

Reconstructing Signaling Pathways from High Throughput Data

by
Dongxiao Zhu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2006

Doctoral Committee:

Professor Alfred O Hero, Chair
Professor Anand Swaroop
Associate Professor Kerby Shedden
Associate Professor Mark Newman

Copyright ©2006
Dongxiao Zhu, All Rights Reserved

ACKNOWLEDGEMENTS

I would like to acknowledge the significant contributions to this work made by my advisor Professor Alfred O Hero. Professor Hero's work using False Discovery Rate Confidence Interval (FDR-CI) for the gene screening problem provided a direct impact on the works reported here, and his guidance and advice on formulating bioinformatics problems into the statistical inference framework formed the basis of my thesis. His patience and understanding make an enjoyable experience of working with him during the last a few years. I also want to acknowledge his graciousness and flexibility as my research and dissertation advisor in giving me much freedom to pursue my academic goals.

I want to acknowledge the generous support and valuable resources from Professor Swaroop and his lab during my graduate studies. I would not have completed my graduate studies without this. The microarray data from Swaroop lab provides an unparalleled opportunity to test my methodology. Many ideas were directly inspired by his penetrating biological insights in our regular microarray meetings.

I would also thank the other committee members, Professor Shedden Kerby, Professor Mark Newman and Professor Zhaohui Qin, for their invaluable advice, guidance and stimulating discussions. Interactions with them are indispensable for completing my graduate studies. The comments, critics and suggestions given by my committee members during my prelim exam, throughout my graduate studies and on my thesis writing have significantly improved the quality of this thesis.

I am grateful to Mike Rabbat and Robert Nowak from the University of Wisconsin at Madison. Their methodology work on estimating transition matrix from incomplete data forms the theoretical basis for the Chapter V of my thesis. The team work with Mike Rabbat in polishing the Chapter V to a book chapter is truly a valuable experience. I am also indebted to Ritu Khanna and Hong Cheng in Swaroop lab, for their collaborations and friendship throughout the years.

I would like to thank my family - my parents, my grandmother and sister for their unconditional love and support throughout the years especially when I was in dilemma. I would like also to extend my cordial thanks to all my friends and colleagues in the University of Michigan for their friendships. I have learned a great deal from them during my life in Ann Arbor. Finally, I would like to express special thanks to my wife, Jun Zhou, for her unselfish love, support and encouragement during the years, which provide me continuous source of energy to complete my graduate studies.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	vii
LIST OF TABLES	x
LIST OF APPENDICES	xii
CHAPTER	
I. Background and Introduction	1
1.1 High Throughput Bio-Molecule Quantification	1
1.1.1 Microarray Transcription Profiling Platforms	1
1.1.2 Image Analysis	4
1.1.3 Low Level Analysis	4
1.2 Screening Differentially Expressed Genes	5
1.3 Gene Clustering	8
1.4 Problem Statement	9
1.4.1 Previous Approaches and Our Challenges	9
1.4.2 Selection of Network Models	11
1.4.3 Constructing Gene Co-expression Networks	12
1.4.4 Estimating Distance between Non-adjacent Genes in the Network	13
1.4.5 Pathway Order Reconstruction	14
1.5 Contributions	14
1.6 Outline of Thesis	15
1.7 List of Relevant Publications and Software	16
1.7.1 Published Journal Papers	16
1.7.2 Published Conference Papers	16
1.7.3 Manuscripts in Preparation	17
1.7.4 Software	17
II. Co-expression Networks Construction - Frequentist Approach	22
2.1 A Two-Stage Algorithm for Constructing Co-expression Networks	23
2.1.1 Measures of the Strength of Association	23
2.1.2 Hypothesis Testing Scheme	24
2.1.3 Two-stage Screening Procedure	26
2.2 Simulation Studies	28
2.2.1 Validating the Two-stage Algorithm	28
2.2.2 Performance Comparisons	30
2.3 Applications in Network Construction and Seeded Clustering	32
2.3.1 Constructing Relevance Networks with Controlled FDR and MAS	32
2.3.2 Seeded Clustering	36

2.4	Discussion	37
III.	Co-expression Networks Construction - Bayesian Approach	47
3.1	The Bayesian Hierarchical Model	47
3.2	Simulation Studies	52
3.2.1	Comparisons in terms of Confidence Interval, Mean Squared Error, and Variance	52
3.2.2	Posterior Predictive Model Checking	57
3.2.3	Evaluation of the Bayesian Hierarchical Model	57
3.3	Applications to Network Construction and Seeded Clustering	60
3.3.1	Constructing Relevance Networks	60
3.3.2	Seeded Clustering	61
3.4	Discussion	63
IV.	Network Constrained Clustering	67
4.1	Network Constrained Clustering - Method	68
4.1.1	Extract the Giant Connected Component	68
4.1.2	Compute "Network Constrained Distance Matrix"	68
4.2	Network Constrained Clustering - Results	70
4.2.1	Sensitivity Analysis	70
4.2.2	Yeast Galactose Metabolism Data	71
4.2.3	Retinal Gene Expression Data	76
4.3	Software Availability	79
4.4	Discussion	80
V.	de Novo Signaling Pathway Reconstruction	84
5.1	Introduction	84
5.2	Methods	90
5.2.1	Mathematical Formulation of the Problem	90
5.2.2	Estimating a Markov Chain from Direct Observations	91
5.2.3	Estimating a Markov Chain from Shuffled Observations via the EM Algorithm	93
5.2.4	Monte Carlo E-Step by Important Sampling	96
5.2.5	Incorporating Prior Information	97
5.3	Results	98
5.3.1	Protein Kinase A Pathway	98
5.3.2	SAPK/JNK Pathway	100
5.3.3	NF κ B Pathway	102
5.3.4	Assembling Signaling Pathways into Signaling Networks	104
5.4	Software Availability	104
5.5	Discussion	104
VI.	Conclusion, Discussion and Future Works	110
APPENDICES		114

LIST OF FIGURES

Figure

1.1	Schematic of SAGE method. (Source: http://www.sagenet.org/findings/index.html) . . .	18
1.2	Schematic of GeneChip expression profiling method. (Source: http://www.affymetrix.com)	19
1.3	Underlying network models and distance matrices for traditional clustering (a)(b) and network constrained clustering (c)(d). Fig. 1.3c is obtained by removing some edges of weak correlations (long distances), e.g. distance longer than 3. The distance between two genes is a decreasing function of their correlation (see Eq. 4.1). (a). Fully connected network, it assumes any two genes interact with each other directly in the network (connected). (b). Part of the distance matrix for the network model (a). (c). Partially connected network, it assumes only two genes with high correlation (e.g. 0.6) directly interact with each other (connected). Red edges represent the shortest-path from A to D. (d). Part of the distance matrix for the network model (c).	20
1.4	The schematic outline of the thesis.	21
2.1	Two-stage direct screening procedure yields a subset \mathcal{G}_2 of all possible gene pairs \mathcal{G} whose strength of association exceeds MAS level <i>cormin</i> at FDR level α	26
2.2	Inverse screening procedure allows the FDR p -value of a gene pair's (λ_0) strength of association to be computed.	28
2.3	Verification of Gaussian null sampling distribution and variance approximation for Pearson correlation coefficient. (a) QQ plot of transformed sampling distribution of Pearson correlation coefficient $\hat{\rho}$ versus Gaussian distribution. (b) Mean squared approximation errors (MSE) of the variances of transformed sample Pearson correlation coefficients $\hat{\rho}$. . .	40
2.4	Verification of two-stage error control procedure based on Pearson correlation coefficient (a) and Kendall correlation coefficient (b). Sample size $N = 20$	41
2.5	ROC curves of “FDR-CI” test procedure and “FDR-only” test procedure based on Pearson correlation statistic	42
2.6	Curves specify lower endpoints (a) and upper endpoints (b) of the 5% FDR-CI's on the positive Pearson correlation coefficients (a) and negative Pearson correlation coefficients (b) for the galactose metabolism study. Only those gene pairs whose FDR-CI's do not intersect $[-cormin, cormin]$ are selected by the second stage of screening. When the MAS strength of association criterion is $cormin = 0.5$, these gene pairs are obtained by thresholding the curves as indicated.	43
2.7	A pair of non-linearly correlated genes.	44

