example, genes ASE1, FAR1 and KIN3 from M/G1 are associated with phase G2/M. The consecutive phases S (blue) and S/G2 (red) are well associated. The genes, HTA2, HTB1 and HTA1 from the S phase are associated with CIK1, CWP2 and CWP1 respectively.

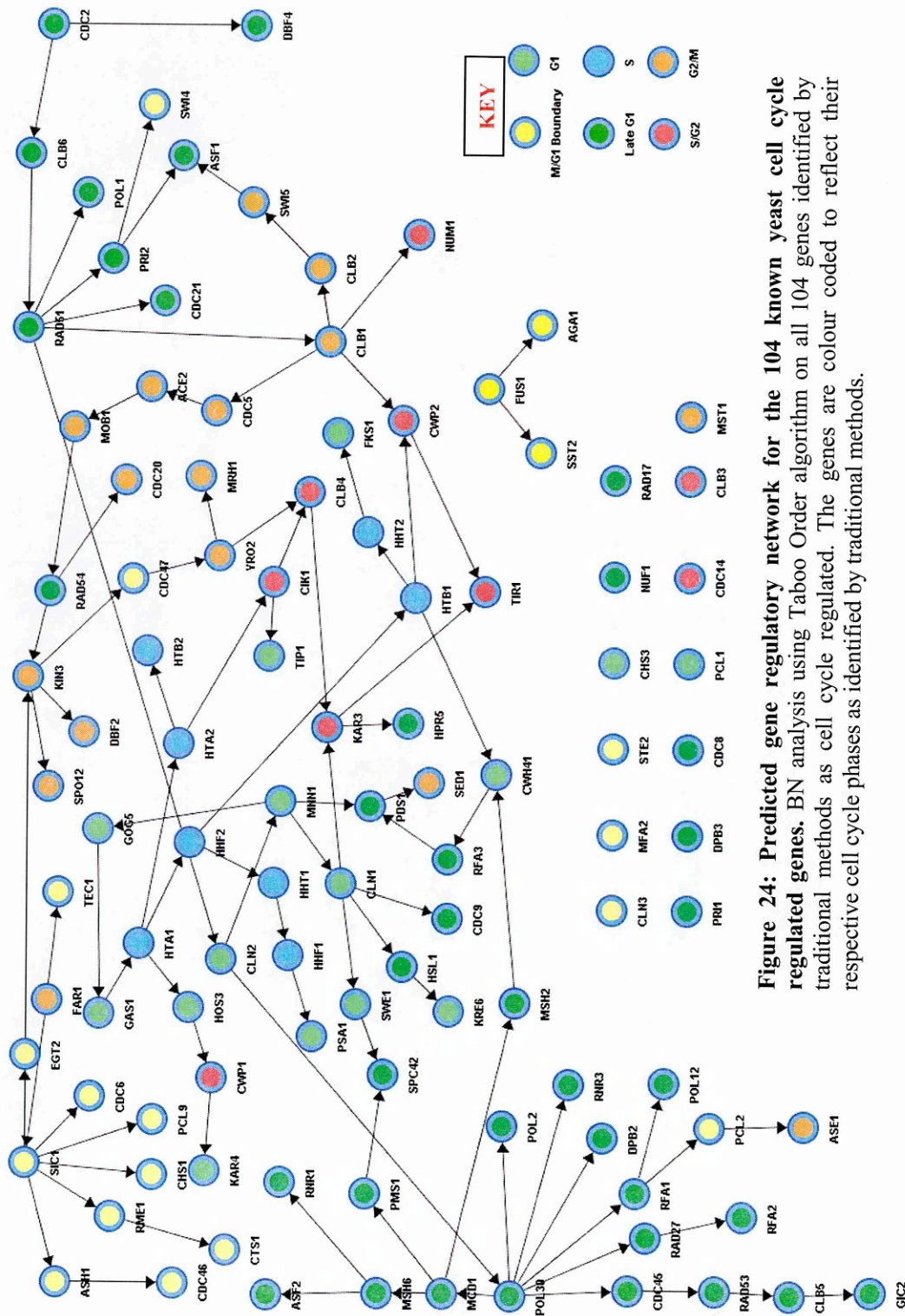


Figure 24: Predicted gene regulatory network for the 104 known yeast cell cycle regulated genes. BN analysis using Taboo Order algorithm on all 104 genes identified by traditional methods as cell cycle regulated. The genes are colour coded to reflect their respective cell cycle phases as identified by traditional methods.

2.3.2.10 GRN for Yeast Cell Cycle, Updated for KEGG

Figure 25 shows the extracted BN (unsupervised) for the 161 genes identified as cell cycle regulated by traditional methods (104 genes) and the KEGG database (57 of the 88 genes in the database are not in the 104 gene set).

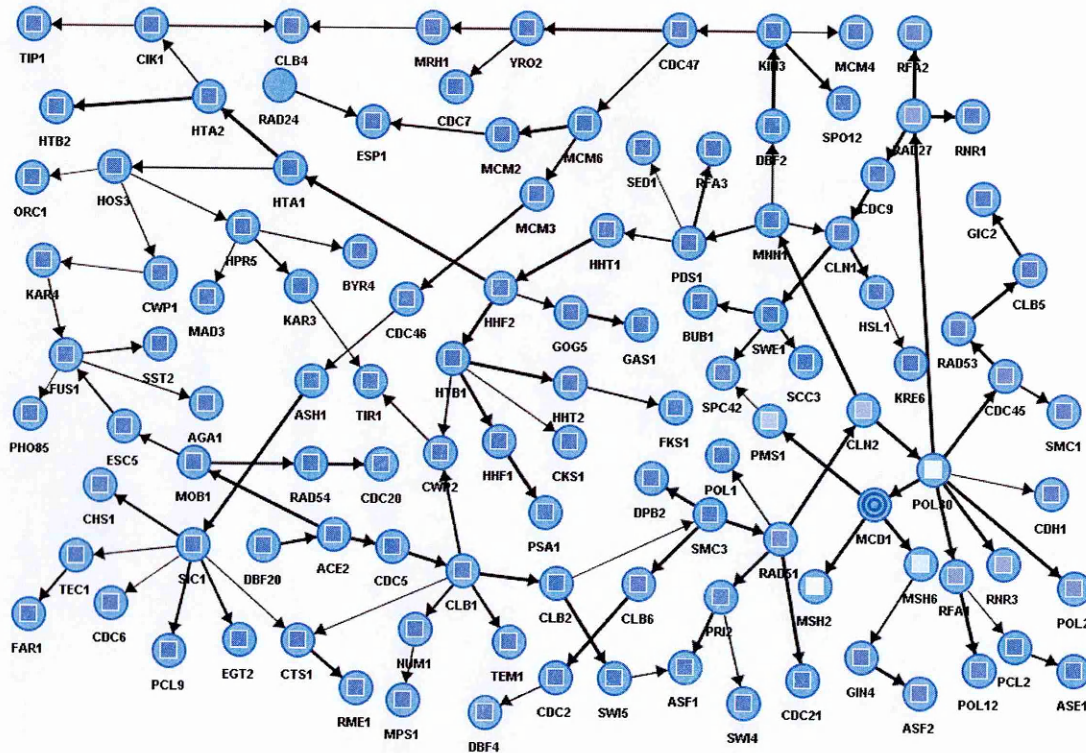


Figure 25: Unsupervised learning, extracting information contribution of the arcs and the nodes to the overall BN of 161 genes. MCD1 is set as the target node (TN) represented by the wavy circular node. The shade of the squares in the centre of the node represents the amount of information brought to the target node. The lighter the shade of the square the more the information contributed to the (TN). The thickness of the arc represents the information contributed to the global structure. The strength of the arc is proportional to the strength of the probabilistic relation.

The derived gene regulatory network captures many known relationships between known cell cycle regulated genes. For example, the immediate cluster of SIC1 i.e. direct child node of SIC1, in **Figure 25** contains ASH1, TEC1, CHS1, EGT2, PCL9, CTS1, and CDC6. This can be compared to: “SIC1 cluster comprises 27 genes, including EGT2, PCL9, TEC1, ASH1, SIC1, CHS1, and CTS1. These genes are strongly cell cycle regulated and peak in late M or at the M/G1 boundary” (Spellman, Sherlock *et al.*, 1998).

Figure 25 shows the further analyses of the strength of the association of other genes relative to gene MCD1 (Mitotic Chromosome Determinant; aka SCC1). MCD1 is an essential gene required for sister chromatid cohesion in mitosis and meiosis (Guacci, Koshland *et al.*, 1997). MCD1 (SCC1) along with IRR1 (SCC3), SMC1, and SMC3 forms the yeast cohesin complex. The shade of the squares in the centre of the node represents the amount of information brought to the target node MCD1 (represented by the wavy circular node). The lighter the shade of the square the more the information contributed to the target node (TN). For example, the shade of the squares in nodes MSH6, POL30 and MSH2 represent indicate close association with MCD1.

This is backed by biological literature: (Friedman, Linial *et al.*, 2000) found that MCD1 and MSH6, both of which bind to DNA during mitosis, are in the top 3 scoring Markov relations (functionally related). “MSH6 forms a complex with MSH2p to repair both single-base & insertion-deletion mispairs (Habraken, Sung *et al.*, 1998). “When the positions of these genes were evaluated in the exit from stationary-phase topography, POL30 and MCD1 clustered with the genes with induction that occurs almost immediately on refeeding, including CLN3 and most of the ribosomal protein genes” (Werner-Washburne, Wylie *et al.*, 2002).

(Friedman, Linial *et al.*, 2000) identified MCD1 as one of the dominant genes of the 800 genes identified as cell cycle regulated by (Spellman, Sherlock *et al.*, 1998), in terms of how they tend to dominate the order relations (i.e. they appear before other genes) in the extracted Bayesian network. (Friedman, Linial *et al.*, 2000) points out that dominant genes in the order relations may be indicative of potential causal sources of the cell-cycle process.

As shown in **Figure 25**, 7 of the Top 13 genes²³ (out of 800 genes from Spellman *et al.*, 1998) listed by (Friedman, Linial *et al.*, 2000) as the dominant genes with the highest scores in the discrete multinomial experiments: MCD1 (2nd most dominant out of the 800), CLN2 (3rd) - a G1 cyclin involved in initiation of cell cycle, RFA2 (5th) – DNA replication factor involved in nucleotide excision repair, POL30 (8th) - required for DNA synthesis and DNA repair, CLN1 (10th) - G1 cyclins involved in initiation of cell cycle, MSH6 (12th) - protein required for mismatch repair in mitosis and meiosis and CDC5 (13th) – involved in cell cycle control. Our results back the findings of (Friedman, Linial *et al.*, 2000), with all 7 of the genes dominating the order.

⁷ The other 6 genes (YLR183C, SRO4, CSI2, YOX1, RSR1, and YBR089W) are not in the 161 genes considered in our study.

2.3.3 Target Node Analysis

The target node analysis report is presented below in **Figure 25**. This shows the *relative* significance of the association of all other genes to SIC1. The top ten significant genes are shown in the first row, and in row 2, some of the genes in **Figure 25** which are not in the top ten are shown for illustration. The target node analysis report shows relative significance of all the other genes in the network (161 genes in this case), including those that are independent (zero significance), with respect to the selected (target) node.

Table 10: Node relative significance with respect to the information gain brought by the node to the knowledge of SIC1.

Gene	ASH1	PCL9	CHS1	EGT2	TEC1	CDC6	CTS1	FAR1	RME1	CLN2
Relative Significance Top10	1.0000	0.8085	0.7522	0.7458	0.5002	0.3652	0.3067	0.2134	0.1425	0.0434
Node Relative significance with respect to SIC1	Gene			CLB5	CLB1	CLB2	SWE1	SWI5	SWI4	CDC20
	Relative significance of other genes			0.0010	0.0005	0.0004	0.0003	0.0002	0.0001	0.0001

The usefulness of target node analysis can be further illustrated by considering **Figure 26** . The network representation shows the strength of the association of other genes in phases M/G1 and G1 to gene SIC1, which plays a key role in cell cycle control and mitosis.

As can be expected, the genes that are cell-cycle regulated in phase G1, are displayed as a main cluster (e.g. around CLN1), and those genes that have no association with this cluster, TIP1, CHS3, and PCL1 are displayed separately.

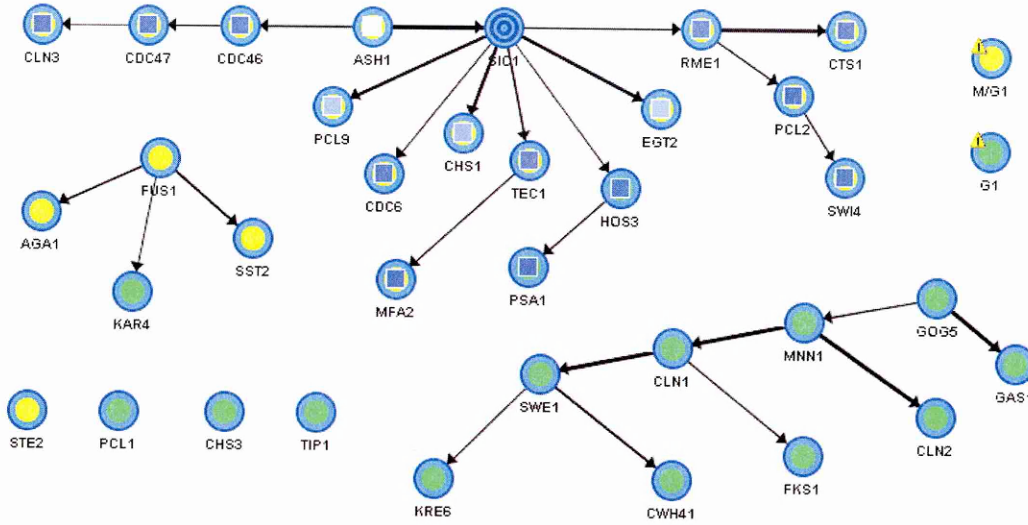


Figure 26: Extracting the information contribution of the nodes in cell cycle phases M/G1 and G1 to target node SIC1. SIC1 is set as the target node (TN) represented by the wavy circular node. The shade of the squares in the centre of the node represents the amount of information brought to the target node. The lighter the shade of the square the more the information contributed to the (TN). The thickness of the arc represents the information contributed to the global structure. The strength of the arc is proportional to the strength of the probabilistic relation.

Within the target node cluster of SIC1, ASH1, whose function is pheromone response and mating-type determination, shows strong association with the target node as identified by the lighter shade of the square. The thickness of the arc also points to the strength of the probabilistic relation. For example, the marginal probability table for SIC1 is (60.80% “Down”, 39.20% UP). However, observing that ASH1 is down regulated (by setting the marginal probability from 53.93% Down to 100% “Down”) increases the probability that SIC1 is also down regulated to 94.34%.

The relative strength of the association can be compared for EGT2, which has a slightly darker shade of the square, by the conditional probability of 91.95% “Down” for a similar exercise. The result for MFA2, which has marginal probability of (47.55% “Down”, 1.37% “Off”, 51.08% “Up”) is 73.43%.

These results show that gene STE2 has no association with other genes in Phase M/G1, which is shown by the lack of arcs linking STE2 to other genes in the extracted gene regulatory network, shown in **Figure 26**. Experimental evidence, for example cell cycle pathway from KEGG, indicates that STE2 is part of the MAPK signalling pathway that links to the cell cycle pathway and is not directly related to the cell-cycle pathway itself.

The Taboo Order algorithm also identifies a possible cluster of 4 genes FUS1, AGA1, SST2, and KAR4. Again, these have no association with the cluster centred on the target node, SIC1, as indicated by the lack of arcs linking the two clusters. Interestingly, the gene KAR4 is identified by traditional methods as cell cycle regulated in phase G1, whereas the others are identified as cell-cycle regulated in M/G1, highlighting possibly the rather crude measures employed in separating the different phases.

2.3.4 Validation of Gene Regulatory Networks

The gene regulatory interactions and networks derived directly from wet lab experiments are updated in gene regulatory pathway databases such as KEGG and MIPS. The genetic pathway of cell cycle regulation of DNA replication in *S. Cerevisae* was downloaded from the MIPS database²⁴, and the pathway map of the yeast cell cycle, genetic regulatory network derived by the traditional method was downloaded from KEGG pathway database²⁵. These are shown in **Figure 27** and **Figure 28** respectively. These pathways and their interactive links to other cell cycle associated pathways online were used to compare the generated cell cycle gene regulatory network. The cell cycle associated pathways include the MAPK signalling pathway, second messenger signal pathway, ubiquitin mediated proteolysis and biosynthesis, shown by the white rounded rectangle in the cell cycle regulatory network in **Figure 66** in **Appendix III**.

²⁴cYGD- Comprehensive Yeast Genome Database of MIPS-Munich information Centre for protein sequences: <http://www.mips.biochem.mpg.de/proj/yeast/pathways/index.html>. As of 26/09/03.

²⁵ Genetic regulatory network for yeast cell cycle is available at <http://www.genome.ad.jp/kegg/>. As of 26/09/03.

Yeast cell growth cell division and DNA synthesis of physiological and genetic pathways of MIPS were created using previously published data. There are altogether 9 small pathways involved in the cell cycle, each one is specific to a particular pathway in cell cycle regulation, as given in the KEGG and MIPS pathways. They are:

- Cell cycle control of DNA replication in *S. Cerevisae*
- Model of DNA replication complexes during the cell cycle
- Multiple roles of MRE11 complex in repair and recombination
- DNA non-homologous-end-joining
- DNA Damage Checkpoint
- Regulation of the cell cycle by the ubiquitin pathway
- model for the role of CDC5p, CDC20p and HCT1p in APC-mediated ubiquitination
- Regulatory pathways involving Cyclin-CDK complexes
- Components of the spindle pole and the SPB

The cell cycle control of DNA replication is shown in **Figure 27**. This was taken from (Huberman J.A., 1996), and has not been updated since. The solid circle indicates a group of proteins thought to be part of a multi-protein complex. Proteins enclosed by dashed ovals are degraded by proteolysis. Hence degradation pathways and energy synthesis pathways are also part of cell cycle mechanism and are cell cycle regulated. The inhibition and stimulation of the gene is indicated by small solid rectangle and arrow heads respectively. The thicker lines indicate the importance of the process for replication control within cell cycle.

The association of genes CLB1, CLB2, CLB3 and CLB4 are from S phase as shown in **Figure 24**. According to MIPS, these genes including a few other unknown genes activate CDC6. The extracted network structure here shows CLB2 has a direct probabilistic association to CLB1, and CLB4 is also indirectly related, where as gene CLB3 is independent of all other 104 genes. MIPS also shows that one of the most important process of cell cycle regulation of DNA replication is simulation of CDC6, which is inhibited by SIC1.

These facts are also shown in our extracted GN, but shown as direct association. This was expected as the cell cycle regulation of DNA replication in yeast from MIPS pathways in **Figure 27** below, only shows how S phase is controlled. The regulatory network in each phase is well interconnected, and especially between the neighbouring phases in cell cycle regulation. The MIPS pathway also has a very limited number of genes compared to the number of genes that were identified as cell cycle regulated.

The **Figure 28** shows genes identified as yeast cell cycle regulated in the KEGG database in green boxes and the matches of 104 genes that were identified as cell cycle regulated by traditional methods. The arrowed line represents stimulation and the line with a dash at the end of the line represents inhibition of the targeted gene. The dotted lines in the cell cycle pathway represent the degradation of proteins involved.

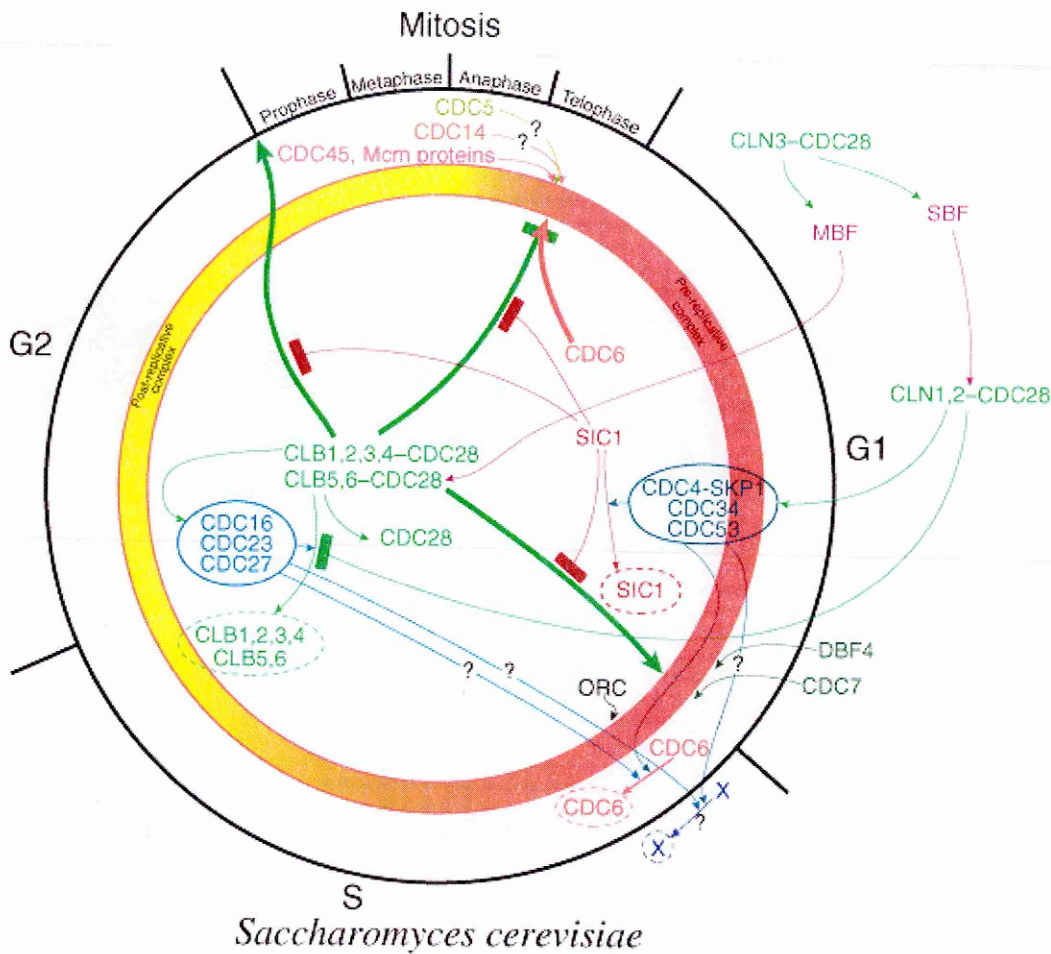


Figure 27: Cell cycle control and DNA replication of S phase in *S. Cerevisiae*. MIPS cYGD interactive physiological and genetic pathway of yeast. The black outer circle represents the phases of the cell cycle. The inner circle represents changes in the proteins bound to a typical replication origin during the cell cycle. Red colour indicates the origin is bound by a "pre-replicative complex" (pre-RC), consisting of ORC plus additional proteins. Yellow indicates the post-replicative state, when only ORC binding is detectable.

The generated gene regulatory network shows strong similarity between the phase groupings of genes traditionally recognised as cell cycle regulated. It also shows a similar architecture compared to the cell cycled pathway of KEGG, shown in **Figure 28**.

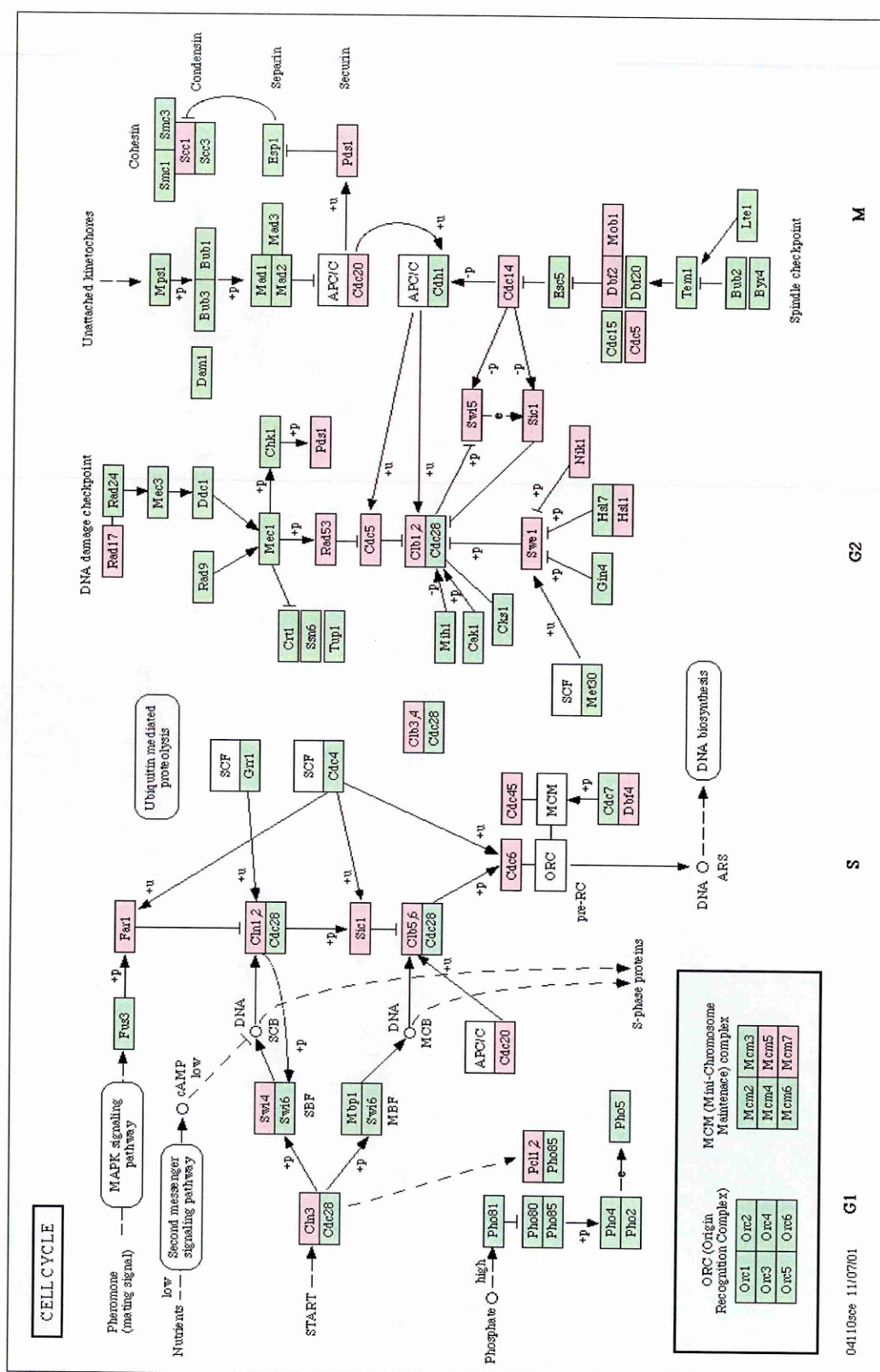


Figure 28: Comparison of KEGG's Cell cycle genetic regulatory pathway of *S. Cerevisiae* and genes identified by traditional methods. The pink colour boxes represents the gene identified traditional methods that is also present in KEGG cell cycle pathway database.

2.3.5 Incorporating prior knowledge

As the gene regulatory network are extracted from data using statistical and computational methods alone, it is vital to validate these networks against literature based on wet lab gene regulation findings as well as the widely available genomic and protein databases. An obvious extension of the analysis presented here would be to add some of the known gene associations in the literature to obtain a richer, more accurate, gene regulatory network. For example, learning gene regulatory network models based on some known gene associations and building the rest of the regulatory network structure around these fixed relationships to strengthen the network e.g. incorporating prior biological knowledge into Bayesian learning.

To demonstrate the application of prior knowledge, we will first consider a GRN extracted purely from data. The dataset considered is 31 genes that are common in both the cell cycle pathway of KEGG (88 genes in total), and 104 genes traditionally identified as cell cycle regulated. The results are presented in **Figure 29**.

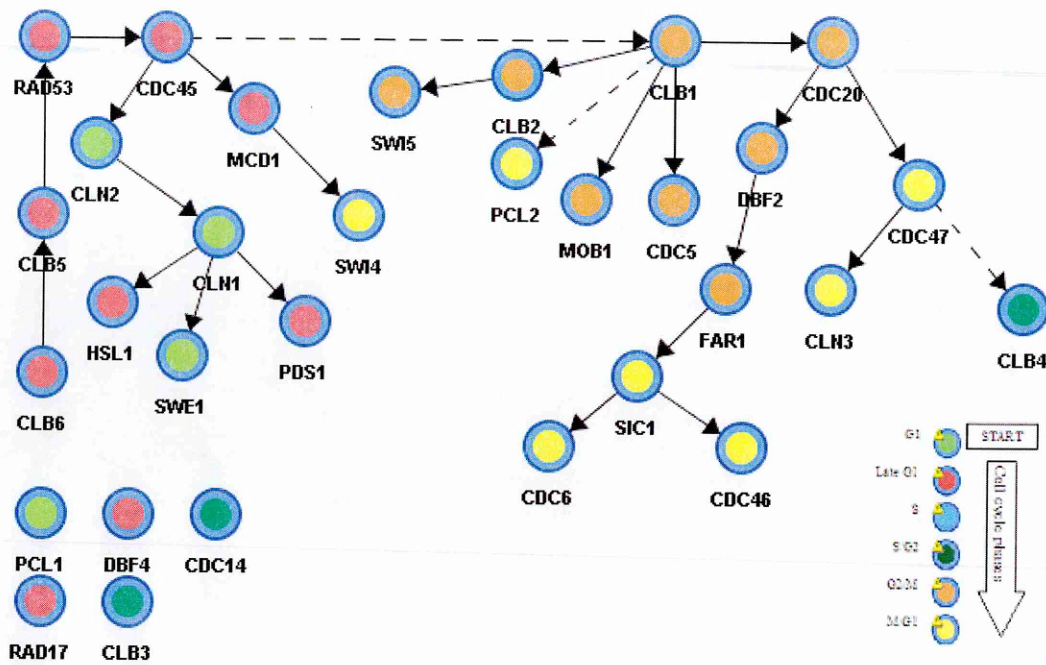


Figure 29: GRN of common genes found in KEGG and yeast cell cycle project. Unsupervised learning of 31 genes that are common in both cell cycle pathway of KEGG and 104 genes traditionally identified as cell cycle regulated using Taboo Search algorithm. The presence of an arc between two genes represent that there is a direct probabilistic relation between them. The relation is identified as positive (solid arc) or negative (dotted arc) i.e. induce or suppress respectively. The Absence of an arc means that gene is independent of all other genes in the network. The cell cycle phases are colour coded according to findings from the traditional methods (<http://genome-www.stanford.edu/cellcycle/data/rawdata/KnownGenes.doc>).

Adding biological knowledge (prior knowledge) to the network means knowing how in cell cycle regulation genes exactly relate. We employed four known relations, SIC1 to CLB5 and CLB6 and CLN1 and CLN2 to SIC1 from KEGG yeast cell cycle pathway database, by fixing the arcs. Purified p40SIC1 inhibits CDC6/CDC28 and CLB5,6/CDC28 interaction in vitro and in vivo. Upon its destruction by programmed proteolysis, cyclin-CDK1 activity is induced. This degradation then triggers the G1-S transition (Lew DJ, 1997; Schwob, Bohm *et al.*, 1994). Phosphorylation of SIC1p by CLN-CDC28p complexes is required for entry into S phase (Andrews and Measday, 1998; Verma, Feldman *et al.*, 1997).

The networks extracted using the Taboo Search algorithm with and without prior knowledge are given in **Figure 30** and **Figure 29** respectively.

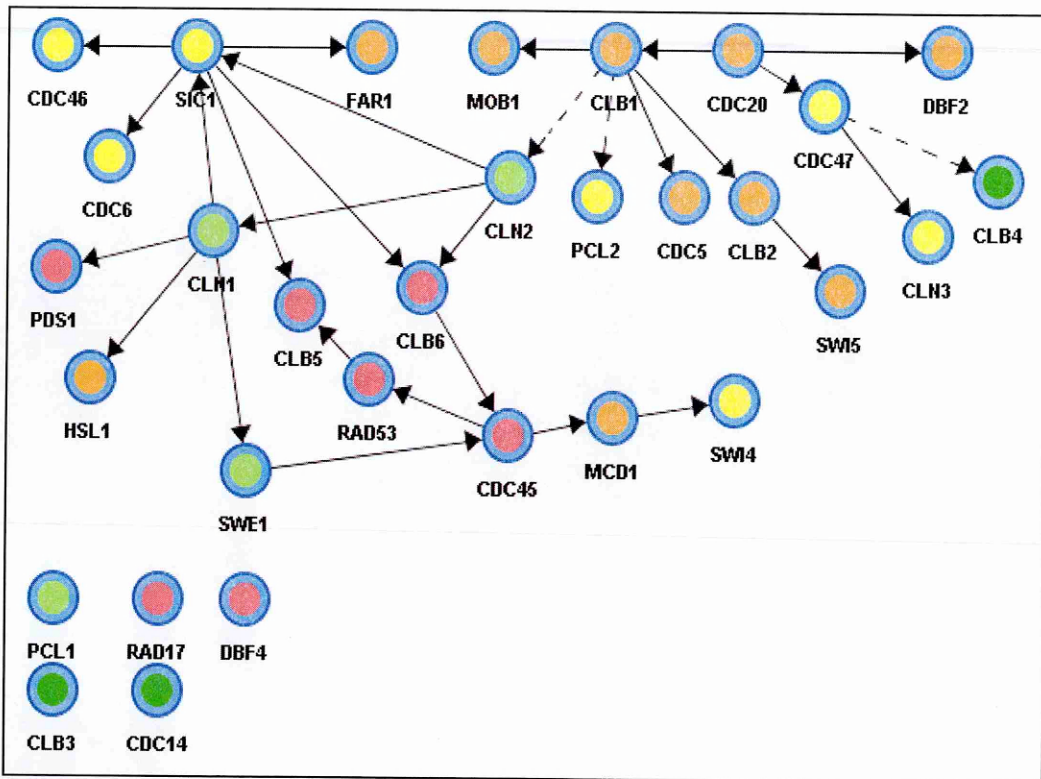


Figure 30: Learning of GRN of common genes of yeast cell cycle project and KEGG after adding biological information. Genes are colour coded according to the cell cycle phases as distinguished by traditional laboratory experiments and it is in yeast cell cycle project site. The lack of an arc means that gene is independent of all other genes in the network. The *direct* probabilistic relation is shown by the presence of an arc. Unless the relation is noted by dotted arc meaning the gene is suppressed i.e. negative and otherwise the relation is between the genes are induced in the direction of the arc i.e. positive. When two genes are associated only through a grandparent gene (no direct arcs), then they are regarded as *indirectly* associated.

Direct relations observed among the 31 genes in **Figure 29** are compared against **Figure 30** (GRN extracted with prior knowledge), and also against the cell cycle pathway of the KEGG database, and published databases and literature. The results are summarised in **Table 11**.

Table 11: Comparison of extracted genetic associations with published information.

Gene Name Gene1 Gene2		Without prior know.	With prior know.	KEGG pathway	Genome databases/ Literature	Reference
FAR1	SIC1	Direct	Direct	Indirect	Indirect	(Tyres, 1996)
CLN2	FAR1	Indirect	Indirect	Direct	Direct	(Valdivesto <i>et al.</i> , 1993)
CLN2	CLN1	Direct	Direct	Complex	Co-regulated	(Wittenberg, Sugimoto <i>et al.</i> ,
SIC1	CDC6	Direct	Direct	Indirect	Indirect	(Calzada, Sacristan <i>et al.</i> , 2001)
SIC1	CDC46	Direct	Direct	Indirect	Indirect	(Sclafani, Tecklenburg <i>et al.</i> ,
CLB6	CLB5	Direct	Indirect*	Complex	Co-regulated	(Smith, Penkner <i>et al.</i> , 2001)
CLB5	RAD53	Direct	Direct	None	None	(Germain, 1997)
CDC45	RAD53	Direct	Direct	Indirect	Indirect	(Jares, Donaldson <i>et al.</i> , 2000)
CDC45	MCD1	Direct	Direct	Indirect	Indirect	(Yen, Gitsham <i>et al.</i> , 2003)
CDC45	CLN2	Direct	Indirect*	Indirect	None	n/a
DBF4	CDC45	None	None	Direct	Direct	(Owens, Detweiler <i>et al.</i> , 1997)
CDC14	DBF2	None	None	Indirect	Indirect	(Visintin, Craig <i>et al.</i> , 1998)
DBF2	MOB1	Indirect	Indirect	Direct	Direct	(Komarnitsky, Chiang <i>et al.</i> ,
CLB1	CLB2	Direct	Direct	Complex	Co-regulated	(Deshaies, 1997)
CLB1	PCL2	Direct	Direct	None	None	(Measday, Moore <i>et al.</i> , 1994)
PCL2	PCL1	None	None	Complex	Complex	(Andrews <i>et al.</i> , 1998)
CLB1	MOB1	Direct	Direct	Indirect	Indirect	(Luca and Winey, 1998)
CLB1	CDC5	Direct	Direct	Direct	Co-regulated	(Althoefer, Schleiffer <i>et al.</i> ,
CLB1	CDC20	Direct	Indirect*	Indirect	Indirect	(Tavormina <i>et al.</i> , 1998)
CDC20	DBF2	Direct	Direct	Indirect	Indirect	(Zachariae <i>et al.</i> , 1999)
CDC20	CDC47	Direct	Direct	Indirect	Indirect	(Spellman, Sherlock <i>et al.</i> , 1998)
CLB3	CLB4	None	None	Complex	Complex	(Lew DJ, 1997)
CLN1	HSL1	Direct	Direct	None	Indirect	(Ciliberto <i>et al.</i> , 2003)
CLN1	SWE1	Direct	Direct	None	Indirect	(Ahn, Tobe <i>et al.</i> , 2001)
CLN1	PDS1	Direct	Direct	None	None	n/a
MCD1	SWI4	Direct	Direct	None	None	(Simon, Barnett <i>et al.</i> , 2001)
CLB2	SWI5	Direct	Direct	Direct	Direct	(Althoefer, Schleiffer <i>et al.</i> ,
DBF2	FAR1	Direct	Indirect*	None	None	n/a
CDC47	CLN3	Direct	Direct	Indirect	Indirect	(McInerny, Partridge <i>et al.</i> ,
CDC47	CLB4	Direct	Direct	None	None	(Chen, 2000)

Table 11 shows that of the 31 relations considered, 27 relations are identical under both GRN without and GRN with prior knowledge. On the rest of the 4 relations (marked by asterisk in the table), we found that all the 4 results have generally been improved by adding prior knowledge in terms of comparison with published results of KEGG and other genome databases and the literature.

A direct comparison (ignoring whether the relationship is direct or indirect) of the extracted GRN with prior knowledge against the KEGG database and other published literature reveals that 21 out of 31 relations (68% of total) are identified in both sets of results. Of the relations identified by the published results, 4 relations, DBF4 and CDC45, CDC14 and DBF2, PCL1 and PCL2, CLB3 and CLB4 were not identified as associated (i.e. ‘independent’) by our GRN. One possible reason for the non-identification of PCL1 and PCL2, CLB3 and CLB4 may be that they are complexes which are co-regulated by other genes and not always present or active during the cell cycle (see the short literature review below). Other possible reasons are the relatively small number (31 genes) of genes considered, the particular choice of arcs fixed for the addition of prior knowledge, strength of the algorithm employed, and the stochastic nature of the data (hidden variables, noise). For example, it is possible that the nature of the expression levels of some genes is such that any association they may form with other genes becomes apparent only when more genes are introduced to the network.

The extracted GRN with prior knowledge identified 6 new relationships, CLB5 and RAD53, CLB1 and PCL2, CLN1 and PDS1, MCD1 and SWI4, DBF2 and FAR1, CDC47 and CLB4, which our searches have shown are not in the cell cycle pathways in the published literature and genome databases. These relationships need to be investigated further through traditional methods. For example, the relationship between CLB5 and RAD53 remains unclear, with the earlier literature of (Germain, Hendley *et al.*, 1997) claiming ‘no association’ while (Meyn Iii and Holloway, 2000) concludes that the evidence is unclear (see the short literature review below).

Finally, a question remains whether the network could be improved upon by the choice of genes that have higher association in terms of their conditional dependencies and/or by choosing more dominant genes of the cell cycle process (in terms of dominating the order of relations), other than those employed above. While SIC1, CLN1 and CLN2, and CLB5 and CLB6 all have core biological functions in the cell cycle process with significant interrelationships, it is possible that an alternative set of genes may have yielded improved results.

Also it is crucial to consider the marginal probability and conditional probability of a relation when deciding which arc to fix. Fixing the arc between two nodes that has a weak relation statistically but biologically important and known, will improve the overall structure by increasing the conditional probability of other genes that are associated with it, and the opposite is true when fixing the arc of non-existing or unknown relation. Efforts should also be made to any fix relations that known to be false positive (i.e. relations that appear on the extracted GRN by chance), as identified by biological knowledge.

Another issue is the choice of optimal of number of relationships to be fixed without introducing biasness into the resulting network. While we are confident that the number of prior relationships used (arcs fixed), 4, is fairly reasonable in a GRN of 31 genes, extending the network to accommodate more genes require careful consideration of the above issues.

The evidence in current literature relating to the nature of associations identified in our GRN with prior knowledge (see **Figure 30**, **Table 11**) is summarised in Appendix I –. The first 5 relations listed in **Table 11** are shown below for illustration.

FAR1 and SIC

Both FAR 1 and SIC1 are CDK inhibitors. Substrates of the CDC28-CLN1 CDK are FAR1p, SIC1p, and the so far not identified proteins p41 and p94 (Andrews and Measday, 1998). FAR1, a pheromone-activated inhibitor of CLN-CDC28 kinases, is dispensable for arrest of CLN⁻ SIC1 cells by pheromone, implying the existence of an alternate FAR1-independent arrest pathway (Tyers, 1996).

CLN2 and FAR1

The FAR1 gene is required for cell cycle arrest but not for a number of other aspects of the pheromone response. Genetic evidence suggests that FAR1 is required specifically for inactivation of the G1 cyclin CLN2 (Valdivieso, Sugimoto *et al.*, 1993). Phosphorylation by the MAP kinase homolog FUS3p may help regulate FAR1p's association with the CDC28p/CLN2p complex (Peter, Gartner *et al.*, 1993).

CLN1 and CLN2

CLN1 and CLN2 are closely related genes with overlapping functions; both are expressed in late G1 phase when they associate with CDC28p to activate its kinase activity. Accumulation of CLN1 and CLN2 mRNA in late G1 is dependent on two transcription factor complexes, MBF (SWI6p-MBP1p) and SBF (SWI6p-SWI4p), which bind to MCB and SCB promoter elements, respectively. In addition, CLN3p has been shown to activate CLN1 and CLN2 transcription (Wittenberg, Sugimoto *et al.*, 1990).

SIC1 and CDC6

Several striking similarities exist between SIC1 and CDC6, a DNA replication factor essential for the formation of pre-replicative complexes at origins of DNA replication. Transcription of both genes is activated during late mitosis by a process dependent on SWI5.

Like SIC1, CDC6 binds CDK complexes *in vivo* and down regulates them *in vitro*.

CDC6, like SIC1, also contributes to inactivation of CDKs during late mitosis in *S. cerevisiae*. Deletion of the CDK-interacting domain of CDC6 does not inhibit the function of origins of DNA replication during S phase, but instead causes a delay in mitotic exit; this delay is accentuated in the absence of SIC1 or of cyclin degradation (Calzada, Sacristan *et al.*, 2001).

SIC1 and CDC46 (MCM5)

Deletion of SIC1, which is an inhibitor of CLB5p (Schwob, Bohm *et al.*, 1994), resulted in an increase of MCM5-BOB1 suppression, as the SIC1 MCM5-BOB1 CDC7ts colonies grew faster at the restrictive temperature. However, this increase of MCM5-BOB1 suppression could not compensate for the loss of CLB5, as the MCM5-BOB1 CDC7ts CLB5 SIC1 strain remained Tsm-. We conclude that CLB5p and CLB2p are required for the MCM5-BOB1 mutation to suppress CDC7ts mutations (Sclafani, Tecklenburg *et al.*, 2002).

2.4 Discussions

As noted at the start of this chapter, traditional laboratory techniques have identified that 104 genes are directly involved in cell cycle regulation in the yeast genome. However it was estimated by Price *et al.*, (1991) that more than 250 genes are cell cycle regulated. Since then Spellman *et al.* (1998) identified 800 genes in the yeast genome that have peak expression levels once per cell cycle period.

Spellman *et al* (1998) applied a Fourier transform algorithm to each gene data series to assess their periodicity, and Pearson correlation was used to correlate the expression profile of each gene to five different profiles (phases) representing genes known to be expressed in M/G1, G1, S, G2, and M. In deciding the significance of the cell cycle regulation of the genes in the genome, they compared the Fourier transform scores among previously known and total gene sets to set a threshold. The identification threshold was set at a level exceeded by 91% (95 of 104) of the genes known to be cell cycle regulated.

When interpreting computationally extracted genome-wide gene regulatory networks, one of the most difficult tasks is distinguishing variables that are cell cycle regulated from those that are not. Some genes are co-regulated, some are regulated periodically or under certain conditions, and some genes are only expressed once. Genes that are co-regulated or periodically regulated might be regulated by a handful of other genes. Thus, in reality, the complexity of gene regulatory network requires filtering or pre-processing of genes using biological knowledge before the final analysis. For example, measuring the gene expression profiles and comparing with them known thresholds, as done by Spellman *et al.* (1998).

Alternatively if the BN algorithm could handle continuous data, we could build in the tools necessary for pre-processing within the algorithm. For example, gene expression profiles can be filtered out that have multiple peaks and a threshold criteria for identifying cell cycle regulating genes could be implemented. It would also provide the scope for further classifying over and under expressed genes. However, using continuous variables would have meant assigning a pre-defined distribution to the variables, which is dependent on the assumed parameter for the dataset such as its mean, for example, linear Gaussian.

We did not consider all 800 genes identified by Spellman *et al.* (1998) because, apart from the lengthy processing time involved in extracting the BN, for our purposes, $n = 800$ is too big.

In consideration of the above, our analysis focussed instead on identifying the gene regulation within and between phases of cell cycle of known cell cycle regulated genes (104). We applied a Bayesian probabilistic learning algorithm to extract the underlying gene network using mRNA microarray expression data of Spellman *et al.* (1998). We further compared the extracted gene regulatory network with the known cell cycle phases to establish whether the gene profiles of each phase matches.

Interpretation of genome wide, microarray high-throughput biological data requires knowledge of the principles underlying the regulatory networks that sustain biological functions of cells. Of particular importance is the genetic network, a set of genes that interact through directed transcriptional regulation. Genes that exert a regulatory role code for dedicated transcription factors, sometimes referred to as regulatory proteins that can bind to specific DNA control regions of regulated genes to activate or inhibit their transcription. Regulated genes may themselves act in a regulatory manner, in which case they participate in a causal pathway. Looping pathways form feedback circuits. Genes that encode for proteins that are involved in many different pathways have several associations and cross-links between other pathways. Thus, in different articles, these genes are often reported to be biologically significant or have a role in many different processes in different pathways making them difficult to validate.

The post transcriptional mechanisms that control the basic timing of the cell cycle, the transacting factors also play a critical role in regulating the production of transcripts. Many cell cycle-regulated genes are involved in processes that occur only once per cell cycle. Such processes include DNA synthesis, budding, and cytokinesis. Additionally many of these genes are involved in controlling the cell cycle itself, although in most cases it is unclear whether their regulated transcription is absolutely required. The cell division cycle is a complex self-regulating program, such that many genes involved in aspects of the cell cycle are also controlled by it.

2.4.1 Issues and limitations

The cell cycle data chosen for this project has been used in other projects e.g. yeast cell cycle project. Therefore, the results obtained are comparable and we are able to validate our findings.

As discussed earlier the continuous data would give more information about the data itself compared to discrete data i.e. a gene that is expressed four fold is discretised as “Up” and a gene two fold is also discretised as “Up”, similarly for under expressed genes. However, certain genes are only expressed when the expression level of genes that they depend on directly or indirectly is very highly expressed e.g. four fold and not when expression level is two fold. In our discretisation, all the expression levels above zero are classed as upregulated, “Up” and all the genes with expression level below zero are classed as down regulated, “Down”. Therefore, the extracted models have to be tested for robustness of the results to various discretisation thresholds.

The problem with continuous data is the inability to identify each gene expression data series. For example, one set of gene data may follow a linear Gaussian (Normal) distribution, another might be non-linear, and a third series may well be best represented by a discrete distribution. Furthermore, these are at best approximations of the underlying data series. So the learning becomes more complex and time consuming whilst the results also require careful interpretation.

One of the central issues faced in validating results is the complexity of the network pathways which makes it difficult for scientists to cross-reference each gene and the discrepancies of updated information across the genome databases. For example, the cell cycle pathway from KEGG is not up to date and only has a portion of the genes in the cell cycle pathway. Also the difficulty in integrating GRNs extracted using alternative methods is an issue that needs to be addressed for improving the learned networks.

As discussed earlier, the use of continuous data would avoid the loss of information from using discretisation thresholds. However, when variables are real valued, there is no single representation that can represent all possible densities. Therefore, use of continuous distributions (e.g. Gaussian, Gamma) for modelling gene expression data is at best an approximation of the underlying data series.

One natural choice for representing multivariate continuous distributions is the use of linear Gaussian (Friedman *et al.*, 2000). Using this representation, Friedman *et al.*, (2000) extracted Bayesian networks, which were reasonably close to those obtained by discrete models. However, as they acknowledged, unlike the discrete models, the linear Gaussian model can only detect dependencies that are close to linear. In particular, it is not likely to discover combinatorial effects (e.g. a gene is over expressed only if several genes are jointly over expressed, but not if at least one of them is not over expressed). Non-linear models have been proposed as a remedy to this problem. For example, Imoto *et al.*, (2003) proposed a statistical model based on Gaussian distribution to handle non-linear as well as linear dependencies between genes. Nevertheless, non-linear as well as dynamic modelling of gene expression data is still relatively at their infancy, and thus presents considerable scope for improvement.

2.4.1.1 Gene Regulatory Networks

It is only in the past few years that the modelling and extraction of gene regulatory networks has become widely employed in analysing gene expression data. This is due to the development of mRNA microarray techniques. However, expression profiling can only reveal gene expression patterns of the experimental conditions studied. But a genetic network map cannot be derived from expression data alone because the transcript data are the product of a large number of variables such as DNA-protein interactions, protein-protein interactions, binding site information, and also noise in the data.

The publicly available GRN information is derived from wet lab experiments and literature. This is not an easy process, as all previously published genetic associations need to be searched through. Moreover the complexity of GRNs of all biological processes makes it difficult for scientists to cross-reference each gene in interactions of genes between regulatory pathways. The KEGG and MIPS interactive pathway databases do hold such data, but are not well updated. In addition, biochemical and signalling pathways are not incorporated well in genetic regulatory pathways. For example, 104 yeast cell cycle regulated genes were discovered by traditional methods before 1998 (Spellman, Sherlock *et al.*, 1998), but only 31 genes are found in the GRN of the yeast cell cycle in KEGG. In MIPS only 20 genes out of these 104 are shown to be cell cycle regulated. This in part, is due to a lack of cross validation and updating of genome databases.

In order for the computationally extracted gene network (such as the BN in this project) to be compared, all established associations among genes, as well as any speculated genetic associations should be included.

The computationally generated GRN model identifies associations/dependencies among genes and also very importantly, the mini clusters within each cell cycle.

However it should be remembered that this GRN was extracted without any prior knowledge. As a result, the gene dependency, e.g. gene A is dependent on gene B or gene A inhibits gene B, is not well characterised. Hence the identified relations between these cell cycle regulated genes are simply called associations. Some of the specific genetic associations shown in our GRN are compared with the KEGG cell cycle pathway and Spellman *et al.* (1998) and MIPS and discussed below.

2.4.2 Yeast Cell Cycle Phases

The eukaryotic cell division cycle generally consists of two characteristic states, G1 and S/G2/M. In G1 replication begins, chromosomes are in pre-replicative state, and in S/G2/M chromosomes are in post-replicative state. However the entire process of replication is initiated by activation of one gene if not a few.

The gene activation in eukaryotes requires chromatin remodelling complexes like SWI/SNF, and histone acetylases like SAGA (Cosma *et al.*, 1999). The underlying processes of gene activation are however unknown. There are a number of ORFs either on the same or opposite strand that overlap. It is unclear whether their co-regulation is because they are simply one big ORF. The ORFs are YLR465C, YLR464W, YLR463C, YLR462W, YLR466W, and YLR466C (Sherlock, 2000). None of these ORFs were however, are in the 104 cell cycle regulated genes.

A more detailed discussion of individual cell cycle phases are given in Appendix I.

2.5 Conclusions

When all 104 genes were analysed together, the resulting network showed strong similarity to the actual classification of the phases of yeast cell cycle. Nodes from a particular phase linking nodes from different phases have revealed a consistency with associations with consecutive phases in cell cycle. The clustering methods carried out in Spellman *et al.* (1998), did not show how genes are associated in clusters or which genes from different phases link to other phases, i.e. which genes the next phase of the cell cycle is dependent on. The findings of this project therefore represent a step forward in relation to Spellman's work.

The application of the methodology developed in this chapter and the findings are summarised below.

- The Bayesian probabilistic framework was employed using the SopLEQ, Taboo Search and Taboo Order Search algorithms to identify the associations and dependencies between genes known to be cell cycle regulated in yeast. These novel algorithms produced highly confidence results, learned entirely statistically and computationally from datasets containing gene expression levels. As demonstrated, these can be improved upon by adding prior biological knowledge from a variety of other sources.
- The derived gene regulatory networks capture many biologically plausible genetic associations between known cell cycle regulated genes, which have been supported by evidence from published literature and genome databases. The order relations of the learned network show remarkable consistency in terms of how certain genes tend to dominate the order relations (i.e. they appear before other genes) in the extracted Bayesian network.

- This methodology allows for better validation of results learned from data using statistical techniques alone, through the inclusion of known associations in the published literature and genome databases. Thus the validation set includes a wider, richer spectrum of results in the literature, obtained using techniques ranging from wet lab methods e.g. (Andrews and Measday, 1998; Gardner and Burke, 2000; Lew DJ, 1997; Measday, Moore *et al.*, 1994; Schwob, Bohm *et al.*, 1994), Fuzzy k-means clustering (Gasch and Eisen, 2002), clustering using a manual decision process (Cho, Campbell *et al.*, 1998), clustering using statistical techniques (Spellman, Sherlock *et al.*, 1998), ANOVA (Kerr, Martin *et al.*, 2000), differential equations (Orr and Scherf, 2002; Scherf, Ross *et al.*, 2000), Boolean networks (D'Haeseleer, Liang *et al.*, 2000), and Bayesian learning (Barash and Friedman, 2001; Drawid and Gerstein, 2000; Friedman, Linial *et al.*, 2000; Lu, Zhang *et al.*, 2004; Murphy and Mian, 1999). The genome databases used for the validation were KEGG, SGD and MIPS.
- These models proved robust to variations in data. For example, the GRNs extracted under individual cell cycle phases such as M/G1, G1 were largely preserved when the data for these two phases were combined together. The new GRN displayed the two sets of genes as two main clusters within the overall structure, with a relatively small degree of overlap. Some overlap can be expected given the fact that the separation of the different cell cycle phases (both traditional methods and Spellman *et al.*, 1998) are based on rather crude cut-off points and the data is subject to noise.
- The target node analysis employed in this study is a novel approach for extracting knowledge hidden within a GRN. In particular it gives the opportunity to visualise the relationships within a GRN: how the gene of interest influences the network given the gene expression data. This feature could be used for further data analysis in gene expression studies, for example, how protein expression (transcription factor) and gene expression data are linked for each gene and also for answering specific questions, such as how important a

particular gene is in a GRN and how other genes in the network influence one gene. Another useful tool we applied is network simulation, which allows us to observe changes in the marginal probabilities of genes in response to user defined interventions. Such a causal interpretation helps determine the effect of an *intervention*, simulated by manually setting the state of a node outside the causal model.

- Validation of the individual associations among ORFs against known relationships described in the literature show that the Taboo Order Search algorithm is the best for discovering genetic relationships among groups of genes, compared to the SopLEQ and Taboo Search algorithms.

The approach developed in this study can be easily and efficiently applied to learning a larger scale network. This is a sensible approach, as at this stage, a number of genes in yeast that are cell cycle regulated are yet to be conclusively identified. For example, (Lu, 2004), using an empirical Bayesian method to identify genes that are periodically expressed (PE), identified 822 genes whose posterior probabilities of being PE are greater than 0.95. Among these 822 genes, only 540 are in the list of 800 genes detected by Spellman *et al.*, 1998.

From the results obtained, it is clear that the methods developed in this chapter can be applied to other data sets for which the gene regulatory network is not known, thereby generating new knowledge. In this chapter the reliability and the usefulness of the approach of using the Taboo Search, and Taboo Order Search algorithms has been demonstrated. In the following two chapters these techniques are applied to two large scale, novel post genomic microarray datasets from *E. coli*, and human genomes.

3 Genetic Network of Human Adipocyte Differentiation

3.1 Introduction

This chapter describes the application of Bayesian GRN extraction techniques developed in Chapter 2 to a novel human adipocyte differentiation gene expression dataset that had only three time points, raising the mathematical issue of high dimensionality. This was tackled by novel pre-processing methods to reduce data dimensionality including generation of artificial data points. Furthermore, gene regulatory networks of the insulin receptor (IR) family, PPAR family, FOX family, CEBP family, MEF2, FABP, ADD1 and KLF, and probes with highly significant change in gene expression level were learned separately. The extracted networks were validated against publicly available information as well as in house interaction and literature databases available from the originators of the data.

The dataset employed in this chapter is a novel dataset, human multipotent adipose-derived stem (hMADS) (Rodriguez, Elabd *et al.*, 2004), from Glaxo Smith Kline (GSK). hMADS cells, isolated from human adipose tissue, are able to respond to drugs under inhibition of adipocyte differentiation, emphasising the potential of these cells for drug screening for the treatment of obesity and other nutrition based disorders. Thus, they represent a novel and valuable tool for studies of human fat tissue development and metabolism (Rodriguez, Elabd *et al.*, 2004).

In this chapter, the issues involved in dimensionality reduction are introduced and briefly discussed, and the use of various pre-processing methods that can be employed for extracting genetic networks when handling data from large scale microarray experiments is demonstrated. The pre-processing methods such as fold change analysis, filtering, clustering, and generation of artificial experimental data were employed to the hMADS dataset to obtain a data matrix that is reduced in dimensionality, statistically sound and one that achieved good predictive performance. Bayesian belief networks were then used to build models for the genetic dependency in biological processes that was captured in the microarray experiments.

In the final section, the relationships captured in the extracted genetic networks are considered, including any novel associations, and their plausibility discussed in terms of validation against many publicly available interaction and literature databases as well as against in-house interaction and literature databases available at GSK.

3.2 Human Adipocyte Differentiation

Advances in molecular biology have highlighted the central role of adipocytes in the development of obesity and other nutrition based disorders. Adipocytes, by virtue of their excellent and sensitive molecular machinery, seem to reflect nutritional alterations very precisely (Erol, 2005). Thus the use of microarray data of adipocytes development is an ideal to study obesity and nutritional based disorders.

Adipogenesis occurs throughout the lifetime of an organism, both in response to normal cell turnover as well as in response to the need for additional fat mass stores that arises when caloric intake exceeds nutritional requirements. **Figure 31** shows the stages of adipogenesis: from egg to mature adipocyte.

The development of obesity involves extensive adipose tissue remodeling at the cellular level and is dependent on the coordinated interplay between adipocyte cell size increase (hypertrophy), increase in adipocyte cell number (hyperplasia), and growth of new blood vessels (angiogenesis) (Hausman, DiGirolamo *et al.*, 2001; Smas and Sul, 1995).

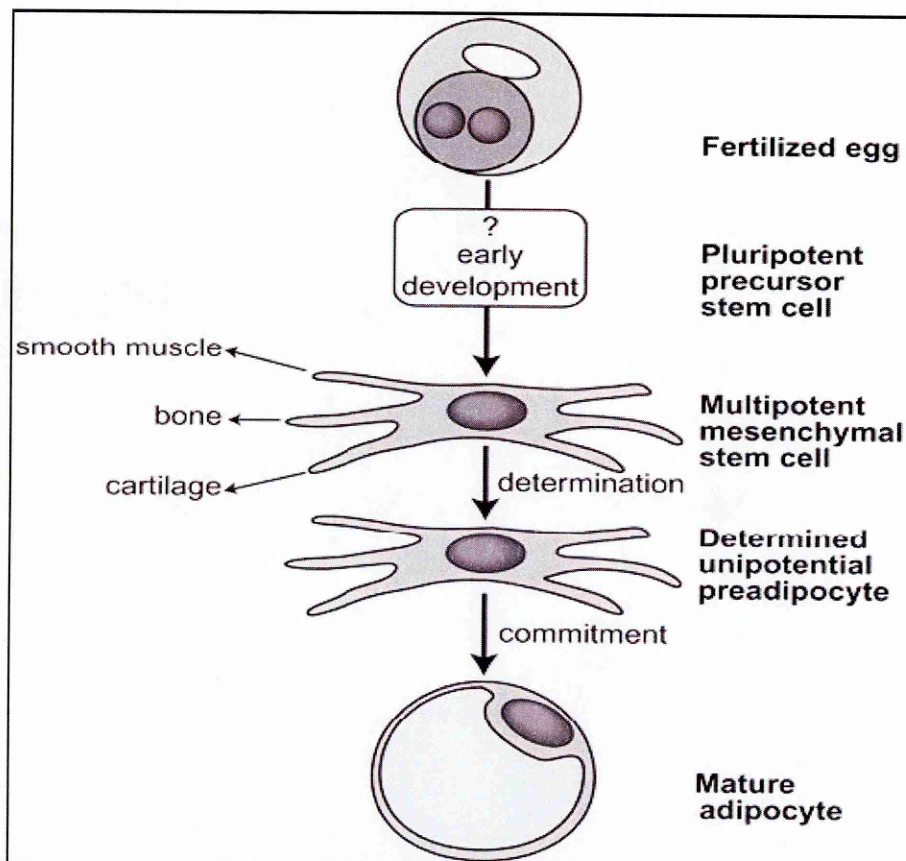


Figure 31: Adipogenesis. The development of mature adipocytes from a fertilized egg. Multipotent mesenchymal stem cells differentiate into cartilage, bone, muscle and adipose tissue. The Dark ovals and circles represent nuclei and pronuclei, respectively. Reproduced from (Avram, Avram *et al.*, 2007).

Adipogenesis occurs throughout the lifetime of an organism, both in response to normal cell turnover as well as in response to the need for additional fat mass stores that arises when caloric intake exceeds nutritional requirements. **Figure 31** shows the stages of adipogenesis: from egg to mature adipocyte.

The development of obesity involves extensive adipose tissue remodeling at the cellular level and is dependent on the coordinated interplay between adipocyte cell size increase (hypertrophy), increase in adipocyte cell number (hyperplasia), and growth of new blood vessels (angiogenesis) (Hausman, DiGirolamo *et al.*, 2001; Smas and Sul, 1995). The mechanism for increased fat mass in obesity was attributed exclusively to adipocyte hypertrophy, that is, the process by which pre-existing fat cells increase in size due to an excessive accumulation of lipids from imbalanced energy intake over energy expenditure (positive energy balance). It is now understood that obesity is associated with both increased fat cell size and number within adipose tissue (Hausman, DiGirolamo *et al.*, 2001; Hirsch and Batchelor, 1976). Studies of adipocytes and their differentiation processes have made a major contribution to the understanding of adipose tissue development and regulation. The sequence of events leading to adipocyte differentiation as well as the role of major transcription factors, i.e. by members of CCAAT enhancer binding protein (C/EBP) and peroxisome proliferator-activated receptor (PPAR) families and adipocyte determination and differentiation factor-1/sterol response element binding protein (ADD1/SREBP1c) (Rodriguez, Elabd *et al.*, 2004) and, adipogenic markers in fat cell differentiation and regulation are important in identifying the associations of genes involved in this process.

Adipogenesis involves two major events, the recruitment and proliferation of preadipocytes followed by their subsequent differentiation into mature fat cells (Gregoire, 2001; Hausman, DiGirolamo *et al.*, 2001). Proliferation refers to the process by which preadipocytes replicate so as to increase fat cell number, whereas differentiation refers to the process by which undifferentiated, proliferating fibroblast-like preadipocytes become permanently cell cycle-arrested, spherical, lipid filled and functionally mature fat cells (Hausman, DiGirolamo *et al.*, 2001). Differentiation is accompanied by alterations in cell shape as well as by molecular changes that lead to

dramatic increase in the ability of the cell for lipid synthesis and increases in hormonal responsiveness specific to the specialized role of the adipocyte in energy homeostasis (Ailhaud, Grimaldi *et al.*, 1992; Gregoire, 2001; Smas and Sul, 1995).

During embryonic development of adipose tissue, the cell-fate of pluripotent precursor stem cells is at first restricted by largely unknown mechanisms to multipotent mesenchymal stem cells (MSCs). Of mesodermal origin, MSCs can differentiate into a number of tissues including adipose tissue, cartilage, bone and muscle tissue (Pittenger, Mackay *et al.*, 1999).

3.3 Reduction of Data dimensionality

Gene-expression data can be thought of in terms of multidimensional space, in which expression vectors are represented as points in that space. High data dimensionality has been a huge problem for microarray data analysis and since the technology was first proposed. However, in recent years, it has become ever more challenging task as more sophisticated and processor-intensive statistical modelling, knowledge extraction and data mining techniques such as Bayesian network learning are applied. The problem is essentially due to the overabundance of data in terms of the number of variables available for each sample in the dataset. This overabundance appears frequently in dependent-independent modelling such as gene regulatory or genetic networks, that is, when one wishes to model the dependence of a response (dependant) variable Y on one or more predictor (independent) variables $X_1, \dots, X_n$. For $n > 2$, this becomes almost impossible to distinguish by eye without allowing some form of dimension reduction. Often the expression of individual genes can also be highly correlated.

This challenging aspect of data management is the so called “*curse of dimensionality*”, which makes it difficult to extract useful information from multidimensional data (Berchtold *et al.*, 1996). Therefore often it is necessary to make low dimensional projections of the data prior to analysis.

Dimensionality reduction is a promising approach because it extracts the most important components of the data and allows for lower complexity in subsequent processing, and also facilitates visualisation of clustered data (Epps and Ambikairajah, 2005). A further argument for dimensionality reduction is, when reducing a sample covariance matrix of n data vectors in a space of p dimensions where the number of variables, p , far exceeds the sample data generated, n , difficulties can arise in the form of data scarcity. This scenario, known as the “ $p > n$ ” problem, can often cause singular estimates of the within-cluster covariance matrices (McLachlan, Bean *et al.*, 2002). Reducing the dimension, p , is one means of mitigating this problem.

One can reduce dimensionality of the data by selecting a subset of interesting genes (*gene selection*), or producing gene components or combinations of related genes (*dimension reduction*), or using a combination of the strategies (Dai, Lieu *et al.*, 2006). Selection of relevant genes for sample classification is a common task in most gene expression studies, where researchers try to identify the smallest possible set of genes that can still achieve good predictive performance. However, this does negate some of the benefits of the large scale genomic coverage afforded by microarrays.

Many gene selection approaches use univariate (gene-by-gene) rankings of gene relevance and arbitrary thresholds to select the number of genes. These include *fold change* (DeRisi, Penland *et al.*, 1996), *unusual ratio* (Schena, 1996), *simple t-test* (Nguyen and Rocke, 2002), and *variable fold-change* method (Mariani, Budhreja *et al.*, 2003).

Gene components can be constructed using multivariate techniques with the premise that, although the microarray data covers numerous genes, there may actually be a small number of underlying variables that account for most of the variation in the data (West, Blanchette *et al.*, 2001). For example, a few linear combinations of genes may explain most of the response variation. Examples of multivariate techniques include *Principal Components Analysis (PCA)* (Datta, 2003; Misra, Schmitt *et al.*, 2002; Raychaudhuri, Stuart *et al.*, 2000), *Independent Component Analysis (ICA)* (Liao, Dasgupta *et al.*, 2002), *Partial Least Squares (PLS)* (Boulesteix, 2004; Nguyen and Rocke, 2004), *Sliced Inverse Regression (SIR)* (Li, 2000), *Discrete Cosine Transform (DCT)* (Epps and Ambikairajah, 2005), and *ANalysis Of VAriance (ANOVA)* (Kerr and Churchill, 2001). Each approach has its own advantages and limitations (Boulesteix, 2004), and a combination of the strategies is often used in practice for classification with gene expression data (Dai, Lieu *et al.*, 2006).

An alternative method for dimension reduction is to classify genes with similar expression profiles into groups, known as *clustering* (Eisen, Spellman *et al.*, 1998). A wide variety of clustering algorithms are employed for this purposes. These include *Hierarchical clustering* (Kohane, Kho *et al.*, 2003), *k-Means clustering* ((DeRisi, Iyer *et al.*, 1997); (Tavazoie, Hughes *et al.*, 1999)), *Self-Organizing Maps (SOM)* (Tamayo, Slonim *et al.*, 1999), *CLICK algorithm* (Sharan and Shamir, 2000; Shamir and Sharan, ; Hartuv and Shamir, 1999), and *Gaussian clustering* using the *ArrayMiner2 algorithm* (Falkenauer and Marchand, 2001).

Some of the most widely used data reduction techniques are considered in detail in the next sections.

3.3.1 Gene Selection Methods

Fold change is the simplest and most intuitive method for identifying differentially expressed genes (Draghici, 2002). Initial methods for interpretation of changes in expression microarray data used a simple fold-change statistic (DeRisi, Penland *et al.*, 1996). It is based on the observed ratio (or average of ratios) between two conditions. An arbitrary cut-off value, commonly 2 to 3 fold, is used to identify differentially expressed genes (Draghici, 2002). The fold change method is subject to bias if the data are not properly normalized and may also be sensitive to heterogeneity in variance across genes.

For example, an excess of low intensity genes may be identified as being differentially expressed due to an excess of variation relative to high intensity genes. The error rates can however be considerably reduced by replicating the experiments. Nevertheless, there are significant statistical limitations to the analysis of replicate data. For instance, multiple pair-wise comparisons of replicate experiments by using a simple fold-change threshold is a significant improvement, but it is still subject to unacceptable error rates (Tusher, Tibshirani *et al.*, 2001).

Another possible approach is to use univariate statistical tests (e.g. the t-test) to select differentially expressed genes (Fox and Dimmic, 2006). If log-ratios (of sample gene expression to its reference) follow a normal distribution with mean μ and standard deviation σ as in **Figure 32**, for a given threshold, and a given distribution, the confidence level or *p*-value is the probability of the measured value being in the shaded area by random chance. The idea is that a gene whose log ratio falls in the shaded area far from the mean log ratio will be classified as differentially regulated (up regulated in this instance).

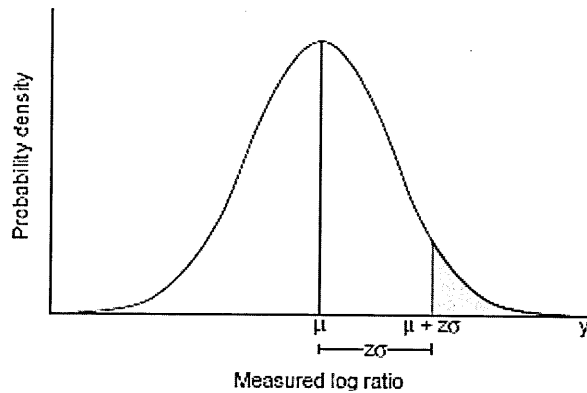


Figure 32: Identifying differentially expressed genes using a t-test. Under this test the log-ratios of gene expression is assumed to follow a normal distribution with mean μ and standard deviation σ . The genes whose log ratios fall under the shaded area ($\mu + z\sigma$, where z is the number of standard deviations) are classified as differentially expressed.

However, such tests applied to expression microarray data are also subject to considerable false discovery rate (FDR)²⁶, and can be effective only if the number of genes on the array is relatively low, i.e. a few hundred (Draghici, 2002).

Intensity specific thresholds, such as the *variable fold-change* method by (Mariani, Budhraja *et al.*, 2003), has been proposed as a remedy for this problem. The *variable fold-change* method recognises that gene expression levels exhibit non-uniform variability at different signal intensities: replicate variability is much higher than commonly used thresholds (two to three fold change) at low intensity, and lower than commonly used thresholds at high intensity. Thus, the method uses a scatter plot smoothing technique called LOESS to assign confidence (p -value) to an intensity-specific fold-change, allowing the calculation of a variable fold-change threshold for any baseline intensity at any given p -value. Hence, for any pair of absolute expression values (x, y), a fold change and p -value are calculated as shown in **Figure 33**.

²⁶ The False Discovery Rate (FDR) of a set of predictions is the *expected* percentage of false predictions in the set of predictions. In this instance the likelihood of such false discovery is given by the p -value (Type 1 Error)

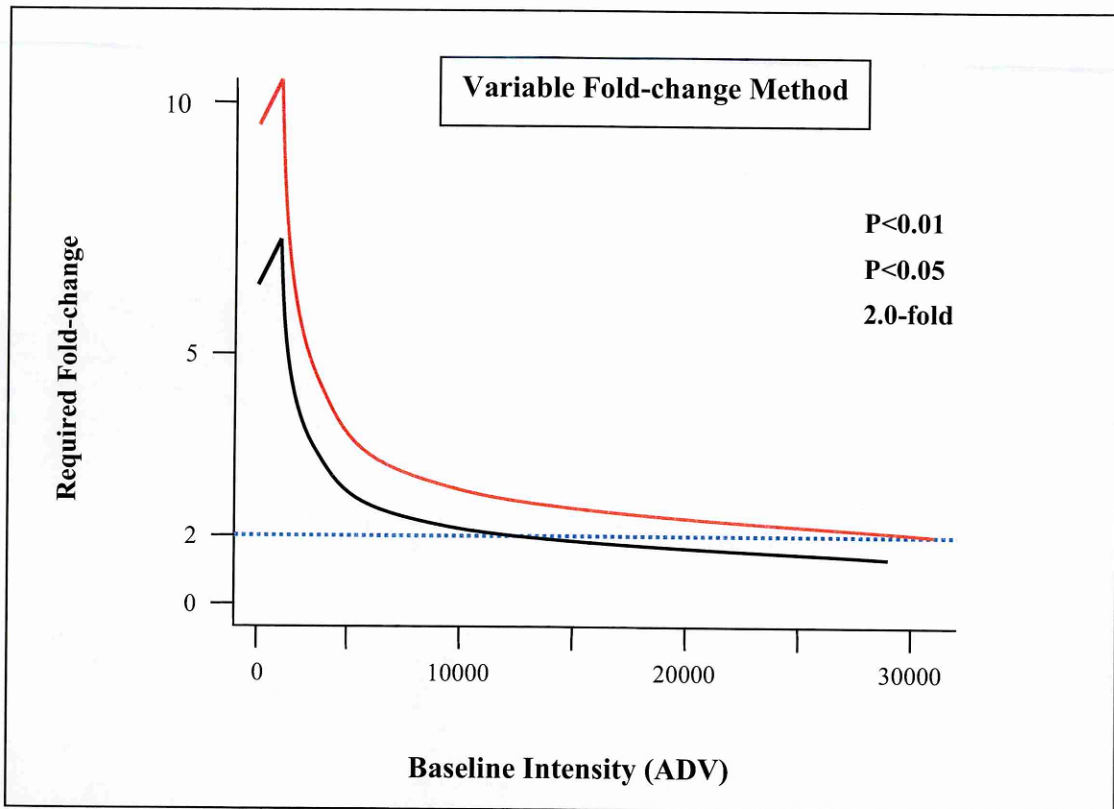


Figure 33: Calculating the variable fold-change. The change in intensity and the fold-change necessary to provide 95% and 99% confidence, that intensity values are indeed significantly different. The base line intensity, calculated as the Average Difference Value (ADV) of the signal intensities, was computed using Affymetrix MAS 4.0 software. Reproduced from (Mariani, Budhraja *et al.*, 2003).

Using a data set²⁷ generated from Affymetrix human U95 GeneChip® set, (Mariani, Budhraja *et al.*, 2003) showed that the variable fold-change threshold exhibited superior false-positive detection rate and removal of intensity-specific bias in determining differential expression compared to the static fold-change method.

²⁷ This dataset consisted of five replicate hybridizations of a single labelled cDNA target from three distinct experimental paradigms.

3.3.2 Dimension Reduction Methods

A popular dimension reduction technique is *Principal Components Analysis (PCA)*. The central idea of PCA is to reduce the dimensionality of a dataset which consists of a large number of interrelated variables by transforming it into a substantially smaller set of uncorrelated variables, while preserving the information contained in the original data. PCA is not a clustering technique; however, projection to lower dimensional spaces can reveal and help visualize the presence of clusters in the data. This method has previously been used with some success on various microarray datasets (Datta, 2003; Misra, Schmitt *et al.*, 2002; Raychaudhuri, Stuart *et al.*, 2000).

Unlike the PCA method, where dimension reduction is achieved by constructing linear combinations that maximize the variance of a linear combination of the original predictor variables, *Partial Least Squares (PLS)*, uses an objective function based on covariance between the response variable matrix Y and the original predictor variables in matrix X . An example of this application would be the prediction of tumor types (response variable) using microarray gene expression data of tissue samples (predictor matrix). The idea of PLS is to estimate w and q by regression, where w and q are the weights of the linear functions of X , Y respectively. Specifically, PLS fits a sequence of bilinear models by least squares, hence the name partial least squares.

Like PCA, PLS reduces the complexity of microarray data analysis by constructing a small number of gene components, which can be used to replace the large number of original gene expression measures. Moreover, since they are obtained by maximizing the covariance between the components and the response variable, the PLS components are generally more predictive of the response variable than the principal components. PLS is computationally very efficient; the number of calculations required by PLS is a linear function of n and p . Thus it is much faster than the other two methods (PCA and SIR) (Dai, Lieu *et al.*, 2006).

Sliced Inverse Regression (SIR) (Li, 2000) is a supervised approach, which utilizes response information in achieving dimension reduction. The idea of SIR is simple: conventional regression models use a forward regression function, $E(Y|X)$, which is a p -dimensional problem and difficult to estimate when p is large. SIR is based on the inverse regression function, $E(X|Y)$, which consists of p one-dimensional regressions and is easier to deal with. The SIR components are linear combinations of the p original predictor variables defined by the weighting vectors v_i , where v_i is the eigenvector for $i = 1, \dots, K$. Geometrically, SIR projects the data from the high p -dimensional space to a much lower K -dimensional space spanned by the projection vectors v .

Dai and Lieu *et al.* (2006) found that the PLS and SIR based classification procedures performed consistently better than the PCA based procedure in prediction accuracy. “The empirical results are consistent with the analysis of the techniques. PLS and SIR construct new predictors using information on the response variable while PCA does not”. They also conclude that based on both predictive accuracy and computational efficiency, the PLS based procedure provided the best performance among the three classification procedures.

3.3.3 Clustering

Eisen, Spellman *et al.* (1998) found that larger groups of genes clustered according to their expression profile tend to share common roles in cellular processes. Hence, a major use of microarray data is to classify genes with similar expression profiles into groups in order to investigate their biological significance. A wide variety of clustering algorithms are employed for these purposes.

In theory, it is possible that a gene or genes can be clustered by error with other genes that don't have similar biological function i.e. false positives. This is the case, when the expression data profile of false positive gene is similar to that of true positive genes by chance. However, false negatives arise for two possible reasons: firstly due to inaccuracies of the clustering algorithm used, and secondly the expression value captured for a particular gene is significantly different from what it actually is. When clustering genes into one single variable, both cases are very misleading and should be taken into account when validating the subsequent knowledge discovery. The clustering results are highly affected by the choice of similarity measure and clustering algorithm, and therefore careful consideration should be given to the data at hand in selecting the method.

The advantages and limitations of the various clustering methods such as hierarchical clustering are well documented (Lodish, Berk *et al.*, 2000). In the following sections we describe the clustering methods used in this study: k-Means, SOM, and Gaussian (ArrayMiner2).

3.3.3.1 k-Means Clustering

The k-Means algorithm is a supervised method used to cluster objects based on attributes into k partitions (MacQueen, 1967). It is a variant of the expectation-maximization algorithm in which the goal is to determine the k means of data generated from Gaussian distributions. The objective it tries to achieve is to minimize total intra-cluster variance. The number of clusters (k) needs to be specified before starting the analysis.

The k-Means algorithm assigns each point to the cluster whose centre, known as the centroid, is nearest. The centre is the average of all the points in the cluster and the method involves minimizing the sum of squares of distances between the data points and their associated centroids.

- The algorithm starts by partitioning the input points into k initial sets or clusters, either at random or using heuristics.
- It then calculates the mean point, or centroid, of each cluster. That is, the expression vector, which forms the centre point of each cluster, is determined.
- All the genes are repartitioned such that each gene is placed in a cluster to whose centroid it is closest to.
- This process continues iteratively until all the genes with similar expression profiles are placed in one cluster such that each cluster is internally homogenous and at the same time, intra-cluster dispersion is large. The process stops when no more partitioning can be done.

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. It maximizes inter-cluster (or minimizes intra-cluster) variance, but does not ensure that the result has a global minimum of variance. Since the algorithm is extremely fast, a common method is to run the algorithm several times and return the best clustering found. It does not yield the same result with each run because the resulting clusters depend on the initial random assignments. Another main drawback of the algorithm is that it has to be told the number of clusters (i.e. k) to find. If the data is not naturally clustered, the user may get some unintended results as the data is forced into unnatural clusters. Also, the algorithm works well only when spherical clusters are naturally available in data.

3.3.3.2 Self-Organizing Map (SOM)

A Self-Organizing Map (SOM) is another supervised clustering approach. It consists of a neural network whose nodes move in relation to the category to which it belongs. A SOM assigns genes to a series of partitions on the basis of the similarity of their expression vectors to reference vectors (seed vectors) that are defined for each partition (see **Figure 34**). This differentiates SOMs from k-Means clustering. The procedure of SOM analysis is:

- The first step in SOM analysis is to assign a geometry of nodes in a k-dimensional (usually two dimensional) grid. The number of nodes represents the number of partitions. Each node is assigned a vector (seed) randomly.
- A gene is then selected at random and its expression vector is compared to each of the seed vectors.
- The seed vector that is most similar to the expression vector of the selected gene is then modified so that it more closely resembles the expression vector of that gene. In addition, the seed vectors of partitions that are physically closest to the already modified vector are also modified, so that they resemble the selected gene's expression vector more closely (Newton, 2002).
- This process is repeated thousands of times. After every iteration, the amount by which the seed vectors are adjusted is decreased and this continues until the reference vectors converge to fixed values.
- In the end, the genes are assigned to the relevant partitions depending on the seed vector to which they are most similar.

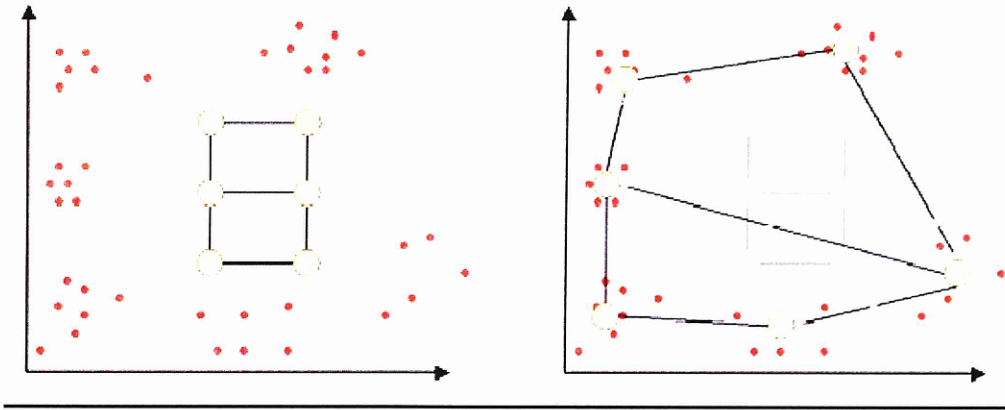


Figure 34: Self-organizing map. The initial, random mapping of nodes is shown on the left; the nodes after the SOM algorithm completed is shown on the right. Reproduced from: (Newton, 2002).

The appeal of this algorithm is that the arrangements of prototypes on a grid allow neighbouring clusters to be interpreted as being more similar (Tamayo, Slonim *et al.*, 1999). One-dimensional SOMs also have interesting uses for microarray data. For example, (Eisen, Spellman *et al.*, 1998)'s cluster package can use SOM with a one-dimensional grid of prototypes to aid the ordering of a hierarchical clustering dendrogram.

3.3.3.3 Gaussian clustering Algorithm

ArrayMiner2 is a commercial package by Optimal Design developed for clustering analysis. It seeks to address the issues encountered in applying k-Means algorithm discussed above. Similar to k-Means, it searches for solutions of minimal total variance in a given number of clusters. However, while the k-Means algorithm implicitly supposes that the variances of the expression levels in the two functional groups are equal, ArrayMiner2 recognises they may be different. Since the genes in two clusters belong to different functional groups, their activity is regulated via different pathways, involving different biological phenomena. There is thus no reason why the variances of the activity levels should be equal in the two groups. On the contrary, most real-world data can be expected to exhibit different variances of activity levels in different functional groups. Consequently, clustering by the proximity principle is inadequate in these conditions.

The fundamental idea behind ArrayMiner2's approach is to realize that the observed expression level values in each cluster can be expected to follow a distribution different from other clusters. The aim of the clustering thus becomes to identify the distribution for each cluster, such that the resulting statistical model of the data matches the observed data. Since numerous random factors are involved in gene activation and the measurement process, ArrayMiner2 assumes that the law of large numbers of statistics applies to the distribution of the observed activity levels. The activity level values are therefore modelled by Gaussian distributions, one for each condition (experiment) and each cluster.

So for instance, identifying ten clusters in a data set over four conditions means finding the mean and variance of 40 (10x4) Gaussians that best fit the observed data.

The standard way of fitting Gaussians to a set of data is to apply the *Expectation Maximization (EM) algorithm*, as in k-Means. The key drawback of the *EM algorithm* is that it is highly sensitive to initialization, often not yielding the same result in each run. ArrayMiner2 seems to have found a solution to the above by employing an algorithm known as *Grouping Genetic Algorithm (GGA)*; starting with an initial pool of solutions generated largely at random, a GA iteratively improves the pool by combining the best members of the pool, yielding solutions of increasing quality (Falkenauer and Marchand, 2001).