2.8	Network topology visualization. The network is discovered by constraining $FDR \leq 5\%$ at a MAS level of 0.9. No significant negative correlation is discovered at this level. The graph is drawn using Pajek (Batagelj and Mrvar 1998).	45
2.9	Diagram of the structural module of the galactose metabolic pathway. The shaded boxes denote the five out of six genes whose gene products lie in the galactose metabolic pathway “rediscovered” by our algorithm.	46
3.1	Bayesian hierarchical model structure (Gelman <i>et al.</i> 2004, Chapter V).	49
3.2	Comparison of average posterior CI’s versus average individual frequentist CI’s over a wide range of significance levels at a small sample size ($N = 4$).	53
3.3	Comparison of average posterior CI’s versus average individual frequentist CI’s over a wide range of significance levels at a larger sample size ($N = 20$).	54
3.4	Mean Squared Errors (MSE’s) and Variances of the Bayesian estimations versus the simple estimations over 500 runs of simulations.	55
3.5	Comparison of average CI’s when the Bayesian model is unsustainable.	56
3.6	Posterior predictive distribution, observed results (red line), and p -value for each of four test statistics.	58
3.7	Evaluation of error control of the Bayesian hierarchical model. Sample size $N = 20$, and $\Lambda = 1000$ correlation coefficients were simulated. Simulations using smaller sample size data yield more stringent error control.	59
4.1	Selection of p using RAND indices.	71
4.2	Traditional clustering: agglomerative hierarchical clustering using all 997 differentially expressed genes. The 14 structural genes are separated into three clusters (red rectangular).	74
4.3	Traditional clustering: agglomerative hierarchical clustering using the GCC selected 772 genes. The 14 structural genes are separated into three clusters (red rectangular). Dots indicate incomplete clusters are shown due to space limitation.	75
4.4	Correlation matrix of 14 structural genes with clustering dendrogram. White to grey corresponds to the low correlations to high correlations.	76
4.5	Network constrained clustering: agglomerative hierarchical clustering using network constrained distance matrix calculated from relevance network (Eq. 4.2).	77
4.6	Clustering comparison - GO vocabulary “visual perception” counts.	79
4.7	Clustering comparison - GO vocabulary “visual perception” separation.	80
4.8	Clustering comparison - GO vocabulary “visual perception” p -values.	81
4.9	A significant update to the open source clustering software (joint work with Ritu Khanna).	82

5.1	The schematic representation of the signaling pathway reconstruction algorithm. The starting pathway component is in red (left), and the ending pathway component is in blue (right). Pathway components in the parenthesis are intermediate and unordered. The solid lines represent the inputs to the algorithm (different sources of pathway information). The dotted lines represent the outputs from the algorithm (the maximum likelihood pathway(s)).	90
5.2	The (unordered) protein kinase A signaling pathway. Membrane receptors are in red (left), and transcription factors are in blue (right). Activation or inhibition information between pathway components are omitted. The pathway is mainly adapted from Van Driessche et al (Van Driessche <i>et al.</i> 2005).	99
5.3	The reconstructed protein kinase A signaling network topology from unordered pathway composition data (Fig. 5.2).	100
5.4	The (unordered) SAPK/JNK signaling pathway. Membrane receptors are in red (left), and transcription factors are in blue (right). Activation or inhibition information between pathway components are omitted. “GF” stands for Growth Factor, “CS” stands for Cellular Stress, “FASL” stands for Fas Ligand”, “OS” stands for Oxidation Stress. The pathway is adapted from http://www.cellsignal.com/ .	102
5.5	Upper panel: The correct SAPK/JNK signaling network topology defined by the probability transition matrix Eq. 5.32 estimated from unordered pathway composition data (Fig. 5.4) improved by incorporating a prior information on gene-gene interactions, in particular the interactions between the two double-circled components. Lower panel: The incorrect SAPK/JNK signaling network topology defined by the probability transition matrix Eq. 5.33 estimated from unordered pathway composition data without incorporating prior information.	106
5.6	The (unordered) NF κ B signaling pathways. Membrane receptors are in red (left), and transcription factors are in blue (right). Activation or inhibition information between pathway components are omitted. “Ag” stands for Antigen, “Ag-MHC” stands for Major Histocompatibility Complex (MHC) Antigen, “IL-1” stands for Interleukemia-1, “dsRNA” stands for double stranded RNA, TNF stands for Tumor Necrosis Factor, “GF” stands for Growth Factor, “LT” stands for heat-labile enterotoxin. “NF κ BC1” and “NF κ BC2” stand for NF κ B complexes 1 and 2. The pathway is adapted from http://www.cellsignal.com/ .	107
5.7	The two possible NF κ B signaling network topologies defined by the probability transition matrix Eq.A.16 and Eq.A.17 estimated from unordered pathway composition data (Fig. 5.6) after incorporating prior information. The relationships between the two double-circled components are disambiguated from prior information. The epistasis relationship labeled with “?” remains ambiguous and deserves further investigation.	108
5.8	The signaling networks assembled from SNK/JNK and NF κ B pathways.	109

LIST OF TABLES

Table

2.1	Performance comparison for three algorithms based on Pearson correlation coefficient for selecting gene pairs with a MAS level of 0.5. Thresholded MAS and thresholded FDR are significantly worse in terms of statistical significance (p -value) than the proposed two-stage FDR-CI algorithm (columns 4 and 5). Furthermore, the average length of the CI's on ρ 's of the discovered gene pairs are shorter for the two-stage FDR-CI algorithm than for the other algorithms (column 6).	31
2.2	Performance comparison for three algorithms based on Kendall's τ statistic for selecting gene pairs with a MAS level of 0.5. Thresholded MAS and thresholded FDR are significantly worse in terms of statistical significance (p -value) than the proposed two-stage FDR-CI algorithm (columns 4 and 5). Furthermore, the average length of the CI's on τ 's of the discovered gene pairs are shorter for the two-stage FDR-CI algorithm than for the other algorithms (column 6).	31
2.3	Top twenty "hub genes" from the two-stage algorithm applied to galactose metabolism data (Ideker <i>et al.</i> 2000). The rank of each gene is the average rank over five different networks. Each of five networks is constrained by a different pair of (FDR,MAS) criteria. The highest ranked gene is the most connected and stable gene under varying constraints of (FDR,MAS).	37
3.1	Top twenty "hub genes" from Bayesian hierarchical model applied to the galactose metabolism data (Ideker <i>et al.</i> 2000). The rank of each gene is the average rank over five different networks with the same set of edge numbers as in Table 2.3. The highest ranked gene is the most connected and stable gene under varying constraints of (FP,MAS).	62
3.2	Comparison of Bayesian estimations versus Marginal estimations using "seeded" clustering at a small and a larger sample sizes. In the former, the ranks were averaged over 100 estimations, in each of which a subset data of sample size $N = 4$ was randomly sampled from the whole data of sample size $N = 20$. In the later, the ranks were obtained using the whole data of sample size $N = 20$	64
A.1	Sample output of screening co-expressed gene pairs based on Kendall correlation coefficient. It was described in section 2.3.1.	122
A.2	Clustering co-expressed genes with controlled FDR (5%) at a MAS level of 0.6 using "GAL10" as the "seed gene". Known genes in the pathway are in bold face. Pearson correlation coefficient was used as metric. It was described in section 2.3.2.	122
A.3	Clustering co-expressed genes with controlled FDR (5%) at a MAS level of 0.5 using "GAL7" as the "seed gene". Known genes in the pathway are in bold face. (a) Pearson correlation coefficient as metric. It was described in section 2.3.2.	122

A.4	Clustering co-expressed genes with controlled FDR (5%) at a MAS level of 0.5 using “GAL7” as the “seed gene”. Known genes in the pathway are in bold face. (b) Kendall correlation coefficient as metric. It was described in section 2.3.2.	123
A.5	Clustering co-expressed genes with controlled FDR (5%) at a MAS level of 0.5 using “GAL1” as the “seed gene”. Known genes in the pathway are in bold face. Pearson correlation coefficient as metric. It was described in section 2.3.2.	123
A.6	Clustering co-expressed genes with Bayesian hierarchical model at the significance level 5% using “GAL10” as the “seed gene”. Known genes in the pathway are in bold face ($N = 20$). It was described in section 3.3.2.	123