In a test involving a number of rat and yeast datasets from GeneSpring, (Falkenauer and Marchand, 2001) found that the solution found by ArrayMiner2 was *identical* in each of the 20 successive runs on each of the data sets, and hence concluded those solutions are probably the *global optima*: “ArrayMiner2 thus *consistently* found the probable best clustering in each case, and it did so within very reasonable execution times”. (Falkenauer and Marchand, 2001) also found the k-Means method, which was relatively faster, however returned inconsistent results over successive runs. Most current clustering tools are unable to handle outliers adequately. Not only are outliers typically not detected at all, they often perturb the algorithm as it tries to include them into clusters with the other genes at any cost. ArrayMiner2 models “outlier noise” as an additional *uniform distribution* that “competes” with the Gaussians. Expression profiles that are found to more probably belong to the uniform “noise” distribution than to any of the Gaussians, and those genes are thus labelled as outliers.

Figure 35 (a) shows the clusters in the first three classifications of the data clustered into two, three and four clusters respectively. When clustering into two clusters (box A), ArrayMiner2 detected the two large groups of data points as the most salient feature (the “big picture”) in the data. The middle group was identified as *not* being part of either of the two, but since only two clusters were requested in this instance, those data points were classified as outliers. When clustering into three clusters (box B), the three groups were correctly identified and no outliers were reported. When clustering into four

clusters (box C), a small number of highly similar data points (green dots at the centre and in the North West corner of the red cluster) were detected inside the large red group and classified as an extra cluster. These three classifications are also depicted in **Figure 35 (b)** together with classifications into four through nine clusters using ArrayMiner2's 'Classification Compare facility'.²⁸

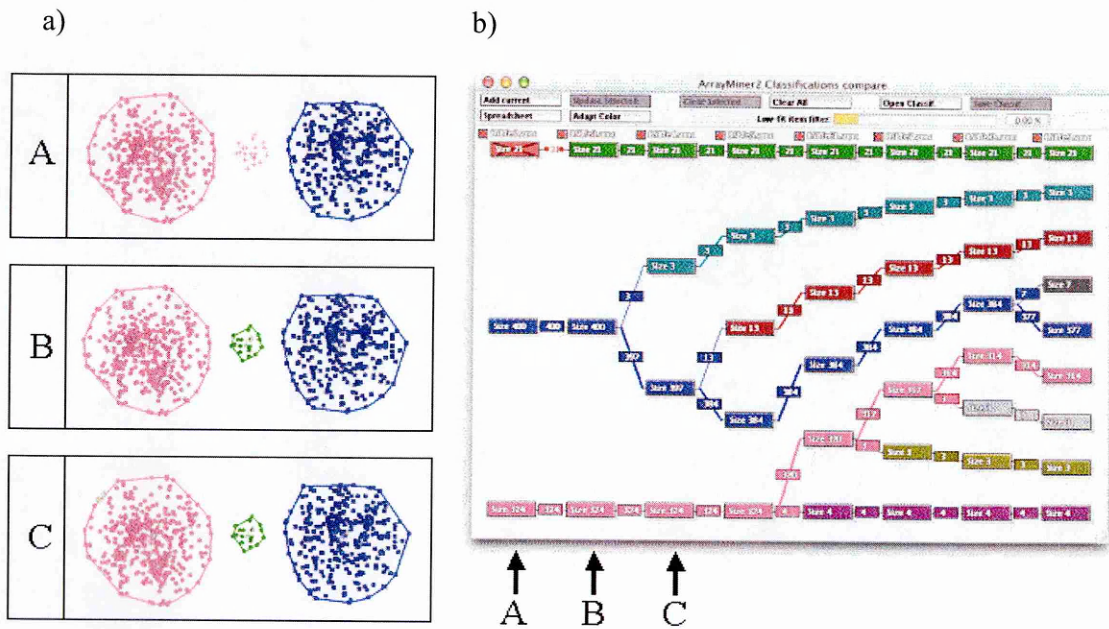


Figure 35: The Gaussian Clustering in ArrayMiner2. Obtained from <http://www.optimaldesign.com>. As of 12/10/2006.

²⁸ It should be noted that ArrayMiner2 is a non-hierarchical clustering method; each of the classifications was computed completely separately from the others. This particular view resembling a hierarchy is constructed to reveal the correspondences among clusters obtained by clustering into different numbers of clusters.

3.4 Materials and Methods

3.4.1 Data Structure

The training dataset used in this chapter is the hMADS dataset from GSK described in section 3.1. hMADS cells have been obtained and characterized from adipose tissue of young donors. These cells can be maintained *ex vivo* using an original procedure. They display extensive self-renewal capacity *in vitro*, exhibit a normal diploid karyotype, and maintain the ability to undergo differentiation into many mesenchymal cell types. (Rodriguez, Elabd *et al.*, 2004).

The dataset contained gene expression values for 54,634 probes measured in triplicate at three different time points: 2 days, 7 days, and 14 days. The normalised data can be visualised as a matrix, with each row consisting of expression values at different time instants for each Affymetrix probe ID.

As the Affymetrix human U133 GeneChip set used for the generation of the hMADS dataset may have multiple distinct oligonucleotide probes that represent a gene, probe IDs are used instead of gene name when extracting GRNs.

3.4.2 Data Pre-processing

3.4.2.1 Artificial Experiments

The low number of data points ($n=3$) compared to the number of probes ($p=54634$) in the hMADS training dataset give rise to the issues relating to high dimensionality discussed above. One way of addressing the problem is to use dimension reduction techniques discussed in section 3.3. This amounts to reducing number of probes (p).

The number of data points, 3, however is inadequate to carry out statistical modelling and is below the minimum number of samples required for by the learning algorithms in the BayesiaLab software used in this study. Nevertheless, alternative options need to be explored as many scientific experimental dataset typically have low number of sample data points due to the associated cost.

In this section, a novel method, termed ‘artificial experiments’, for increasing data points (n) is used to address this issue. For each time point, in triplicate, the reference and the sample data point combinations ratios gave rise to nine possible experiments, hence giving rise to 27 experimental data points in total, which were used for the extraction of genetic networks through out this chapter. Data was generated from the triplicate time points at day 2, 7 and 14 as shown in **Figure 36**.

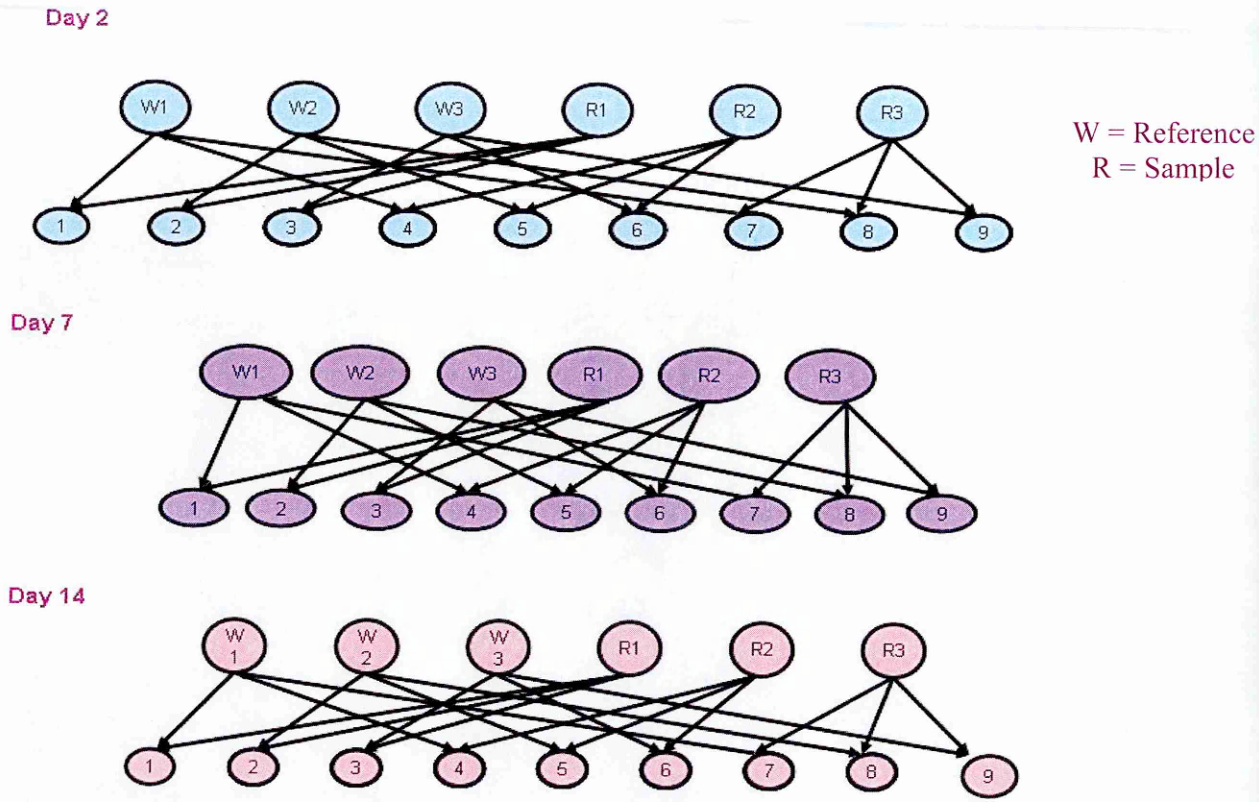


Figure 36: Generation of artificial experiments. This shows, for day 2, 7 and 14, colour coded blue, purple and pink respectively, each sample data (denoted by R) was divided by its triplicate reference (denoted by W).

3.4.2.2 Selection of Pre-processing Method

In a normal scenario (for example, modelling of Yeast cell cycle described in **Chapter 2**), the method for pre-processing would have been to take the log ratio of the average sample expression to its reference, and then apply the normal methods for replacing missing values (MV), filtering, and discretisation described in the previous sections of this thesis. However, due to the high dimensionality issue in this case, an alternative methodology as summarised in **Figure 37** was used.

First, the ratios of each data point in triplicate to its reference were calculated as described previously. Second, a variation filter to remove the invariant genes was applied. Third, clustering analysis was carried out to reduce the number of probes further: the probes were replaced by clusters of probes, and the average values of the expression levels of probes contained in each cluster was used as the data value for the cluster. Fourth, logarithms of these values were taken²⁹. Finally, data was discretised using the methodology described in section 3.4.3.

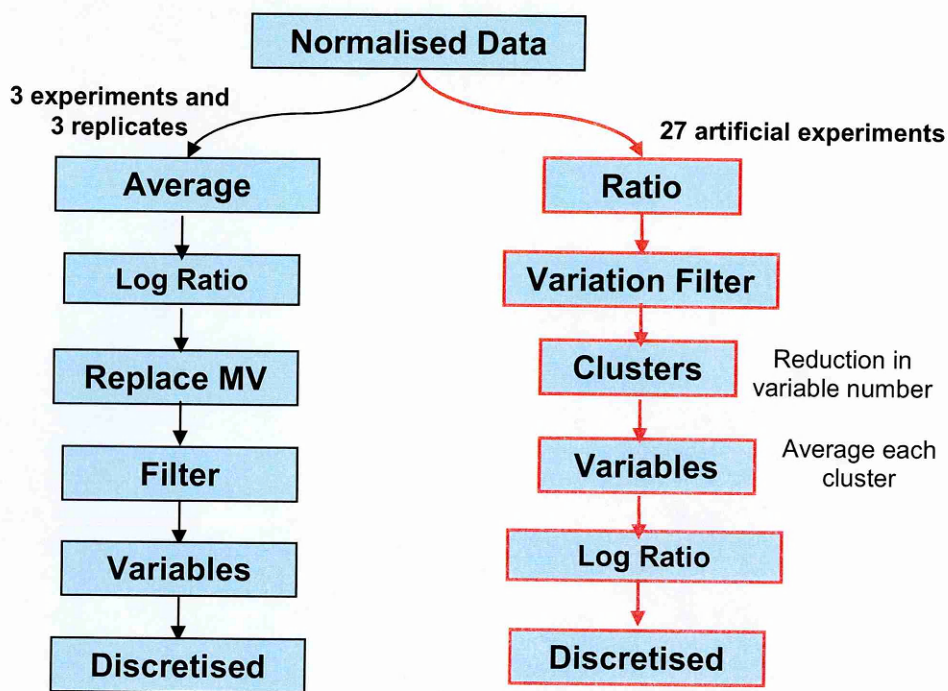


Figure 37: Data pre-processing methods with and without artificial experiments. The method highlighted in red on the right shows the pre-processing method A and on the left the normal method that would have been followed if more experimental data was available.

²⁹ The log transformation of the data is performed so the data is more normally distributed.

3.4.2.3 Gene Selection Methods

As mentioned earlier, a crucial step in data analysis is reduction of the size of the data set by selection of subsets of interesting genes from the initial set of genes. In many cases, especially when comparing genes expressed in a specific condition to relative to a reference condition, the genes of interest are those which are differentially regulated.

Most genes remain invariant, within a single microarray experiment. Since genes that are not differentially expressed have a true log ratio of zero, the observed log ratios for these genes, after an appropriate transformation, are modelled as a group with a Gaussian density. The differentially expressed genes have log ratios that are, for the most part, in some sense "far" from the other group. So these genes can be viewed as outliers from the main distribution of non-differentially expressed genes.

To analyse the high-dimensional data effectively, the invariant genes were removed, and only the differentially-expressed genes were retained for further analysis. The fold change filter was applied for filtering genes that varied significantly in expression level and this was also essential to overcome the limitation of memory and computation time in analysis.

The hMADS microarray data was pre-processed in three different filtration processes in order to identify the best method for GRN extraction as shown in **Figure 37** (method A) and **Figure 38** (methods B and C).

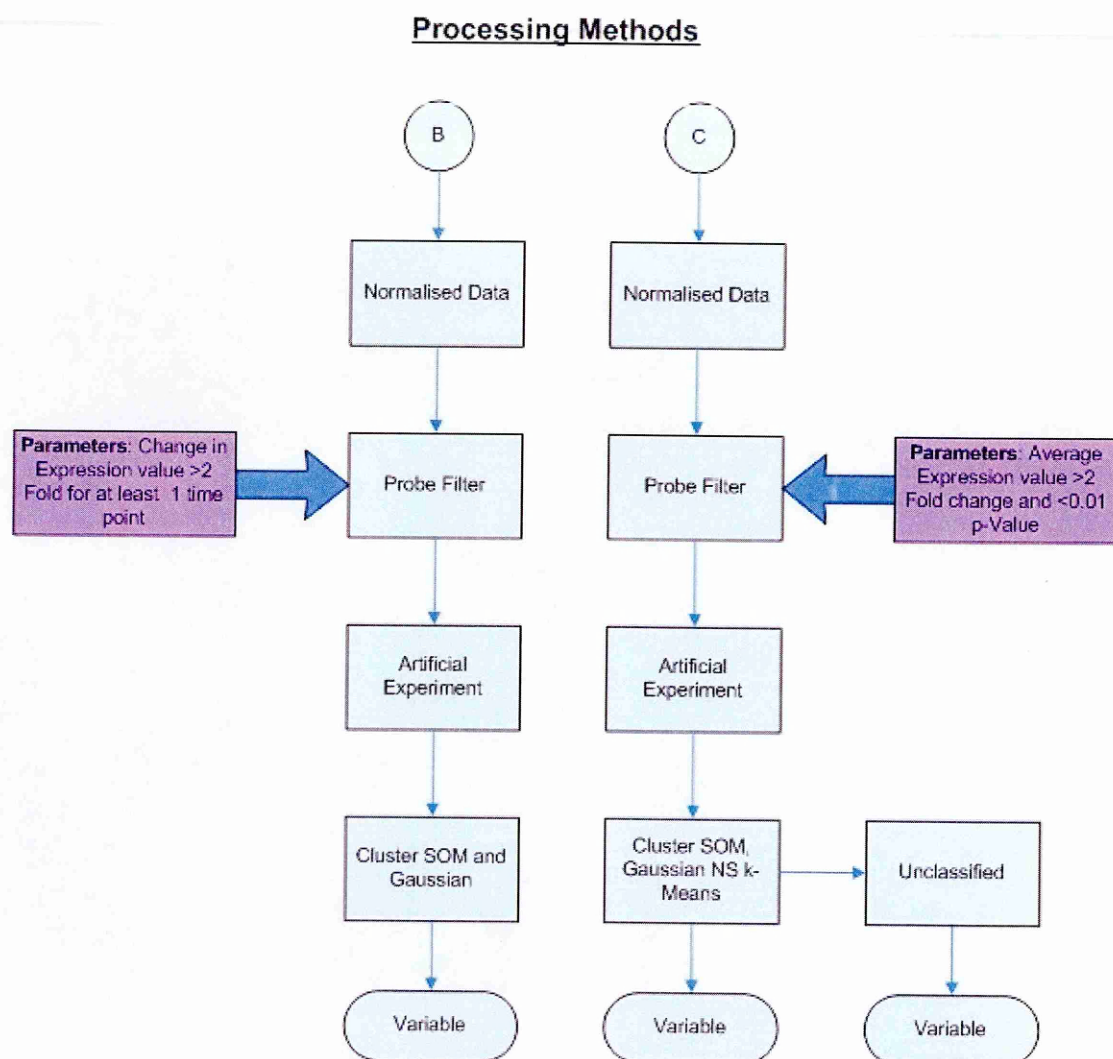


Figure 38: Pre-processing methods, B and C. The parameters for probe filtering set the criteria followed for pre-processing methods B and C.

The pre-processing method C was chosen for GRN extraction and knowledge extraction from it as it uses a combination of fold change parameter and a confidence measure (p-value) whereas the method B is based only on a simple fold change parameter (see section 3.3.1 for the comparison of the two methods).

The average 2 fold change in expression values and the p values for pre-processing method C as shown in **Figure 38** were calculated in the PADI (Process and Analysis Data Integration) bespoke system at GSK. The fold change of average of the entire row (normalised) across all the time points for reference and sample were used in method C as shown in **Figure 38**. The fold change and the statistical significance of hMADS data was calculated using the Cyber t-test, as previously described by (Hatfield, Wesley *et al.*, 2003) and (Baldi and Long, 2001). The p -values of cyber t-test were thus extracted and the data was imported into Spotfire DecisionSite 8.1.1, (Spotfire, USA) to filter out probes that varied in expression level more than 2 fold and with p value <0.01 .

The normalised hMADS data set contained some missing gene expression values (0.64% - 350 probes). Those probe IDs that had missing values didn't produce any fold change or p -values in PADI and could not be filtered. Hence these data were included later in the probe list for clustering.

Clustering analysis was then performed. The methodology used is described in the next section in detail.

3.4.2.4 Clustering Genes

As discussed in section 3.3.3, the Gaussian clustering using ArrayMiner2 provides more consistent results compared to the k-Means and SOM clustering methods and also it targets minimization of total intra-cluster variance (global minimum). However, the main drawback of ArrayMiner2 is the lengthy times it takes to analyse large datasets such as the hMADS dataset used in this chapter.

ArrayMiner2 took 6-7 hours for each run when applied to the hMADS dataset. This is compared to less than one minute for SOM and k-Means methods. ArrayMiner2, similar to SOM and k-Means, also requires specification of the number of clusters before the start of each run. Thus, to find the optimal number of clusters it needs to be run several times until no further improvement is made to the variance based score, which is highly time consuming.

Therefore, the SOM and k-Means algorithms were employed initially to indicate the number of clusters for this dataset. The ArrayMiner2 clustering algorithm was then applied by using the initial size of clusters indicated by the SOM, k-Means methods and then incrementally increasing the number of clusters until an optimum number of clusters is reached.

The SOM and k-Means clustering were performed using Spotfire DecisionSite and the Gaussian clustering was performed using ArrayMiner2 version 5.3 (Optimal Design, USA). The missing values for SOM and k-Means clustering were replaced with row interpolation in Spotfire DecisionSite. ArrayMiner2 method employed the Euclidean distance similarity measure, with data centroid³⁰ based search as cluster initialisation.

The distance measure used by ArrayMiner2 can be best illustrated by referring to **Figure 39**: the closer the distance of a gene from the centre of its own cluster, the better the measure of fitness; and the further the distance away from the centre of closest cluster, the better the measure of fitness.

³⁰ The centroid of a cluster is the weighted average point in the multidimensional space. In a sense, it is the centre of gravity for the respective cluster.

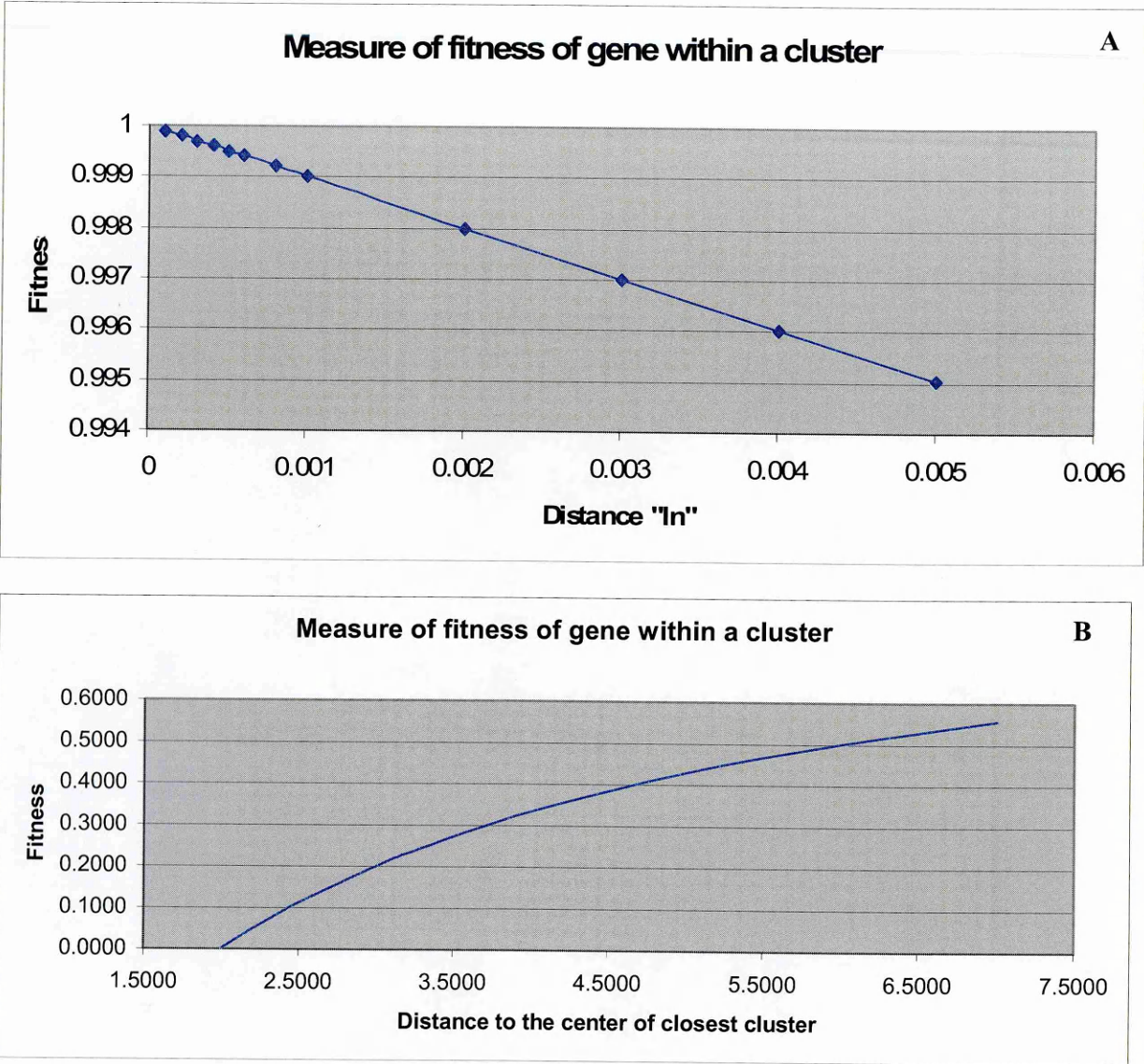


Figure 39: Measure of Fitness of a Gene within a Cluster. A) The closer the distance from the center of the own cluster, the better the fitness. B). The further the distance away from the center of closest cluster, the better the fitness.

Each extracted cluster was represented as a single variable by taking the average values of all the probes in that cluster at each time point. The clusters were assigned a cluster ID “Cluster_”, for example, Cluster_1, Cluster_2 *etc*, during the clustering process by the ArrayMiner2 algorithm.

3.4.3 Discretisation

The ratio of induction/repression magnitude is indicated by the intensity of the colours displayed. Hence the expression levels, $\log_2$ (Sample)/(Reference) ratio of continuous data were discretised into three classes: the (log) expression level between ≤ 0.1 and ≥ -0.1 was classified as 'Normal', > 0.1 as 'up', and < -0.1 as 'down'. The results were also tested to ensure robustness to variations in the discretisation thresholds.

3.4.4 Extraction of GRNs

GRN extraction was carried out using the techniques validated with Yeast data in **Chapter 2** and described in detail in sections 2.2.2-2.2.6.

3.5 Results and Discussions

3.5.1 Gene Selection

As discussed earlier, the hMADS dataset contained 54634 probes. This not only creates computational issues due to high dimensionality, but also many of the genes in the dataset are irrelevant for the purposes of this study in that their expression levels are invariant relative to the reference condition and hence may not contribute any additional information.

In addition, some low expressed, or extremely highly expressed probe sets are data noise which is harmful to statistical data analysis. Such genes need be excluded from subsequent statistical analysis.

The choice of processing method has a major impact on differential expression analysis of microarray data. Use of False Discovery Rate (FDR) to quantify the sensitivity of a processing method in terms of its ability to distinguish differentially expressed genes from genes having invariant expression is one way of addressing this issue. This is a highly relevant property as differential expression analysis is the most common application of microarray data.

The cyber t-test method employed in this chapter meets this criterion: it tests the significance of the differential expression level of a gene measured by its fold change, by evaluating against a confidence measure with a p -value; the likelihood of false discovery in this test is given by its p -value (Type 1 Error). The lower the p -value, the lower the FDR, and the higher the confidence level placed on the test. The p -value used in this study was 0.01, giving a high confidence level of 99.0%.

In addition, the expression profiles of the genes were analysed, and the impact of different threshold values on gene selection was considered prior to the application of the filtration process to ensure an appropriate value for the threshold is chosen. A fold-change of two was chosen as the filtration threshold based on this analysis.

Figure 40 shows the number of probes that had at least one experiment value outside each class width.

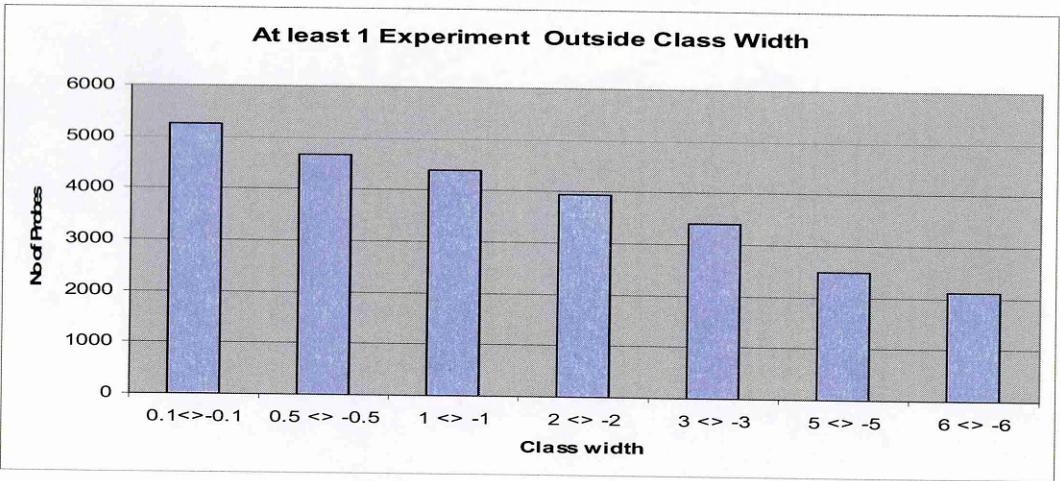


Figure 40: Fold Change Filtration. Number of differentially expressed probes with at least one experiment value outside for the fold change class widths.

The directional fold change and the associated p -values for the entire probe IDs in the dataset are shown in **Figure 41**. This is known as the ‘Volcano’ plot due its shape. Of the 54634 probes, only 54284 were visible as 350 probe IDs had negative values. Hence the cyber t-test and fold change for these 350 IDs weren’t generated in PADI. **Figure 42** shows the volcano plot for filtering based only on 2-fold change, which identified 4995 probes (9.1% of the original dataset)) as differentially regulated.

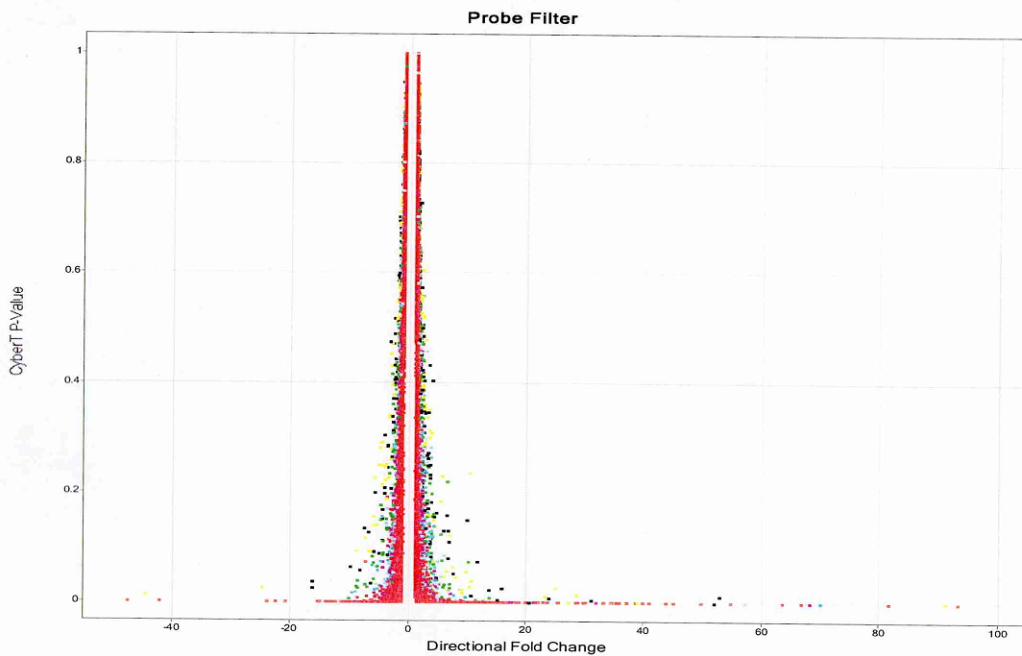


Figure 41: Volcano plot of directional fold change and its significance p-value.

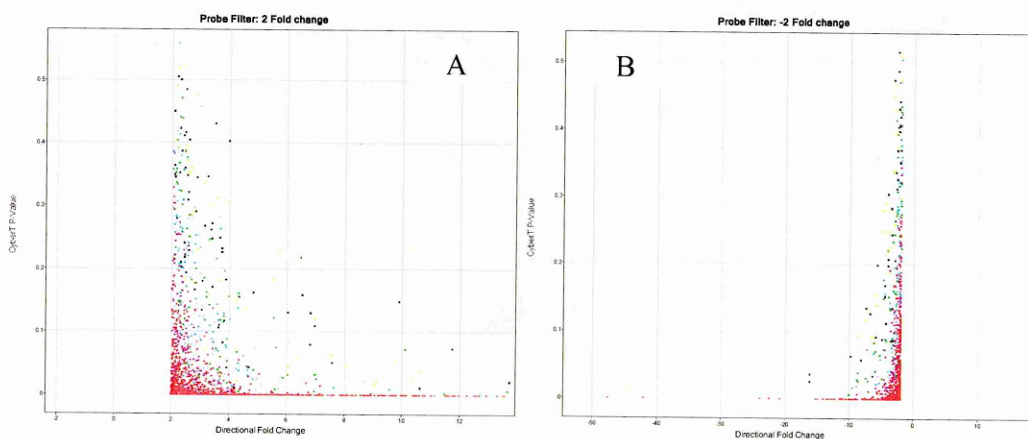


Figure 42: Directional fold change. The figures A and B show the up-regulated probes by 2 fold: 2729 probes (4.99%) and the down-regulated by 2 fold: 2266 probes (4.15%) respectively.

The joint filtration criteria of 2-fold change and p -value of 0.01 (i.e. 99% confidence level) chosen as the final gene selection criterion in this analysis yielded a smaller subset of 3516 probes (6.4%). This is shown in **Figure 43**.

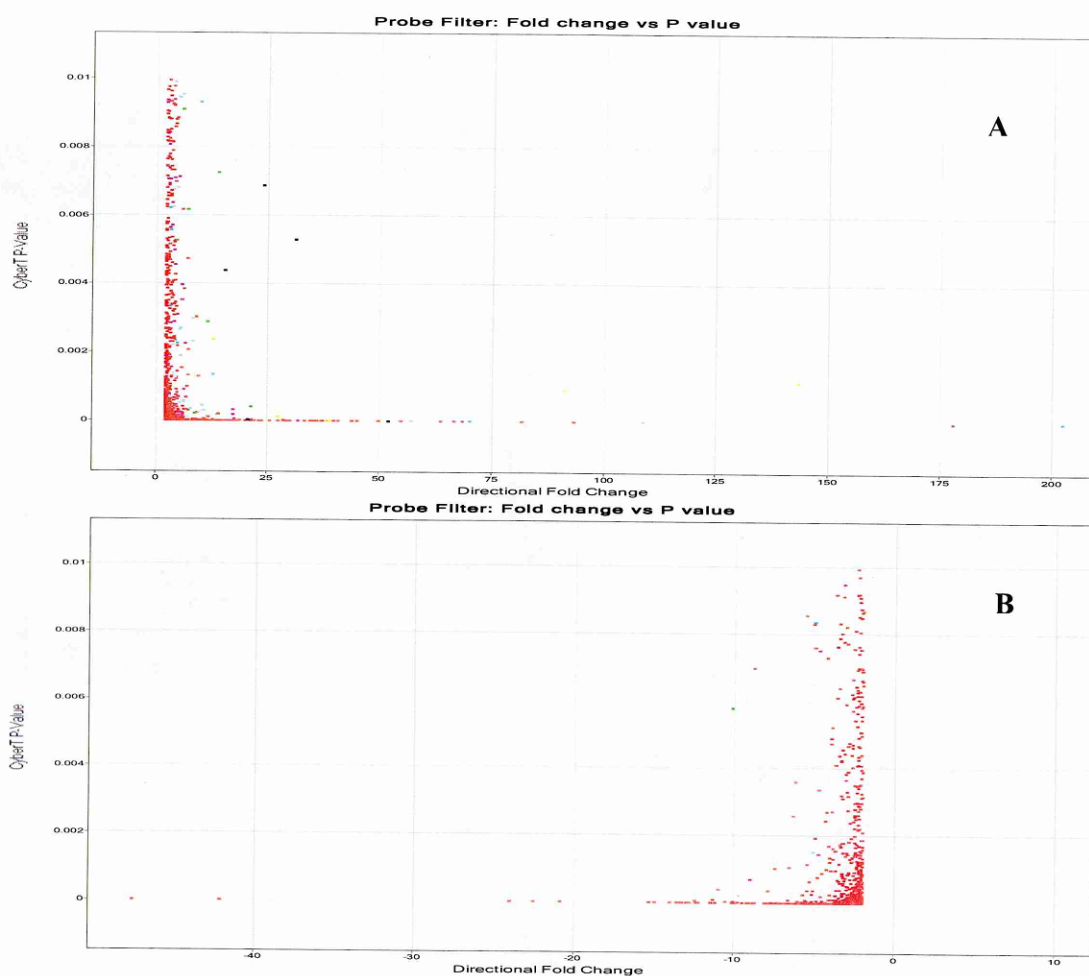


Figure 43: Significant change in gene expression. A cut off point of 0.01 p value generated 1976 probes with significance change (3.62%) for up-regulated probes and 1540 probes with significance change (2.82%) for down-regulated probes as shown in A and B respectively.

As discussed in section 3.4.2, the 350 probe IDs which could not be filtered using the above method due to missing values, were added to the probe set prior to clustering.

3.5.2 Clustering Analysis

A total of 3866 probes were selected for clustering as discussed above. 558 probes were however rejected by the ArrayMiner2 clustering algorithm in the test runs due to incomplete data. The remaining 3308 probes were then clustered. SOM and k-Means, which are fast and thus allow multiple runs within a short period of time, were used to indicate the number of clusters. The initial SOM clustering was performed with a 6×6 grid size. This indicated the presence of around 22 clusters, while the k-Means method indicated the presence of 27 clusters. Hence, 25 clusters were chosen as a starting number for Gaussian clustering in ArrayMiner2.

However, with the choice of 25 clusters, there were still a significant number of probes shown as unclassified in the ArrayMiner2 analysis. Hence the SOM and k-Means clustering was repeated by raising the cluster/grid size incrementally. With the 10×10 grid size, clustering analysis indicated the presence of 42 clusters as shown in **Figure 44**. Clustering using k-Means however, indicated the presence of 60 clusters as shown in **Figure 45**.

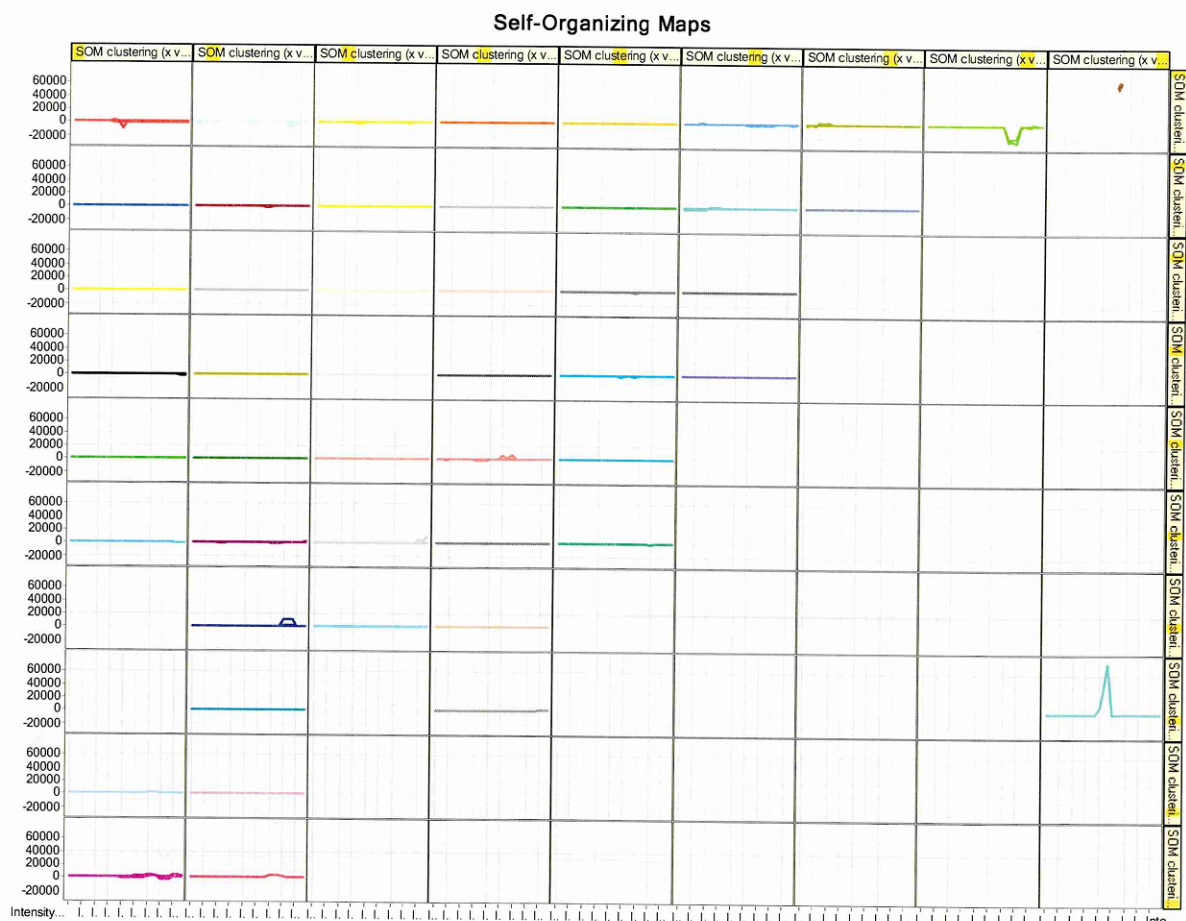


Figure 44: Cluster optimisation: SOM 10x10. Each grid shows if cluster is present or not. The empty grids means no cluster was found. Each cluster is shown in different colour with the profile it follows. In this instance, the grid shows the presence of 42 clusters. The 10th column of the grid was omitted as it did not contain any clusters, and hence is not shown above.

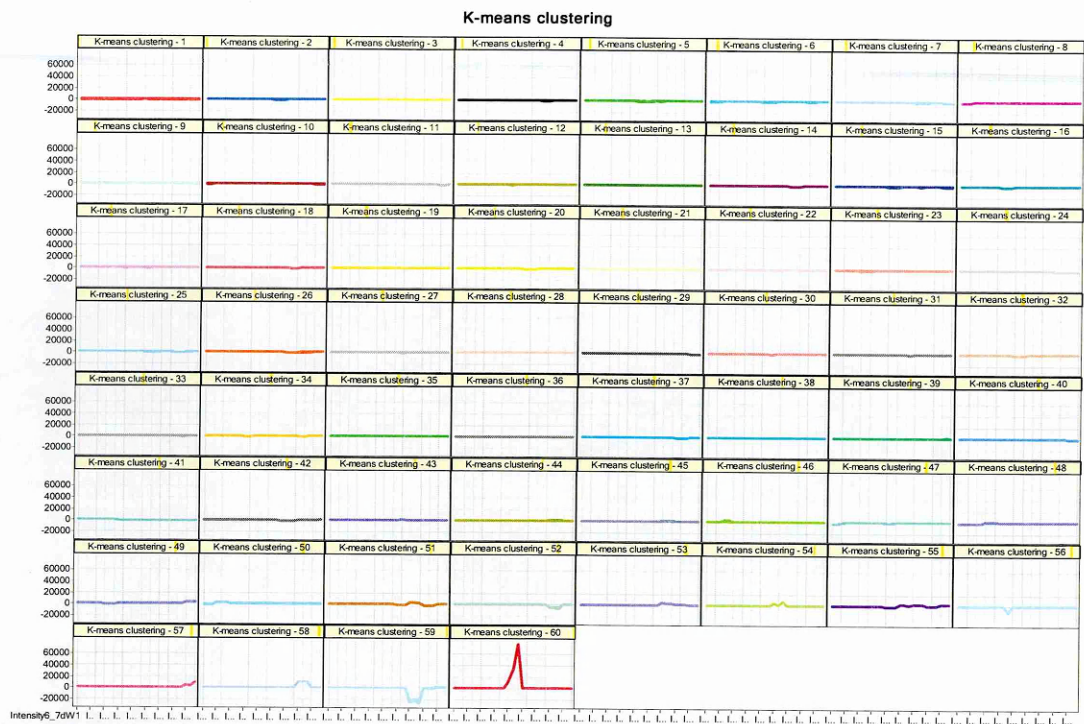


Figure 45: k-Means Clustering. This visualization shows the profiles in each of 60 calculated clusters. Number of iterations: 29, Total score: 6.718e+007.

It was noted that the output parameters for the SOM method, such as mapping precision and topology preservation were improved from 3.544×10^4 to 3.565×10^4 and 1.526×10^{-2} to 2.587×10^{-2} from 6×6 to 10×10 grids respectively.

As the results from SOM, and k-Means regarding the number of clusters were inconclusive, 36 clusters, which was lower than the number of clusters indicated by the two methods was used as the new starting point for the Gaussian clustering. The process was repeated with incremental increase in the cluster size, and the total variance score was monitored. At 45 clusters, the total variance was minimised, and there was only slight improvement in reduction in outliers beyond 45 clusters, indicating that an optimum number of clusters have been reached. At this stage there were still a group of 437 unclassified probes or outliers. The unclassified probes were clustered separately to yield another 25 clusters without any further outliers.

A screen shot of cluster output from ArrayMiner2 is shown in **Figure 46**. The main graph window shows the profiles of all 66 probes in cluster 19 over the 27 data points. The fitness of each probe belonging to its cluster is given in percentage in the bottom panel.

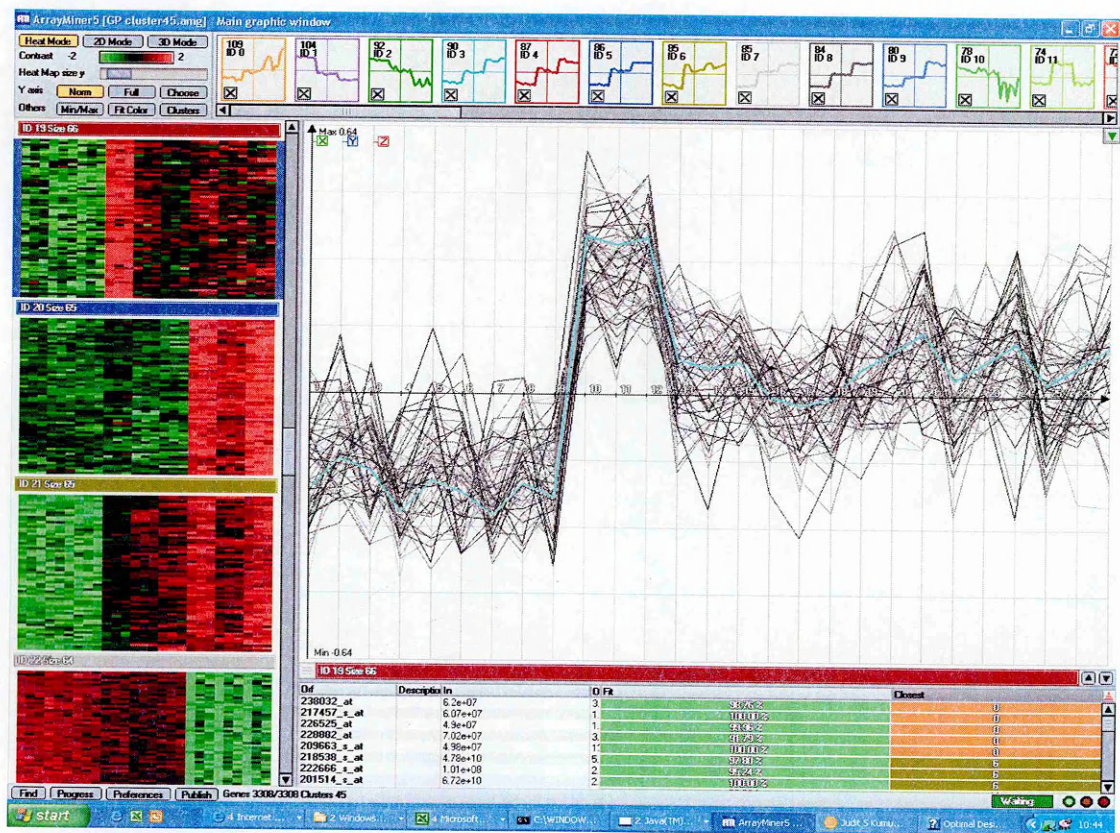


Figure 46: Gaussian Clustering. The grids on top panel show the cluster ID, number of probes within that cluster and the overall profile of the cluster. On the left panel the heat map of individual clusters are shown.

Figure 47 shows the PCA of clusters to project three selected attributes of each probe in a 3D space, which gives a better appreciation of mutual profile distances than a 2D mode.

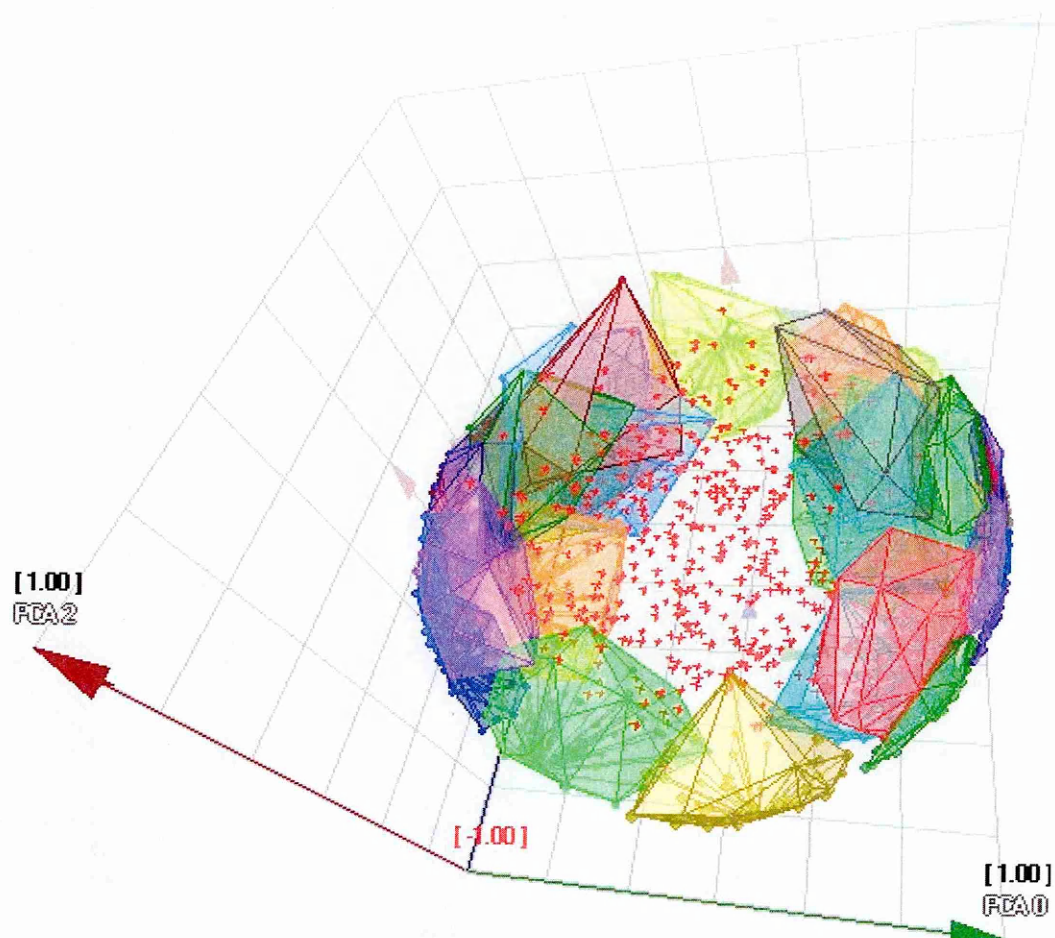


Figure 47: PCA view of Gaussian clusters. Each clusters are shown in different colours with hulls. The red crosses in the middle represent the outliers.

The average profile of all clusters, a view useful for assessing the differences between clusters is shown in **Figure 48**.

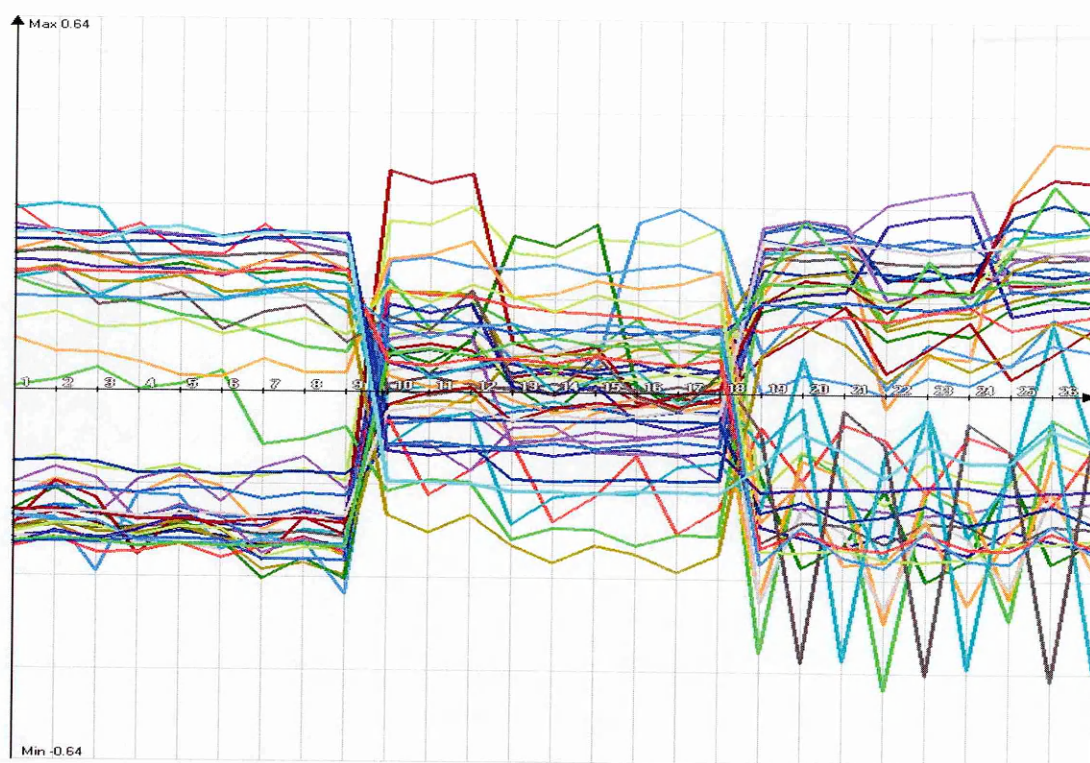


Figure 48: Average cluster profile. Profile of all 45 clusters over the 27 data points. Each colour represents different clusters. The vertical axis shows the average expression level in logarithms, and the horizontal axis (at the centre) shows the experiment time points 1-27.

The strength of the Affymetrix, human, U133 GeneChip set used for the generation of the hMADS dataset in this study is that multiple distinct oligonucleotide probes may represent a gene. However the signal intensity from the different probes for the same gene may not be the same. For example, probes within a probe set may map to various gene loci in the genome. Moreover, different probes directed to one transcript can have high variability in hybridization, resulting in high variability in signal strength even if melting temperatures are calculated carefully. Consider a PCR experiment where different primers were designed for a specific amplification but may give completely different amplification rates. The same holds true for microarray probes.

A close examination of the derived clusters in this study also highlights this issue. For example: the probe IDs 203548_s_at, 203549_s_at for gene LPL (lipoprotein lipase) were clustered into two separate clusters, Cluster_7 and Cluster_40, respectively; the probe IDs 203980_at, 235978_at for gene FABP4 (Fatty acid-binding protein 4) were clustered into two separate clusters, Cluster_28 and Cluster_39, respectively. **Figure 49** shows the varying signal intensities of different probes for the genes LPL, and FABP4. The significant difference in the expression profiles of these probes in the dataset meant they are ('correctly') clustered into separate clusters.

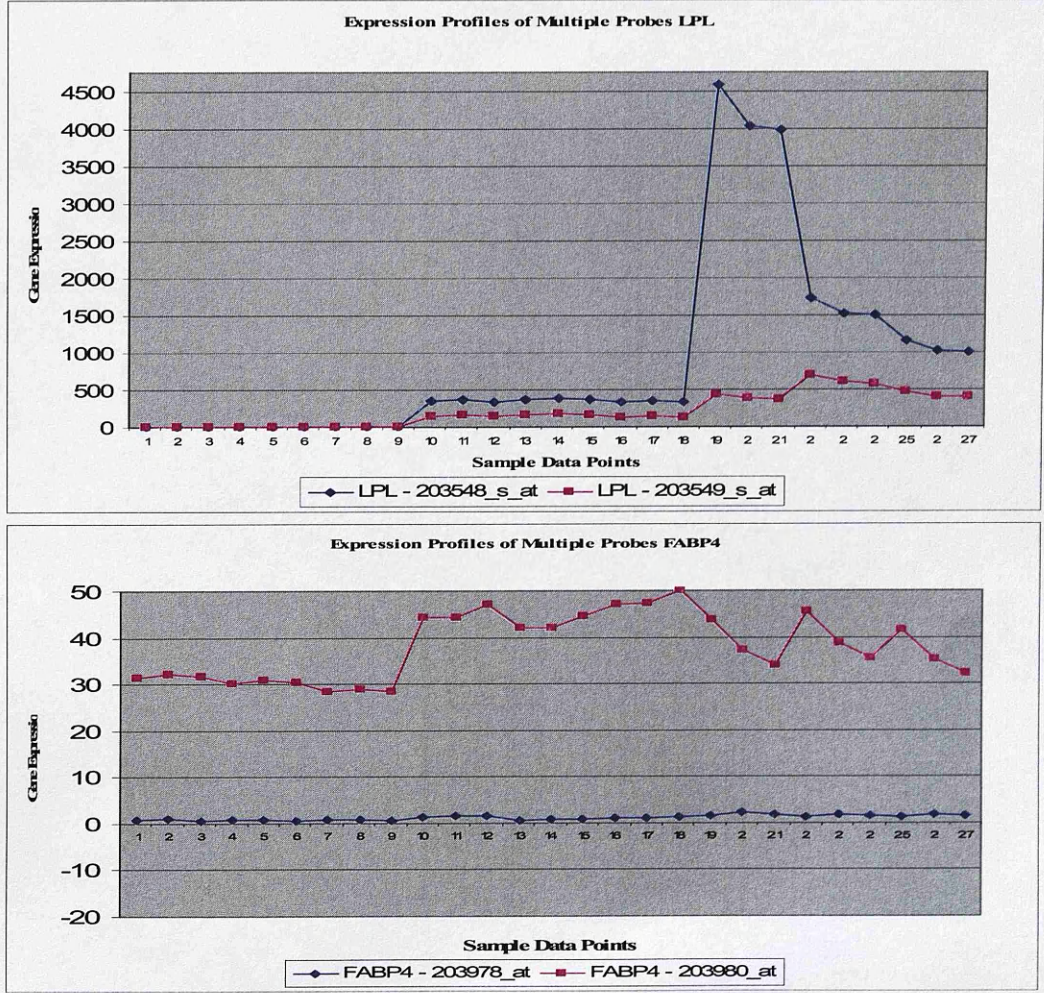


Figure 49: Gene expression profiles of multiple probes from LPL and FABP4.

Thus a single probe approach (one probe per gene) may yield a more focused picture than probe set approaches. Due to the single probe approach, ChipInspector finds more significant genes and reduces the false positive rate by an order of magnitude. In the case of probe set approach with multiple probes, the question remains however, how signals from the many probes for a gene, could be combined into a single estimate of the abundance of that gene. The variations of probe-level intensity with the same exon might suggest a new splice variant site or the presence of SNPs. Further research work is needed to eliminate all probes with known SNPs, and must be expanded to the all probes on U133 to evaluate the effectiveness of using absolute cut-off value for alternative splicing analysis.

3.5.3 Extraction of Gene Regulatory Networks

The final dataset (3308 probes) after the application of the dimension reduction methods included 45 clusters containing 2871 probes, and a further 25 ‘clusters’ of the outliers, containing 437 probes.

In order to validate that the resulting dataset can still achieve good predictive performance, a selection of 87 gene probes from the TIGR database with known functionality were selected from the original hMADS dataset and a Bayesian network was extracted from them. Gene probes were selected from the insulin receptor family (IR), PPAR family, C/EBP family, FABP family, ADD1, FOX family, KLF and MEF2.

3.5.3.1 Extraction and Validation of Selected Genes of Interest

The GRN of 87 probes using the Taboo search algorithm, tagged according to the group or family of gene of interest is shown in **Figure 50**. There is more than one probe per gene in the hMADS data and therefore the probe ID name is used instead of gene name. The presence of an arc between two genes represent that there is a direct probabilistic relation between them.

Out of 87 individual probes of interest representing each gene, the extracted GRN shows association amongst 36 probes; the rest of the probes that were found to be independent are not shown for simplification. Validation of the extracted GRN presented below shows that these methods have been able to identify many biologically plausible relationships between genes despite the small number of time points over which the data was generated.

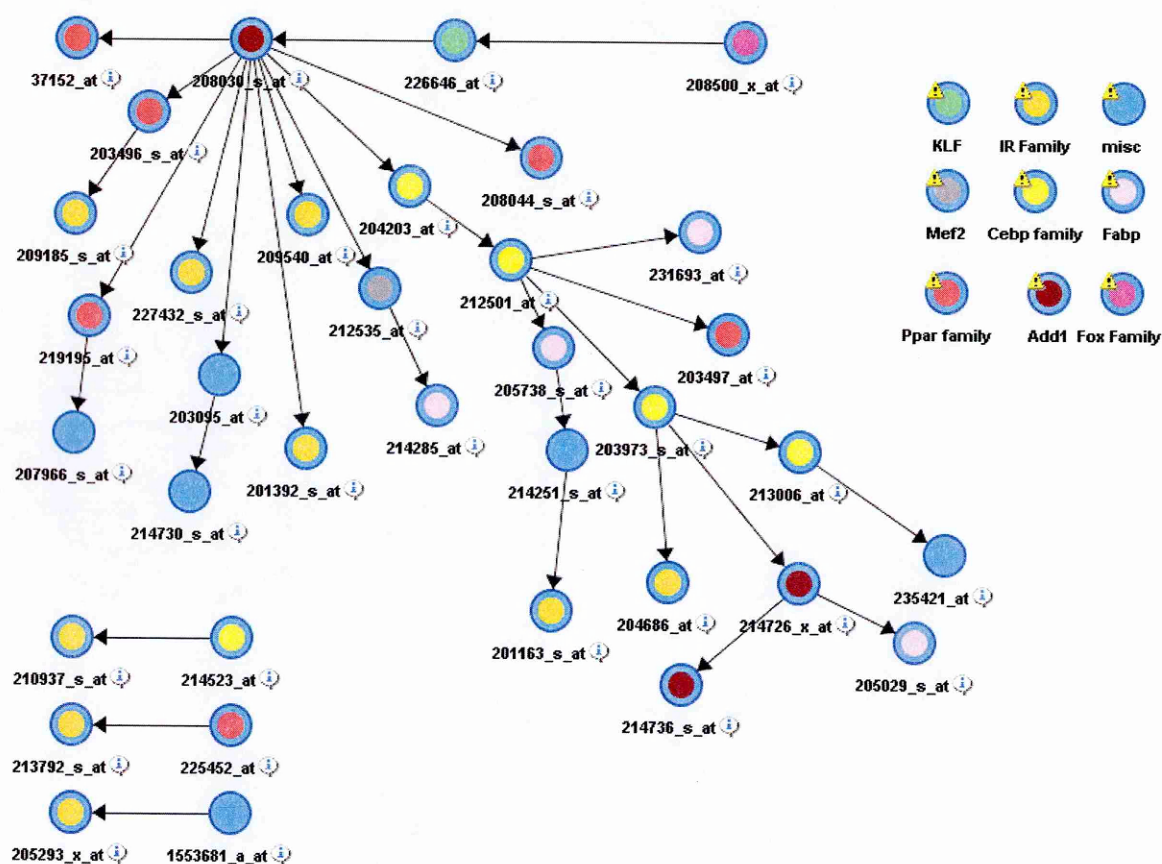


Figure 50: GRN of genes of interest, colour coded according to the biological family they belong to. The respective Affymetrics probe ID annotation from TIGR database indicated by the sign “i”, is only available on the modelling view in BayesiaLab.

This also shows that the probe 208030_s_at from the gene ADD1 (Adipocyte determination and differentiation factor 1), colour coded in brown, is a parent gene to eleven other genes in the network, and a common ancestor to most other genes in the network. Hence, it could be regarded as playing a central role within the GRN of interest. This is supported by evidence from published literature: ADD1/sterol regulatory element binding protein-1c (SREBP-1c), which is the main transcription factor, regulates the characteristic features of adipocyte, senses the glucose and fat excess and draws the excess into the adipocyte to preserve energy, and maintains the blood biochemistry within physiological ranges. ADD1/SREBP-1c has regulatory functions via transactivation over the other important mature adipocyte markers such as leptin, peroxisome proliferator-activated receptor γ (PPAR γ) and lipogenic enzymes (Erol, 2005). The other two probes from the same gene ADD1, 214726_x_at and 214736_s_at are shown as directly associated with each other and indirectly associated with 208030_s_at.

PPAR family of ligand-inducible transcription factors³¹ are of major relevance in lipid and cholesterol metabolism. For example, PPARGC1 is a transcriptional co-activator of many nuclear receptors including PPAR- γ , involved in the regulation of fatty acid oxidation, skeletal muscle fiber type specificity, and gluconeogenesis (Vohl, Houde *et al.*, 2005). The extracted GRN shows that 37152_at (PPAR δ), 203496_s_at (PPARBP), 219195_at (PPARGC1), and 208044_s_at (PPAR δ) are directly associated with ADD1, while 203497_at is indirectly related. Another probe from gene PPARBP, 225452_at, is however, shown as independent of ADD1 and rest of the members of the PPAR family.

Experiments show that under the combined hormonal activities (IDM induction) of insulin, the glucocorticoid dexamethasone, and the phosphodiesterase inhibitor isobutylmethylxanthine, most of the cells undergo a cycle of post-confluent proliferation, and then initiate a cascade of steps with several key regulatory genes involved.

³¹ A transcription factor binds upstream or downstream to either enhance or repress transcription of a gene by assisting or blocking RNA polymerase binding.

These genes are found necessary and/or sufficient for the transition from preadipocytes to adipocytes in an in vitro culture model, including the CCAAT/enhancer binding proteins (C/EBPs) α , β , and δ , as well as the peroxisome proliferator-activated receptor (PPAR) (Hung, Chang *et al.*, 2004).

The extracted GRN also identified positive associations between the probes 204023_at (C/EBP γ), 212501_at (C/EBP β), 203973_s_at (C/EBP δ), and 213006_at (C/EBP δ) from the C/EBP (CCAAT/enhancer binding protein) family, which are colour coded in yellow. Members of the C/EBP family of transcriptional factors have been implicated in the regulation of genes in a variety of tissues (Antonson, Stellan *et al.*, 1996). The extracted GRN, however, did not find any association between 214523_at (C/EBP ϵ) and the other members of the family.

The members of the insulin receptor family (IR), color coded in deep yellow in **Figure 50**, plays a vital role in adipogenesis. Increased expression of the insulin receptor and insulin-dependent glucose transporters (e.g. GLUT4), during late differentiation makes adipocytes sensitive to insulin. Hormones, including growth hormone (GH), glucocorticoids and insulin, stimulate induction pathways that commit cells toward terminal differentiation (Avram, Avram *et al.*, 2007). Insulin-like growth factor-I (IGF-I), an exception, stimulates both proliferation and differentiation (Smith, Wise *et al.*, 1988). Before expression of insulin receptors during later stages of differentiation, insulin mediates its early adipogenic effect via IGF-I receptors which are expressed by preadipocytes. In vitro and in vivo studies both show that insulin exposure stimulates the expression of PPAR γ in human preadipocytes from subcutaneous fat, thereby suggesting a possible mechanism for insulin action in vivo (Rieusset, Andreelli *et al.*, 1999).

The extracted GRN shows three of the key members of the IR family as directly associated with ADD1 (208030_s_at): the IGF-1 (209540_at) gene described above, IGF-II/mannose 6-phosphate receptor (IGF2R; 201392_s_at), and the insulin receptor (INSR; 227432_s_at). Three genes are shown as indirectly associated: insulin receptor substrate 2 (IRS2, 209185_s_at), Insulin-like growth factor-binding protein 7 (IGFBP7; 201163_s_at) and IRS1 (204686_at). Three other genes are shown independent of the main network of genes. Insulin promoter factor 1 (IPF1; 210937_s_at), a key pancreatic β -cell gene, is shown as associated with C/EBP ϵ (214523_at); another probe of INSR (213792_s_at) is shown as associated with PPARBP (225452_at); and finally, the brain-specific angiogenesis inhibitor 1-associated protein 2 (BAIAPA2) is shown as associated with Perforin gene (PRF1; 1553681_a_at), which regulates perforin expression in NK cells and CTL.

Fatty acid binding proteins (FABPs) are a family of small, highly conserved, cytoplasmic proteins that bind long-chain fatty acids and other hydrophobic ligands. It is thought that FABPs roles include fatty acid uptake, transport, and metabolism. (Staloch, Divine *et al.*, 2005) found that C/EBP and Cdx family factors regulate liver fatty acid binding protein (FABP) transgene expression in the small intestinal epithelium of rats. The extracted GRN for human genome supports this finding: 212501_at (C/EBP β) is directly associated with 205738_s_at (FABP3) and 231693_at (FABP1), and indirectly associated with 205029_s_at (FABP7) and 214285_at (FABP3). This is found to be highly conserved in both of the species.

3.5.3.2 Analysis of Clusters

The names of genes contained within each cluster generated by the ArrayMiner2 algorithm and the outliers are provided in Appendix VII in a CD due to its sheer size. Affymetrix probe IDs are provided, where gene names are not available; The probe IDs can be identified by their suffix: `_at`, `_a_at`, `_s_at`, and `_x_at`, for example 1560662_s_at. As discussed in section 3.5.2, some of the genes in the filtered dataset have multiple probes, in which case a gene name may appear more than once in the data.

In this section, the relationships between the main 45 clusters generated are studied. Detailed analysis on a selection of clusters based on this analysis is presented in the following section.

Figure 51 shows the target node analysis of all 45 clusters without outliers. Cluster_1 was chosen as a target node as this cluster is predominantly the father node for many other clusters. The shade of the squares in the centre of the node represents the amount of information brought to the target node. The lighter the shade of the square the more the information contributed to the (TN). Visual inspection suggests, that apart from Cluster_1, Cluster_27 also dominates the order, albeit to a lesser extent (see bottom right hand corner).

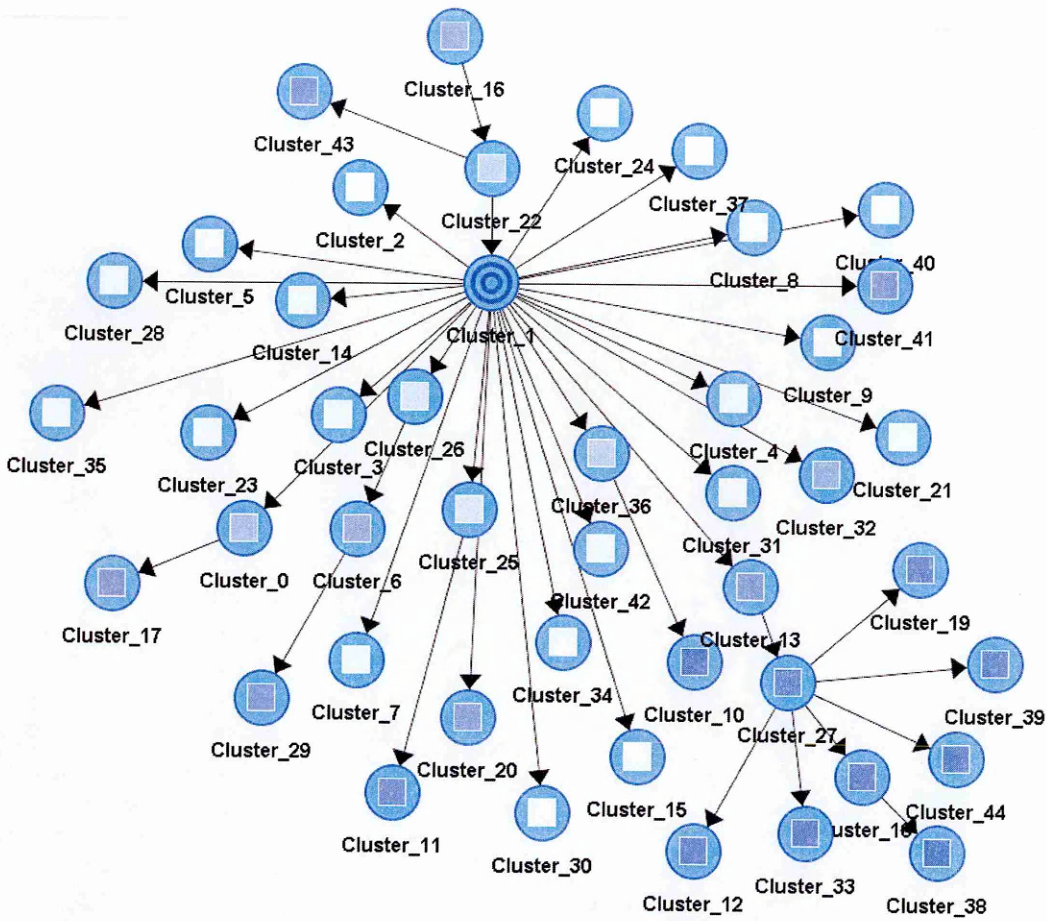


Figure 51: Unsupervised learning, extracting information contribution of the arcs and the nodes to the overall BN. Target node (TN) represented by the wavy circular node. The shade of the squares in the centre of the node represents the amount of information brought to the target node. The lighter the shade of the square the more the information contributed to the (TN).

Figure 52 shows the partition of the GRN extracted in **Figure 51** into 20 colour coded clusters. Partitioning is generated with the Taboo algorithm. This algorithm creates groups of variables that are close in terms of their high correlation. The clusters are calculated based on the Kullback-Leibler divergence score³² (see section 1.6.4, **Chapter 1**); the nodes with similar score are grouped in one cluster.

³² In probability theory and information theory, the Kullback–Leibler divergence is a natural distance measure from a "true" probability distribution P to an arbitrary probability distribution Q . Typically P represents data, observations, or a precise calculated probability distribution. The measure Q typically represents a theory, a model, a description or an approximation of P .

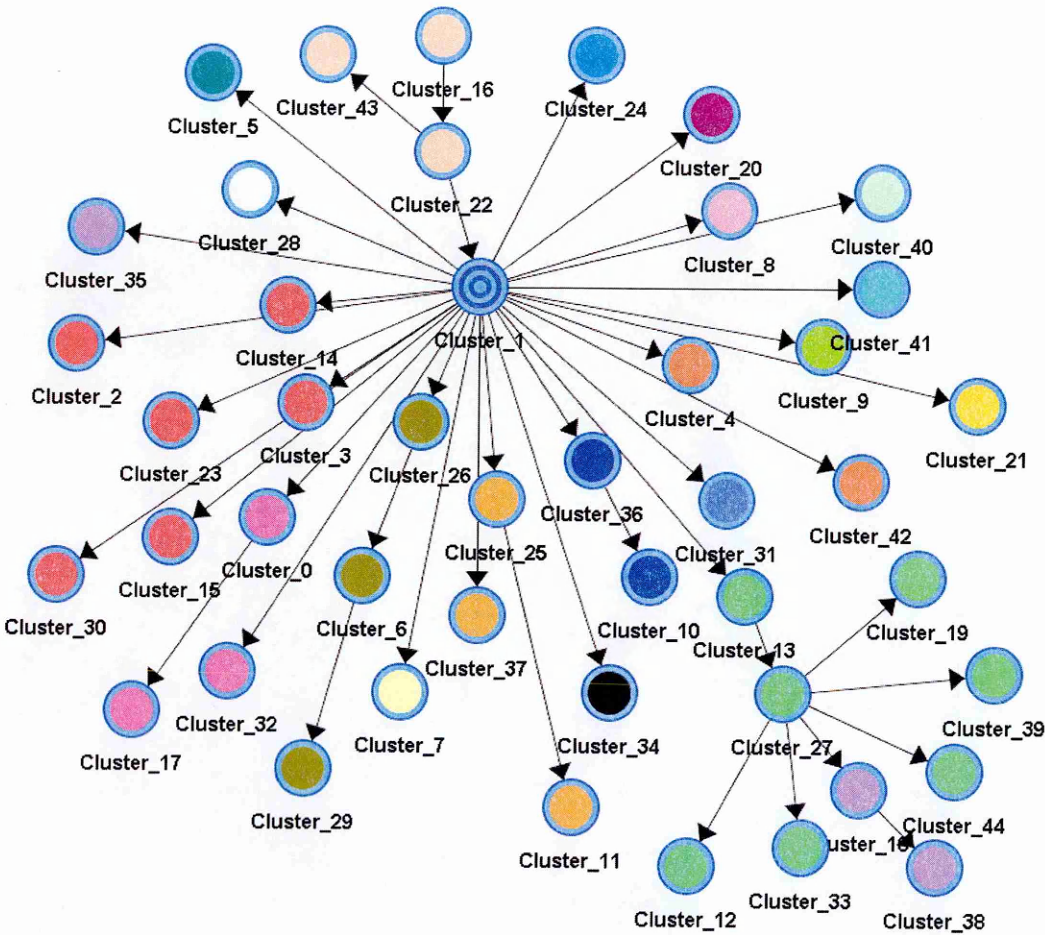


Figure 52: Natural partitioning of GRN based on target node Cluster_1. The nodes in the GRN are grouped into sets of colour coded clusters based on their closeness as measured by the Kullback-Leibler divergence.

This analysis computes the Kullback-Leibler divergence (a.k.a. information divergence, or information gain, or relative entropy) for each arc of the network. This distance is a measure of the differences between the probability distributions represented by the network with the arc corresponding to the relation, and without this arc. In this case, this measure is equal to the mutual information between the parent and the child. It thus provides a visual tool for identifying nodes that may also be functionally related.

Another useful tool for analysing a GRN is the Target Node (TN) analysis presented in **Figure 53**. Again the GRN shown is the same as **Figure 51** and **Figure 52**. However, this analysis shows the strength of the relationship between nodes; the thickness of the arcs represents the strength of the correlation between two nodes, and the length of the arc is proportional to how close the node is to the TN.



Figure 53: Arc analysis of network of all 45 clusters. TN represented by the wavy circular node. The thickness of the arc represents the information contributed to the global structure. The strength and the length of the arc are proportional to the strength of the probabilistic relation and how close the node is to the TN respectively.

The thickness of an arc is directly proportional to the absolute value of Pearson's linear correlation coefficient (see 1.5.2.1, Chapter 1), r . The Pearson's linear correlation coefficient between two cluster nodes, r , is computed by taking into consideration the marginal probabilities of the nodes modalities (outcomes).

A blue arrow indicates positive correlation and the red arrow a negative correlation as shown in **Figure 53**. The exact value of the correlation for each arc is displayed in the comment of the arc as indicated by the sign “ $\pm$ ”. The information is however, only visible in BayesiaLab modelling mode.

This analysis shows that Cluster_36, Cluster_23, Cluster_13 and Cluster_34 are much closer in terms of mutual information brought in to the TN than other Clusters. This means that probes in these clusters are similar in profile as the TN cluster_1. The nodes, Cluster_29 and Cluster_9 have the strongest association represented by the thickness of the arcs.

3.5.3.3 Analysis of Clusters with Outliers

The probabilistic dependency between the 45 cluster variables and the 437 individual probes classified as outliers in ArrayMiner2 is shown in **Figure 54**. It shows that the GRN from the previous section, learned only on the 45 clusters was generally preserved; but some of the outliers (31 of the 437 probes) have now shown to be associated with this network of clusters around the centre of **Figure 54**. A further 87 probes form small clusters around this. Two of the 45 clusters, Cluster_10 and Cluster_38 however are now independent of the extracted GRN along with the rest of the outliers.

Figure 55 shows the GRN of the 45 clusters together with the outliers now clustered into 25 clusters separately by the ArrayMiner2 algorithm. Again, as could be expected, the GRN structure is consistent with the previous result shown in **Figure 54**, Cluster_1 and Cluster_27 again form the central clusters and Cluster_10 and Cluster_38 remain independent of the extracted GRN. 19 of the 25 clusters of outliers are also shown as independent of the main GRN.

This analysis raises two interesting points. Firstly, it is no surprise that Cluster_10 and Cluster_38 no longer show any association with the extracted GRN; target node analysis presented in **Figure 51** showed these two clusters were weakly associated with the rest of the clusters as represented by the dark blue square inside the nodes. A plausible explanation could be that the addition of new information in the form of outliers re-enforced their conditional independence properties, i.e. reduced their likelihood of being associated with the GRN given the new information.

Secondly, while the clustering method used classified 437 of the probes as outliers, a sizeable proportion of them showed associations with some of the clusters. This demonstrates the differences on the methodology used, as Bayesian learning is based on the conditional probabilities of the two sets of nodes, while clustering is based on similarity measure of the gene expression profiles.

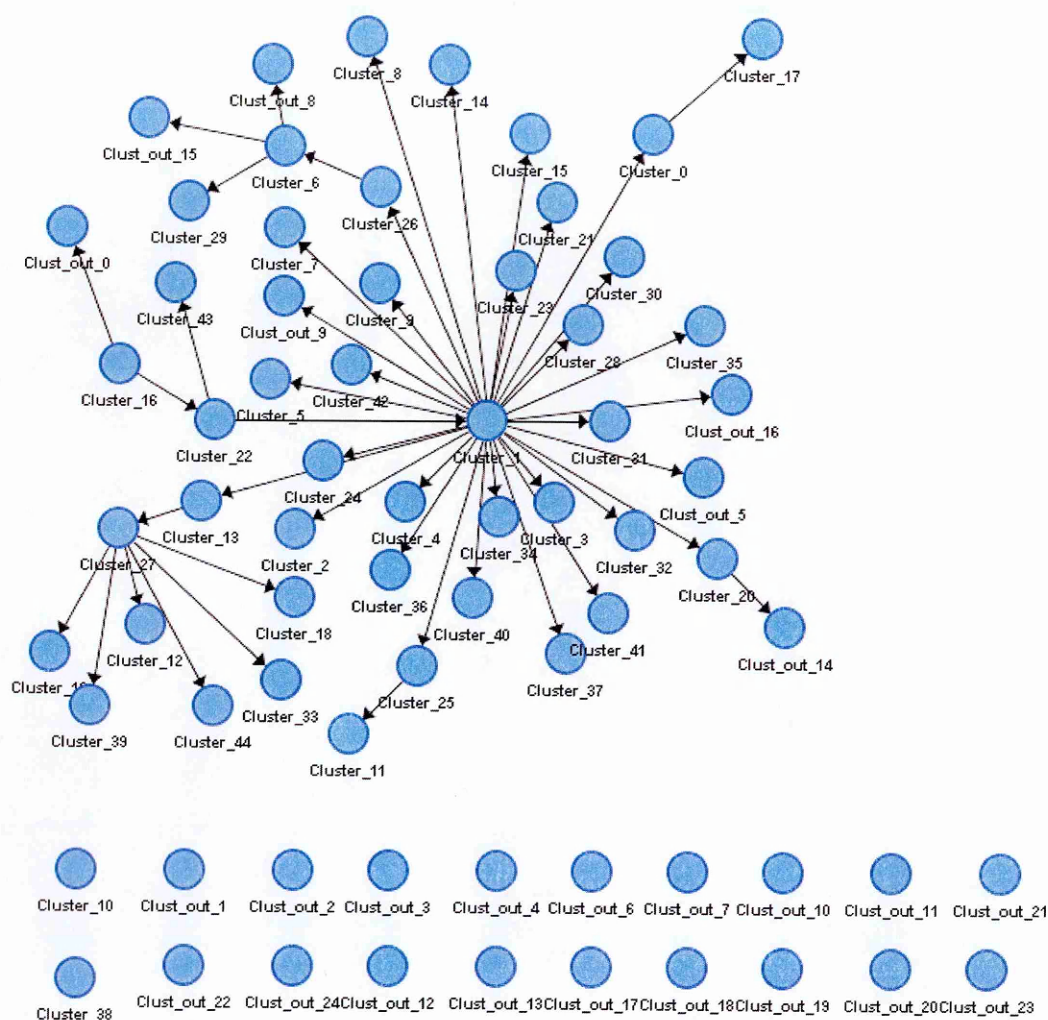


Figure 55: GRN of all 45 clusters and 25 clusters of Outliers. The 25 clusters of outliers are named as Clust_out_1, Clust_out_2, etc., in order to differentiate them from the other 45 clusters.

3.5.3.4 GRN Extraction of Specific Clusters

The GRNs of clusters 25 and 37 are chosen here to demonstrate how the coarse grained view of individual clusters can be used for visualising gene/probe interactions. Thus, the global view of the hMADS dataset as shown in **Figure 51** can be interpreted for deeper understanding of the genome as a whole.

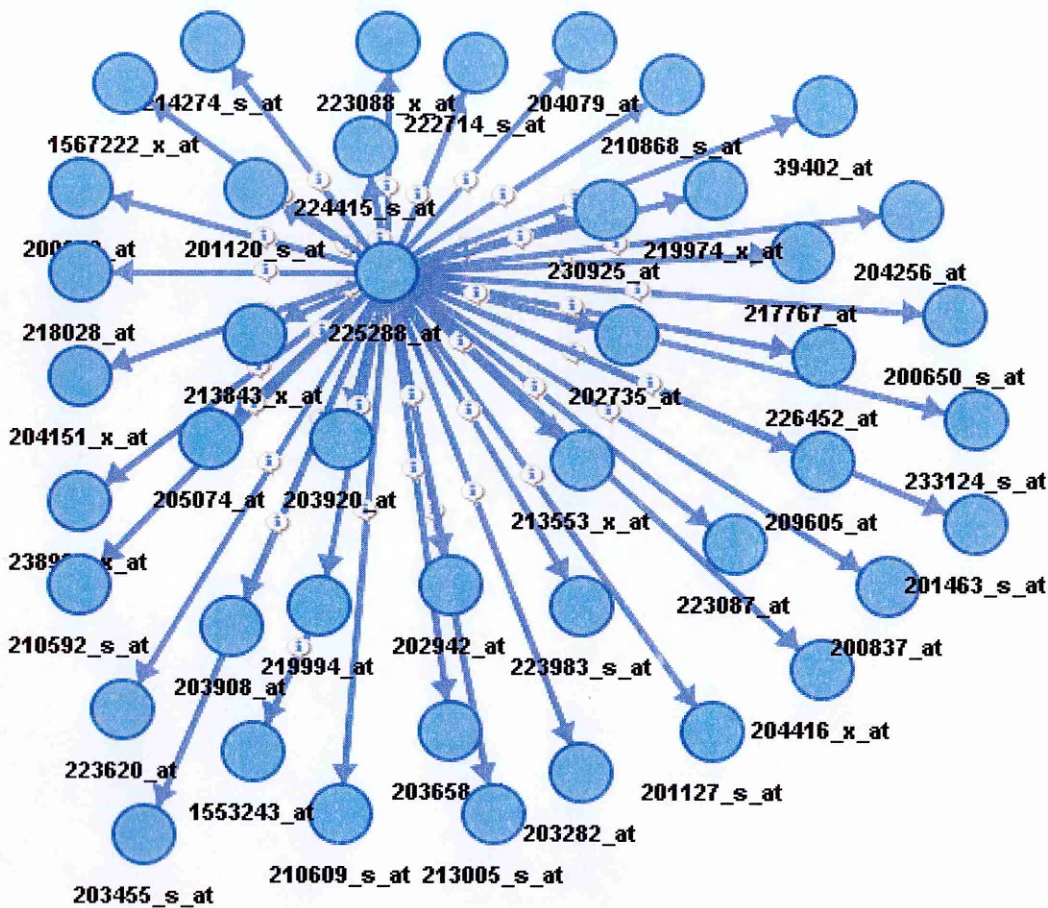


Figure 56: GRN of Cluster 37. Arc analysis of Cluster_37 showing the strength of the relationships between the probes; blue arc represents a positive relationship.