LIST OF APPENDICES

Appendix

A.	Technical Details and Supplemental Tables	115
A.1	Construct PCER-CI for ρ	115
A.2	Construct PCER-CI for τ	115
A.3	Simulating Bivariate Data Based on Pre-specified Population Covariances . .	116
A.4	Selecting Prior Distribution	117
A.5	Deriving Posterior Distribution $p(\beta y)$	118
A.6	Two Equally Likely Probability Transition Matrices for NF κ B Pathway . . .	120
B.	*	124

CHAPTER I

Background and Introduction

1.1 High Throughput Bio-Molecule Quantification

1.1.1 Microarray Transcription Profiling Platforms

The complete genome sequence of human and other species provides a new starting point for understanding our basic genetic makeup and how variations in genetic instructions result in human disease or other individual variations. The biological research in post-genomic era has shifted from the traditional single component analysis that focuses on a single gene or protein to the simultaneous analysis of thousands of biomolecules analyzed/identified by high throughput quantification techniques, e.g. gene expression microarrays (Lockhart *et al.* 1996, DeRisi *et al.* 1997) .

The abundance levels of biomolecules are tightly regulated to ensure the proper functions of the biological system. Abnormal variations at each level can correlate with many diseases, e.g. genomic DNA copy number changes are hallmarks of cancer (LaFramboise *et al.* 2005, Zhao *et al.* 2004). Simultaneous quantification of the abundance levels of these biomolecules on the genomic scale and follow-up data analyses provide a potential source of profound knowledge. Some of the more successful applications are: cancer classification and prediction using gene expression arrays (Alizadeh *et al.* 2000, Golub *et al.* 1999), discovery of differentially expressed genes, functionally related genes, and gene regulation networks using expression ar-

rays (Tusher *et al.* 2001, Butte and Kohane 2000, Zhu *et al.* 2005a), detecting transcription factor binding sites using CHIP-on-chip technology (Harbison *et al.* 2004, Lee *et al.* 2002), high throughput genotyping and DNA quantification using Single Nucleotide Polymorphism (SNP) arrays (Kennedy *et al.* 2003), discovery of disease bio-markers using Liquid Chromatography (LC) coupled with Mass Spectrum (MS) for protein and lipid quantification (Patterson and Aebersold 2003, Goodacre *et al.* 2004).

Recent development of high throughput bio-molecule quantification techniques makes data acquisition less of a challenge. The primary challenge lies in the analytical side imposed by noisy nature of the data and so-called “small N , large p ” paradigm, which includes thousands of variables (bio-molecules, denoted as p) with only a few of observations (denoted as N). High throughput data analysis has raised a number of statistical and computational questions in diverse traditional areas, such as image processing, generalized linear models, linear mixed effect models, discriminative analysis, machine learning, multiple testing and Bayesian statistics (Lee 2004, Zarepari *et al.* 2004). We use gene expression array data throughout this thesis to illustrate our data analysis schemes. However the proposed techniques can also be applied to data acquired through other platforms.

There are two types of microarray gene expression profiling techniques, i.e. sequencing-based (Fig. 1.1) and hybridization-based (Fig. 1.2)(Lee 2004). For the sequencing-based techniques, the strategy is to attach a double-stranded DNA tag to each copy of cDNA, and the number of tags of each cDNA read from sequencing correspond to its abundance. The representative example is Serial Analysis of Gene Expression (SAGE)(Velculescu *et al.* 1995). For the hybridization-based techniques, the strategy is to immobilize a large number of DNA clones (probes) with known sequences

on the solid support. The pool of examined RNA (targets) is then labeled with fluorescence tags and hybridized to the probes. There are three major hybridization-based microarray technology platforms, namely, spotted cDNA array (DeRisi *et al.* 1997), spotted oligonucleotide array and *in situ* oligonucleotide arrays (Affymetrix GeneChip, Lockhart *et al.* 1996). These three technology platforms quantify targets based on fluorescence signal intensity. The first two techniques are similar except that the former exploits cDNAs as a probe and the latter exploits synthetic oligonucleotides as a probe. These two techniques differ significantly from the third in many aspects such as the hybridization method and the chip design:

- For spotted cDNA arrays, the reference sample and the treated sample, labeled with different dyes, e.g. cy3 (green) and cy5 (red), are competitively hybridized to the same chip, while in Affymetrix GeneChip arrays, the reference and treated samples are hybridized to two different chips.
- For the spotted cDNA arrays, one gene is represented by a long probe, while for Affymetrix GeneChip, each gene is typically represented by 11-20 pairs of shorter oligonucleotide probes. The first component of these pairs is referred to as a Perfect Match (PM) probe and is designed to hybridize only with transcripts from the intended gene (specific hybridization). Nevertheless, hybridization to other sequences (non-specific hybridization) is unavoidable. The second component of these pairs is referred to as a Mismatch (MM) probe and is designed to measure the noise introduced by non-specific hybridization. Recent studies tend to use PM probe intensities only, since MM probe intensities also measure specific hybridization and sometime are larger than PM intensities (Irizarry *et al.* 2003).

- Compared to the spotted arrays, Affymetrix GeneChips enjoy the feature of high density, on which the whole human genome transcription can be profiled.

1.1.2 Image Analysis

Following hybridization of a microarray and the readout of gene expression levels, the data is stored as 16-bit images. Image analysis is the first important step, and the accuracy of extracted intensities can have a large impact on subsequent data analysis. For two-color cDNA array images, the processing of scanned microarray images can be separated into three tasks (Yang *et al.* 2002):

- Addressing. Estimate location of spots centers.
- Segmentation. Classify pixels as foreground (signal) or background (noise).
- Information extraction. This step includes calculating, for each spot on the array, red and green foreground fluorescent intensity pairs (R, G), background intensities, and possibly, quality measures.

Affymetrix GeneChip image processing follows a similar procedure but only to Addressing (Step I) and Information extraction (Step III).

1.1.3 Low Level Analysis

The raw intensity data output from image scanner is usually subjected to a series of pre-processing analysis, e.g. background correction, normalization, summarization (Irizarry *et al.* 2003, Yang *et al.* 2002). The background correction is to minimize non-specific hybridization noise. The default adjustment, provided as part of the Affymetrix system, is based on the difference between PM and MM probe intensities (MAS4) or its robust estimation (MAS5). This approach can be improved via the use of estimators derived from a statistical model that incorporates probe sequence

information (Wu *et al.* 2004, Zhang *et al.* 2003). The normalization is to adjust microarray data for effects which arise from variation in the technology rather than from biological differences between the RNA samples or between the printed probes. The ultimate goal of normalization is to minimize the technical (systematic) variation. Popular normalization methods include quantile normalization (Bolstad *et al.* 2003) and invariant-set normalization (Tseng *et al.* 2001).

Summarization as used in Affymetrix GeneChip, provides an estimate of mRNA abundance, called a score, from a number of Perfect Match (PM)/Miss Match (MM) intensities. Popular summarization methods include: Robust Multichip Average (RMA)(Irizarry *et al.* 2003), Li-Wong's model (Li *et al.* 2001), gcRMA (Wu *et al.* 2005), and trimmed mean (Rickman *et al.* 2001). There is still no single approach that outperforms the others in all reported test cases (Bolstad *et al.* 2003, Shedden *et al.* 2005), but RMA seems to be the most popular for general purposes, while gcRMA has been reported to be more sensitive for low abundance probe intensities.

After these steps, the gene expression data is usually available in the format of data matrix, in which rows correspond to gene names, and columns correspond to relevant physiological/genetic conditions under which the gene expression levels are quantified.

1.2 Screening Differentially Expressed Genes

Initial efforts to gain biological insight from gene microarray data focused on ranking genes according to some ranking statistic, followed by examination of a handful of genes on the top or bottom of the ranked list by biological experts. Identifying truly differentially expressed genes in microarray studies is a major statistical challenge that has received much attention. Existing approaches can be roughly divided

into three categories, univariate approaches (Wolfinger *et al.* 2001), multivariate approaches (Efron *et al.* 2001, Tusher *et al.* 2001), and multicriterion approaches (Hero and Fleury 2004, Hero *et al.* 2004, Fleury *et al.* 2002).