Figure 56 shows the extracted GRN for Cluster 37. This cluster contains 44 gene probes all of which are shown as directly dependent on probe 225288_at (COL27A1; collagen, type XXVII, alpha 1) shown at the centre of the GRN. It also shows that there is no interdependency between the child nodes.

In contrast, the Cluster_25 shown in **Figure 57** has 12 independent probes, which reflect the weak contribution they bring to the Bayesian score. Also a negative relationship is observed between 228731_at and 205193_at as shown by the red arc.

A plausible explanation for the negative association found within this cluster is that 205193_at may have been a false positive, i.e. it was clustered into Cluster_25 by chance.



Figure 57: GRN of Cluster 25. Arc analysis of Cluster_25 showing the strength of the relationships between the probes; blue arc represents a positive relationship, while a red arc represents a negative relationship

3.6 Conclusions

The hMADS microarray dataset used in this chapter, is a large scale genome-wide dataset, containing 54,634 gene probes. However, the sample data contained gene expression levels measured at only three time points. This represented previously uncharted territory for modeling gene regulatory networks using Bayesian methods firstly because, the dataset did not meet the minimum required sample size for statistical modeling, and secondly, the large number of probes introduced computational and interpretational challenges associated with high dimensionality; and finally, Bayesian structural learning has not been applied for extracting such large scale GRNs previously. The results presented in this work demonstrate that the dimension reduction techniques used in this study together with the application of Bayesian learning, can successfully extract biologically sensible relationships despite the challenges faced with high dimensionality and small sample size. However an alternative solution for the limited number of sample points would be to be less ambitious in terms of number of variables used to model the network.

Dimensionality reduction is an important process for learning genetic interaction or dependency from large sets of microarray data. A vary large p (number of genes or probes) can be an impediment to the studies due to large computational time and other challenges. Appropriate assessments need to be undertaken to avoid over-fitting in complex highly parameterised models such as biological systems. A usual approach to this is to select a subset of genes of interest and use this subset for model comparisons. Thus, gene subset selection is often included into the procedure and subsets of genes used in the validation studies. However, careful consideration should be given to the expression profiles of the genes under consideration to ensure that the resulting subset after the filtration remains a true representation of the original dataset.

The gene selection methodology used in this chapter meets this criterion: it combines a threshold for identifying differentially expressed genes based on a 2 fold-change with a statistical confidence measure based on p -value of 0.01. The low p -value used in this test means that the likelihood of false discovery is small.

In addition, the expression profiles of the genes were analysed, and the impact of different threshold values on gene selection was considered prior to the application of the filtration process to ensure that the results are not sensitive to the fold change criteria chosen. Through the application of this technique, the final dataset was reduced to 3866 probes (7% of the original data).

However, even $p = 3,866$ is too big for structure learning and also for the interpretation of extracted GRNs; hence further pre-processing methods were employed, through clustering genes with similar properties. Application of clustering analysis to this dataset showed that while the new Gaussian clustering algorithm, ArrayMiner2, produced consistent results over successive runs compared to SOM and k-Means methods, it was considerably slower than the latter two methods when analysing large datasets. These results presented also support (Falkenauer and Marchand, 2001), who found that the solution found by ArrayMiner2 was identical for each of the 20 successive runs on each of the data sets used³³, compared to the k-Means method which returned varying results.

All three methods require specification of the number of clusters before the start of each run. Hence, the SOM and k-Means clustering methods, which are fast and thus allow multiple runs within a short period of time, were used to provide an indication of the number of clusters present in the dataset. The strategy was to use this cluster size in the theoretically more superior ArrayMiner2 algorithm (see section 3.3.3.3) initially, and then incrementally increase the number of clusters until an optimum score based on minimum total intra-cluster variance is achieved. The SOM and k-Means methods however, produced somewhat differing number of clusters: SOM method indicated the presence of 42 clusters, while k-Mean produced 60 clusters. While the two algorithms do differ in their criteria and approach (see section 3.3.3), perhaps the presence of outliers as also confirmed by the ArrayMiner2 algorithm, may have also contributed to the differing number of clusters identified.

³³Based on a test involving a selection of rat and yeast datasets from GeneSpring®.

The results from ArrayMiner2 indicated an optimum number of 45 clusters. At this stage there were still a group of 437 unclassified probes or outliers. The unclassified probes were clustered separately to yield another 25 clusters without any further outliers.

A close examination of the derived clusters found that, although the clustering methods used generally produces groups of genes that are similar in terms of expression profile and thus are highly correlated, the presence of multiple probes for some of the genes in the hMADS dataset, may lead to inconsistent results due to the different signal intensities of the probes. For example, the probe IDs 203548_s_at, 203549_s_at for gene LPL (lipoprotein lipase) were ('correctly') clustered into two separate clusters by ArrayMiner2, Cluster_7 and Cluster_40, respectively due to their different signal intensities (see **Figure 49**, section 3.5.2).

This suggests that use of single probe approach to generating microarray data (one probe per gene), may yield more accurate results, for example, reduction of false positives, than probe set approaches.

The issue of small sample size was addressed by the use of a novel method (see section 4.4.2.1), termed artificial experiments, which used the combinations of sample data points in triplicate to their reference to generate 27 data points in total.

The resulting data was discretised as described in section 4.4.3, and the results were tested to ensure robustness to variations in the discretisation thresholds.

Bayesian belief networks were then used to build models for the genetic dependency in biological processes that was captured in the microarray experiments. Gene regulatory networks of the insulin receptor (IR) family, PPAR family, FOX family, CEBP family, MEF2, FABP, ADD1 and KLF, and probes with highly significant change in gene expression level were learned separately.

The derived GRN shown in **Figure 50** identified many biologically plausible relationships between genes despite the small number of time points over which the data was generated. Validation of these results is presented in section **3.5.3.1**. The results thus also confirm the effectiveness of the dimension reduction techniques used in this analysis.

The GRN extraction of the 45 clusters and outliers enable a global view of the network structure, while analysis of individual clusters demonstrated in this study facilitates deeper understanding of the genetic interactions.

4 GRN Extraction of Metabolic Stress of *E. coli*

4.1 Introduction

In this chapter, the genome-wide genetic networks of a prokaryotic expression system were extracted to identify the associations among stress related genes in recombinant protein production. The cDNA oligo, dye swap microarray data from *E. coli*, K12, genome generated as part of the research undertaken at the institute of Applied Biology at the University of BOKU, Vienna, which primarily concerns the production of recombinant proteins using *E. coli* as the host/vector system.

Optimisation is among the key objectives of the recombinant protein production on industrial scale; high yield and the quality of the product are the most important issues. For achieving high quantities of recombinant protein strong host/vector systems are utilised. The use of such strong host/vector systems rapidly lead to an overburden of the host cell metabolism. Thereby an endogenous nutrient starvation is triggered resulting in a collapse after a short time. The microarray was one of the technologies applied to monitor this process, in terms of endogenous influences of the recombinant protein as well as exogenous impacts on host cell metabolism.

This dataset used, however, was characterised by experimental noise, small sample size, and high dimensionality. These issues were tackled through the use of dimension reduction techniques, and application of multiple discretisation thresholds.

Bayesian Networks were employed for extraction of the genetic network, and further knowledge extraction techniques such as relationship analysis, variable clustering, and arc analysis were carried out to identify the influential genes. Target node analysis, and network simulation were then employed for learning the stimulatory effect these genes have on other related genes.

The extracted GRNs captured many biologically sensible relationships that were in agreement with published literature, KEGG and *E. coli* genome database EcoCyc. The good predictive performance achieved in this thesis show that the novel relationships identified in this work can be used with high confidence for understanding the workings of the *E. coli* metabolic stress pathway and for simulation.

4.2 Recombinant Protein Production

E. coli is one of the leading hosts for the efficient, cost-effective and large-scale plasmid-based expression and production of heterologous proteins (Baneyx, 1999). The extent of the bacterial stress response is determined by the specific properties of the recombinant protein, and by the rates of transcription and translation. Taking into account the capacities of the host for protein processing and physiological adaptation, production schemes can be developed that enhance volumetric productivity and sustainability of the process.

In order to attain the highest possible exploitation of the host cell's potential, host and vector need to be balanced by tuning the formation of recombinant protein. Therefore appropriate induction and cultivation strategies to keep the recombinant gene expression rate below the critical threshold level along with highest possible product formation over a prolonged production phase have been developed. Thereby the critical threshold level of recombinant gene expression for a particular product needs to be considered. Due to its complex nature, metabolic burden is not a significant variable and no readily measurable variables are yet available to deliver appropriate information about the synthesis potential of the cell.

4.2.1 Genome-wide GRN Extraction of *E. coli*

The features of the production strain *E. coli* HMS174 (DE3) (pET11a) on transcriptome and proteome level were analysed as part of ongoing research to find the key marker genes and proteins using two incorporated recombinant proteins, superoxide dismutase (SOD) and the green fluorescent protein (GFP) variant GFPmut3.1.

The monitoring process of the host organism and detection of marker proteins and genes were aimed to optimise the recombinant protein production. Due to the incomplete understanding of the interaction between the recombinant protein expression and the host cell metabolism many industrial processes are still far from optimal. It is thus essential to widen the understanding of the functional genomics of the host cell metabolism through stressed conditions in recombinant protein production and the genes associated with it.

4.2.2 The Role of Microarray Technology

Microarray analysis provides an efficient tool for screening transcriptional changes due to recombinant protein production and is therefore used to select highly upregulated promoters for the design of transcriptional fusions. Genes with similar function or genes situated on the same transcription unit often displayed similar expression profiles (Karin Dürschmid, 2003). The recording of transcription profiles on microarrays is a highly time and space efficient tool to screen the whole *E. coli* genome for stress responsive genes under process conditions of recombinant gene expression (VanBogelen, Greis *et al.*, 1999).

Microarray experiments are however, rather costly in terms of materials (RNA sample, reagents, chip, etc), laboratory manpower, and data analysis effort. As a consequence, such experiments often involve only a small number of samples. In many cases, the selected sample size may not be adequate to provide reliable statistical answers to the scientific question of interest, as explained in the previous Chapter 3.

4.3 Materials and Methods

4.3.1 Data Acquisition and Structure

The oligo cDNA, microarray data of *E. coli*, K12 genome, generated at the Institute of Applied Biology at the University of BOKU, Vienna, Austria, is used as the training dataset in this chapter. This dataset contained gene expression values in triplicate for all the 4416 open reading frames of the *E. coli* K12 genome.

All experiments were performed with *E. coli* HMS174 (DE3) (pET11a_hSOD) and *E. coli* HMS174 (DE3) (pET11a_GFPmut3.1), expressing the model proteins human superoxide dismutase (SOD) and the green fluorescent protein variant GFP mut3.1.

The data for the GFP experiment contained 5 time points, taken 12, 32, 90, 150 and 180 minutes after induction and the data for the SOD experiment contained 8 time points taken 15, 22, 45, 68, 90, 150, 180 minutes after induction.

Pre-processing of the microarray data was done using Bioconductor package limma. The data was normalized using print-tip LOWESS normalization within arrays and A-quantile normalization between arrays.

The LOWESS method was used in order to reliably discriminate regulation from non-specific fluctuations in the data set. The statistical evaluation revealed transcripts which, compared to the steady state sample, had highly significance difference in expression levels after the induction of recombinant protein.

The samples were labelled using the fluorescent dyes Cyanine 5 (Cy5), and Cyanine 3 (Cy3). After hybridization, the slides were scanned at a 10 μ m resolution. Probes labelled with Cy3 were scanned at 532 nm, and probes labelled with Cy5 were scanned at 635 nm.

4.3.2 Data Pre-processing

The low number of data points ($n=5$) for GFP and ($n=8$) for SOD compared to the number of genes ($p=4416$) in the training dataset described above raises issues relating to high dimensionality (see Chapter 3), and required application of various pre-processing methods prior to Bayesian structure learning.

4.3.2.1 Missing Values and Data Noise

The cDNA microarray data used for the generation of the *E. coli* data is usually processed by subtracting the background from the fluorescence values of the spots (foreground).

The exclusion of background intensities from the calculation of gene expression levels may however lead to inaccurate expression data as measured intensity values are affected by numerous sources of fluctuation and noise, and hence foreground data alone cannot be relied on to provide a value that is representative of the expression value of a gene. (Ting Lee, 2000) found that for spotted cDNA arrays, there is a ~5% probability that the hybridization of any single spot containing complementary DNA will not reflect the presence of the mRNA, and a ~ 10% probability that a single spot will provide a signal even if the mRNA is not present.

Hence, the foreground minus the background values was calculated and this was combined with the dataset containing just the foreground data. 223 genes that had more than 4 missing values in the foreground minus background dataset were removed and the rest of the dataset containing 4193 genes were then filtered as described in the next section.

Missing values were replaced using EM algorithm as encoding the assumptions as described in section 1.6.3.2.

4.3.2.2 Filtering

A crucial step in data analysis involving gene expression data is identifying those genes which are differentially regulated. The fold change filter described in Chapter 3 provides a simple and intuitive approach to filtering out invariant genes enabling the selection of subset of genes of interest. It is measured as the ratio between the sample expression and its control. Typically, a difference is considered significant if at least two-to three-fold is observed (Draghici, 2002).

In this study, a fold change of 2 was selected based on analysis of the data including sensitivity of gene selection to various fold change criteria as well as alternative methods.

The fold change analysis was carried out using Gene Spring, which identified 931 genes as differentially regulated from the dataset containing 4193 genes.

The number of missing values after filtration was, GFP: 216 values (4.6%), and SOD: 1018 values (13.7%), highlighting the issue of noise present in microarray data generation.

4.3.2.3 Discretisation

Log transformation of the expression ratios, $\log_2(\text{Sample})/(\text{Reference})$, was carried out prior to discretisation for the final dataset containing 931 genes.

The log ratios of the expression levels were then discretised into three classes. For example, the ($\log_2$) expression level between ≤ 0.1 and ≥ -0.1 was classified as ‘normal’ (‘no change’), > 0.1 as ‘up’, and < -0.1 as ‘down’.

However, the results obtained when learning structure might be sensitive to the threshold value used in the discretisation method, and hence detailed checks needs to be carried to ensure this is not so. An example illustrating this potential issue is discussed below.

Over a number of sample data points, the marginal probabilities of being upregulated, downregulated and no change (normal) would be calculated, and this information is used for calculating the dependencies (conditional probabilities) when extracting the GRNs using Bayesian networks. The issue arises, when the conditional probabilities calculated are a reflection of the discretisation threshold used, and not the true dependencies between genes.

Consider the example presented in Table 12, which looks at the discretisation of gene expression ratios for a single sample point. The expression ratios of genes 1, 2, and 3 are classified differently when a discretisation threshold of 0.2 is used instead of, say 0.1. Under discretisation method using threshold value of 0.1, gene 2 is upregulated, while gene 3 is downregulated. However, under discretisation method using 0.2, genes 2 and 3 would be classified as ‘normal’ (no change); only in the case of Gene 1 the classifications are the same.

	Log Expression Ratio	Discretisation Threshold	
		0.1	0.2
Gene 1	1.30	Up	Up
Gene 2	0.19	Up	Normal
Gene 3	-0.15	Down	Normal

Table 12: Impact of Discretisation Thresholds. Discretisation of gene expression values using thresholds of $\log_2$ 0.1 and $\log_2$ 0.2. Under threshold of $\log_2$ 0.1, expression level between ≤ 0.1 and ≥ -0.1 is classified as ‘normal’, > 0.1 as ‘up’, and < -0.1 as ‘down’; under threshold of $\log_2$ 0.2, expression level between ≤ 0.2 and ≥ -0.2 is classified as ‘normal’, > 0.2 as ‘up’, and < -0.2 as ‘down’.

The sensitivity to discretisation would be less when larger samples are used, provided that the expression profiles of the genes vary across the sample data/time points such that the impact is counter balanced. However, this is not guaranteed even for larger samples. Hence, the results need to be tested for robustness to various discretisation thresholds.

This issue may be more pertinent for data analysis involving a smaller sample such as the *E. coli* dataset considered in this study. Figure 58 shows significant changes in the number of genes classified as ‘normal’ under the discretisation thresholds ($\log_2$) 0.1, 0.2, 0.4, and 0.5 for each of the sample points for both GFP and SOD proteins. Consequently, similar results were also obtained for ‘up’ and ‘down’ classifications. Discretisation of the original dataset of 4,416 genes also confirmed these results. The GRNs extracted under these different discretisation thresholds showed significant differences in the network structure.

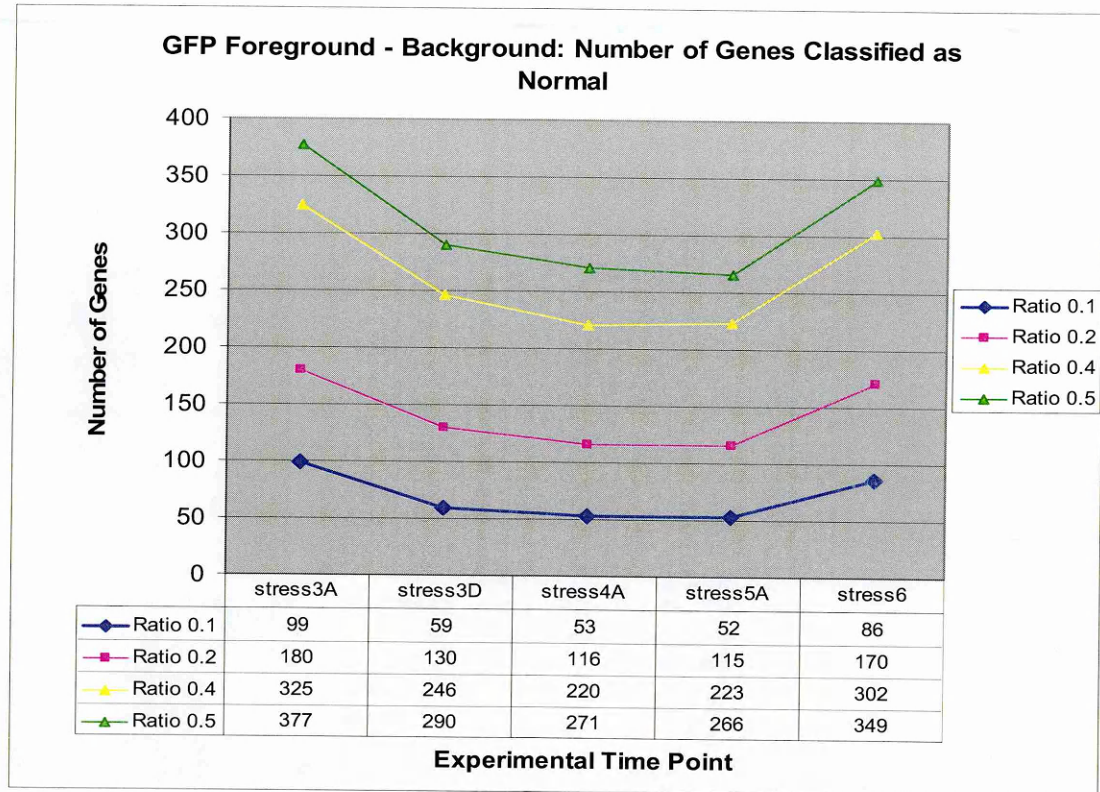
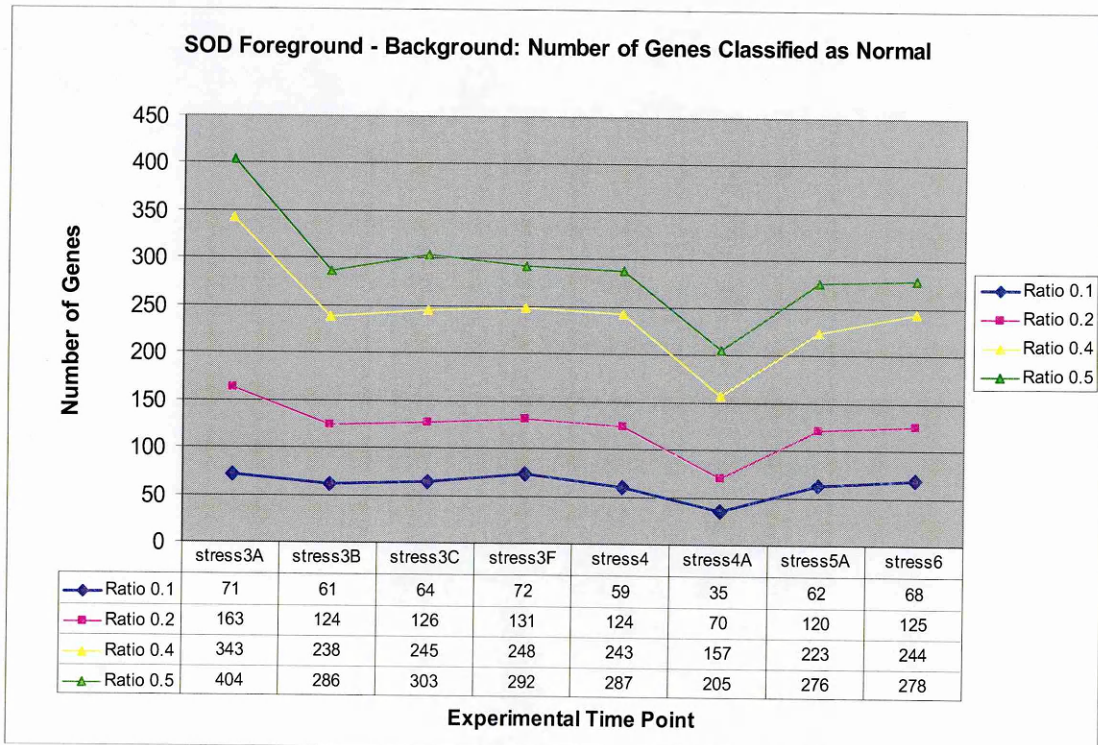


Figure 58: Discretisation of gene expression of GFP and SOD proteins under different thresholds. The number of genes classified as ‘Normal’ under discretisation thresholds (log) 0.1, 0.2, 0.4, and 0.5 for *E. coli* proteins GFP and SOD.

This suggests that the expression profiles of the *E. coli* genes are highly sensitive to the value of discretisation threshold used. Hence it would be reasonable to conclude no single discretisation threshold could be used to represent the data profile accurately in this instance, unlike the cases of yeast cell cycle regulation and hMADS data considered in chapters 2 and 3 respectively, where the impact was not significant.

One solution is to model the data using continuous distributions. However, as discussed in section 1.6.1 (Chapter 1) and in 2.4.1 (Chapter 2), use of continuous distributions has many drawbacks. Firstly, it can only detect dependencies that are close to linear. It is not likely to discover combinatorial effects, for example, it portrays that a gene is over expressed only if several genes are jointly over expressed, but not when just one of them is not over expressed (Friedman, 2000). Secondly, when the real values of the variables are used, there is no single representation that can represent all possible densities, and hence use of continuous distribution (e.g. Gaussian, Gamma) for modelling gene expression data at best represents approximation of the underlying data series. Finally, in modelling a small sample dataset such as this, it would be unreasonable to assume that the sample parameters such as mean and standard deviation calculated represent the true population parameters; use of a continuous distribution such as Gaussian (Normal) distribution requires specification of these parameters.

In this chapter, a novel approach was employed to address the issues raised above: the use of multiple discretisation thresholds for the discretisation of data. The data was discretised using thresholds (log) 0.1, 0.2, 0.4 and 0.5 and these four independent sets of data were combined for GRN extraction.

4.3.2.4 Extraction of GRNs

GRN extraction was carried out using the techniques validated with Yeast data in **Chapter 2** and described in detail in sections 2.2.2-2.2.6. The latest version of BayesiaLab, 4.1.9, was used for the Bayesian structure learning.

4.4 Results and Discussions

4.4.1 Data Pre-Processing

The 2-fold change filtering method shown in

Figure 59 identified 931 genes as differentially regulated from the dataset containing 4193 genes. The differentially regulated genes selected for discretisation and structure learning are shown by the red lines and the genes which were filtered out are shown by the yellow lines. The few red lines shown in the middle part containing the yellow lines are due to, some of the expression values out of the five sample points of the differentially regulated genes, falling within the 2-fold change.

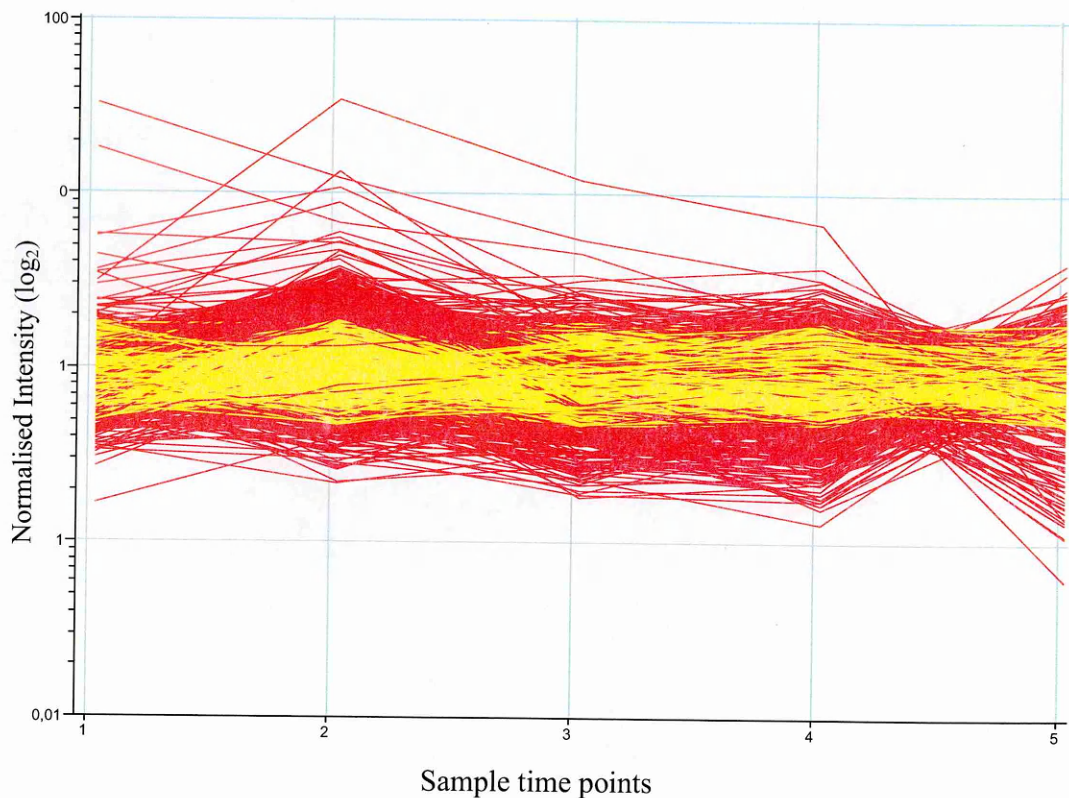


Figure 59: Application of Fold Change Method for Gene Filtration of CFP. Vertical axis measures the fold change in logarithms; a 2-fold change is equivalent to $\log_2 0.5$. Horizontal axis shows the sample time points for the CFP experiment: taken 12, 32, 90, 150 and 180 minutes after induction. The red lines show the genes that are differentially regulated and the yellow lines at the centre shows the genes that are filtered out.

In the following sections, GRNs of the *E. coli* genome are extracted using the Bayesian structure learning and the results are validated by agreement with published literature and genome databases KEGG and EcoCyc.

4.4.2 Extracting Gene Regulatory Networks

In this section genome-wide GRNs of *E. coli* are extracted, interpreted and validated from the filtered dataset containing 931 differentially expressed genes. GRNs are extracted for the GFP and SOD protein experiments.

4.4.2.1 GRN Extraction for the CFP Protein Experimental Data

Figure 60 shows an overview of the extracted GRN for the GFP experiment. 311 genes are shown as associated; the rest are found to be independent. In order to analyse the specific genetic relationships, the circular cluster of genes shown on the centre right of the GRN is considered. This cluster, containing 16 genes, is chosen as it appears to be dominated in terms of the number of direct associations by a single gene, *pflB*, shown at the centre of the circle; hence it would be interesting to consider the functional role of this gene in relation to the other genes.

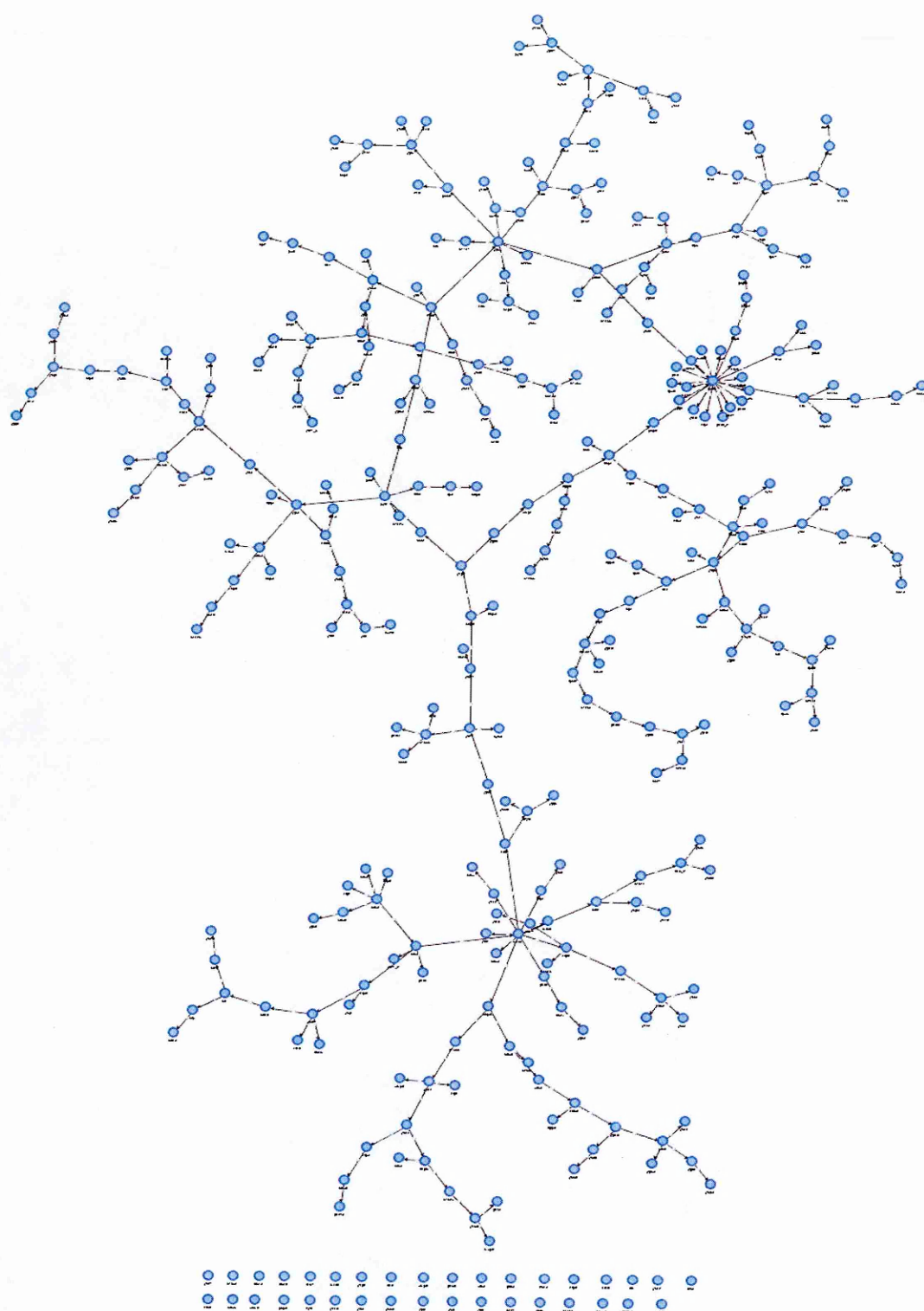


Figure 60: Extraction of Genome-wide GRN for the GFP Experiment. The nodes represent the genes and the arcs represent the relationship between the genes. Genes that are conditionally independent of the GRN are shown at the bottom of the figure. Not all independent genes are shown, and gene names shown in small letters due to space restrictions.

This circular cluster is shown in **Figure 61**. Gene *pflB* (pyruvate formate-lyase type B) appears to be the father or dominant gene of this small cluster. All the genes in this cluster are shown as directly dependent on *pflB*.

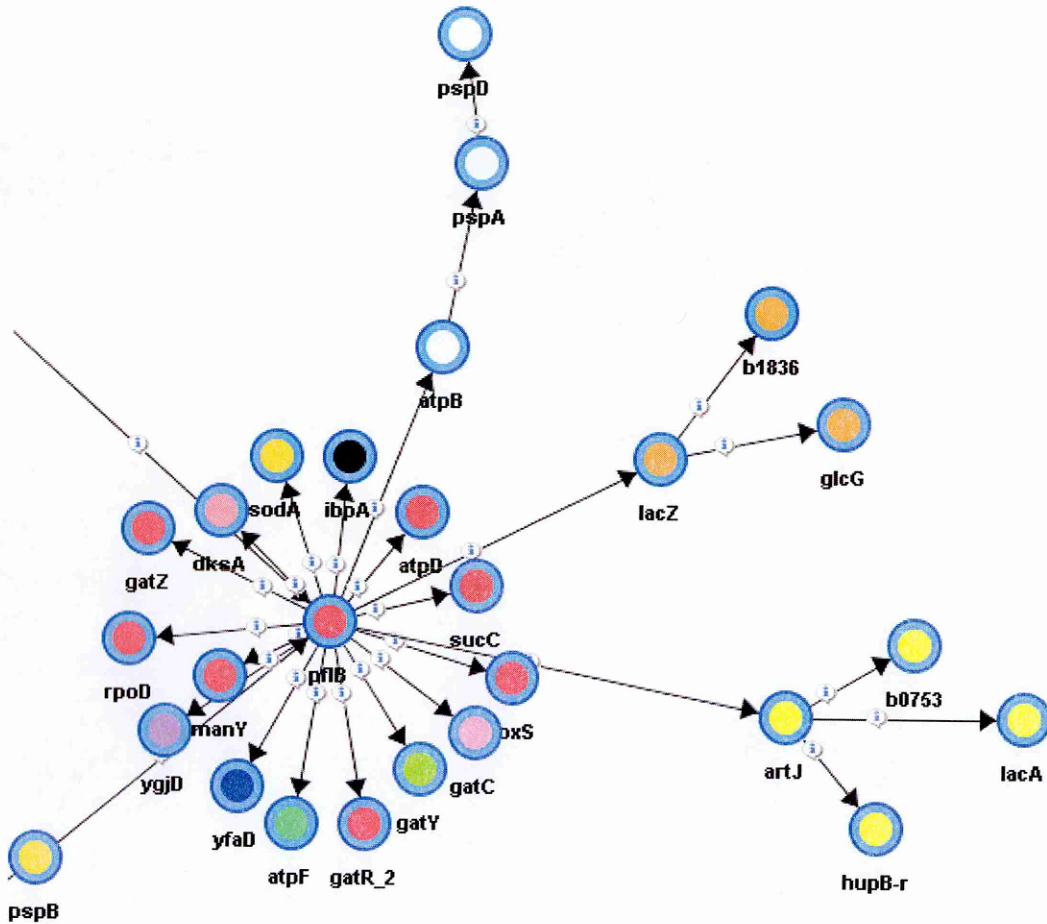


Figure 61: Inset view of the Circular Cluster of Genes in the GRN for GFP. The genes that are directly associated with the gene *pflB*; *pflB* is the ‘father’ node of this cluster. The colour coded sets of genes are generated using the Taboo algorithm, which creates groups of variables that are close in terms of their mutual information as measured by Kullback-Leibler divergence (Eguchi and Copas, 2006). This technique is known as variable clustering.

The colour coding of the genes within the GRN is based on the Kullback-Leibler divergence score (Eguchi and Copas, 2006). This technique is a technique known as variable clustering. It is a measure of the differences between the probability distributions represented by the network with the arc corresponding to the relation, and without this arc. Hence, this measure is equal to the mutual information between the parent and the child (see section 1.6.4).

For example, *pflB* is closed to genes *sucC*, *soxS*, *gatZ*, *gatR_2*, *atbB*, *manY*, *rpoD*, and *dksA* based on this criterion³⁴; these genes are colour coded in red. KEGG pathway database shows for example *pflB* and *sucC* (succinyl-CoA synthetase) play important roles in propanoate metabolism, and *gatC*, *gatY*, *gatZ*, and *lacZ* shown as associated with *pflB* in this cluster are involved in galactose metabolism. *pflB* also plays important roles in butanoate, and pyruvate metabolism pathways. KEGG also shows that *manY* is a phosphotransferase system (PTS) gene and that operons *atpD* and *atpF* play vital roles in oxidative phosphorylation (Ziemke and McCarthy, 1992). However, KEGG did not find any direct associations with other genes in this cluster such as *rpoD*, *dksA*, *sodA*, *ibpA*, *soxS*, *gatR_2*, *yfaD*, and *ygjD*.

The pathways and biological functions for the *E. coli* K12 genes obtained from the KEGG pathway database are shown in Appendix V.

In summary, *pflB* has a well known function and plays an important role in many metabolic pathways. The extracted GRNs show that it is also a parent gene for many others genes. The many associations captured in this cluster shown in **Figure 61**, but not found in previously published literature and genome databases suggests that these may have been due to the stress conditions of the experiments conducted. This indicates a potential role for this gene as a marker in recombinant protein production.

An arc analysis of the relationships between the genes in this cluster is shown in **Figure 62**. It shows that all the genes in the circular cluster except for *idpA* and *yfaD*, with red arcs, are negatively associated with *pflB*. The thickness of an arc is directly proportional to the absolute value of Pearson's linear correlation coefficient.

³⁴ Mutual information is closely linked to the Pearson correlation coefficient.

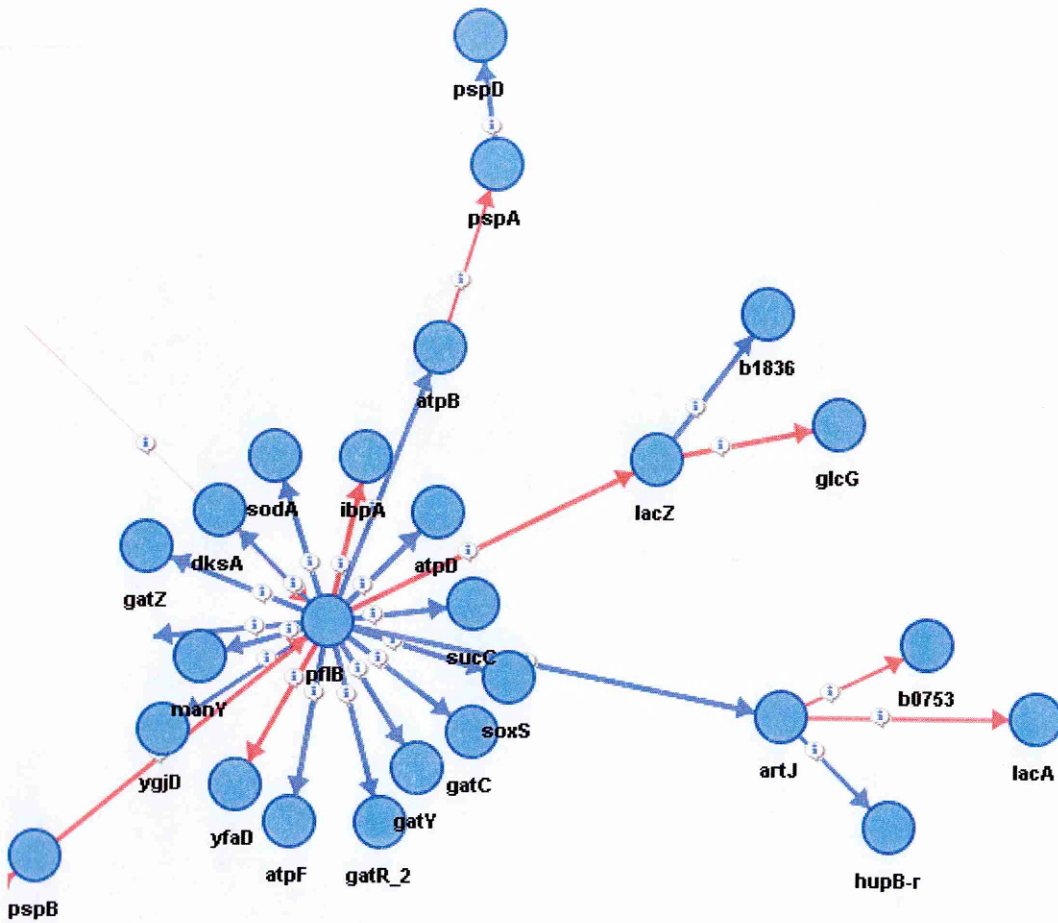


Figure 62: Arc analysis of the Circular Cluster of Genes in the GRN for GFP. The strength and the length of the arcs are proportional to the strength of the probabilistic relation and closeness of the relationship respectively. Blue arc represents a positive relationship, while a red arc represents a negative relationship.

4.4.2.2 GRN Extraction for the SOD Protein Experimental Data

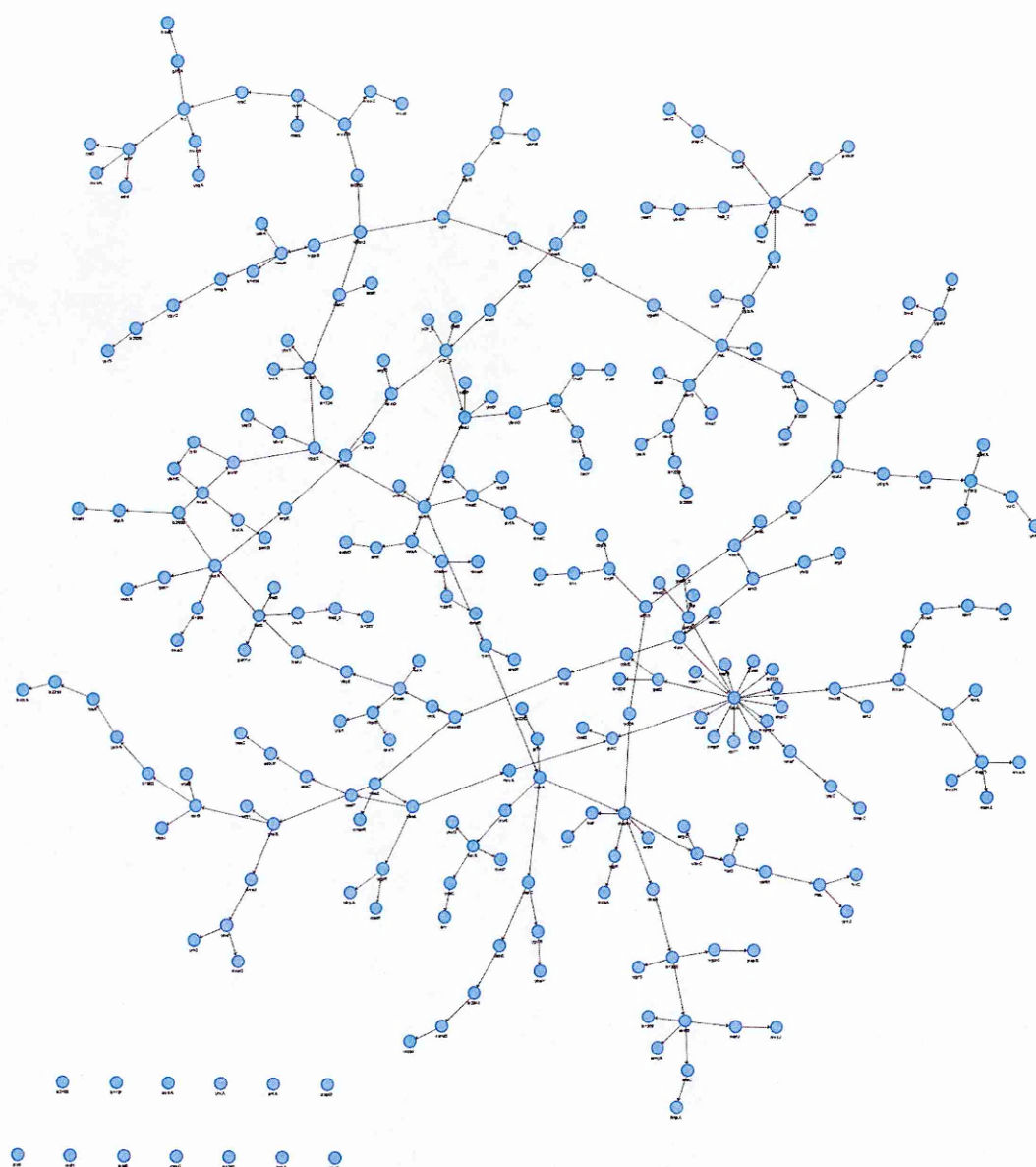


Figure 63: Extraction of Genome-wide GRN for the SOD Experiment. Genes that are conditionally independent of the GRN are shown at the bottom of the figure. Not all independent genes are shown due to space restrictions.