Many univariate approaches are based on the classical Analysis of Variance (ANOVA) model and its extensions such as linear mixed effect models. These type of approaches fit models for each gene expression one-at-a-time. In the standard application of ANOVA, one would proceed to test whether the sequential sum of squares for experimental condition is large enough to reject the null hypothesis. The standard model assumptions require the error term ϵ to be independently and identically distributed (i.i.d.) Gaussian random variables. Other approaches use permutation test to avoid assumptions about the error distribution. The ANOVA model assumes that the main effects and interaction effects of the model are fixed, not random. In some studies, however, it may be quite reasonable to treat some of these effects as random, and more specifically, as Gaussian distributed. For example, sometimes array effects in microarray experiments may be modeled as random effects and mixed model inference can outperform standard ANOVA (Zhu and Hero 2005b).

Univariate approaches allow flexible and powerful modeling choices, however, they ignore the underlying covariance structure of gene regulation networks (termed “network constraint” in the later chapters). Multivariate approaches study a set of genes in one model, such as Empirical Bayes (EB) (Efron *et al.* 2001) and Significant Analysis of Microarray (SAM) (Tusher *et al.* 2001), in which the ordinary t-test statistic is modified by adding a fudge factor (estimated from a large set of genes) to the denominator. The fudge factor is estimated in different ways to avoid the situation where tiny variances can create large t -statistics even for a very small fold change.

Both univariate and multivariate statistics have been successfully applied to screen-

ing differentially expressed genes. However, no single statistic is universally optimal and there is seldom any basis or guidance for selecting a particular statistic. To circumvent this difficulty, Hero and Fleury, 2004, (Hero and Fleury 2004) described a novel gene screening approach in which they ranked genes using a multi-dimension plot, called *multicriterion scattergram*. Genes that are maximal in the component-wise ordering in the P -dimensional scatterplot corresponding to P screening statistics are defined as *Pareto fronts*, on which the genes were selected as the most differentially expressed. The multicriterion optimization method had also been generalized to rank genes based on a set of pre-defined criteria, not limited to differential expression, such as strong monotonic increase, high end-to-end slope and low slope deviation (Fleury *et al.* 2002, Speed 2003). More recently, Yang et al proposed a distance synthesis scheme to integrate multiple statistics for screening differentially expressed genes. Using the Affymetrix spike-in data in which the magnitude of differential expressions were known, they reported that the integrated statistic compares favorably with the best individual statistics, while achieving robustness properties lacked by the individual statistics (Yang *et al.* 2005).

The practice of screening differentially expressed genes plays a key role in understanding underlying biological mechanisms, and discovering disease bio-markers. However, it has a number of major limitations:

- The process is subjective and often requires considerable biological expertise.
- The biological findings are often discrete and sporadic.
- Single gene analysis only reveals the most differentially expressed genes in the pathway while missing out many less differentially expressed genes with concordant changes.

- There is poor overlap among different approaches, e.g. less than half of genes are in common between the top 1000 differentially genes declared by ordinary t-test and SAM t-test.

1.3 Gene Clustering

To overcome these limitations, the recent efforts have been focused on studying a set of functionally related genes (so-called signaling pathway). We define signaling pathway as a series of gene interactions leading to a specific biological endpoint function. The interactions among genes can be interpreted as co-regulation or chemical modification, e.g. phosphorylation, acetylation, and methylation. Therefore, gene interactions are often inferred through calculating the correlation between gene expression profiles over multiple relevant physiological/genetical conditions. Gene pairs with high correlation (e.g. greater than 0.6) are hypothesized to be biologically relevant and to interact directly in the signaling pathways.

It is typical that only a few genes are experimentally confirmed to be in a signaling pathway. Gene clustering is a widely used approach that attempts to group all the genes in the pathway into a cluster such that functional prediction of unknown genes can be made based on the functionally known genes. Some of the more popular clustering methods include: hierarchical clustering (Eisen *et al.* 1998), K -means type clustering (Hartigan and Wong 1979), mixture model-based clustering (Yeung *et al.* 2001) and Hidden Markov Model (HMM) based clustering (Schliep *et al.* 2003). The hierarchical clustering and K -means type clustering are heuristic approaches with major distinction that the former belongs to unsupervised learning while the latter belongs to supervised learning. These methods have been successful in inferring many signaling pathways from gene microarray data. A new set of methods, called

Gene Set Enrichment Analysis (GSEA) have been recently developed to address the statistical significance of a given gene set (Subramanian *et al.* 2005, Mootha *et al.* 2003, Kim *et al.* 2005). Instead of analyzing a few differentially expressed genes, the method inspects all the genes on the chips. Since it requires a pre-defined gene set that is assumed to be functionally related, the method has not been very effective in reconstructing gene pathways in unsupervised manner.

1.4 Problem Statement

We divide the signaling pathway reconstruction problem into two sub-problems: discovery of pathway components and ordering the pathway components. Briefly, we solve the first sub-problem using an innovative network constrained clustering approach (Zhu *et al.* 2005c, Zhu and Hero 2005d, Zhu and Hero 2005e), and we solve the second sub-problem applying a first-order Markov model approach (Rabbat *et al.* 2006). We first describe the previous approaches and their shortcomings.

1.4.1 Previous Approaches and Our Challenges

The ultimate goal of all gene clustering approaches is to group genes with similar functions into one single cluster. These functional related genes are likely to be in the same signaling pathway. In practice, most approaches simply group genes with similar expression profiles (Eisen *et al.* 1998, Stuart *et al.* 2003, Lee *et al.* 2004), denoted as “traditional clustering” throughout this thesis. However, many genes in the same functional pathway may not have similar expression profiles as measured by correlation statistics or other pairwise expression similarity measure. This is especially true for pairs of genes that are not in the same region of a signaling pathway. These genes will not be discoverable using the traditional clustering methods. Thus, a well-known limitation of the traditional clustering approaches is that it only groups

functional related genes with similar expression profiles, but misses out many others with dissimilar expression profiles.

In a gene regulation network, graph vertices represent genes, and edges represent biological relationships between genes, such as co-regulation, chemical modification. Different network models are able to infer many kinds of relationships among genes (Butte and Kohane 2000, Butte *et al.* 2000, Zhou *et al.* 2002, Friedman *et al.* 2000, Perrin *et al.* 2003, Yu *et al.* 2004, Rao *et al.* 2005). The traditional methods of clustering assume that the underlying network is fully connected, i.e. any biological function is executed through a direct interaction (relationship) between a pair of genes (Fig. 1.3). Direct pairwise gene interactions, represented by the fully connected subgraph (clique), only describes a small subset of gene interactions. In many cases, an endpoint biological function is more commonly executed through a series of interconnected gene interactions (see Fig. 1.3c, gene A, B, C, D, E, F). Consequently, for genes lying in a single pathway traditional clustering approaches often group these genes into several different clusters, e.g., each cluster determined by a similarly co-expressed clique. This breaking of a pathway across several clusters makes it more difficult for biologists to identify groups of genes having common function. Thus approaches that are able to go beyond pairwise interactions to group the whole pathway into a single tight cluster are highly desirable.

A more realistic assumption for gene clustering may be that the underlying gene regulation network is only partially connected, i.e. biological function is executed through either direct interaction or through indirect interaction via one or more intermediate genes (Fig. 1.3). A gene clustering algorithm that accounts for such realistic network constraints is likely to be more powerful (Zhou *et al.* 2002, Zhou and Gibson 2004). There are several challenges to developing such an approach. What

kind of network model is most appropriate? How to reliably extract the relevance network from noisy high throughput data? How to estimate the distance between two non-adjacent genes (genes that do not have similar expression profiles) in the network?

1.4.2 Selection of Network Models

Different network models have been applied to high throughput data to gain insight into regulatory function. Popular network models include the Boolean network (Liang *et al.* 1998, Szallasi *et al.* 1998, Wuensche 1998), the Bayesian network (Friedman *et al.* 2000), the Relevance network (Butte and Kohane 2000, Basso *et al.* 2005, Fuente *et al.* 2004, Magwene and Kim 2004) and the Dynamic network (Rao *et al.* 2005, Perrin *et al.* 2003, Yu *et al.* 2004). The advantages and disadvantages of each method are becoming increasingly clear. Boolean networks feature conceptual and computational simplicity but have the problem of choice of threshold for the one bit quantification of the expression scores. Bayesian networks and Dynamic networks enable one to draw causal inference and/or to infer time varying network topologies, but currently are only practical for reconstructing very small-size networks due to the high computational complexity in the dimension of the topology search space. Gene co-expression networks such as relevance and dependency networks, provide satisfactory approximations to large-scale networks, however they do not allow for causality and directionality of interactions. We choose to use the gene co-expression network model for network based signaling pathway discovery for the following reasons. First, the large scale network reconstructions can be used as a filter to select a small number of significant interactions, regardless of directionality, prior to implementing a Bayes network reconstruction. Second, it is too ambitious to reconstruct signaling pathways in great detail from small numbers of replicates of the expression

data alone. Third, co-expression network models enable computationally tractable large-scale network construction with error control (false positives and false negatives) (Zhu *et al.* 2005a, Zhu and Hero 2005f).

1.4.3 Constructing Gene Co-expression Networks

Gene co-expression networks typically use correlation statistics as pairwise similarity measures (a decreasing function of the distance for clustering) between gene expression profiles, followed by either direct correlation thresholding (Zhou *et al.* 2002) or a combination of significance level tests with correlation thresholding (Lee *et al.* 2004). While direct thresholding is useful in many cases it only controls biological significance but not error rate. Combining correlation thresholding with a level of significance test does not allow one to control biological and statistical significance in a systematic and reliable manner.

In estimating pairwise gene correlation from high throughput data, the estimates are subject to high variances due to the noisy nature of the data and “small N , large p paradigm” (Dobra *et al.* 2004, Schafer and Strimmer 2005a, Schafer and Strimmer 2005b). One way to account for these variances is to construct confidence interval(CI) for each correlation parameter, and threshold on the upper/lower bounds rather on sample estimate directly (Hero *et al.* 2004). Both frequentist statistics and Bayesian statistics provide constructions of CI’s. In frequentist approaches, the confidence interval on a scalar parameter is constructed by finding a “pivot” whose distribution is independent of the parameter. In Bayesian approaches, the confidence interval of a ‘parameter’ is constructed from its posterior distribution. We present a pair of complementary network construction approaches for the small sample problem and the larger sample problem using Bayesian and frequentist confidence interval thresholding.

For the relatively large sample problem (e.g. $N = 20$), we propose an approach based on application of False Discovery Rate Confidence Intervals (FDR-CI) recently proposed to control biological and statistical significance simultaneously (Hero *et al.* 2004, Zhu *et al.* 2005a). This approach is able to identify both linearly and non-linearly co-expressed genes using the Pearson correlation coefficient and the Kendall correlation coefficient. The employment of Kendall’s correlation is important when functionally related gene expression profiles may be non-linearly correlated. Non-linear correlation can occur, for example, when gene expressions of different subunits of a whole enzyme are differentially regulated due to different enzyme efficiencies (Berg *et al.* 2006).

For the relatively small sample problem (e.g. $N < 10$), the frequentist approach may suffer from “overfitting” and low discriminating power (Ledoit and Wolf 2004). To solve this problem, we propose a Bayesian hierarchical model approach that is able to globally estimate the correlation parameters (Zhu and Hero 2005g).

1.4.4 Estimating Distance between Non-adjacent Genes in the Network

Assume a reasonably well constructed co-expression network is available. The shortest-path distance between two non-adjacent genes represents a natural and parsimonious representation of biological interaction since genes along the shortest-path are likely to have similar function (Zhou *et al.* 2002). Based on our network reconstruction algorithms and our shortest-path distance measure, we present a new clustering approach, called “network constrained (NC) clustering”. NC clustering is able to group more functionally related genes into a single tight cluster even if their expression profiles are dissimilar.

1.4.5 Pathway Order Reconstruction

While the network constrained clustering method may be able to group the whole pathway into a single tight cluster, it does not directly address the ordering of genes in the pathway. This raises the following sub-problem. Assuming the pathway components are known, can we optimally infer the order of the genes in the pathway? We propose solution to this problem by applying a first-order Markov model based approach that was originally developed and applied to a network tomography problem in telecommunication networks (Rabbat *et al.* 2006). The key advantage of this model is that it takes full advantage of pathway composition information and network constraints, allowing for a high level data integration and knowledge extraction.

1.5 Contributions

In this thesis, we present a unified framework for reconstructing signaling pathways from high throughput data by accounting for underlying network constraints. The following lists the principal contributions of this thesis and the chapters in which they are covered.

- Full frequentist statistical treatment of the co-expression network construction problem. We hypothesize that only highly correlated gene pairs are biologically relevant, and we extract an estimate of the network by combining correlation thresholding and control of statistical significance. (Chapter II)
- One of the first full Bayesian treatments of the co-expression network construction problem. This approach provides a natural and seamless combination of correlation strength thresholding, and variance regularization to small sample size. (Chapter III)

- Using both linear and nonlinear correlation statistics in screening network edges. Most of previous approaches use only a linear correlation statistic. Nonlinear correlations are commonly seen in many biological scenarios. (Chapter II)
- The first network constrained clustering approach that is able to group functional related genes according to shortest path expression similarity. Other clustering approaches do not exploit the underlying network structure. (Chapter IV)
- A more objective way to evaluate clustering performance using Gene Ontology. Instead of comparing clustering performance at one specified cluster number, we compared it over a wide range of cluster numbers. (Chapter IV)
- Application of a novel model based pathway ordering algorithm to reconstruct the ordering of pathway components from multiple data sources. (Chapter V)

1.6 Outline of Thesis

The goal of this thesis is to explore a long-standing problem in high throughput data analysis, i.e. signaling pathway reconstruction. The problem is addressed in a multi-stage process (Fig. 1.4). Chapter II introduces a full frequentist treatment of the co-expression network construction problem for the reasonably large sample data. Complementary to Chapter II, Chapter III introduces a full Bayesian treatment of the co-expression network construction problem for problems having relatively small sample size data. Based on the networks constructed in either Chapter II or Chapter III, Chapter IV introduces the network constrained clustering approach. Chapter V demonstrates the application of the first-order Markov model based method to the pathway order reconstruction problem. Finally, in Chapter VI, we conclude with a discussion of important issues and future work.

1.7 List of Relevant Publications and Software

1.7.1 Published Journal Papers

Zhu, D., Hero, A.O., Cheng, H., Khanna, R. and Swaroop, A. 2005. Network constrained clustering for gene microarray data. *Bioinformatics*, **21**(21), 4014-4021.

Zhu, D., Hero, A.O., Qin, Z.S. and Swaroop, A. 2005. High throughput screening of co-expressed gene pairs with controlled False Discovery Rate (FDR) and Minimum Acceptable Strength (MAS). *J. Comput. Biol.*, **12**, 1029-1045.

Zhu, D. and Qin, Z.S. 2005. Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics*, **6**:8.

Akimoto, M., Cheng, H., Zhu, D. et al. 2006. Targeting of green fluorescent protein to new-born rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors. *Proc. Natl. Acad. Sci. USA*, **103**(10), 3890-3895.

1.7.2 Published Conference Papers

Zhu, D. and Hero, A.O. 2005. Gene co-expression network discovery with controlled statistical and biological significance. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, USA, March 2005. (Finalist for Best Student Paper Award).

Zhu, D. and Hero, A.O. 2005. Network constraint clustering for gene microarray data. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, USA, March 2005.

Zhu, D., Hero, A.O. and Swaroop, A. 2005. An unsupervised posterior analysis of signaling pathways from gene microarray data. *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS'05)*, New Port, Rhode Island, USA, May 2005.

Zhu, D. and Hero, A.O. 2005. Identifying differentially expressed genes from probe level intensities in longitudinal Affymetrix microarray experiments. *IEEE International Workshop on Statistical Signal Processing (SSP'05)*, Bordeaux, France, July 2005.