The extracted GRN for the SOD experiment, which displays associations for 267 genes, is shown in **Figure 63**. Analysis of this GRN reveals that a circular cluster on the centre right of the GRN containing 11 genes are all positively associated with the father gene *cstA*, which codes for carbon starvation protein; this is shown by the blue arrows between nodes. However, both genome databases KEGG and EcoCyc did not show any associations for this gene. Nevertheless, given the experiments being studied is the metabolic stress system of *E. coli*, a carbon starvation protein such as *cstA* could be expected to be activated, and our results indicate that the associations shown in this GRN are perhaps indication of novel relationships; these should be investigated further through laboratory experiments. In relation to the others genes in this cluster, KEGG identifies that both *gatB* and *manY* are phosphotransferase system (PTS) genes; both *gatB* and *gatD* play vital roles in galactose metabolism; *ompF* is a shown as a two-component system gene; and *rpsB* is shown as playing an important role in ribosome.

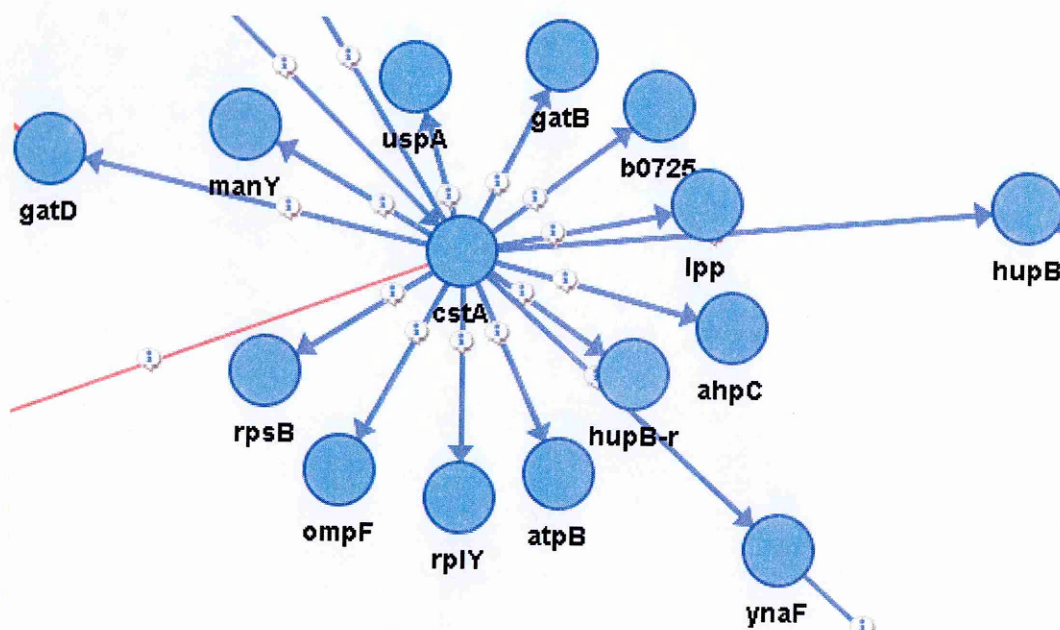


Figure 64: Inset view of the Small Circular Cluster of Genes in the GRN for SOD. The genes that are directly associated with the gene *cstA*; *cstA* is the ‘father’ node of this cluster. Blue arc represents a positive relationship, while a red arc represents a negative relationship.

KEGG did not find any associations for *uspA*, *b0725*, *lpp*, *hupB*, *hupB-r*, *ahpC*, *rpiY* and *ynaF*. However, the strong correlations shown between the genes from this cluster, and presented in the relationship analysis report in **Table 13** indicate that these relationships may be novel relationships.

Relationship Analysis Report: SOD			
Parent	Child	K-L	PMCC
<i>cstA</i>	<i>ahpC</i>	0.499	0.966
<i>cstA</i>	<i>atpB</i>	0.499	0.966
<i>cstA</i>	<i>b0725</i>	0.504	0.901
<i>cstA</i>	<i>gatA</i>	0.503	0.910
<i>cstA</i>	<i>gatB</i>	0.499	0.966
<i>cstA</i>	<i>gatD</i>	0.503	0.923
<i>cstA</i>	<i>hupB</i>	0.402	0.895
<i>cstA</i>	<i>hupB-r</i>	0.499	0.966
<i>cstA</i>	<i>lpp</i>	0.497	0.894
<i>cstA</i>	<i>manY</i>	0.499	0.966
<i>cstA</i>	<i>ompF</i>	0.499	0.966
<i>cstA</i>	<i>prlC</i>	0.397	-0.598
<i>cstA</i>	<i>rplY</i>	0.469	0.919
<i>cstA</i>	<i>rpsB</i>	0.499	0.966
<i>cstA</i>	<i>uspA</i>	0.506	0.954
<i>cstA</i>	<i>ynaF</i>	0.506	0.904
<i>dps</i>	<i>cstA</i>	0.498	0.838
<i>hupB</i>	<i>artJ</i>	0.455	0.961
<i>hupB</i>	<i>hns-r</i>	0.358	0.883
<i>ynaF</i>	<i>yajC</i>	0.462	0.739

Table 13: Relationship Analysis Report for *cstA* cluster from the GRN for SOD experiments. The nature and strength of the relationships between genes from relationship analysis report. The higher the value of Kullback-Leibler (K-L) divergence and Pearson correlation coefficient (PCC), the stronger is the association.

4.4.2.3 Relationship Analysis and Validation of extracted GRNs

Table 14 shows the nature and strength of the relationships for some of the genes in the extracted GRNs for GFP and SOD. This data was obtained from the BayesiaLab’s Relationship analysis report, which shows how any two genes in the learned networks are related and their strength of association as measured by the Kullback-Leibler (K-L) divergence and Pearson correlation coefficient (PCC) described in the previous sections. These relationships identified purely from data through the application of Bayesian learning, are validated below against the published literature and the *E. coli* genome database EcoSyc.

Bayesian Modelling of <i>E. coli</i> : Evidence for Genetic Relationships from Data							
GFP				SOD			
Parent	Child	K-L	PCC	Parent	Child	K-L	PCC
artQ	oxyR	0.834	0.668	dnaK	mopB	0.756	0.881
atpA	nuoC	0.912	0.730	dnaK	ibpB	0.412	0.375
atpA	ybeZ	0.842	-0.938	gatA	cstA	0.503	0.910
atpC	nuoL	0.647	0.444	gatA	gatR_2	0.618	0.563
dksA	pflB	0.398	0.743	gatD	cstA	0.503	0.923
gatD	nuoL	0.776	0.881	gatY	ibpA	0.540	-0.837
gatY	pflB	0.600	0.961	grpE	ybeZ	0.754	0.875
gatZ	pflB	0.607	0.891	grpE	dnaJ	0.753	0.686
grpE	ycfF	0.704	0.564	ibpB	rpoS	0.368	0.335
ibpA	pflB	0.600	-0.961	manY	cstA	0.499	0.966
manY	pflB	0.605	0.931	nuoG	nuoI	0.539	0.491
nuoH	nuoF	0.413	0.844	nuoG	nuoM	0.439	0.518
nuoL	nuoF	0.647	0.864	ybeZ	mopB	0.760	0.849
ybeZ	atpA	0.842	0.675	ybeZ	grpE	0.754	0.686

K-L: Kullback-Leibler divergence

PCC: Pearson's Correlation Coefficient

Table 14: Relationship analysis report for GFP and SOD GRNs. The nature and strength of the relationships between genes; genes considered include operons atp, nuo and gat family and heat shock response genes and stress induced genes.

The results show that the methods employed in this study were successful in identifying many biologically plausible relationships despite the noise, missing values present in the data and the small number of experiments used to generate the data.

In some organisms, such as *E. coli*, there are many sets of genes that are already known to be transcribed together and hence are strongly co-regulated. These genes that are transcribed into mRNA together and expressed into protein are known as operons.

Other operons such as the *nuo* and *atp* operon are found to be down regulated in GFP and SOD experiments, which strongly suggests a global regulation taking place during stress response. It was also found in the GFP experiment that *atpA* and *atpC* have direct dependency with four other genes as shown in **Table 14**, including *atpA* influencing *nuoC*, *atpC* influencing *nuoL*, and *atpQ* influencing *oxyR* and *ybeZ* influencing *atpA*.

Table 14 also shows *gatD* influences *nuoL* and two other direct dependencies among *nuo* operons: *nuoH* influencing *nuoF* and *nuoL* influencing *nuoF*. The results for the SOD data also show two direct dependencies between *nuo* operons: *nuoG* influences both *nuoL* and *nuoM*.

(Baneyx, 1999) found that the ATP-dependent *dnaK*-*dnaJ*-*grpE* and *groEL*-*groES* systems interact closely. “*dnaJ*, which independently binds folding intermediates, activates *dnaK* for tight substrate binding and might direct it to high-affinity sites. The nucleotide exchange factor *grpE* mediates complex resolution: released proteins may either fold into a proper conformation, be recaptured by *dnaK*-*dnaJ* for additional cycles of interaction or be reversibly”. The experiment for SOD shows that *grpE* and *dnaJ* are positively associated with a correlation of 0.69.

(Nobelmann and Lengeler, 1995) found that the genes *gatA*, *gatB*, *gatC*, *gatD*, *gatR*, *gatY*, and *gatZ* are from the galactitol (*gat*) family of operons. Many of these relationships were correctly captured by both GFP and SOD sample experiments as shown in **Table 14**.

Figure 65 shows the validation of the results extracted against one of the pathways of the *E. coli* K12 genome from the KEGG database: the oxidative phosphorylation pathway. Pink cells represent the genes identified by the extracted GRNs that are also present in this pathway. As mentioned previously, a summary showing all the pathways and biological functions for the genes from the KEGG pathway database are shown in Appendix V.

Overall the validation of the results for the extracted GRNs in this study shows they are supported by biological evidence. Further, this study found strong evidence for many complex and previously unknown relationships, which warrant further investigation through laboratory experiments.

4.5 Conclusions

In this chapter, the genome-wide genetic networks of *E. coli* prokaryotic expression system was extracted to identify the associations among stress related genes in recombinant protein production. *E. coli* is one of the leading hosts for the efficient, cost-effective and large-scale plasmid-based expression and production of heterologous proteins (Baneyx, 1999). Taking into account the capacities of the host for protein processing and physiological adaptation, production schemes can be developed that enhance volumetric productivity and sustainability of the process.

- The multiple discretisation was applied as dimension reduction technique to extract SOD and GFP cDNA oligo microarray experiment data of *E. coli*, K12, genome containing ORFs for 4416 genes.
- Bayesian Networks were used for the extraction of the genetic network, and further knowledge extraction techniques such as relationship analysis, variable clustering, arc analysis, and target node analysis were employed for the analysis and interpretation of the genetic associations found.
- The GRNs extracted from the GFP and SOD experiments provide a more comprehensive view of the metabolic stress system of the recombinant protein production. The extracted GRNs captured many biologically sensible relationships that were in agreement with published literature and genome databases KEGG and EcoCyc.
- Several of genes that have highly significant change in expression level, for example, *gat* family, *ibpA*, *ibpB*, *mopA*, *mopB*, *ybeZ* and *artQ*, are shown to regulate each other in the GFP and SOD experiments (see Table 14), and a comparison with KEGG pathway database (Appendix V) shows many of the

associations extracted by the GRNs in this work are in agreement with those findings from laboratory experiments. Further, this study found strong evidence for many complex and previously unknown relationships, such as the strong associations shown by the cSTA gene with several other genes in the extracted GRN for SOD (see **Table 13**), which warrant further investigation through laboratory experiments.

Finally, the analysis presented in 4.4.2.1 indicates a possible role for gene pflB as marker for recombinant protein production. However, this analysis was performed only on a subset of the extracted GRNs and it is beyond the scope of this study to extend the analysis in such detail for the rest of the network.

Overall, this study has demonstrated that the extracted GRNs in this work will help gain better understanding of *E. coli* as a host/vector system for the optimisation of recombinant protein production process.

5 Conclusions

5.1 Summary

The aim of the work described in this thesis was to develop methods for the extraction of gene regulatory networks from expression data, and use these methods to capture complex and previously unknown relationships between genes in specific biological applications. This has been accomplished through the application of a Bayesian probabilistic approach for learning the structure of gene expression from three large scale microarray gene expression datasets from *S. cerevisiae*, *E. coli* and human stem cells.

The models extracted have rich probabilistic semantics, which better fits the stochastic nature of both the biological processes and the noisy nature of microarray gene expression data, and can handle hidden variables and missing data in a structured way. They are also capable of being extended to incorporate additional information such as experimental conditions, time/stage of the sample which can be used to identify cyclic regulations, background variables (e.g. clinical procedure employed in obtaining the sample), and exogenous cellular conditions.

The work succeeded in extracting many biologically plausible genetic relationships from the experimental data, relationships which have been supported by evidence from publicly available literature and databases such as KEGG, MIPS, SGD, and, PubMed. It has been demonstrated that the methods developed are capable of discovering causal relationships, order relations (e.g. discovering dominant genes), and finer intra-cluster structure.

The Taboo search algorithm employed throughout this study is a dynamic neighbourhood search technique that utilises memory and heuristic techniques in its own operating rules to obtain a global optimum. Meeting of this criterion ensures that the network which best describes the data is extracted. Many computational experiments have shown that *Taboo Search* has now become an established optimization technique which can compete with almost all known techniques and which, by its flexibility, can beat many classical procedures such as maximum likelihood (Hertz and de Werra, 1990).

Two novel knowledge extraction techniques, target node analysis and network simulation, were applied in this study, which demonstrate how quantitative and qualitative information captured within the Bayesian networks (BN) can be used to simulate the nature of interaction between genes, for example to test if a gene is influencing or suppressing another and the strength of these relations in the network. The GRNs we extracted are intuitive, and provide excellent visual interpretation of the direction and strength of the associations between the genes.

The pre-processing methods used exhibit, how methods such as gene selection and dimension reduction techniques, multi-discretisation, clustering, and generation of artificial experimental data can be employed to truly large scale datasets to obtain a data matrix that is reduced in size, statistically sound and with minimal loss of biological information.

It was also demonstrated how the addition of prior knowledge to the learnt network can improve the network structure extracted purely from data.

5.2 Extraction of Gene Regulatory Networks

5.2.1 Yeast Cell Cycle Regulation

A much studied microarray gene expression data set from a well characterised organism, yeast cell cycle regulation data, was used to develop and validate the BN technique. From this data (Spellman, Sherlock *et al.*, 1998) had putatively identified 800 genes as cell cycle regulated. The Bayesian learning described in this thesis focussed on a discretized subset of 161 known cell cycle regulated genes identified by traditional wet lab methods (104 genes) and KEGG database (additional 57 genes; 31 genes overlapped).

The derived gene regulatory networks captured many biologically plausible genetic associations between known cell cycle regulated genes, which have been supported by biological evidence. For example, the immediate cluster of SIC1 i.e. direct child node of SIC1, in our extracted GRN (Figure 12, Chapter 2) contained ASH1, TEC1, CHS1, EGT2, PCL9, CTS1, and CDC6. This can be compared to: “SIC1 cluster comprises 27 genes, including EGT2, PCL9, TEC1, ASH1, SIC1, CHS1, and CTS1” (Spellman, Sherlock *et al.*, 1998).

The order relations of the learned network showed remarkable consistency in terms of how certain dominant genes tend to dominate the order relations (i.e. they appear before other genes) in the extracted Bayesian networks. For example, our extracted GRN (Figure 12, Chapter 2) included seven of the top 13 genes (out of 800 genes listed by (Friedman, Linial *et al.*, 2000) as the dominant genes with the highest scores in the discrete multinomial experiments: MCD1 (2nd most dominant out of the 800), CLN2 (3rd), RFA2 (5th), POL30 (8th), CLN1 (10th), MSH6 (12th) and CDC5 (13th). The other six genes (YLR183C, SRO4, CSI2, YOX1, RSR1, and YBR089W) were not in the 161 genes considered in our study.

Employing target node analysis, it was possible to study the *relative* significance of the association of all other genes to a gene of interest, for example, MCD1 (an essential gene required for sister chromatid cohesion in mitosis and meiosis), and SIC1 (Phosphorylation of SIC1p by CLN-CDC28p complexes is required G1-S transition). This feature could be used for further data analysis in gene expression studies, for example, how protein expression (transcription factor) and gene expression data are linked for each gene and also for answering specific questions, such as how important a particular gene is in a GRN and how other genes in the network influence one gene.

The models learned proved robust to variations in data. For example, the GRNs extracted under individual cell cycle phases such as M/G1, G1 were largely preserved when the data for these two phases were combined together. The new GRN displayed the two sets of genes as two main clusters within the overall structure, with a relatively small degree of overlap. When all 104 genes were analysed together, the resulting network showed strong similarity to the actual classification of the phases of yeast cell cycle. The clustering methods carried out in (Spellman, Sherlock *et al.*, 1998), did not show how genes are associated in clusters or which genes from different phases link to other phases, i.e. which genes the next phase of the cell cycle is dependent on. The findings in this thesis therefore represent a step forward in relation to Spellman's original work.

As an extension of the work, some known gene associations from the in literature were added to the GRN to evaluate the effect of prior knowledge on network extraction. A comparison of results (Table 5, Chapter 2) showed that of the 31 relations considered, 27 relations were identical under both GRN without and GRN with prior knowledge. It was found that the other four results (marked by asterisk in the table), have generally been improved by adding prior knowledge in terms of comparison with published results.

A direct comparison (ignoring whether the relationship is direct or indirect) of the extracted GRN with prior knowledge against the KEGG database and other published literature reveal 21 out of 31 relations (68% of total) are identified in both sets of results. Of the relations identified by the published results, four relations, DBF4 and CDC45, CDC14 and DBF2, PCL1 and PCL2, CLB3 and CLB4 were not identified as associated (i.e. ‘independent’) by the extracted GRN. One possible reason for the failed identification of PCL1 and PCL2, CLB3 and CLB4 is that they are complexes which are co-regulated by other genes and not always present or active during the cell cycle (see the short literature review in Appendix V). Other possible reasons are the relatively small number of genes considered (31 genes), the particular choice of arcs fixed for the addition of prior knowledge, strength of the algorithm employed, and stochastic nature of the data (hidden variables, noise). For example, it is possible that the nature of the expression levels of some genes are such that any associations they may form with other genes becomes apparent only when more genes are introduced to the network. Overall, this exercise demonstrated the value of adding prior biological knowledge from other sources into the learning.

5.2.2 Human Adipocyte Differentiation

The dataset employed in this analysis is a novel dataset, human multipotent adipose-derived stem (hMADS), (Rodriguez, Elabd *et al.*, 2004), from Glaxo Smith Kline (GSK), which present a novel and valuable tool for studies of human fat tissue development and metabolism (Erol, 2005; Rodriguez, Elabd *et al.*, 2004).

The hMADS dataset is a truly large scale genome wide dataset, containing 54,634 gene probes, but with samples generated at only three time points, and thus represented previously uncharted territory for modelling gene regulatory networks using Bayesian methods. Solutions for the reduction of data dimensionality were therefore considered,

and implemented with a view to facilitating extraction of genetic networks from large scale microarray data sets.

The number of time points, three, was below the required minimum sample size for carrying out statistical modelling. This was addressed by the use of a novel method (see section 3.4.2.1), termed artificial experiments, which used the combinations of sample data points in triplicate with their reference to generate 27 data points in total. The gene selection technique employed for filtering differentially expressed genes, to reduce the number of genes considered, combines the fold-change method with a statistical confidence measure known as cyber t-test and hence delivers superior false-positive detection rates.

A new clustering algorithm, ArrayMiner2, which is based on Gaussian distributions, was then applied to yield a final dataset of 45 clusters and 437 outliers, containing 3308 probes in total. (Falkenauer and Marchand, 2001), based on a test involving rat and yeast datasets from GeneSpring, found that ArrayMiner2 produced identical solutions in each run and concluded those solutions are probably the global optima.

A close examination of the derived clusters found that, although the clustering methods used generally produces groups of genes that are similar in terms of expression profile and thus are highly correlated, the presence of multiple probes for some of the genes in the hMADS dataset, may lead to inconsistent results due to the different signal intensities of the probes. This suggests that use of single probe approach to generating microarray data (one probe per gene), may yield more accurate results, for example, reduction of false positives, than probe set approaches.

The data was discretised after clustering as described in section 4.4.3, and the results were tested to ensure robustness to variations in the discretisation thresholds. Bayesian belief networks were then used to build models for the genetic dependency in biological processes that was captured in the microarray experiments.

Gene regulatory networks of the insulin receptor family, ppar family, fox family, cebp family, mef2, fabp, add1 and klf, and probes with highly significant change in gene expression level were learned separately.

The derived GRN (Figure 17, Chapter 4) identified many biologically plausible relationships between genes despite the small number of time points over which the data was generated. For example, it was found (in Figure 50) that the gene ADD1 (Adipocyte determination and differentiation factor 1), is a parent gene to eleven other genes in the network, and a common ancestor to most other genes in the network. Hence, it could be regarded as playing a central role within the GRN of interest. This is supported by evidence from published literature: ADD1/sterol regulatory element binding protein-1c (SREBP-1c), which is the main transcription factor, regulates the characteristic features of adipocyte, senses the glucose and fat excess and draws the excess into the adipocyte to preserve energy, and maintains the blood biochemistry within physiological ranges (Erol, 2005).

The extracted GRN also identified positive associations between C/EBP γ , C/EBP β , (C/EBP δ from the C/EBP (CCAAT/enhancer binding protein) family. Members of the C/EBP family of transcriptional factors have been implicated in the regulation of genes in a variety of tissues (Antonson, Stellan *et al.*, 1996) The extracted GRN, however, did not find any association between C/EBP ϵ and the other members of the family.

Discussion of the results obtained is presented in section 4.5.3.1.

GRNs were also extracted for the 45 clusters with and without outliers to provide a global view of the network structure and then some of the individual clusters were analysed further to obtain a deeper understanding of the genetic interactions.

5.2.3 Metabolic Stress System of *E. Coli*

In this chapter, the genome-wide genetic networks of metabolic burden of *E. coli* was learnt to identify the associations among stress related genes in recombinant protein production. *E. coli* is one of the leading host for the efficient, cost-effective and large-scale plasmid-based expression and production of heterologous proteins (Baneyx, 1999).

The SOD and GFP cDNA microarray experiment data of *E-coli*, K12, genome containing ORFs for 4416 genes, generated as part of the research undertaken at the institute of Applied Biology at the University of BOKU, Vienna, was used as the training set for the extraction of genetic networks in this study.

This dataset used, however, was characterised by experimental noise, small sample size, and high dimensionality. These issues were tackled through the use of dimension reduction techniques, and application of multiple discretisation thresholds.

The final data set contained 931 genes, and thus presented a difficult challenge in terms of both computation and interpretation of large networks extracted. To overcome this challenge, Bayesian networks were used for the extraction of the genome-wide genetic network, and further knowledge extraction techniques such as relationship analysis, variable clustering, arc analysis, and target node analysis were employed for the analysis of subsets of the extracted GRNs.

The GRNs extracted provide a more comprehensive view of the metabolic stress system of the recombinant protein production.

A comparison with KEGG pathway database (Appendix V) shows many of the associations extracted by the GRNs in this work are in agreement with those findings from laboratory experiments. Several of genes that have highly significant change in

expression level, for example, *gat* family, *ibpA*, *ibpB*, *mopA*, *mopB*, *ybeZ* and *artQ*, are shown to regulate each other in the GFP and SOD experiments (see **Table 14**, Chapter 4).

Further, this study found strong evidence for many complex and novel relationships, such as the strong associations shown by the *cstA* gene with several other genes in the extracted GRN for SOD (see Table 13, Chapter 4), which warrant further investigation through laboratory experiments.

The analysis presented in on a subset of the extracted GRNs (see 4.4.2.1), also indicates that possible roles for genes such as *pflB* as marker for recombinant protein production.

The leanings from this study show that the dual approach of extracting genome-wide GRNs to provide an overview of the biological system being modelled and applying the analytical techniques such as TN analysis to understand the finer inner-structure and genetic interactions. This is probably the way forward for understanding complex biological systems.

Finally, the consistently good predictive performance achieved in this thesis in extracting GRNs for three distinct and complex genomes show that the models extracted can be used for identifying novel genetic relationships and for simulation.

5.3 Current Issues Limiting the Performance of GRN Extraction

The modelling of gene regulatory networks has made great strides in recent years, and the work described in this thesis has demonstrated the ability to extract networks from data in some novel applications. Nevertheless, there are important issues that still remain to be addressed.

5.3.1 Algorithm to extract Bayesian network

The current algorithms available such as Taboo order search are relatively slow in terms of processing time when extracting larger networks. However, improving the efficiency of the search algorithm, whilst maintaining its ability to find the best solution, is a considerable challenge.

5.3.2 Pre-Processing Issues

As discussed in Chapters 3 and 4, when analysing large scale genome-wide data, a very large p (number of genes) can be an impediment to the studies due to large computational time and other challenges. Second, it is often the case that only a subset of genes is of interest in practice. In this thesis, gene subset selection has been included in the network extraction procedure, with networks being generated from subsets of genes. This involves considerable pre-processing of data prior to analysis. Hence, testing the robustness of the learned networks and features to various pre-processing methods employed is highly critical, if we to have confidence in the extracted GRNs.

Current software such as the BayesiaLab package used throughout this thesis requires discretization of data prior to structure learning, which severely restricts the ability to test the sensitivity of the results to changing expression thresholds when dealing with continuous data such as gene expression levels. For example, classifying over-expressed as above 0.5 instead of above zero, may well have an impact on the resulting network. Due to the noise present in microarray datasets used, which are generated under different conditions, at different time points, using different techniques, tools for robustness tests for cases such as the above are very important.

If there is integrated software available for both pre-processing and structure learning, it would be possible to build-in the tools necessary for pre-processing within the algorithm.

For example, gene expression profiles can be studied for the detection of abnormal observations, for example multiple peaks in cell cycle regulation, and build-in threshold criteria (either static or variable) for identifying significantly expressed or cell cycle regulated genes. It would also provide the scope for further classing of over, and under expressed genes. A built-in statistical tool for testing the confidence of various features of the learned model is also crucial. For example, one could employ computer-based methods such as the bootstrap method, which generates perturbed versions of the original dataset and learns from them.

5.3.3 Incorporating prior knowledge from heterogeneous sources

The high cost of microarray experiments often prevents gathering of enough samples for a reliable analysis in a single laboratory. A complementary approach for overcoming the problems created by the small number of samples is to incorporate known biological information into the network learning process, such as by incrementally adding additional genes into the network using existing knowledge about gene interactions.

Employing existing microarray datasets from other studies can be another efficient way of improving the reliability of a study. As the number of publicly available datasets grows rapidly on public data depositories (particularly the Gene Expression Omnibus and ArrayExpress), it is clear that these datasets should be combined to generate a more comprehensive understanding of underlying biology, despite the lack of standardization across the experiment procedures/platforms.

The Bayesian networks can readily handle the above scenarios for combining prior knowledge with learning from data as we demonstrated in Chapter 2. However, biological knowledge from sources other than gene expression data, for example, protein-protein interactions, protein-DNA interactions, binding site information, and so on can also be exploited in learning the GRNs, provided these can be incorporated into the learning process in a structured manner. “Since information contained in microarrays is

limited by their quality, noise, and experimental errors, using only microarray data is not enough for estimating gene regulatory networks accurately. Therefore, the use of additional biological knowledge data is considered as a key to microarray data analysis” (Nariai, Tamada *et al.*, 2005).

There are several contributions to the literature combining microarray data with biological knowledge, such as localization data ((Bar-Joseph, Gerber *et al.*, 2003); (Hartemink, Gifford *et al.*, 2002)), DNA promoter sequence data ((Pilpel, Sudarsanam *et al.*, 2001); (Tamada, Kim *et al.*, 2003)), and Transcriptional bindings of regulators ((Bernard and Hartemink, 2005); (Segal, Shapira *et al.*, 2003); (Imoto, Higuchi *et al.*, 2004)), and protein-protein interaction networks ((Segal, Wang *et al.*, 2003); (Nariai, Tamada *et al.*, 2005)).

However, there is currently no general framework available for combining all the various biological knowledge available with microarray data in practice. For example, the efforts in the past have been focussed on combining one source of biological knowledge or another with gene expression data, which meant other potential sources of gene regulation have been ignored. In addition, it is important to specify how to handle any errors in the biological knowledge given the evidence in data (a detection mechanism), and the weighting of information between prior knowledge and microarray data in structure learning. Constructing such a general framework would be a very challenging task given the sheer amount of complexity involved.

5.3.4 Handling Continuous Data and Non-linear Dependencies

As discussed in Chapter 1, the use of continuous data would avoid the loss of information from using discretisation thresholds. Currently, the ability to handle continuous data without the need for discretisation in identifying Bayesian networks is severely limited. For example, most software can only estimate BNs with discrete nodes (variables). One reason is that when real values are used, there is no single representation that can

represent all possible densities, and hence it is necessary to use continuous distribution (e.g. Gaussian, Gamma) for modelling gene expression data, to approximate the underlying data series.

One possible choice for representing multivariate continuous distributions is the use of linear Gaussian distribution. Even though, this representation has yielded Bayesian networks which were reasonably close to those obtained by discrete models (e.g. (Friedman, Linial *et al.*, 2000)), unlike the discrete models, the linear Gaussian model can only detect dependencies that are close to linear. In particular, it is not likely to discover combinatorial effects (e.g. a gene is over expressed only if several genes are jointly over expressed, but not if at least one of them is not over expressed).

Non-linear models have been proposed as a remedy to this problem. For example, (Imoto, Kim *et al.*, 2003) proposed a statistical model based on Gaussian distribution to handle non-linear as well as linear dependencies between genes. Nevertheless, non-linear modelling of gene expression data is yet to be fully explored, and thus presents considerable scope for improvement.

5.3.5 Dynamic Models

A shortcoming of the standard acyclic Bayesian network model is that this model cannot construct cyclic networks such as the feedback loops in pathways, while a real gene regulation mechanism has cyclic regulations. Although Bayesian networks work well in many cases, when there is a cyclic gene regulation, it tends to estimate some false positives (Kim, Imoto *et al.*, 2004). A temporal extension, known as Dynamic Bayesian networks (DBNs), which enables the feedback loop to be represented by unrolling it over time, has been proposed as a remedy to this problem (Bilmes, 2000; Friedman, Murphy *et al.*, 1998; Murphy and Mian, 1999; Zou and Conzen, 2005).

DBNs refer to the application of Bayesian networks to time series data, which looks at how the random variable X evolves over time, i.e. the *inter-slice* connections. Again the first-order Markov assumption is employed: the future is independent of the past given the present, i.e. X_{t+1} is independent of X_{t-1} given X_t .

In addition to connections *between* time slices, one can also allow connections *within* a time slice (intra-slice relations). For example, we may want to measure the expression levels of genes simultaneously, as they change over time and react to external stimuli, as well as any co-occurrence effects. Therefore, the techniques for learning DBNs are similar to those developed above for BNs (Murphy and Mian, 1999).

As an alternative to discrete DBN models (Friedman, Murphy *et al.*, 1998); (Murphy and Mian, 1999); (Bilmes, 2000)), two continuous dynamic models have been put forward: one is an ordinary differential equation model (Chen, He *et al.*, 1999) and the other is the linear dynamical systems (LDS)³⁵ approach (Rangel, Wild *et al.*, 2001). However, both models assume that the relationship between genes is linear, which is not guaranteed.

The non-linear DBN model of (Kim, Imoto *et al.*, 2004) allows for such complexity, and demonstrated its application to yeast cell cycle pathway with a good degree of success in terms of validation against the KEGG database. However, their attempts to model the metabolic pathway was less successful suggesting that the application of the first order Markov assumption is rather restrictive.

A solution is to allow for correlations between non-adjacent time slices by adding *lag variables* to switch to a higher-order Markov chain. In addition, despite its name, a DBN assumes that the model structure and parameters do not change, i.e. the model is time-invariant. However, one can add extra hidden nodes to represent the ‘current regime’, thereby creating mixtures of variables to capture periodic non-stationarities (e.g. peaks in

³⁵ These are undirected graphical models, where all the nodes are assumed to have linear-Gaussian distributions.

cell cycle expression); also the size of the state space (structure), can be allowed to change over time (Murphy and Mian, 1999).

5.4 Future Work

The next logical step in analysing gene expression data is to address the issues discussed above: improving the efficiency of search and inference algorithms, implementing an integrative approach to Bayesian data analysis and learning, examine ways of incorporating prior biological knowledge from heterogeneous sources in a structured manner, and building Bayesian network models that can analyze continuous data, capture non-linear dependencies between genes, and handle cyclic regulation embodied in time series gene expression data.

Tackling these present a formidable, but exciting challenge for the future.

References

- (2001) Gene Network Sciences, Inc.
- Aerne, B. L., Johnson, A. L., Toyn, J. H. and Johnston, L. H. (1998) *Swi5 controls a novel wave of cyclin synthesis in late mitosis*, *Molecular Biology Of The Cell*, **9**, 945-956.
- Ahn, S. H., Tobe, B. T., Fitz Gerald, J. N., Anderson, S. L., Acurio, A. and Kron, S. J. (2001) *Enhanced cell polarity in mutants of the budding yeast cyclin-dependent kinase Cdc28p*, *Molecular Biology Of The Cell*, **12**, 3589-3600.
- Ailhaud, G., Grimaldi, P. and Negrel, R. (1992) *Cellular and molecular aspects of adipose tissue development*, *Annual Review of Nutrition*, **12**, 207-33.
- Althoefer, H., Schleiffer, A., Wassmann, K., Nordheim, A. and Ammerer, G. (1995) *Mcm1 is required to coordinate G2-specific transcription in Saccharomyces cerevisiae*, *Molecular And Cellular Biology*, **15**, 5917-5928.
- Andrea Ciliberto, Bela Novak and Tyson, J. J. (2003) *Mathematical model of the morphogenesis checkpoint in budding yeast*, *The Journal of Cell Biology*, **163**, 1243-1254.
- Andrews, B. and Measday, V. (1998) *The cyclin family of budding yeast: abundant use of a good idea*, *Trends In Genetics: TIG*, **14**, 66-72.
- Antonson, P., Stellan, B., Yamanaka, R. and Xanthopoulos, K. G. (1996) *A Novel Human CCAAT/Enhancer Binding Protein Gene, C/EBP ϵ , Is Expressed in Cells of Lymphoid and Myeloid Lineages and Is Localized on Chromosome 14q11.2 Close to the T-Cell Receptor α/δ Locus*, *Genomics*, - **35**, - 38.
- Avram, M. M., Avram, A. S. and James, W. D. (2007) *Subcutaneous fat in normal and diseased states: 3. Adipogenesis: From stem cell to fat cell*, *Journal of the American Academy of Dermatology*, **56**, 492.
- Baldi, P. and Hatfield, G. W. (2002) *DNA Microarrays and Gene Expression: from Experiments to Data Analysis and Modeling*, Cambridge University Press, Cambridge.
- Baldi, P. and Long, A. D. (2001) *A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes*, *Bioinformatics*, **17**, 509-19.
- Baneyx, F. (1999) *Recombinant protein expression in Escherichia coli*, *Current Opinion in Biotechnology* **10**, 411-421.
- Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A. and Gifford, D. K. (2003) *Computational discovery of gene modules and regulatory networks*, *Nature Biotechnology*, **21**, 1337-1342.
- Barash, Y. and Friedman, N. (2001) *Context-specific Bayesian clustering for gene expression data*, *RECOMB'01*.
- Bardin, A. J. and Amon, A. (2001) *Men and sin: what's the difference?*, *Nature Reviews. Molecular Cell Biology*, **2**, 815-826.
- Bernard, A. and Hartemink, A. (2005) *Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data*, *Pac. Symp. Biocomput.*, **10**, 459-470.

- Bertone, P. and Gerstein, M. (2001) *Integrative data mining for genomics and proteomics: the new direction in bioinformatics*, *IEEE Eng Med Biol Mag* **20**, 33-40.
- Bilmes, J. A. (2000) In *Proc. 16th Conf. on Uncertainty in Artificial Intelligence*, pp. 38-45.
- Boulesteix, A. (2004) *PLS Dimension reduction for classification with microarray data*, *Statistical applications in genetics and molecular biology*, **3**, 1-33.
- Brazhnik, P., de la Fuente, A. and Mendes, P. (2002) *Gene networks: how to put the function in genomics*, *Trends in Biotechnology*, **20**, 467-472.
- Brazma, A. and Vilo, J. (2000) *Gene expression data analysis*, *FEBS Letters*, **480**, 17-24.
- Brazma, A. and Vilo, J. (2001) *Gene expression data analysis*, *Microbes and Infection*, **3**, 823-829.
- Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. J. and Haussler, D. (1997) In *National Academy of Science*, Vol. 1, pp. 262-267.
- Brown, T. A. (2002) *Genomes 2*, Bios.
- Calzada, A., Sacristan, M., Sanchez, E. and Bueno, A. (2001) *Cdc6 cooperates with Sic1 and Hct1 to inactivate mitotic cyclin-dependent kinases*, *Nature*, **412**, 355-358.
- Chen, T., He, H. L. and Church, G. M. (1999) In *Pacific Symposium of Biocomputing*, Vol. 4, pp. 29-40.
- Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, 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*, *Molecular Cell*, **2**, 65-73.
- Ciosk, R., Zachariae, W., Michaelis, C., Shevchenko, A., Mann, M. and Nasmyth, K. (1998) *An ESP1/PDS1 Complex Regulates Loss of Sister Chromatid Cohesion at the Metaphase to Anaphase Transition in Yeast*, *Cell*, **93**, 1067-1076.
- Cover, T. M. and Thomas, J. A., , , . (1991) *Elements of Information Theory*, John Wiley & Sons, New York.
- D'Haeseleer, P., Liang, S. and Somogyi, R. (2000).
- Dai, J. J., Lieu, L. and Rocke, D. (2006) *Dimension Reduction for Classification with Gene Expression Microarray Data*, *Statistical Applications in Genetics and Molecular Biology*, **5**.
- Datta, S. (2003) *Statistical Techniques for Microarray Data: A Partial Overview*, *Communications in Statistics - Theory and Methods*, **32**, 263-280.
- DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. and Trent, J. M. (1996) *Use of a cDNA microarray to analyse gene expression patterns in human cancer*, *Nature Genetics*, **14**, 457-460.
- DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997) *Exploring the metabolic and genetic control of gene expression on a genomic scale*, *Science*, **278**, 680-686.
- Deshaies, R. J. (1997) *Phosphorylation and proteolysis: partners in the regulation of cell division in budding yeast*, *Current Opinion in Genetics & Development*, **7**, 424.
- Dewey, T. G. (2002) *From microarrays to networks: mining expression time series*, *Drug Discovery Today*, **7**, s170-s175.

- Draghici, S. (2002) *Statistical intelligence: effective analysis of high-density microarray data*, *Drug Discov Today*, **7**, S55-S63.
- Drawid, A. and Gerstein, M. (2000) *A Bayesian system integrating expression data with sequence patterns for localising proteins: comprehensive application to the yeast genome*, *Journal of Molecular Biology*, **301**, 1059-1075.
- Eguchi, S. and Copas, J. (2006) *Interpreting Kullback-Leibler Divergence with the Neyman-Pearson Lemma*, *Multivariate Analysis*, **97**, 2034-2040.
- Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) *Cluster analysis and display of genome-wide expression patterns*, *Proceedings Of The National Academy Of Sciences Of The United States Of America*, **95**, 14863-14868.
- Epps, J. and Ambikairajah, E. (2005) In *Proceedings of the 2005 Asia-Pacific symposium on Information visualisation* Vol. 45 Australian Computer Society, Inc, Sydney, Australia.
- Erol, A. (2005) *Adipocyte insensitivity syndromes – novel approach to nutritional metabolic problems including obesity and obesity related disorders*, *Medical Hypotheses*, - **64**, - 832.
- Espinoza, F. H., Ogas, J., Herskowitz, I. and Morgan, D. O. (1994) *Cell cycle control by a complex of the cyclin HCS26 (PCL1) and the kinase PHO85*, *Science*, **266**, 1388-1391.
- Falkenauer, E. and Marchand, A. (2001) In *Proceedings of the 2001 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'2001)* Las Vegas, USA.
- Fox, R. J. and Dimmic, M. W. (2006) *A two-sample Bayesian t-test for microarray data*, *BMC Bioinformatics*, **7**, 126.
- Friedman, N. and Goldszmidt, M. (1996) In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*.
- Friedman, N., Goldszmidt, M. and Wyner, A. *Data analysis with Bayesian networks: a bootstrap approach*.
- Friedman, N., Linial, M., Nachman, I. and Dana, P. e. (2000) *Using Bayesian networks to analyse expression data*, *Journal of Computational Biology*, **7**, 601-620.
- Friedman, N., Murphy, K. and Russell, S. (1998) In *Proc. Conf. on Uncertainty in Artificial Intelligence*, pp. 139-147.
- Gaasterland Terry and Bekiranov, S. (2000) *Making the most of microarray data*, *Nature Genetics*, **24**, 204-206.
- Gardiner-Garden, M. and Littlejohn, T. G. (2001) *A comparison of microarray databases*, *Briefings in Bioinformatics*, **2**, 143-158.
- Gardner, R. D. and Burke, D. J. (2000) *The spindle checkpoint: two transitions, two pathways*, *Trends in Cell Biology*, **10**, 154-158.
- Gasch, A. P. and Eisen, M. B. (2002) *Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering*, *Genome Biology*, **3**, 1-22.
- Germain, D., Hendley, J. and Futcher, B. (1997) *DNA damage inhibits proteolysis of the B-type cyclin Clb5 in S. cerevisiae*, *Journal Of Cell Science*, **110 (Pt 15)**, 1813-1820.
- Goto, S., Kawashima, S., Okuji, Y., Kamiya, T., Miyazaki, S., Numata, Y. and Kanehisa, M. (2000) *KEGG/Expression: a database for browsing and analysing microarray expression data*, *Genome Informatics*, **11**, 222-223.