Zhu, D. and Hero, A.O. 2005. Bayesian hierarchical model for estimating gene association networks from microarray data. *IEEE International Workshop on Genomic signal processing and statistics (GENSIPS'06)*, College Station, Texas, USA, May 2006.

Rao, A., Hero, A.O., Engel, J.D., States, D.J. and Zhu, D. 2005. Inferring Time-varying Network Topologies from Gene Expression Data. *IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS'05)*, Newport, May 2005.

1.7.3 Manuscripts in Preparation

Zhu, D., Rabbat, M.G., Hero, A.O., Nowak, R.D. and Figueiredo, M.A.T. De Novo Signaling Pathway Reconstruction From Multiple Data Sources. to appear in a chapter of the book “New Research on Signal Transduction”, Nova Science Publishers, Hauppauge, NY.

Zhu, D. and Hero, A.O. Bayesian hierarchical model for estimating gene association networks from microarray data. In preparation.

Zhu, D., Li, Y., and Hero, A.O. Estimating gene expression correlation from replicated microarray data - A multivariate approach. In preparation.

1.7.4 Software

- R package GeneNT [available from, <http://cran.r-project.org/>]
- Gene clustering software with Graphic User Interface (GUI)[available from, http://www-personal.umich.edu/~zhud/cluster_31.htm]

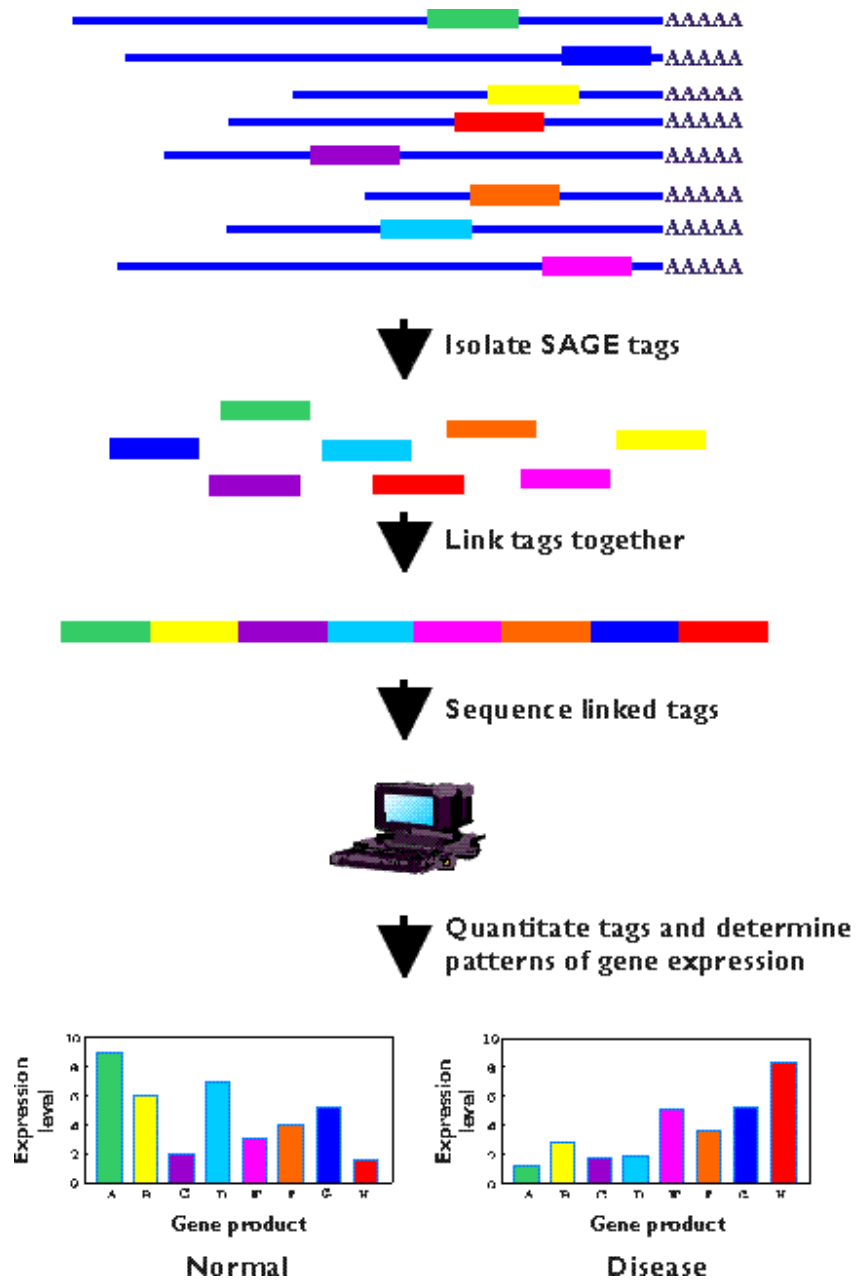


Figure 1.1: Schematic of SAGE method. (Source: <http://www.sagenet.org/findings/index.html>)

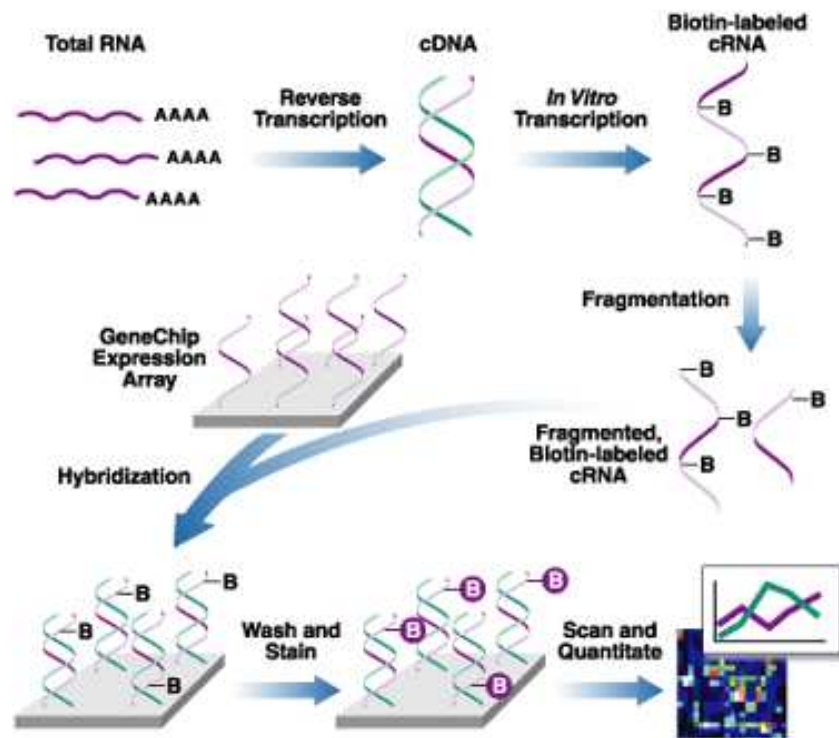


Figure 1.2: Schematic of GeneChip expression profiling method. (Source: <http://www.affymetrix.com>)

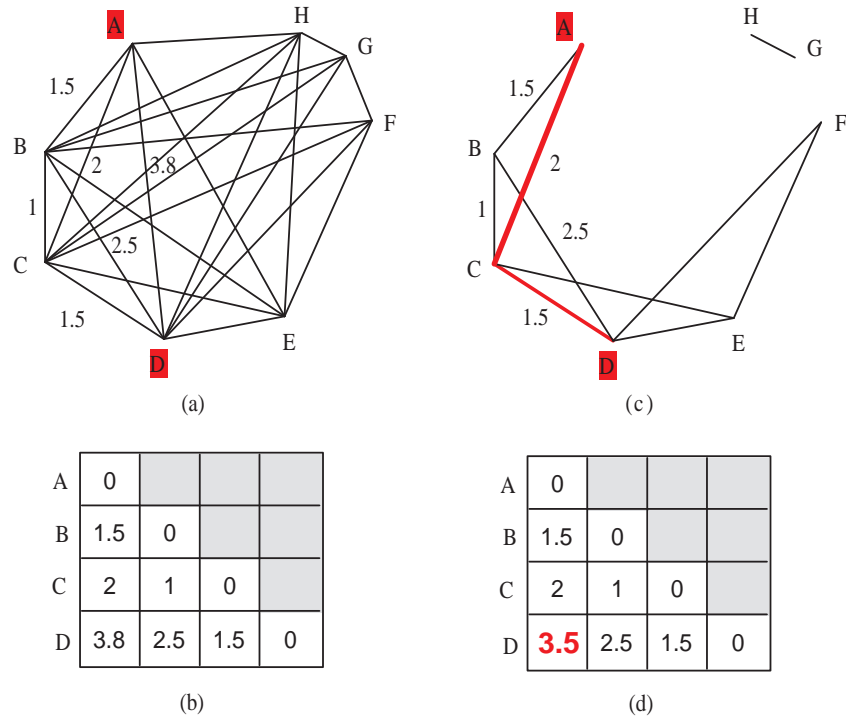


Figure 1.3: Underlying network models and distance matrices for traditional clustering (a)(b) and network constrained clustering (c)(d). Fig. 1.3c is obtained by removing some edges of weak correlations (long distances), e.g. distance longer than 3. The distance between two genes is a decreasing function of their correlation (see Eq. 4.1). (a). Fully connected network, it assumes any two genes interact with each other directly in the network (connected). (b). Part of the distance matrix for the network model (a). (c). Partially connected network, it assumes only two genes with high correlation (e.g. 0.6) directly interact with each other (connected). Red edges represent the shortest-path from A to D. (d). Part of the distance matrix for the network model (c).

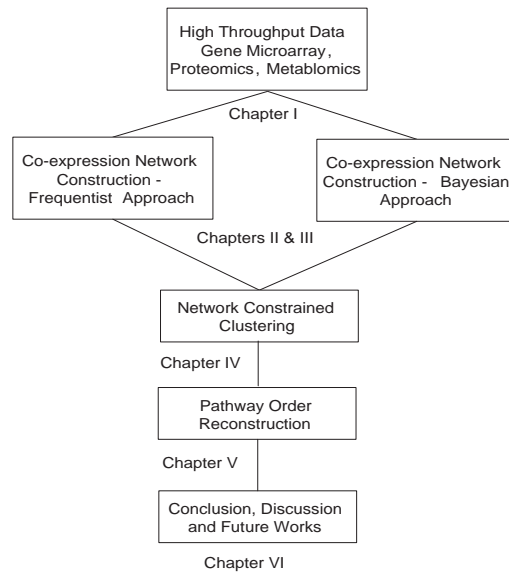


Figure 1.4: The schematic outline of the thesis.

CHAPTER II

Co-expression Networks Construction - Frequentist Approach

Gene co-expression networks provide a good approximation to the complicated web of gene functional associations (Butte and Kohane 2000, Butte *et al.* 2000). In constructing the gene co-expression networks, a most frequently assumed hypothesis is that the gene pairs of high expression correlation indicate functional relevancy (Eisen *et al.* 1998). Thus, the gene co-expression network can be viewed as a cluster of functional modules, in which the pairwise gene expression correlations above the threshold or below the threshold correspond to the presence or absence of co-expression network edges. In order to construct the co-expression networks, statistically, we need to simultaneously draw inference on a large number of correlation parameters.

A full frequentist statistical treatment of the co-expression network construction problem can be translated into either rejecting null hypotheses or accepting null hypotheses at a certain significance level, in which the former declares the presence of network edges, and the later declares the absence of network edges. Similarly, a full Bayesian statistical treatment of the co-expression network construction problem can be translated into simultaneously thresholding the posterior distributions of many correlation ‘parameters’.

2.1 A Two-Stage Algorithm for Constructing Co-expression Networks

2.1.1 Measures of the Strength of Association

There are many possible discriminants for strength of association between two variables, which we generally denote as a real number Γ . Under a Gaussian linear hypothesis, the Pearson correlation coefficient ρ is an appropriate metric. A robust distribution-free alternative is the Kendall rank correlation coefficient (Kendall's τ). The Pearson (Bickel and Doksum 2000) and Kendall correlation coefficients (Hollander and Wolfe 1999) are special cases of the generalized correlation coefficient (Daniel 1944). We define $\{g_p\}_{p=1}^G$ as the indices of G gene probes on the microarray; $\{X_{g_p}\}_{p=1}^G$ as normalized probe responses (random variables); and $\{\{x_{g_p(n)}\}_{p=1}^G\}_{n=1}^N$ as realizations of $\{X_{g_p}\}_{p=1}^G$ under N i.i.d. microarray experiments.

Pearson Correlation Coefficient.

The population Pearson correlation coefficient between random variables X_{g_i} and X_{g_j} (defined as long as $\text{var}(X_{g_i}), \text{var}(X_{g_j})$ are positive) is:

$$(2.1) \quad \rho(X_{g_i}, X_{g_j}) = \frac{\text{cov}(X_{g_i}, X_{g_j})}{\sqrt{\text{var}(X_{g_i})\text{var}(X_{g_j})}}.$$

The sample Pearson correlation coefficient $\hat{\rho}$ is an asymptotically consistent unbiased estimator of ρ :

$$(2.2) \quad \hat{\rho}_{i,j} = \frac{S_{X_{g_i}, X_{g_j}}}{\sqrt{S_{X_{g_i}, X_{g_i}} S_{X_{g_j}, X_{g_j}}}},$$

where $S_{X_{g_i}, X_{g_i}}$, $S_{X_{g_j}, X_{g_j}}$, and $S_{X_{g_i}, X_{g_j}}$ are sample variances and covariance given by

$$S_{X_{g_i}, X_{g_i}} = (N-1)^{-1} \sum_{n=1}^N (X_{g_i(n)} - \overline{X_{g_i}})^2,$$

$$S_{X_{g_j}, X_{g_j}} = (N-1)^{-1} \sum_{n=1}^N (X_{g_j(n)} - \overline{X_{g_j}})^2,$$

$$S_{X_{g_i}, X_{g_j}} = (N-1)^{-1} \sum_{n=1}^N (X_{g_i(n)} - \overline{X_{g_i}})(X_{g_j(n)} - \overline{X_{g_j}}),$$

and $\overline{X_{g_i}} = N^{-1} \sum_{n=1}^N X_{g_i(n)}$, $\overline{X_{g_j}} = N^{-1} \sum_{n=1}^N X_{g_j(n)}$ are sample means.

Kendall Rank Correlation Coefficient.

Kendall's τ statistic is a measure of correlation that captures both linear and non-linear associations. The parameter τ is defined as $\tau = P_+ - P_-$, where, for any two independent pairs of observations $(x_{g_i(n)}, x_{g_j(n)})$, $(x_{g_i(m)}, x_{g_j(m)})$ from the population: $P_+ = P[(x_{g_i(n)} - x_{g_i(m)})(x_{g_j(n)} - x_{g_j(m)}) \geq 0]$ and $P_- = P[(x_{g_i(n)} - x_{g_i(m)})(x_{g_j(n)} - x_{g_j(m)}) < 0]$. An unbiased estimator of τ is given by the Kendall τ statistic:

$$(2.3) \quad \hat{\tau}_{i,j} = 2 \sum \sum_{1 \leq n \leq m \leq N} \frac{K_{nm}}{N(N-1)},$$

here K_{nm} is a indicator variable defined as $K_{nm} = \text{sgn}(x_{g_i(n)} - x_{g_i(m)}) \text{sgn}(x_{g_j(n)} - x_{g_j(m)})$ for each set of pairs drawn from $\{X_{g_i}\}_{i=1}^G$ and $\{X_{g_j}\}_{j=1}^G$.

2.1.2 Hypothesis Testing Scheme

To screen the strongly co-expressed pairs of G genes on each microarray, we will simultaneously test the $\Lambda = \binom{G}{2}$ pairs of composite hypotheses: $\{H_\lambda, K_\lambda : \lambda = (g_i, g_j)\}$.