- Green, C. D., Simons, J. F., Taillon, B. E. and Lewin, D. A. (2001) *Open systems: panoramic views of gene expression*, *Journal of Immunological Methods*, **250**, 67-79.
- Gregoire, F. (2001) *Adipocyte differentiation: from fibroblast to endocrine cell*, *Experimental Biology and Medicine*, **226**, 997-1002.
- Grunwald, P. (2000) *Model Selection Based on Minimum Description Length*, *Journal of Mathematical Psychology*, **1**, 133-152.
- Grunwald, P. D., I.J., M. and M.A., P. (2004) *Advances in Minimum Description Length: Theory and Applications*, MIT Press.
- Guacci, V., Koshland, D. and Strunnikov, A. (1997) *A Direct Link between Sister Chromatid Cohesion and Chromosome Condensation Revealed through the Analysis of MCD1 in S. cerevisiae*, *Cell*, **91**, 47-57.
- Habraken, Y., Sung, P., Prakash, L. and Prakash, S. (1998) *ATP-dependent assembly of a ternary complex consisting of a DNA mismatch and the yeast MSH2-MSH6 and MLH1-PMS1 protein complexes*, *The Journal Of Biological Chemistry*, **273**, 9837-9841.
- Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. and Young, R. A. (2002) *Combining location and expression data for principled discovery of genetic regulatory network models*, *Pacific Symposium on Biocomputing*, **7**, 437-449.
- Hartuv, E. and Shamir, R. (1999) *A clustering algorithm based on graph connectivity*.
- Hatfield, G., Wesley, Hung, S. P. and Baldi, P. (2003) *Differential analysis of DNA microarray gene expression data*, *Molecular Microbiology*, **47**, 871-77.
- Hausman, D. B., DiGirolamo, M., Bartness, T. J., Hausman, G. J. and Martin, R. J. (2001) *The biology of white adipocyte proliferation*, *Obesity Rev*, **2**, 239-54.
- Heckerman, D., Meek, C. and Cooper, G. (1999) In *Computation, Causation, and Discovery*(Ed, Cooper, C. G. a. G.) MIT Press, Cambridge, MA.
- Herrero, J., Valencia, A. and Dopazo, J. (2001) *A hierarchical unsupervised growing neural network for clustering gene expression patterns*, *Bioinformatics*, **17**, 126-136.
- Hertz, A. and de Werra, D. (1990) *The Tabu Search Metaheuristic: how we used it*, *Annals of Mathematics and Artificial Intelligence*, **1**, 111-121.
- Hirsch, J. and Batchelor, B. (1976) *Adipose tissue cellularity and human obesity*, *Journal of Clinical Endocrinology & Metabolism*, **5**, 299-311.
- Hung, S.-C., Chang, C.-F., Ma, H.-L., Chen, T.-H. and Low-Tone Ho, L. (2004) *Gene expression profiles of early adipogenesis in human mesenchymal stem cells*, *Gene*, **340**, 150.
- Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2004) *Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks*, *J. Bionformatics Comput. Biol.*, **2**, 77-98.
- Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S. and Miyano, S. (2003) *Bayesian network and nonparametric heteroscedastic regression for nonlinear modelling of genetic network*, *J. Bionformatics Comput. Biol.*, **1**, 231-252.
- Jares, P., Donaldson, A. and Blow, J. J. (2000) *The Cdc7/Dbf4 protein kinase: target of the S phase checkpoint?*, *EMBO Reports*, **1**, 319-322.
- Jeong-Ho Chang, Sung Wook Chi and Zhang, B.-T. (2003) *Gene Expression Pattern Analysis via Latent Variable Models Coupled with Topographic Clustering*, *Genomics & Informatics*, **1**, 32-39.

- Karin Dürschmid, H. R., Norbert Auer and Karl Bayer (2003) BOKU, Vienna.
- Kelsuke Iida, I. N. (2002) *Gene Expression Profiling By DNA Microarray Technology, Critical Reviews in Oral Biology and Medicine*, **1**, 35-50.
- Kerr, M. K. and Churchill, G. A. (2001) *Statistical design and the analysis of gene expression microarray data, Genet. Res., Camb.*, **77**, 123-128.
- Kerr, M. K., Martin, M. and Churchill, G. A. (2000) *Analysis of variance for gene expression microarray data, Journal of Computational Biology*, **7**, 819-837.
- Kim, S. Y., Imoto, S. and Miyano, S. (2004) *Dynamic Bayesian Network and Nonparametric Regression for Nonlinear Modeling of Gene Networks from Time Series Gene Expression Data, Biosystems*, **75**, 57-65.
- Kohane, I. S., Kho, A. T. and Butte, A. J. (2003) *Microarrays for an Integrative Genomics*, The MIT Press, Massachusetts.
- Komarnitsky, S. I., Chiang, Y. C., Luca, F. C., Chen, J., Toyn, J. H., Winey, M., Johnston, L. H. and Denis, C. L. (1998) *DBF2 protein kinase binds to and acts through the cell cycle-regulated MOB1 protein, Molecular And Cellular Biology*, **18**, 2100-2107.
- Kurata, H., Matoba, N. and Shimizu, N. (2003) *CADLIIVE for constructing a largescale biomedical network based on a simulation-detected notation and its application to yeast cell cycle, Nucleic Acids Research*, **31**, 4071-4084.
- Kurella, M., Hsiao, L., Yoshida, T., Randall, J. D., Chow, G., Sarang, S. S., Jensen, R. V. and Gullans, S. R. (2001) *DNA microarray analysis of complex biologic processes, J Am Soc Nephrol*, **12**, 1072-1078.
- Lee, Y., Lin, Y. and Wahba, G. (2002) University of Wisconsin, Madison.
- Lew DJ, e. a. (1997) In *Molecular and Cellular Biology of the Yeast Saccharomyces: Cell Cycle and Cell Biology*(Ed, Pringle JR, B. J. a. J. E.) Cold Spring Harbor Laboratory Press, New York, pp. 607-695.
- Li, K. C. (2000), Vol. 2004 <http://www.stat.ucla.edu/keli/sir-PHD.pdf>.
- Liao, X., Dasgupta, N., Lin, S. M. and Carin, L. (2002) In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, Vol. 4, pp. IV-3880-3
- Lodish, H., Berk, A., Zipursky, L. S., Matsudaira, P., Baltimore, D. and Darnell, J. (2000) *Molecular Cell Biology*, W.H Freeman and Company, New York.
- Loy, C. J., Lydall, D. and Surana, U. (1999) *NDD1, a high-dosage suppressor of cdc28-1N, is essential for expression of a subset of late-S-phase-specific genes in Saccharomyces cerevisiae, Molecular And Cellular Biology*, **19**, 3312-3327.
- Lu, X., Zhang, W., Qin, Z. S., Kwast, K. E. and Liu, J. S. (2004) *Statistical resynchronization and Bayesian detection of periodically expressed genes, Nucleic Acids Research*, **32**, 447-455.
- Luca, F. C. and Winey, M. (1998) *MOB1, an essential yeast gene required for completion of mitosis and maintenance of ploidy, Molecular Biology Of The Cell*, **9**, 29-46.
- MacQueen, J. B. (1967) *Some Methods for classification and Analysis of Multivariate Observations*, University of California Press, Berkeley.
- Mariani, T. J., Budhreja, V., Mecham, B. H., Gu, C. C., Watson, M. A. and Sadovsky, Y. (2003) *A variable fold-change threshold determines significance for expression microarrays, The FASEB Journal*, **17**, 321-3.
- McInerney, C. J., Partridge, J. F., Mikesell, G. E., Creemer, D. P. and Breeden, L. L. (1997) *A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and*

- CDC47 promoters activates M/G1-specific transcription*, *Genes & Development*, **11**, 1277-1288.
- McLachlan, G. J., Bean, R. W. and Peel, D. (2002) *A Mixture Model-based Approach to the Clustering of Microarray Expression Data*, *Bioinformatics*, **18**, 413-422.
- Measday, V., Moore, L., Ogas, J., Tyers, M. and Andrews, B. (1994) *The PCL2 (ORF1)-PHO85 cyclin-dependent kinase complex: a cell cycle regulator in yeast*, *Science*, **266**, 1391-1395.
- Meyn Iii, M. A. and Holloway, S. L. (2000) *S-phase cyclins are required for a stable arrest at metaphase*, *Current Biology*, **10**, 1599-1602.
- Misra, J., Schmitt, W., Hwang, D., Hsiao, L. L., Gullans, S., Stephanopoulos, G. and Stephanopoulos, G. (2002) *Interactive exploration of microarray gene expression patterns in a reduced dimensional space*, *Genome Research*, **12**, 1112-1120.
- Murphy, K. and Mian, S. (1999) Computer Science Division, University of California, California.
- Murphy, K. P. (2001), Vol. 2003
<http://www.cs.berkeley.edu/~murphyk/Bayes/bayes.html>.
- Nariai, N., Tamada, Y., Imoto, S. and Miyano, S. (2005) *Estimating Gene Regulatory Networks and Protein-Protein interactions of Saccharomyces Cerevisiae From Multiple Genome-wide Data*, *Bioinformatics*, **21**, 206-212.
- Nasmyth, K. and Dirick, L. (1991) *The role of SWI4 and SWI6 in the activity of G1 cyclins in yeast*, *Cell*, **66**, 995-1013.
- Newton, J. (2002) <http://www.cs.ualberta.ca/~newton/academic/ml.pdf>.
- Nguyen, D. V. and Rocke, D. M. (2002) *Tumor classification by partial least squares using microarray gene expression data*, *Bioinformatics*, **18**, 39-50.
- Nguyen, D. V. and Rocke, D. M. (2004) *On partial least squares dimension reduction for microarray-based classification: a simulation study*, *Computational Statistics & Data Analysis*, **46**, 407-425.
- Nishizawa, M., Kawasumi, M., Fujino, M. and Toh-e, A. (1998) *Phosphorylation of sic1, a cyclin-dependent kinase (Cdk) inhibitor, by Cdk including Pho85 kinase is required for its prompt degradation*, *Molecular Biology Of The Cell*, **9**, 2393-2405.
- Nobelmann, B. and Lengeler, J. W. (1995) *Sequence of the gat operon for galactitol utilization from a wild-type strain EC3132 of Escherichia coli*, *Biochimica Et Biophysica Acta*, **1262**, 69-72.
- Oehlen, L. J., McKinney, J. D. and Cross, F. R. (1996) *STE12 and MCM1 regulate cell cycle-dependent transcription of FAR1*, *Mol Cell Biol*, **16**, 2830-7.
- Ogas, J., Andrews, B. J. and Herskowitz, I. (1991) *Transcriptional activation of CLN1, CLN2, and a putative new G1 cyclin (HCS26) by SWI4, a positive regulator of G1-specific transcription*, *Cell*, **66**, 1015-1026.
- Orr, M. S. and Scherf, U. (2002) *Large-scale gene expression analysis in molecular target discovery*, *Leukemia*, **16**, 473-477.
- Owens, J. C., Detweiler, C. S. and Li, J. J. (1997) *CDC45 is required in conjunction with CDC7/DBF4 to trigger the initiation of DNA replication*, *Proceedings Of The National Academy Of Sciences Of The United States Of America*, **94**, 12521-12526.

- Pasero, P. and Schwob, E. (2000) *Think global, act local -- how to regulate S phase from individual replication origins*, *Current Opinion In Genetics & Development*, **10**, 178-186.
- Pellman, D., Bagget, M., Tu, Y. H., Fink, G. R. and Tu, H. (1995) *Two microtubule-associated proteins required for anaphase spindle movement in Saccharomyces cerevisiae*, *The Journal of Cell Biology*, **130**, 1373-1385.
- Pennington, S. R. and Dunn, M. J. (2001) *Proteomics: from protein sequence to function*, BIOS Scientific Publishers Ltd, Oxford.
- Peter, M., Gartner, A., Horecka, J., Ammerer, G. and Herskowitz, I. (1993) *FAR1 links the signal transduction pathway to the cell cycle machinery in yeast*, *Cell*, **73**, 747-760.
- Pilpel, Y., Sudarsanam, P. and Church, G. M. (2001) *Identifying regulatory networks by combinatorial analysis of promoter elements*, *Nature Genetics*, **29**, 153-9.
- Pittenger, M. F., Mackay, A. M., Beck, S. C., Jaiswal, R. K., Robin, D., Mosca, J. D., Moorman, M. A., Simonetti, D. W., Craig, S. and Daniel, R. (1999) *Multilineage Potential of Adult Human Mesenchymal Stem Cells*, *Marshall Science*, **284**.
- Rangel, C., Wild, D. L. and Falciani, F. (2001) In *Proc. 2nd International Conference on Systems Biology (ICSB)* (Ed, Technology, C. I. o.) ICSB, Pasadena, pp. 249-258.
- Raychaudhuri, S., Stuart, J. M. and Altman, R. B. (2000) *Principal components analysis to summarize microarray experiments: application to sporulation time series*, *Pacific symposium on Biocomputing*, 452-463.
- Rieusset, J., Andreelli, F., Auboeuf, D., Roques, M., Vallier, P. and Riou, J. P. (1999) *Insulin acutely regulates the expression of the peroxisome proliferator-activated receptor-gamma in human adipocytes*, *Diabetes*, **4**, 699-705.
- Rodriguez, A. M., Elabd, C. and Delteil, F. (2004) *Adipocyte differentiation of multipotent cells established from human adipose tissue*, *Biochemical and Biophysical Research Communications*, **315**, 255-263.
- Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C. E., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D. and Brown, P. O. (2000) *Systematic variation in gene expression patterns in human cancer cell lines*, *Nature Genetics*, **24**, 227-235.
- Rung, J., Schlitt, T., Brazma, A., Freivalds, K. and Vilo, J. (2002) *Building and analysing genome-wide gene disruption networks.*, *Bioinformatics*, **2**, S202-10.
- Schena, M. (1996) *Genome analysis with gene expression microarrays*, *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, **18**, 427-431.
- Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. and Davis, R. W. (1996) *Parallel human genome analysis: microarray-based expression monitoring of 1000 genes*, *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 10614-10619.
- Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O. and

- Weinstein, J. N. (2000) *A gene expression database for the molecular pharmacology of cancer*, *Nature Genetics*, **24**, 236-244.
- Schwob, E., Bohm, T., Mendenhall, M. D. and Nasmyth, K. (1994) *The B-type cyclin kinase inhibitor p40SIC1 controls the G1 to S transition in S. cerevisiae*, *Cell*, **79**, 233-244.
- Sclafani, R. A., Tecklenburg, M. and Pierce, A. (2002) *The mcm5-bob1 bypass of Cdc7p/Dbf4p in DNA replication depends on both Cdk1-independent and Cdk1-dependent steps in Saccharomyces cerevisiae*, *Genetics*, **161**, 47-57.
- Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003) *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data*, *Nature Genetics*, **34**, 166-76.
- Segal, E., Taskar, B., Gasch, A., Friedman, N. and Koller, D. (2001) *Rich probabilistic models for gene expression*, *Bioinformatics*, **1**, 1-9.
- Segal, E., Wang, H. and Koller, D. (2003) *Discovering molecular pathways from protein interaction and gene expression data*, *Bioinformatics*, **19**, i264-i272.
- Shamir, R. and Sharan, R. *CLICK: a clustering algorithm for gene expression analysis*.
- Sharan, R. and Shamir, R. (2000) *CLICK: a clustering algorithm with applications to gene expression analysis*, *proc.ISMB*, 307-316.
- Sherlock, G. (2000) *Analysis of large-scale gene expression data*, *Current Opinion in Immunology*, **12**, 201-205.
- Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C., Dwight, S. S., Kaloper, M., Weng, S., Jin, H., Ball, C. A., Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. and Cherry, M. J. (2001) *The Stanford microarray database*, *Nucleic Acids Research*, **29**, 152-155.
- Shirayama M., Toth A., Galova M. and K., N. (1999) *APC(Cdc20) promotes exit from mitosis by destroying the anaphase inhibitor Pds1 and cyclin Clb5*, *Nature*, **402**, 203-7.
- Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S. and Young, R. A. (2001) *Serial Regulation of Transcriptional Regulators in the Yeast Cell Cycle*, *Cell*, **106**, 697-708.
- Smas, C. M. and Sul, H. S. (1995) *Control of adipocyte differentiation*, *Biochem Journal*, **309**, 697-710.
- Smith, K. N., Penkner, A., Ohta, K., Klein, F. and Nicolas, A. (2001) *B-type cyclins CLB5 and CLB6 control the initiation of recombination and synaptonemal complex formation in yeast meiosis*, *Current Biology*, **11**, 88-97.
- Smith, P. J., Wise, L. S., Berkowitz, R., Wan, C. and Rubin, C. S. (1988) *Insulin-like growth factor-1 is an essential regulator of the differentiation of 3T3-L1 adipocytes*, *J Biol Chem*, **263**, 9402-8.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization*, *Molecular Biology Of The Cell*, **9**, 3273-3297.
- Staloch, L. J., Divine, J. K., Witten, J. T. and Simon, T. C. (2005) *C/EBP and Cdx family factors regulate liver fatty acid binding protein transgene expression in*

- the small intestinal epithelium, BIOchimica et Biophysica Acta (BBA) - Gene Structure and Expression* - **1731**, - 178.
- Stoeckert, C. J., Jr., Causton, H. C. and Ball, C. A. (2002) *Microarray databases: standards and ontologies*, *Nature Genetics*, **32**, 469-473.
- Strachan, T. and Read, A. P. (1999) *Human Molecular Genetics* 2, BIOS Scientific Publishers Ltd, Oxford.
- Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. and Miyano, S. (2003) *Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection*, *Bioinformatics*, **19**, i227-i236.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999) *Interpreting patterns of gene expression with self-organising maps: methods and application to hematopoietic differentiation*, *Proc.Natl.Acad.Sci.USA*, **96**, 2907-2912.
- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M. (1999) *Systematic determination of genetic network architecture*, *Nature Genetics*, **22**, 281-5.
- Tavormina, P. A. and Burke, D. J. (1998) *Cell cycle arrest in cdc20 mutants of Saccharomyces cerevisiae is independent of Ndc10p and kinetochore function but requires a subset of spindle checkpoint genes*, *Genetics*, **148**, 1701-1713.
- Taylor, R. Center for Computational Pharmacology at the University of Colorado, Denver.
- Taylor, R. (2002), Vol. 2003 Center for Computational Pharmacology at the University of Colorado, Denver.
- Ting Lee, M. L. (2000) *Importance of replication microarray gene expression studies: statistical methods and evidence from repetitive cDNA microarray hybridizations.*, *Proc Natl Acad Sci U S A*, **97**, 9834-9839.
- Tuma, R. S. (2002), Vol. 2003 http://www.bio-itworld.com/archive/071102/horizons_microarray.html.
- Tusher, V. G., Tibshirani, R. and Chu, G. (2001) *Significance analysis of microarrays applied to the ionizing radiation response*, *Proceedings Of The National Academy Of Sciences Of The United States Of America*, **98**, 5116-5121.
- Tyers, M. (1996) *The cyclin-dependent kinase inhibitor p40SIC1 imposes the requirement for Cln G1 cyclin function at Start*, *Proceedings Of The National Academy Of Sciences Of The United States Of America*, **93**, 7772-7776.
- Valdivieso, M. H., Sugimoto, K., Jahng, K. Y., Fernandes, P. M. and Wittenberg, C. (1993) *FAR1 is required for posttranscriptional regulation of CLN2 gene expression in response to mating pheromone*, *Molecular And Cellular Biology*, **13**, 1013-1022.
- VanBogelen, R. A., Greis, K. D., Blumenthal, R. M., Tani, T. H. and Matthews, R. G. (1999) *Mapping regulatory networks in microbial cells*, *Trends in Microbiology*, **7**, 328.
- Verma, R., Feldman, R. M. and Deshaies, R. J. (1997) *SIC1 is ubiquitinated in vitro by a pathway that requires CDC4, CDC34, and cyclin/CDK activities*, *Molecular Biology Of The Cell*, **8**, 1427-1437.
- Vilo, J. and Kivinen, K. (2001) *Regulatory sequence analysis: application to the interpretation of gene expression*, *European Neuropsychopharmacology*, **11**, 399-411.

- Visintin, R., Craig, K., Hwang, E. S., Prinz, S., Tyers, M. and Amon, A. (1998) *The Phosphatase Cdc14 Triggers Mitotic Exit by Reversal of Cdk-Dependent Phosphorylation*, *Molecular Cell*, **2**, 709-718.
- Vohl, M.-C., Houde, A., Lebel, S., Hould, F.-S. and Marceau, P. (2005) *Effects of the peroxisome proliferator-activated receptor- γ co-activator-1 Gly482Ser variant on features of the metabolic syndrome*, *Molecular Genetics and Metabolism*, **86**, 306.
- Wasch, R. and Cross, F. R. (2002) *APC-dependent proteolysis of the mitotic cyclin Clb2 is essential for mitotic exit*, *Nature*, **418**, 556-562.
- Werner-Washburne, M., Wylie, B., Boyack, K., Fuge, E., Galbraith, J., Weber, J. and Davidson, G. (2002) *Comparative analysis of multiple genome-scale data sets*, *Genome Research*, **12**, 1564-1573.
- West, M., Blanchette, C., Fressman, H., Huang, E., Ishida, S., Spang, R., Zuan, H., Marks, J. R. and Nevins, J. R. (2001) In *Proceedings Of The National Academy Of Sciences Of The United States Of America*, Vol. 98, pp. 11462-7.
- Wittenberg, C., Sugimoto, K. and Reed, S. I. (1990) *G1-specific cyclins of S. cerevisiae: Cell cycle periodicity, regulation by mating pheromone, and association with the p34CDC28 protein kinase*, *Cell*, **62**, 225-237.
- Woolf, P. J. and Wang, Y. (2000) *A fuzzy logic approach to analyzing gene expression data*, *Physiological Genomics*, **3**, 9-15.
- Xu, H. Q., Wu, P. R., Wu, C. F. J., Tidwell, C. and Wang, Y. X. (2002) *A smooth response surface algorithm for constructing a gene regulatory network*, *Physiological Genomics*, **11**, 11-20.
- Yen, K., Gitsham, P., Wishart, J., Oliver, S. G. and Zhang, N. (2003) *An improved tetO promoter replacement system for regulating the expression of yeast genes*, *Yeast (Chichester, England)*, **20**, 1255-1262.
- Zachariae, W. and Nasmyth, K. (1999) *Whose end is destruction: cell division and the anaphase-promoting complex*, *Genes & Development*, **13**, 2039-2058.
- Zhu, G., Spellman, P. T., Volpe, T., Brown, P. O., Botstein, D., Davis, T. N. and Fitcher, B. (2000) *Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth*, *Nature*, **406**, 90-94.
- Ziemke, P. and McCarthy, J. E. G. (1992) *The control of mRNA stability in Escherichia coli: manipulation of the degradation pathway of the polycistronic atp mRNA*, *Gene Structure and Expression* **1130**, 297-306.
- Zou, M. and Conzen, S. D. (2005) *A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data*, *Bioinformatics*, **21**, 71-9.

Appendices

Appendix I –

Discussions of Yeast Cell Cycle Phases

Phase M/G1

The SopLEQ and Taboo Search algorithms failed to associate STE2 with other genes in phase M/G1 from the known regulated genes. The cell cycle pathway from KEGG also confirms that STE2 is not directly associated but it plays a main role in MAPK signalling pathway as a pheromone mating signal which precedes cell cycle, see **Figure 67** in Appendix III. The biological function of pheromone is important for overall structure of the cell. Response of mating pheromone causes physiological changes, arrest of cell cycle, morphological changes and induction of variety of genes involved in the mating process. Hence pheromone acts as model for external signals regulating cell cycle decisions. Therefore the role of STE2 in cell cycle regulation is critical.

TEC1, FAR1 from M/G1 phase and G2/M phase respectively are also involved in MAPK signalling pathway³⁶. MAPK pathway from KEGG shows that STE2 activates (a positive impact) on FAR1 indirectly, through handful of genes, CDC42, STE20, STE11, STE7, FUS3 and FAR1. However FAR1 inhibits cell cycle i.e. cell cycle arrest through CLN2. Our extracted network shows FAR1 activates SIC1 and TEC1. With this prior knowledge, the arcs between FAR1, SIC1 and TEC1 could be appropriately fixed with right dependency, enabling us to extract the actual GRN with more correct gene dependency. More network connections and cross links association prior knowledge between genes needs to be added to network structure to extract better and more accurate gene regulatory network for the genes and experiments under investigation. This would be true with any case.

Phase G1

Taboo Order Search identified CLN1 as a ‘father’ node; Strong association identified with FKS1, MNN1, SWE1, and GAS1. KEGG however, did not identify GAS1 as a known regulated gene. It also shows that the other associations as indirect, with their main roles in activities associated with the cell cycle process itself; FKS1 (MAPK signalling pathway), MNN1 (N-Glycous Biosynthesis), and SWE1 (Inositol Phosphate Metabolism).

³⁶ http://www.genome.ad.jp/kegg-bin/mark_pathway_www?/opt/local/WWW/pub/tmp/search_pathway8694/sce04010.args. As of 10/11/03.

The Taboo Search did not find any correlation with CHS3/CSD2, PCL1, TIP1, and KAR4. This contradicts the results from traditional laboratory techniques and KEGG. Only in the case of KAR4, the results are similar. KEGG identifies that PCL1 is cell cycle regulated. The association with TIP1 (Butanoate Metabolism), and CHS3/CSD2 are identified as indirect.

Phase S

The genes involved in S phase showed very strong association among them. HTA2 and HTA1 from S phase were not found in KEGG cell cycle pathway. However SCPD³⁷ identified the function of both genes as mRNA synthesis and transcriptional control, coding for identical proteins. The computationally generated GRN also shows there is a direct strong connection between these genes.

DOT1 is one those genes that is not included in 104 cell cycle regulated genes and in KEGG, but however it is identified as nucleosomal histone H3-Lys79 methylase, involved in silencing at telomeres, most likely by directly modulating chromatin structure and Sir protein localisation. Both the publications of Spellman *et al.* (1998) and the last update of the KEGG pathway database pre-dates its identification. Along with this, also many other genetic interactions have also been discovered since, e.g. STE1, HHT1 and HHT2 (Krogan *et al.*, 2003). It would be interesting to see the interactions of these genes in computationally derived GRN.

Cell cycle-regulated chromatin assembly pathway in budding yeast requires a specific pattern of posttranslational modification of histones H3 and/or H4 (encoded by gene HHT1 and HHT2, and HHF1) which is distinct for H3/H4 present in S phase versus M and G1 phase. Yeast BDF1 bromodomains recognise endogenous acetyl-histone H3/H4 as a mechanism for chromatin association in vivo. Deletion of BDF1 or a BDF1 mutation that abolishes histone binding leads to transcriptional down regulation of genes located at heterochromatin-euchromatin boundaries. BDF1 competes with the SIR2 deacetylase for binding to acetylated histone H4 (Ladurner *et al.*, 2003). These genetic regulations were not shown in KEGG or MIPS, and it is also not included in the 104 cell cycle regulated genes. There are many other genetic associations that were discovered or confirmed since the last update. One of the highly documented genetic interactions in cell cycle regulation is DBF4 and CDC7 (Leatherwood, 1998; Dutta *et al.*, 1997 and Toone *et al.*, 1997). DBF4 is an essential gene that is expressed in a cell cycle-dependent manner, encodes for the regulatory subunit of protein kinase.

The activity of the CDC7p-DBF4p complex is required throughout S phase for the initiation of DNA synthesis at replication origins (Jackson *et al.*, 1993).

³⁷ <http://cgsigma.cshl.org/jian/HTML/functiongroup.html>. As at 10/11/03.

In all eukaryotes, the initiation of DNA synthesis requires the formation of pre-replicative complexes (pre-RCs) on replication origins, followed by their activation by two S-T protein kinases, an S-phase cyclin-dependent kinase (S-CDK) and a homologue of yeast DBF4-CDC7 kinase (DBF4p-dependent kinase (DDK).

Yeast DDK activity is cell cycle regulated, though less tightly than that of the S-CDK CLB5-CDK1, and peaks during S phase in correlation with DBF4p levels. DBF4p is short-lived throughout the cell cycle, but its instability is accentuated during G1 by the anaphase-promoting complex. Down regulating DDK activity is physiologically important, as joint CDC7p and DBF4p over expression is lethal. Because pre-RC formation is a highly ordered process, one needs to establish whether S-CDK and DDK need also to function in a specific order for the firing of origins. Zou *et al.* (1998), found that both kinases are activated independently, but they shown that DDK can perform its function for DNA replication only after S-CDKs have been activated. CDC45p, a protein needed for initiation, binds tightly to chromatin only after S-CDK activation (Zou *et al.*, 1998). Zou *et al* also show that CDC45p is phosphorylated by DDK in vitro, suggesting that it might be one of DDK's critical substrates after S-CDK activation. Nougarede R, *et al.* (2000), found that linking the origin-bound DDK to the tightly regulated S-CDK in a dependent sequence of events may ensure that DNA replication initiates only at the right time and place.

Cyclins are the regulatory subunits that activate CDKs at the appropriate time in the cell cycle; cyclical accumulation during particular phases of the cell cycle. CLB6 is one the many genes that encodes a cyclin. In particular it codes for a B-type cyclin that activates CDC28p to promote initiation of DNA synthesis. Progression through the cell cycle is a carefully regulated process that is conserved throughout eukaryotes. Periodic activation of cyclin-dependent kinases (CDKs) are required for this process and the critical CDK involved in cell cycle progression in yeast is CDC28p (Lew, *et al.*, 1997). CLN1, CLN2, and CLN3 encode the yeast G1 cyclins while there are 6 B-type cyclins (CLB) genes involved in activation of S, G2, and M phases of the cell cycle (Lew, *et al.*, 1997). CLB5 and CLB6 transcripts are expressed periodically throughout the cell cycle and are most abundant during late G1 (Schwob, *et al.*, 1993; Spellman, *et al.*, 1998).

The combination of genetic, biochemical, molecular and cell biological techniques give us clues of what is known of the molecular control of cell growth processes. The checkpoint regulatory mechanism has an important role in maintaining the integrity of the genome. This is particularly important in S phase of the cell cycle, when genomic DNA is most susceptible to various environmental hazards. When chemical agents damage DNA, activation of checkpoint signalling pathways results in a temporary cessation of DNA replication.

Eukaryotic cells coordinate chromosome duplication by the assembly of protein complexes at origins of DNA replication by sequential binding of member proteins of the origin recognition complex

(ORC), CDC6, and mini chromosome maintenance (MCM) proteins. These pre-replicative complexes (pre-RCs) are activated by cyclin-dependent kinases and DBF4/CDC7 kinase.

A short literature review of extracted genetic associations

FAR1 and SIC1

Both FAR 1 and SIC1 are CDK inhibitors. Substrates of the CDC28-CLN1 CDK are FAR1p, SIC1p, and the so far not identified proteins p41 and p94 (Andrews and Measday, 1998). FAR1, a pheromone-activated inhibitor of CLN-CDC28 kinases, is dispensable for arrest of CLN⁻ SIC1 cells by pheromone, implying the existence of an alternate FAR1-independent arrest pathway (Tyers, 1996).

CLN2 and FAR1

The FAR1 gene is required for cell cycle arrest but not for a number of other aspects of the pheromone response. Genetic evidence suggests that FAR1 is required specifically for inactivation of the G1 cyclin CLN2 (Valdivieso, Sugimoto *et al.*, 1993). Phosphorylation by the MAP kinase homolog FUS3p may help regulate FAR1p's association with the CDC28p/CLN2p complex (Peter, Gartner *et al.*, 1993).

CLN1 and CLN2

CLN1 and CLN2 are closely related genes with overlapping functions; both are expressed in late G1 phase when they associate with CDC28p to activate its kinase activity. Accumulation of CLN1 and CLN2 mRNA in late G1 is dependent on two transcription factor complexes, MBF (SWI6p-MBP1p) and SBF (SWI6p-SWI4p), which bind to MCB and SCB promoter elements, respectively. In addition, CLN3p has been shown to activate CLN1 and CLN2 transcription (Wittenberg, Sugimoto *et al.*, 1990).

SIC1 and CDC6

Several striking similarities exist between SIC1 and CDC6, a DNA replication factor essential for the formation of pre-replicative complexes at origins of DNA replication. Transcription of both genes is activated during late mitosis by a process dependent on SWI5. Like SIC1, CDC6 binds CDK complexes in vivo and downregulates them in vitro. CDC6, like SIC1, also contributes to inactivation of CDKs during late mitosis in *S. cerevisiae*. Deletion of the CDK-interacting domain of CDC6 does not inhibit the function of origins of DNA replication during S phase, but instead causes a delay in mitotic exit; this delay is accentuated in the absence of SIC1 or of cyclin degradation (Calzada, Sacristan *et al.*, 2001).

SIC1 and CDC46 (MCM5)

Deletion of SIC1, which is an inhibitor of CLB5p (Schwob, Bohm *et al.*, 1994), resulted in an increase of MCM5-BOB1 suppression, as the SIC1 MCM5-BOB1 CDC7ts colonies grew faster at the restrictive temperature. However, this increase of MCM5-BOB1 suppression could not compensate for the loss of CLB5, as the MCM5-BOB1 CDC7ts CLB5 SIC1 strain remained Tsm-. We conclude that CLB5p and CLB2p are required for the MCM5-BOB1 mutation to suppress CDC7ts mutations (Sclafani, Tecklenburg *et al.*, 2002).

SIC1 and CLB5

The complementary C-terminal segment of SIC1 binds to the S-phase cyclin CLB5, indicating a modular structure for SIC1 (Verma, Feldman *et al.*, 1997). Purified p40SIC1 inhibits CDC6/CDC28 and CLB5,6/CDC28 interaction in vitro and in vivo. Upon its destruction by programmed proteolysis, cyclin-CSK1 activity is induced. This degradation then triggers the G1-S transition (Lew DJ, 1997).

CLB5 and CLB6

B-type cyclins CLB5 and CLB6 control the initiation of recombination and synaptonemal complex formation in yeast meiosis. CLB5 and CLB6 transcripts are expressed periodically throughout the cell cycle and are most abundant during late G1. Both CLB5 and CLB6 promoters contain MCB (MluI cell cycle box) motifs, which are elements found in several DNA synthesis genes (Smith, Penkner *et al.*, 2001).

CLB5 and RAD53

In addition to the cell-size dependent Start check point (CLN3), there are additional checkpoints that arrest cells during the cell cycle to avoid DNA damage or cell death due to events occurring out of order. These control points are situated at the G1-S and G2-M boundaries and can be considered as internal regulatory systems that arrest the cell cycle if prerequisites for progression are not met. In the yeast *Saccharomyces cerevisiae* a principal DNA damage checkpoint delay occurs at the G2/M stage of mitosis. This delay occurs through the activation of two MEC1 dependent pathways. In one pathway, the anaphase inhibitor PDS1 binds ESP1 and prevents its ability to activate cleavage of the sister chromatid 'glue' SCC1 (Ciosk, Zachariae *et al.*, 1998). The other pathway includes RAD53, DUN1, and CDC5 and inhibits sister chromatid separation by an unknown mechanism (Gardner and Burke, 2000). "Though the precise position of CLB5 and CLB6 in this regulatory network is unclear, one testable model is that these two cyclins function jointly with PDS1 to regulate

ESP1 activation of SCC1 degradation during a prolonged metaphase arrest. Whether they do this independently of the RAD53 pathway remains to be determined” (Meyn Iii and Holloway, 2000). “The stabilisation of B-type cyclin is independent of the MEC1/RAD53 pathway suggesting that the signalling pathway responsible for the proteolytic response is distinct from this DNA damage checkpoint pathway. The inhibition of CLB5 proteolysis is independent of the MEC1 and RAD53 genes“(Germain, Hendley *et al.*, 1997).

CDC45 and RAD53

CDC45 functions interdependently with CDC7p and DBF4p (Owens, Detweiler *et al.*, 1997). As a consequence of RAD53/CDS1 activation, CDC7/DBF4 becomes hyper-phosphorylated and is displaced from chromatin (Jares, Donaldson *et al.*, 2000).

CDC45 and MCD1

MCD1 also known as SCC1 is known to be associated with PDS1 (separase inhibitor called securin) (Yen, Gitsham *et al.*, 2003), which is directly influenced by CDC20 and APC/C in M/G1 (Shirayama M., Toth A. *et al.*, 1999). This complex is known to activate CLB5 and CLB6 (Shirayama M., Toth A. *et al.*, 1999; Wasch and Cross, 2002), which in turn binds to CDC6, CDC45 with Orc and Mcm complex known as Pre-Replicative Complex (Pre-RC) (Pasero and Schwob, 2000).

DBF4 and CDC45

Following CDK activation in late G1, CDC45p becomes associated with the origin. Removal of CDC6p may stimulate this event. It is a DNA replication initiation factor, recruited to MCM pre-RC complexes at origins of replication. It promotes disengagement of MCM from its anchor, MCM10p, and recruits elongation machinery. Functions interdependently with CDC7p and DBF4p. CDC45p together with ORC and MCM proteins is component of prereplication complexes (Owens, Detweiler *et al.*, 1997).

CDC14 and DBF2

CDC14 is self regulated i.e. expression of CDC14 is controlled by itself. When CDC14 is over expressed, the CDC15 and DBF2 are induced and this in turn suppresses CDC14 expression (Bardin and Amon, 2001). CDC14 regulates hyperphosphorylation and hyperactivity of DBF2p; required directly or indirectly to inhibit the kinase activity of DBF2p (Visintin, Craig *et al.*, 1998).

DBF2 and MOB1

MOB1, an essential protein required for a late mitotic event in the cell cycle, genetically and physically interacts with DBF2. DBF2 binds MOB1 *in vivo* and can bind it *in vitro* in the absence of other yeast proteins.

The expression of MOB1 is also cell cycle regulated, its expression peaking slightly before that of DBF2 at the G2/M boundary (Komarnitsky, Chiang *et al.*, 1998).

CLB1 and CLB2

CLB1 and CLB2 both promote cell cycle progression into mitosis. In yeast cell cycle pathway of KEGG CLB1 and CLB2 are given as genetic complex in G2 phase. The CLBs are regulated both transcriptionally and post-translationally (Deshaies, 1997). CLB1 and CLB2 transcripts accumulate during G2 and M, and their transcription is repressed by the end of mitosis and their proteins are degraded at the end of mitosis as well (Lew DJ, 1997). Transcription of CLB2 is autocatalytic, because CLB2/CDC28 activates its own transcription factor (MCM1/SFF) (Ross, Scherf *et al.*, 2000).

PCL2 and PCL1

PHO85 kinase was shown to associate with G₁ cyclin homologues, including PCL1, PCL2, and PCL9 (Nishizawa, Kawasumi *et al.*, 1998; Espinoza, Ogas *et al.*, 1994; Measday, Moore *et al.*, 1994). Expression of the genes coding for these PCL proteins is periodical: that of PCL1 is peaked in G₁ and is regulated by SBF as CLN1 and CLN2 are (Nasmyth and Dirick, 1991; Ogas, Andrews *et al.*, 1991). PCL2 reaches the maximum between those of PCL1 and PCL9, being regulated by SBF and SWI5 } (Measday, Moore *et al.*, 1994; Aerne, Johnson *et al.*, 1998). Either PCL1 or PCL2 required for G₁ progression when both CLN1 and CLN2 are deleted (Measday, Moore *et al.*, 1994).

CLB1 and MOB1

(Jeong-Ho Chang, Sung Wook Chi *et al.*, 2003) found that 21 genes including ACE2, CLB1, CLB2, and MOB1 among 24 in cluster 31 with typical expression characteristics similar to the dominant patterns in the phase groups.

CLB1 and CDC5

Genes that normally exhibit a G2-to-M-phase-specific expression pattern, such as CLB1, CLB2, CDC5, SWI5, and ACE2, remain uninduced in the absence of functional MCM1. A gene fusion that provides MCM1 with a strong transcriptional activation domain causes

transcription of SWI5, CLB1, CLB2, and CDC5 at inappropriate times of the cell cycle (Althoefer, Schleiffer *et al.*, 1995).

CLB1 and CDC20

CDC20 can be suppressed by overproduction of CDC5p. CDC5 directly inhibits CLB1 and CLB2 (Tavormina and Burke, 1998).

CDC20 and DBF2

The APC/C regulates both transitions by targeting different substrates for proteolysis at two different times. This is accomplished by two accessory factors: CDC20p targets PDS1p to the APC/C for proteolysis at the metaphase-to-anaphase transition, and CDH1p/HCT1p targets the major B-type cyclin (CLB2p) for destruction after telophase when cells exit mitosis (Zachariae and Nasmyth, 1999). Late regulation of the APC/C in the cell cycle requires a number of genes that form the mitotic exit network (MEN), including CDC5p, CDC15p, DBF2p, DBF20p, LTE1p, MOB1p and TEM1p (Gardner and Burke, 2000).

CDC20 and CDC47 (MCM7)

(Spellman, Sherlock *et al.*, 1998) clustered the cell cycle regulated genes using the clustering algorithm of (Eisen, Spellman *et al.*, 1998). The “CLB2” cluster contains 35 genes and includes many genes involved in mitosis such as CLB2, CDC5, CDC20, and SWI5. (Spellman, Sherlock *et al.*, 1998) compared the upstream regions of genes in the CLB2 cluster and certain other co-regulated genes (e.g., ASE1, also thought to be a possible target of SFF (Pellman, Bagget *et al.*, 1995) and found that most of them contain an easily recognizable MCM1 1 SFF motif. “Of the 35 genes in the cluster, only 9 genes (KIP2, MOB1, NUM1, YCL012W, BUD3, CHA1, YCL063W, YLR057W, and YML033W) did not have an easily recognizable near match to the MCM1 1. The “MCM” cluster contains 34 genes, including all six MCM genes that are directly involved in DNA replication (MCM2, MCM3, CDC54, CDC46, MCM6, and CDC47 as well as FAR1, DBF2, SPO12, and KIN3. These genes peak late in the cycle, at about the M/G1 bound-ary, and are induced by CLB2 and somewhat repressed by CLN3. This cluster has similarities to the CLB2 cluster, except that peak expression is slightly later”.

CLB3 and CLB4

Genetic interactions have shown that CLB3 and CLB4 may both been involved in DNA replication and spindle assembly as well as the G2/M-phase transition. CLB3 and CLB4

transcripts accumulate during S phase and G2. Like the other CLB proteins, CLB3p and CLB4p contain a destruction box motif in their amino termini, which may target them for ubiquitin-mediated degradation by the proteasome (Lew DJ, 1997).

CLN1 and HSL1

HSL1 and HSL7 promote proper timing of cell cycle progression by coupling septin ring assembly to alleviation of SWE1-dependent inhibition of CDC28. HSL1p is degraded in an APC-dependent manner. Bud formation, needed for HSL1 activation, is due to CDC28-CLN1-2 (Andrea Ciliberto, Bela Novak *et al.*, 2003).

CLN1 and SWE1

SWE1 (*Saccharomyces weel*) encodes a protein kinase involved in regulating the G2/M transition 1. SWE1p inhibits the kinase activity of the main cell-cycle cyclin-dependent kinase CDC28p through phosphorylation of a conserved tyrosine residue (Tyr 19). SWE1p localizes to the nucleus and also to the daughter-side of the mother-bud neck where it may be marked for degradation. *Glu12Gly*, a mutation that alters a residue involved in SWE1p inhibition of CDC28p, was unique in being suppressed by deficiency of SWE1 or CLN1. With wild-type CDC28, filament formation induced by CLN1 overexpression was markedly decreased in a SWE1 deletion. These results suggest that SWE1p, via inhibition of CLB2p/CDC28p, may mediate much of the effect of CLN1p on filamentous morphogenesis (Ahn, Tobe *et al.*, 2001).

MCD1 and SWI4

SBF (SWI4 and Swi6) regulates genes involved in morphological changes - for example, chromatin state, cell-wall growth, budding and cytokinesis - such as histone genes (HTA1-3, HTB1-2 and HHO1), genes for cell-wall proteins (SCW4, CIS3, CWP1, MSB2), budding proteins (GIC2, BUD9, RSR1, CRH1), and enzymes involved in cell-wall synthesis (OCH, MNN1, GLS1, KRE6, GAS1, PSA1). In contrast, MBF (MBP1 and SWI6) mainly regulates genes involved in DNA replication and repair (RNR1, RAD27, CDC21, IRR1, MCD1, DUN1, MCM2, CDC45) (Simon, Barnett *et al.*, 2001).

CLB2 and SWI5

SWI5p, NAP1p, SWI4p and probably CDC6p are substrates for CDC28-CLB2 Cyclin CDK (Andrews and Measday, 1998). Overproduction of NDD1 also enhances the expression of

SWI5, whose transcription, like that of CLB1 and CLB2, is activated in the late S phase. NDD1 is essential for the expression of CLB1, CLB2, and SWI5, since none of these genes are transcribed in its absence (Loy, Lydall *et al.*, 1999). In vivo footprinting experiments show that MCM1, in conjunction with an MCM1-recruited factor, binds to the promoter regions of SWI5 and CLB2 at sites shown to be involved in cell cycle regulation. (Althoefer, Schleiffer *et al.*, 1995). The 'CLB2' cluster contains 33 genes whose transcription peaks early in mitosis, including CLB1, CLB2, SWI5, ACE2, CDC5, CDC20 and other genes important for mitosis (Zhu, Spellman *et al.*, 2000).

CLN3 and CDC47 (MCM7)

(McInerny, Partridge *et al.*, 1997) identified a promoter element that confers M/G1-specific transcription in *S. Cerevisiae*. This element, called an ECB (early cell cycle box), was first identified in the SWI4 promoter, but it is also present in the promoter of a G1 cyclin CLN3, as well as in the promoters of three DNA replication genes: CDC6, CDC47, and CDC46 (McInerny, Partridge *et al.*, 1997). CLN3 is present at low and nearly constant levels throughout the cell cycle (Wittenberg, Sugimoto *et al.*, 1990).

Appendix II

Table 15: Gene regulatory reference databases

Name	Data Type	Description	WWW Link
TRANSPATH	Db/XML	Signal transduction pathways and cellular mechanisms regulating transcription	http://www.biobase.de/pages/products/transpath.html
KEGG	XML based	Molecular interactions: metabolic pathways, regulatory pathways and molecular assemblies.	http://www.kegg.com
BIND	Interactive	Mutated component of signal transduction pathways	http://www.bind.ca/index.phtml?page=databases
PathoPath		mutated components of signal transduction pathways, pathological defects due to impairment of GR	currently under development
GO		provides three structured networks of defined terms to describe gene product attributes.	http://www.geneontology.org/
Biocarta		Dynamic graphical models for molecular relationship in pathways	http://www.biocarta.com/
PubGene		Text and sequence homology search for genes and expression data	http://www.pubgene.org
TRANSFAC		Regulating genome elements, transcription factors	http://www.gene-regulation.com/

GeNet		Functional organisation of regulatory gene network at embryogenesis	http://www.esa.ru:82/Inst/gorb_dep/inbios/genet/genet.htm
GATO	Software	visualises algorithms on graphs	http://algorithms.molgen.mpg.de/gato.html
UrchiNet	Java applets	The database contains information on molecular mechanisms of embryogenesis in sea urchin	http://www.esa.ru/Inst/gorb_dep/inbios/genet/Urchin/u0urmt.htm
MedMiner	Online	Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling	http://discover.nci.nih.gov/textmining/filters.html
GoMiner		biological interpretation of 'omic' data	http://discover.nci.nih.gov/gominer/
FatiGO		A web tool for finding significant associations on GO terms with groups of genes	http://fati.go.bioinfo.cnio.es/
Geome Object Net (GON)	XML based	environment for simulating and representing biological systems. It is an integrated tool of several software tools for describing biopathways, visualising simulation results, evaluating a hypothesis, biopathways recreation from biopathway databases.	http://www.genomicobject.net/member3/index.html
GenePath	Web enabled	An intelligent assistant to discovery of Genetic Pathways	http://genepath.org/
CYGD	XML	present information on the molecular structure and functional network of the entirely sequenced, well-studied model eukaryote, the budding yeast <i>Saccharomyces Cerevisiae</i> .	http://www.mips.biochem.mpg.de/proj/yeast/pathways/index.html
The YoungLab		transcriptional regulatory circuitry of eukaryotic cells	http://web.wi.mit.edu/young/

Table 16: Cross-reference databases of biological data. Adapted from (Gardiner-Garden and Littlejohn, 2001).

Cross-reference Database	Cross-referencing For	URL
dBEST	sequence data and Expressed Sequence Tags (ESTs) from a number of organisms.	www.tigr.org
DDBj	genetic sequence data.	www.ddbj.nig.ac.jp
EMBL	nucleotide sequence data.	www.ebi.ac.uk
Ensemble	automatic annotation on metazoan genomes.	www.ensembl.org
GenBank	nucleotide sequences from > 130,000 organisms.	www.ncbi.nlm.nih.gov/
GeneMap99	DNA sequence data and an index locations of more than 30,000 genes of the Human Genome.	www.ncbi.nlm.nih.gov/genemap/
GUS	human and mouse transcript index.	www.allgenes.org
KEGG	genes data, ligand data on chemical reactions and a pathway database.	www.genome.ad.jp/kegg/
LocusLink	curated sequence and descriptive information about genetic loci.	www.ncbi.nlm.nih.gov/LocusLink/refseq.html
MGD	mouse genome database.	www.informatics.jax.org/mgihome/MGD/aboutMGD.shtml
PathDB	cellular networks, pathways and phenotypes.	www.ncgr.org/software/pathdb/
PubMed	literature database (medicine)	www.nlm.nih.gov/databases/free Medl.html

SGD	database of the molecular biology and genetics of yeast	genome-www.stanford.edu/Saccharomyces/
SWISS-PROT	annotated protein sequence database.	www.ebi.ac.uk/swissprot/
Taxonomy	contains the names of all organisms that are represented in the genetic databases with at least one nucleotide or protein sequence.	www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/
TIGR	data on genomes and gene products for a variety of organisms including viruses, eubacteria, archaea, and eukaryotes.	www.tigr.org/
UniGene	an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters.	www.ncbi.nlm.nih.gov/UniGene/
YPD	Provides scientific references including data on protein classification and function for the yeast genome.	www.proteome.com/

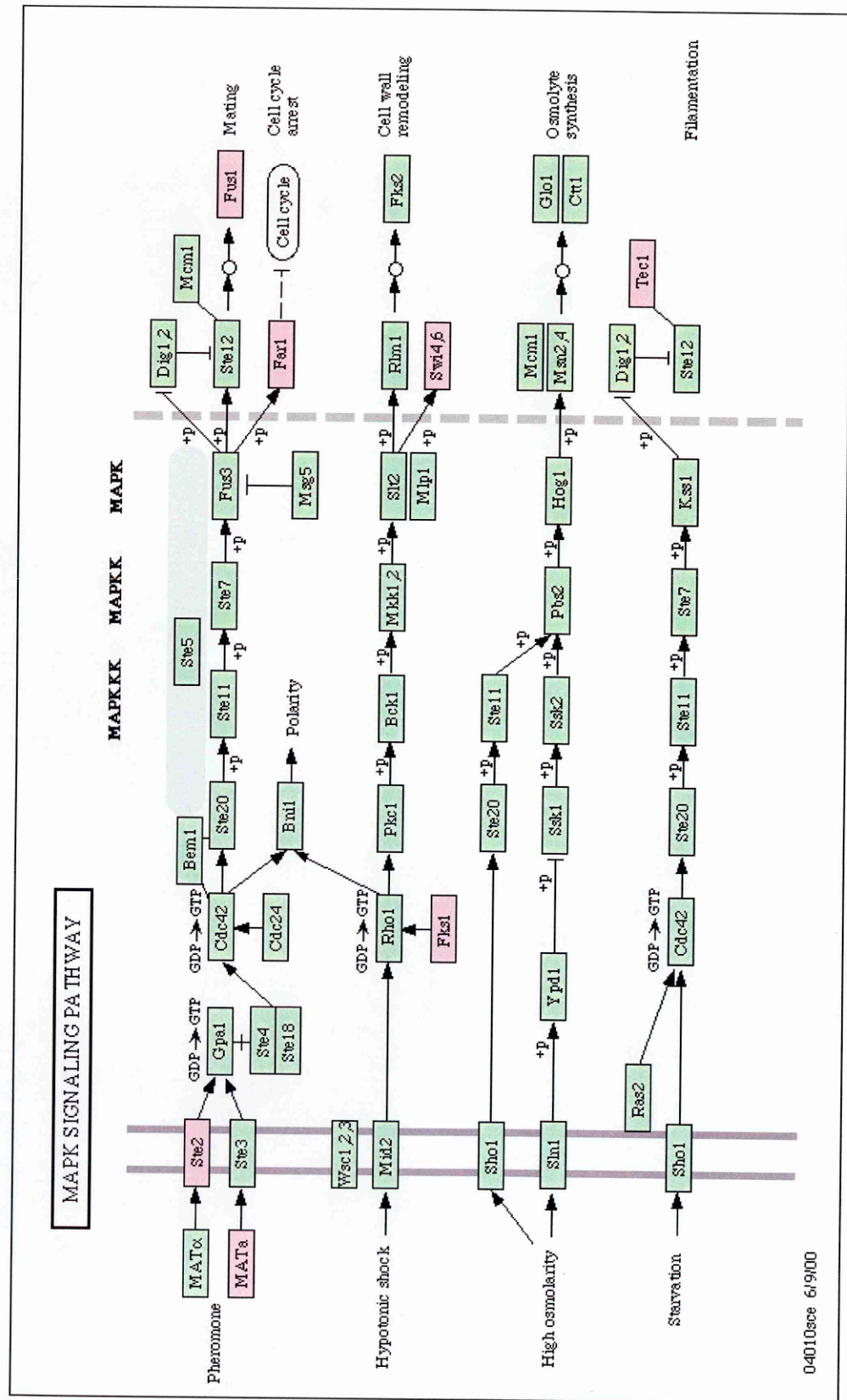


Figure 67: MAPK signalling pathway in yeast. The pink box represents the gene that is identified as cell cycle reulated by traditional methods. Appendix IV

Table 17: Gene classification. All 104 traditionally determined to be yeast cell cycle regulated genes names, ORFs and their biological processes. (based on data from SGD)

Description of yeast genes identified as cell cycle regulated by traditional methods				
Cell Cycle Phase				
ORF	Gene Name	Traditional	Spellman	Biological Process
YJR057W	CDC8	Late G1	-	DNA repair, DNA replication
YOR074C	CDC21	Late G1	G1	DNA replication, dTMP biosynthesis
YER070W	RNR1	Late G1	G1	DNA replication
YIL066C	RNR3	Late G1	G1	DNA replication
YNL192W	CHS1	M/G1	M/G1	Budding;cytokinesis
YLR286C	CTS1	M/G1	G1	Cytokinesis, Cell wall biogenesis
YLR342W	FKS1	G1	G1	Beta-1,3 glucan biosynthesis; Cell wall organization
YGL027C	CWH41	G1	G1	cell wall organization and biogenesis
YPR159W	KRE6	G1	S	beta-1,6 glucan biosynthesis; cell wall organization
YER001W	MNN1	G1	G1	N-glycan processing,protein glycosylation
YDL055C	PSA1	G1	G1	mannose metabolism; GDP-mannose biosynthesis.
YBR023C	CHS3	G1	-	cell wall chitin biosynthesis; cytokinesis;response to osmotic stress
YLR131C	ACE2	G2/M	G2/M	G1-specific transcription in mitotic cell cycle
YDR309C	GIC2	Late G1	G1	bud emergence; Rho protein signal transduction; axial budding.

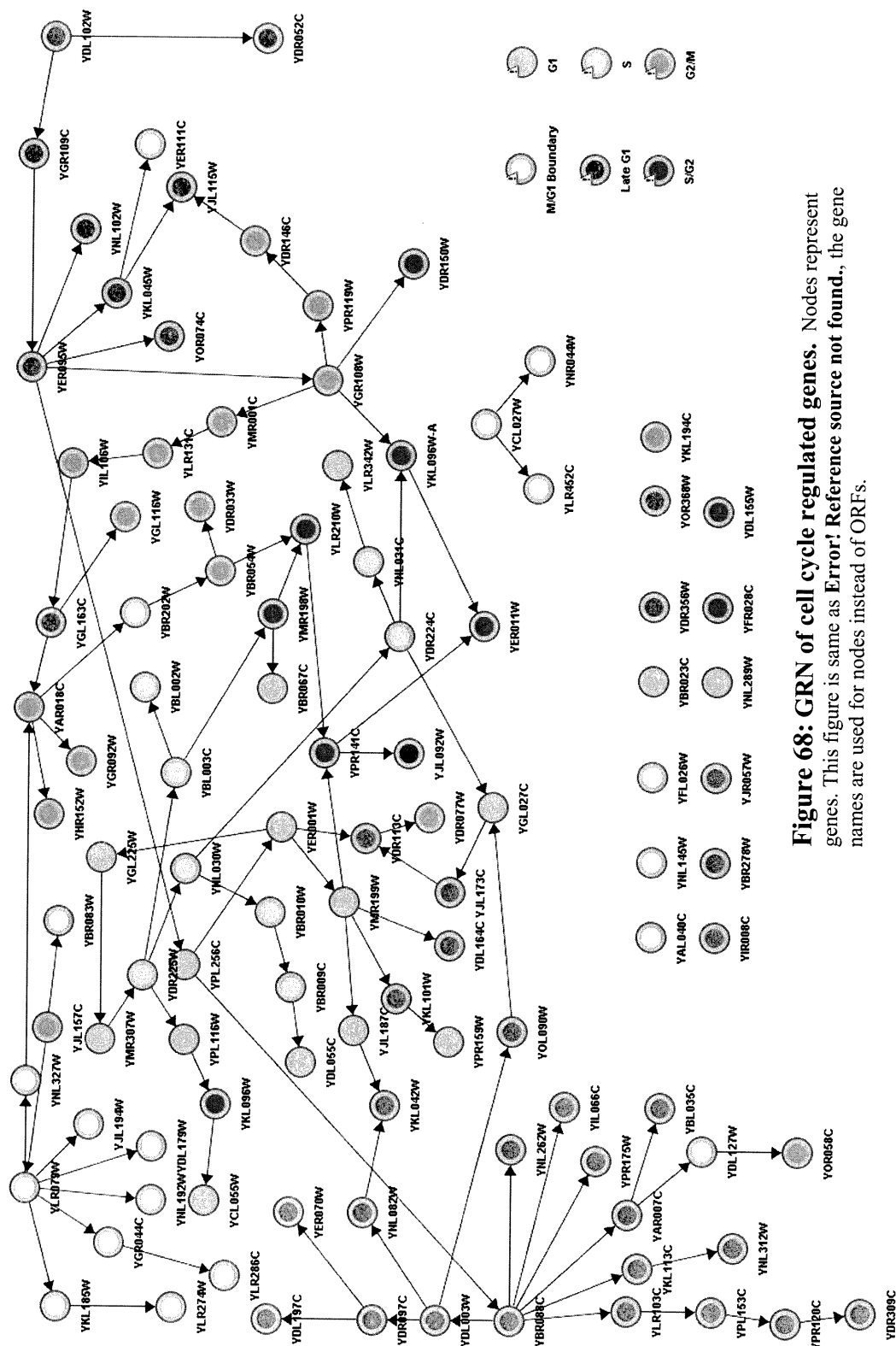
YCL027W	FUS1	M/G1	M/G1	conjugation (mating) with cellular fusion.
YLR452C	SST2	M/G1	M/G1	adaptation to pheromone during conjugation (mating) with cellular fusion; signal transduction.
YFL026W	STE2	M/G1	G2/M	cell projection biogenesis; response to pheromone during conjugation (mating) with cellular fusion.
YJL157C	FAR1	G2/M	G2/M	cell cycle arrest; signal transduction during conjugation with cellular fusion
YPL153C	RAD53	Late G1	G1	DNA repair; nucleobase, nucleoside, nucleotide and nucleic acid metabolism
YBR067C	TIP1	G1	M/G1	stress response (putative); cell wall organization and biogenesis
YPR141C	KAR3	S/G2	G1	karyogamy during conjugation with cellular fusion; meiosis; mitosis
YER011W	TIR1	S/G2	-	response to stress
YBR054W	YRO2	G2/M	G2/M	Unknown
YDR033W	MRH1	G2/M	G2/M	Unknown
YDR077W	SED1	G2/M	G2/M	cell wall organization and biogenesis
YDL164C	CDC9	Late G1	G1	DNA ligation; DNA recombination.
YNL262W	POL2	Late G1	G1	DNA repair synthesis.
YDL102W	CDC2	Late G1	G1	base-excision repair; mismatch repair; mutagenesis; nucleotide-excision repair;
YER095W	RAD51	Late G1	G1	chromatin remodeling; DNA repair and recombination
YGL163C	RAD54	Late G1	G1	chromatin remodeling; DNA repair
YJL092W	HPR5	Late G1	S	DNA repair; Meiosis
YOR368W	RAD17	Late G1	-	DNA damage response, signal transduction resulting in cell cycle arrest; meiotic recombination
YNL327W	EGT2	M/G1	M/G1	cytokinesis
YIL106W	MOB1	G2/M	G2/M	protein amino acid phosphorylation; regulation of exit from mitosis
YBR083W	TEC1	M/G1	M/G1	positive regulation of transcription from Pol II promoter; pseudohyphal growth
YPL256C	CLN2	G1	G1	Cell cycle; re-entry into mitotic cell cycle after pheromone arrest; regulation of CDK activity
YPR119W	CLB2	G2/M	G2/M	G2/M transition of mitotic cell cycle; regulation of CDK activity

YNR044W	AGA1	M/G1	M/G1	agglutination during conjugation (mating) with cellular fusion
YKL185W	ASH1	M/G1	M/G1	pseudohyphal growth; regulation of transcription, mating-type specific
YNL145W	MFA2	M/G1	G2/M	signal transduction during conjugation (mating) with cellular fusion
YCL055W	KAR4	G1	M/G1	karyogamy during conjugation with cellular fusion; meiosis; mitosis.
YAR007C	RFA1	Late G1	G1	DNA replication, priming; DNA recombination; DNA strand elongation; DNA repair
YDR146C	SWI5	G2/M	G2/M	G1-specific transcription in mitotic cell cycle
YHR152W	SPO12	G2/M	G2/M	meiosis I; mitotic cell cycle; regulation of exit from mitosis
YGR044C	RME1	M/G1	G1	meiosis; negative regulation of transcription from Pol II promoter
YLR274W	CDC46	M/G1	M/G1	DNA replication initiation; pre-replicative complex formation and maintenance
YBR202W	CDC47	M/G1	G2/M	DNA replication initiation; DNA unwinding; establishment of chromatin silencing
YJL194W	CDC6	M/G1	M/G1	pre-replicative complex formation and maintenance
YLR103C	CDC45	Late G1	G1	DNA replication initiation; DNA strand elongation; pre-replicative complex formation and maintenance
YPR120C	CLB5	Late G1	G1	G1/S and G2/M transition of mitotic cell cycle; premeiotic DNA synthesis; regulation of CDK activity
YGR109C	CLB6	Late G1	G1	G1/S transition of mitotic cell cycle; premeiotic DNA synthesis; regulation of CDK activity
YPR175W	DPB2	Late G1	G1	lagging strand elongation; leading strand elongation; mismatch repair; nucleotide-excision repair
YNL102W	POL1	Late G1	G1	DNA replication initiation; DNA repair synthesis; DNA replication, priming; lagging strand elongation
YBL035C	POL12	Late G1	G1	DNA replication initiation; DNA replication, priming; lagging strand elongation
YBR088C	POL30	Late G1	G1	base-excision repair; mismatch repair; nucleotide-excision repair;
YKL045W	PR12	Late G1	G1	DNA replication initiation; DNA repair synthesis; DNA replication, priming; lagging strand elongation
YKL113C	RAD27	Late G1	G1	DNA replication; DNA repair; DNA repair synthesis; replicative cell aging.
YNL312W	RFA2	Late G1	G1	DNA recombination; DNA replication; DNA strand elongation.
YJL173C	RFA3	Late G1	G1	DNA recombination; DNA replication; DNA strand elongation.

YDR052C	DBF4	Late G1	-	DNA replication initiation; protein amino acid phosphorylation; regulation of DNA replication
YBR278W	DPB3	Late G1	-	lagging strand elongation; leading strand elongation; mismatch repair; nucleotide-excision repair
YIR008C	PR1	Late G1	-	DNA replication initiation; DNA replication, priming; lagging strand elongation
YFR028C	CDC14	S/G2	-	protein amino acid dephosphorylation; regulation of exit from mitosis
YMR001C	CDC5	G2/M	G2/M	DNA dependent DNA replication; protein amino acid phosphorylation
YOL090W	MSH2	Late G1	G1	DNA recombination; mismatch repair; mitotic recombination; removal of nonhomologous ends
YDR097C	MSH6	Late G1	G1	mismatch repair
YNL082W	PMS1	Late G1	G1	meiosis; mismatch repair
YAL040C	CLN3	M/G1	G2/M	G1/S transition of mitotic cell cycle; regulation of CDK activity
YDL127W	PCL2	M/G1	G1	cell cycle
YDL179W	PCL9	M/G1	M/G1	cell cycle
YLR079W	SIC1	M/G1	M/G1	G1/S transition of mitotic cell cycle; regulation of CDK activity
YER111C	SWI4	M/G1	M/G1	G1/S transition of mitotic cell cycle; cell cycle; transcription
YMR199W	CLN1	G1	G1	regulation of CDK activity
YNL289W	PCL1	G1	G1	cell cycle
YJL187C	SWE1	G1	G1	G2/M transition of mitotic cell cycle; regulation of CDK activity; regulation of meiosis
YDL003W	MCD1	Late G1	G1	mitotic chromosome condensation; mitotic sister chromatid cohesion
YKL101W	HSL1	Late G1	G1	G2/M transition of mitotic cell cycle; protein amino acid phosphorylation; regulation of cell cycle
YDR113C	PDS1	Late G1	G1	mitotic sister chromatid separation; protein-nucleus import
YDR356W	NUF1	Late G1	G1	microtubule nucleation
YKL042W	SPC42	Late G1	G1	microtubule nucleation; spindle pole body duplication (sensu Saccharomyces)
YMR198W	CIK1	S/G2	S/G2	meiosis; mitotic chromosome segregation; mitotic spindle assembly; mitotic spindle orientation

YLR210W	CLB4	S/G2	S/G2	G1/S transition of mitotic cell cycle; G2/M transition of mitotic cell cycle ; regulation of CDK activity
YDR150W	NUM1	S/G2	S/G2	microtubule cytoskeleton organization and biogenesis ; nuclear migration (sensu Saccharomyces)
YDL155W	CLB3	S/G2	-	G1/S transition of mitotic cell cycle ; G2/M transition of mitotic cell cycle ; regulation of CDK activity
YOR058C	ASE1	G2/M	G2/M	mitotic anaphase B ; mitotic spindle assembly
YGL116W	CDC20	G2/M	G2/M	cyclin catabolism ; mitotic chromosome segregation ; mitotic metaphase/anaphase transition ;
YGR108W	CLB1	G2/M	G2/M	G2/M transition of mitotic cell cycle ; regulation of CDK activity
YGR092W	DBF2	G2/M	G2/M	nuclear division ; protein amino acid phosphorylation
YJL115W	ASF1	Late G1	G1	DNA damage response ; signal transduction resulting in induction of apoptosis
YDL197C	ASF2	Late G1	G1	chromatin silencing at HML and HMR (sensu Saccharomyces)
YBR010W	HHT1	S	S	chromatin assembly/disassembly
YNL031C	HHT2	S	S	chromatin assembly/disassembly
YBR009C	HHF1	S	S	chromatin assembly/disassembly
YNL030W	HHF2	S	S	chromatin assembly/disassembly
YDR225W	HTA1	S	S	chromatin assembly/disassembly
YBL003C	HTA2	S	S	chromatin assembly/disassembly
YDR224C	HTB1	S	S	chromatin assembly/disassembly
YBL002W	HTB2	S	S	chromatin assembly/disassembly
YKL194C	MST1	G2/M	-	threonyl-tRNA aminoacylation
YGL225W	GOG5	G1	S	N-linked glycosylation ; lipid glycosylation ; nucleotide-sugar transport
YMR307W	GAS1	G1	G1	cell wall organization and biogenesis
YPL116W	HOS3	G1	S/G2	histone deacetylation
YKL096W	CWP1	S/G2	S/G2	cell wall organization and biogenesis
YKL096W-	CWP2	S/G2	S/G2	cell wall organization and biogenesis ; regulation of pH

A						
YAR018C	KIN3	S/G2	S/G2		chromosome segregation	
YDL227C	HO	G1	G1		gene conversion at MAT locus; mating-type switching/recombination	



Appendix V

Table 18: The pathways and biological functions for the *E. coli* K12 genes obtained from the KEGG .

Pathways and Biological Functions of <i>E. coli</i> K12 from KEGG	Gene Names
<u>Oxidative phosphorylation</u> NADH dehydrogenase I chain	nuoM, nuoL, nuoK, nuoJ, nuoI, nuoH, nuoG nuoF, nuoE, nuoC, nuoB, nuoA, nuoD, atpC, atpD, atpG, atpA, atpF, atpE, atpB frdA, fuma
membrane-bound ATP synthase, F1 sector membrane-bound ATP synthase, F0 sector rate reductase, anaerobic, catalytic and NAD/flavoprotein subunit	
<u>Ubiquinone biosynthesis</u> NADH dehydrogenase I chain	nuoM, nuoL, nuoK, nuoJ, nuoI, nuoH, nuoG nuoF, nuoE, nuoC, nuoB, nuoA, nuoD ubiC ubiA
chorismate lyase cyr; 4-hydroxybenzoate-octaprenyltransferase	
<u>ABC transporters</u> outer membrane pore protein, receptor for ferrichrome, colicin M lasmic putrescine-binding protein; permease protein arginine 3rd transport system periplasmic binding protein arginine 3rd transport system permease protein ATP-binding component of 3rd arginine transport system sulfate transport system permease W protein ATP-binding component of sn-glycerol 3-phosphate transport system sn-glycerol 3-phosphate transport protein (ABC superfamily, membrane) sn-glycerol 3-phosphate transport protein (ABC superfamily, peri_bind) D-xylose transport protein (ABC superfamily, peri_bind)	fhuA potF artJ artM, artQ artP cysW ugpC ugpE ugpB xylF, xylT

Purine metabolism

DNA polymerase III
adenosine deaminase
adenylylsulfate kinase
ATP-sulfurylase (ATP:sulfate adenylyltransferase), subunit 1, probably a GTPase
ATP-sulfurylase, subunit 2 (ATP:sulfate adenylyltransferase)
(p)ppGpp synthetase I/GTP pyrophosphokinase
DNA polymerase III, beta-subunit

dnaQ, dnaE, dnaX, dnaZ, mutH, mutD
add
cysC
cysN
cysD
relA
dnaN

Galactose metabolism

beta-D-galactosidase
galactitol-1-phosphate dehydrogenase
PTS family enzyme II, galactitol-specific
putative tagatose 6-phosphate kinase 1
tagatose 1,6-diphosphate aldolase

lacZ
gatD
gatC, gatB, gatA
gatZ
gatY, yegF

Two-component system

outer membrane pore protein
nitrate extrusion protein
response regulator (sensor, EnvZ) affecting transcription of OmpC and OmpF
putative aminotransferase
fumarate reductase, anaerobic, catalytic and NAD/flavoprotein subunit
negative response regulator of genes in aerobic pathways, (sensors, ArcB and CpxA)

ompF, ompC
narU
ompR
yfbE
frdA
arcA

Phosphotransferase system (PTS)

PTS enzyme IIA_B, mannose-specific
PTS family enzyme II, galactitol-specific
PTS family enzyme IIBC, trehalose(maltose)-specific

manX, manY, manZ, ptsB
gatC, gatB, gatA
treB

Pentose phosphate pathway

putative isomerase
fructose-bisphosphate aldolase class I

ybhE
fbaB, b2097

transaldolase A

transketolase 2, thiamin-binding

ribose 5-phosphate isomerase B

Urea cycle and metabolism of amino groups

putrescine:2-oxoglutaric acid aminotransferase, PLP-dependent

argininosuccinate synthetase

acetylornithine deacetylase

N-acetyl-gamma-glutamylphosphate reductase

ornithine carbamoyltransferase 1

talA

tktB

rpiB

ygjG

argG, argD, arg6, argE

argE, argA, arg4, argG

argC, argH, arg2

argI

Alanine and aspartate metabolism

pyruvate dehydrogenase (decarboxylase component)

glutamate decarboxylase, PLP-dependent

argininosuccinate synthetase

aspartate carbamoyltransferase, catalytic subunit

aceE

gadB, gadA

argG, argD, arg6, argE

pyrB, carB

Butanoate metabolism

pyruvate dehydrogenase (decarboxylase component)

pyruvate formate lyase I

glutamate decarboxylase B, PLP-dependent

glutamate decarboxylase A, PLP-dependent

fumarate reductase, anaerobic, catalytic and NAD/flavoprotein subunit

aceE

pflB

gadB

gadA

frdA

Fructose and mannose metabolism

PTS enzyme IIB, mannose-specific

bfructose-bisphosphate aldolase class I

manX, manY, manZ, ptsB

fbaB

Pyrimidine metabolism

DNA polymerase III

aspartate carbamoyltransferase, catalytic subunit

dnaQ, dnaX, dnaZ, dnaN, mutH, mutD

pyrB, carB

Selenoamino acid metabolism

adenylylsulfate kinase	
ATP-sulfurylase (ATP:sulfate adenylyltransferase), subunit 1, probably a GTPase	cysC
ATP-sulfurylase, subunit 2 (ATP:sulfate adenylyltransferase)	cysN
cystathionine beta-lyase	cysD
	metC
Starch and sucrose metabolism	
periplasmic trehalase	treA
trehalose-6-phosphate phosphatase, biosynthetic	otsB
cytoplasmic trehalase	treF, treB
trehalose-6-P hydrolase	treC
cytoplasmic alpha-amylase	amyA
<u>Sulfur metabolism</u>	
adenylylsulfate kinase	cysC
ATP-sulfurylase (ATP:sulfate adenylyltransferase), subunit 1, probably a GTPase	cysN
ATP-sulfurylase, subunit 2 (ATP:sulfate adenylyltransferase)	cysD
cystathionine beta-lyase	metC
<u>Ribosome</u>	
ribosomal protein S2	rpsB
ribosomal subunit protein S22; stationary phase-induced ribosome-associated protein	rpsV
ribosomal protein	rplY
ribosomal protein S21	rpsU
<u>Glycolysis / Gluconeogenesis</u>	
pyruvate dehydrogenase (decarboxylase component)	aceE
alcohol dehydrogenase	adhP
fructose-bisphosphate aldolase class I	fbaB
<u>Citrate cycle (TCA cycle)</u>	
2-oxoglutarate decarboxylase, thiamin-binding	sucA
succinyl-CoA synthetase, beta subunit	sucC
umarate reductase, anaerobic, catalytic and NAD/flavoprotein subunit	frdA

Glutamate metabolism

putative glutaminase
glutamate decarboxylase, PLP-dependent
succinate-semialdehyde dehydrogenase I, NADP-dependent
glutathione oxidoreductase

ybaS
gadA, gadB
gabD
gor

Pyruvate metabolism

pyruvate dehydrogenase (decarboxylase component)
pyruvate dehydrogenase/oxidase: FAD- and thiamine PPi-binding
pyruvate formate lyase I

aceE
poxB
pflB

Carbon fixation

fructose-bisphosphate aldolase class I
transketolase 2, thiamin-binding
ribose 5-phosphate isomerase B

fbaB, b2097
tktB
rpiB

DNA polymerase

DNA polymerase III

dnaQ, dnaZ, dnaX, dnaN, mutH, mutD

Fatty acid metabolism

alcohol dehydrogenase
3-ketoacyl-CoA thiolase; (thiolase I, acetyl-CoA transferase), in complex with FadB catalyzes

adhP
fadA

Bile acid biosynthesis

alcohol dehydrogenase
3-ketoacyl-CoA thiolase; (thiolase I, acetyl-CoA transferase), in complex with FadB catalyzes

adhP
fadA

Arginine and proline metabolism

argininosuccinate synthetase
ornithine carbamoyltransferase 1

argG, argD, arg6, argE
argI

Tryptophan metabolism

2-oxoglutarate decarboxylase, thiamin-binding catalase; hydroperoxidase HPII (III); RpoS-dependent	sucA katE
<u>beta-Alanine metabolism</u> glutamate decarboxylase B, PLP-dependent glutamate decarboxylase A, PLP-dependent	gadB gadA
<u>Taurine and hypotaurine metabolism</u> glutamate decarboxylase, PLP-dependent	gadB, gadA
<u>Glycerolipid metabolism</u> beta-D-galactosidase alcohol dehydrogenase	lacZ adhP
<u>Propanoate metabolism</u> succinyl-CoA synthetase, beta subunit pyruvate formate lyase I	sucC pfIB
<u>Reductive carboxylate cycle (CO2 fixation)</u> succinyl-CoA synthetase, beta subunit fumarate reductase, anaerobic; catalytic and NAD/flavoprotein subunit	sucC frdA
<u>Nitrogen metabolism</u> putative glutaminase cystathionine beta-lyase	ybaS metC
<u>Metabolism of xenobiotics by cytochrome P450</u> hypothetical GST-like protein alcohol dehydrogenase	yncG adhP
<u>Pentose and glucuronate interconversions</u> ribulokinase	araB

<u>Methionine metabolism</u> cystathionine beta-lyase	metC
<u>Cysteine metabolism</u> cystathionine beta-lyase	metC
<u>Valine, leucine and isoleucine degradation</u> 3-ketoacyl-CoA thiolase; (thiolase I, acetyl-CoA transferase), in complex with FadB catalyzes	fadA
<u>Valine, leucine and isoleucine biosynthesis</u> pyruvate dehydrogenase (decarboxylase component)	aceE
<u>Lysine degradation</u> 2-oxoglutarate decarboxylase, thiamin-binding	sucA
<u>Tyrosine metabolism</u> alcohol dehydrogenase	adhP
<u>Gamma-Hexachlorocyclohexane degradation</u> putative dienelactone hydrolase	ysgA
<u>Benzoate degradation via hydroxylation</u> 3-ketoacyl-CoA thiolase; (thiolase I, acetyl-CoA transferase), in complex with FadB catalyzes	fadA
<u>D-Glutamine and D-glutamate metabolism</u> putative glutaminase	ybaS
<u>Glutathione metabolism</u> hypothetical GST-like protein	yncG
<u>N-Glycan degradation</u> beta-D-galactosidase	lacZ

<u>Glycosaminoglycan degradation</u> beta-D-galactosidase	lacZ
<u>Lipopolysaccharide biosynthesis</u> O-antigen ligase	rfaL
<u>Glycerophospholipid metabolism</u> cytoplasmic glycerophosphodiester phosphodiesterase	ugpQ
<u>Sphingolipid metabolism</u> beta-D-galactosidase	lacZ
<u>Glycosphingolipid biosynthesis - ganglioseries</u> beta-D-galactosidase	lacZ
<u>1- and 2-Methylnaphthalene degradation</u> alcohol dehydrogenase	adhP
<u>1,4-Dichlorobenzene degradation</u> putative dienelactone hydrolase	ysgA
<u>Benzoate degradation via CoA ligation</u> fumarate reductase, anaerobic, catalytic and NAD/flavoprotein subunit	frdA
<u>C5-Branched dibasic acid metabolism</u> succinyl-CoA synthetase, beta subunit	sucC
<u>Methane metabolism</u> catalase; hydroperoxidase HPII (III), RpoS-dependent	katE
<u>Glycan structures - degradation</u> beta-D-galactosidase	lacZ

Protein export
hypothetical protein

yajC

Appendix VI

Pearson Correlation Coefficient

The *Pearson product-moment correlation coefficient* (PMCC), r , is a measure of the tendency of two variables X and Y , measured on the same object or organism to increase or decrease together. It is defined as the sum of the products of the standard scores of the two measures divided by the degrees of freedom:

$$r = \frac{\sum ZxZy}{n-1} = \frac{Cov(x,y)}{V(x)V(y)}$$

The result obtained is equivalent to dividing the covariance between the two variables by the product of their standard deviations. In general the correlation coefficient is the square root (can take either positive or negative) of the coefficient of determination (r^2), which is the ratio of explained variation to total variation:

$$r^2 = \frac{\sum (Y' - Y'')^2}{\sum (Y - Y'')^2}$$

Where, Y = a score on a random variable Y , Y' = corresponding predicted value of Y , given the correlation of X and Y and the value of X , and Y'' = sample mean of Y .

The correlation coefficient adds a sign to show the direction of the relationship, and applies *only* when the relationship is *linear*.

The coefficient ranges from -1 to 1. A value of 1 show that a linear equation describes The relationship perfectly and positively, with all data points lying on the same line and with Y increasing with X . A score of -1 shows that all data points lie on a single line but that Y increases as X decreases. A value of 0 shows that a linear model is inappropriate - that there is no linear relationship between the variables.

Support Vector Machines (SVM)

SVMs (Lee, Lin *et al.*, 2002; Brown, Grundy *et al.*, 1997; Gaasterland Terry and Bekiranov, 2000) use an initial training set of positive examples (genes that are known to belong to a class) and negative examples (genes of unknown class) to learn to distinguish between members and non-members of the class based on expression data.

As stated previously, gene expression data can be thought of as an m -dimensional space and expression vectors are represented as points in that space. SVMs map this data from m -dimensional space (*input space*) into a higher dimensional *feature space*. Then using a mathematical function called a **kernel function** (Φ), distances are calculated and the training data is separated into two classes by inserting a hyperplane in the feature space. After this, genes from an unknown sample can be identified as a member or non-member of the class based on whether it is on the member or non-member side of the hyperplane. Determination of a proper kernel function so that the data can be separated by a hyperplane is an important step in SVMs.

SVMs are particularly useful for the classification of unknown samples. For example, in the case of tumours, they can be used to classify previously unclassified samples if an initial set of known expression patterns is available. This can be used to develop a '*molecular expression fingerprint*' classification of diseases (Lodish, Berk *et al.*, 2000). This kind of classification is extremely useful in devising novel treatment methods and predicting the prognosis.

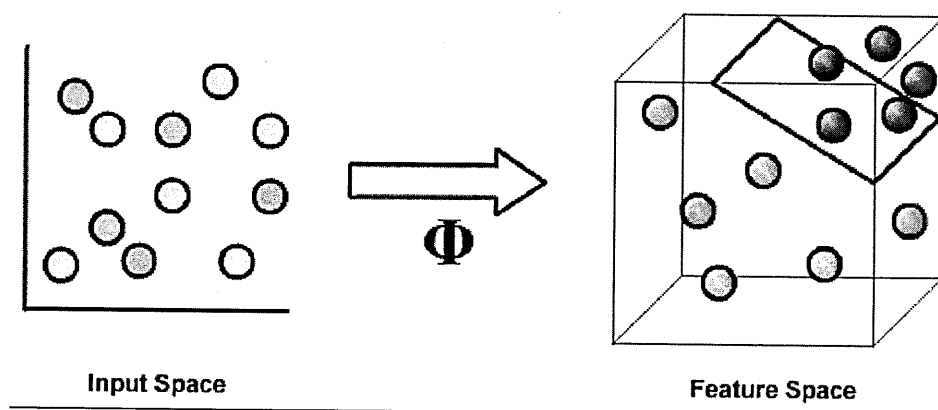


Figure 69: Support vector machine: mapping non-separable data from input space to a higher-dimensional feature space, where a separating hyperplane can be constructed. Reproduced from: Newton J., Analysis of Microarray Gene Expression Data Using Machine

Learning Techniques. Available on the World Wide Web:
<http://www.cs.ualberta.ca/~newton/academic/ml.pdf>

Modelling with differential equation

Differential equation model (Orr and Scherf, 2002; Scherf, Ross *et al.*, 2000) allows to model gene expression data measured at different time, using two different methods. Minimum Weight Solution to Linear Equation (MWSLE) determines the gene regulation, i.e. genes control the production of mRNA and mRNA that codes for protein, as the protein concentration of the cell goes down more proteins are synthesised by feedback control mechanisms that regulate these processes. The second statistical method, Fourier Transform for Stable Systems (FTSS), refines the model with cell cycle constraints. This model is used to check if small set of data is sufficient enough to construct the model at genome level.

Fuzzy logic algorithm

There are three main advantages of applying fuzzy logic algorithm (FLA) to the analysis of gene expression data: it takes into account for the noisy data, because it extracts trends not exact values; Fuzzy logic provides a systematic and unbiased way to perform transformation of quantitative, precise values (descriptors) into heuristic descriptions, such process of transformation is called “fuzzification”. Thereby fuzzy logic removes the need for expert knowledge of the systems; Fuzzy logic algorithm uses day-t-day language used in daily life and hence the results are easier to interpret e.g. gene expression level is either high medium or low; Fuzzy logic technique can be applied to unlimited data for one system. Thus they are able to recognise patterns in large gene expression data. (Woolf and Wang, 2000)

Gene expression data is fuzzified by normalising it from 0 to 1, and then the normalised values are broken up into various membership classes, e.g. HI, MEDIUM and LOW, as shown in Figure 70.

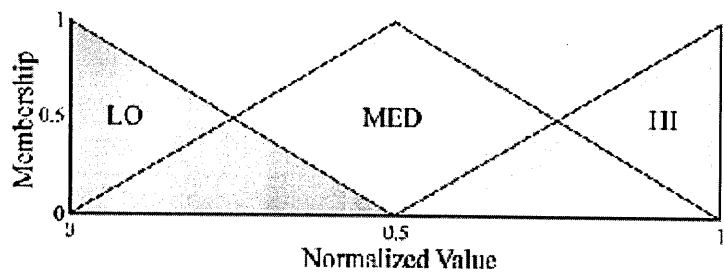


Figure 70: Fuzzy membership as a function of a normalised expression level

Then heuristic rules applied to triplet data sets in the form of decision matrix. The triplets (data set) are defined as the expression level of different protein, i.e. “A” is activator, “B” is repressor control “C”, the target, as shown in Figure 71. Hence, when the concentration of activator is high and repressor is low the target would be high. Conversely, the concentration of target would be expected to be low when repressor is high and concentration of activator is low. (Woolf and Wang, 2000)

		IF “B” IS		
		HI	MED	LO
IF “A” IS	HI	is MED	“C” is HI	“C” is HI
	MED	is LO	“C” is MED	“C” is HI
	LO	is LO	“C” is LO	“C” is MED

Figure 71: Decision matrix describing an activator (A) and repressor (B) acting on a target

The fuzzified values get transformed into numerical value. Hence the predicted expression value for “C” is calculated for each time point. Predicted vales of target “C” are then compared with observed measurements. A residual, r^2 , calculated between calculated and observed “C”, reflecting the goodness of strength between the expected and the predicted “C”.

Smooth Response Surface Algorithm (SRS)

The SRS algorithm (Xu, Wu *et al.*, 2002) is aimed to improve the fuzzy logic approach. Response surface methodology uses the input factors and response model. The underlying mechanism is widely employed by manufacturing and high-tech industries.

The smooth mathematical model describes qualitative biological model as relationships between an activator gene, a repressor gene and target gene. This approach is aimed to extract information from gene expression data. The SRS algorithm uses a 3D response, given by $S(A, B)$ as a graphical representation of high-dimensional decision matrix, i.e. $n \times n$ decision matrix as n tends to infinity, and leads to quantitative gene expression data. In contrast, heuristic rules, in a decision matrix are used in fuzzy logic approach, and stepwise process is used instead of direct plug in process..

Appendix VII

The gene list for hMADS is provided in a CD.