(2.4)

$$H_\lambda : \Gamma_{g_i, g_j} \leq \text{cormin} \text{ versus } K_\lambda : \Gamma_{g_i, g_j} > \text{cormin}, \text{ for } g_i \neq g_j, \text{ and } g_i, g_j \in (1, 2, \dots, G)$$

where *cormin* is the specified minimum acceptable strength of correlation. The sample correlation coefficient $\hat{\Gamma}_{i,j}$ ($\hat{\rho}_{i,j}$ or $\hat{\tau}_{i,j}$) could be thresholded to decide on pairwise dependency of two genes in the sample. When we must decide between the null hypothesis H_λ and the alternative hypothesis K_λ based on such a threshold test,

there will generally be decision errors in the form of false positives (Type I errors: decide K_λ when H_λ is true) and false negatives (Type II errors: decide H_λ when K_λ is true). The Per Comparison Error Rate (PCER) is defined as the number of Type I errors over the number of independent trials, i.e. the probability of Type I error. The p -value is the probability that a more improbable sample could have been drawn from the population(s) being tested given the assumption that the null hypothesis is true.

For N realizations of any pair of gene probe responses, $\{x_{g_i(n)}, x_{g_j(n)}\}_{n=1}^N$, we first calculate $\hat{\tau}_{i,j}$ or $\hat{\rho}_{i,j}$ respectively. For large N , the PCER p -values for $\rho_{i,j}$ or $\tau_{i,j}$ are:

$$(2.5) \quad p_{\rho_{i,j}} = 2 \left(1 - \Phi \left(\frac{\tanh^{-1}(\hat{\rho}_{i,j})}{(N-3)^{-1/2}} \right) \right)$$

$$(2.6) \quad p_{\tau_{i,j}} = 2 \left(1 - \Phi \left(\frac{K}{N(N-1)(2N+5)/18^{1/2}} \right) \right)$$

where Φ is the cumulative density function of a standard Gaussian random variable, and $K = \sum \sum_{1 \leq n \leq m \leq N} K_{nm}$. The above expressions are based on asymptotic Gaussian approximations to $\hat{\rho}_{i,j}$ (Bickel and Doksum 2000) and to $\hat{\tau}_{i,j}$ (Hollander and Wolfe 1999).

The PCER p -value refers to the probability of Type I error incurred in testing a single pair of hypothesis for a single pair of genes g_i, g_j . It is the probability that purely random effects would have caused g_i, g_j to be erroneously selected based on observing correlation between this pair of genes only. When considering the Λ multiple hypotheses for all possible pairs, two adjusted error rates have frequently been considered in microarray studies. These are family-wise error rate (FWER) and false discovery rate (FDR)(Benjamini and Hochberg 1995). The FWER is the probability that the test of all Λ pairs of hypotheses yields at least one false positive in

the set of declared positive responses. In contrast, the FDR is the average proportion of false positives in the set of declared positive responses. The FDR is dominated by the FWER and is therefore a less stringent measure of significance. As in previous studies (Reiner *et al.* 2003), we adopt the FDR to control statistical significance of the selected gene pair correlations in our screening procedure.

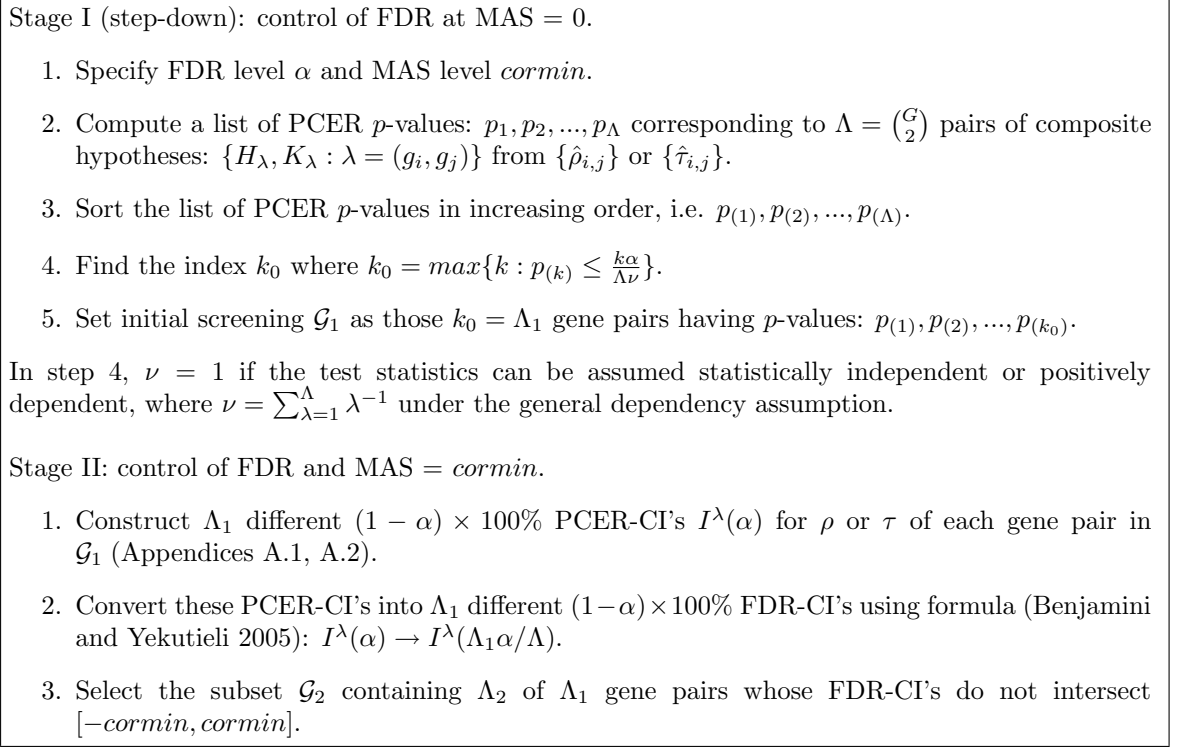


Figure 2.1: Two-stage direct screening procedure yields a subset \mathcal{G}_2 of all possible gene pairs \mathcal{G} whose strength of association exceeds MAS level *cormin* at FDR level α .

2.1.3 Two-stage Screening Procedure

One would proceed to test the hypothesis (Eq. 2.4) in one stage if the distributions of Pearson and Kendall correlations under the specific null hypothesis were known. Under the null hypothesis of zero correlation, the distributions and p -values of Pearson and Kendall correlations can be conveniently derived using the asymptotic approximations (Fisher 1923, Hollander and Wolfe 1999). While deriving p -value under the null hypothesis of non-zero correlation remains an open problem. For this

reason, we test the composite hypothesis (Eq. 2.4) using a two-stage procedure.

Select a level α of FDR and a level *cormin* of MAS significance levels. We use a modified version of the two-stage screening procedure proposed for gene screening by (Hero *et al.* 2004). This procedure consists of two stages, summarized in Fig. 2.1.

Stage I. For each gene pair $\lambda = (g_i, g_j)$ in the set \mathcal{G} of all $\Lambda = \binom{G}{2}$ gene pairs, test the simple null hypothesis:

$$(2.7) \quad H_\lambda : \Gamma_{g_i, g_j} = 0 \quad \text{versus} \quad K_\lambda : \Gamma_{g_i, g_j} \neq 0, \text{ for } g_i \neq g_j, \text{ and } g_i, g_j \in (1, 2, \dots, G)$$

at FDR level α . The step-down procedure of Benjamini and Hochberg (Benjamini and Hochberg 1995) is used to accomplish this.

Stage II. Suppose a number Λ_1 pairs of genes, denoted by the set $\mathcal{G}_1 \subset \mathcal{G}$, pass the Stage I procedure. In Stage II, we first construct asymptotic PCER Confidence Intervals (PCER-CI's): $I^\lambda(\alpha)$ for each Γ (ρ or τ) in subset \mathcal{G}_1 . We convert these PCER-CI's into FDR Confidence Intervals (FDR-CI's): $I^\lambda(\alpha) \rightarrow I^\lambda(\Lambda_1\alpha/\Lambda)$ using the procedure in (Benjamini and Yekutieli 2005). A gene pair in subset \mathcal{G}_1 is declared to be both statistically significant and biologically significant if its FDR-CI does not intersect the MAS interval $[-cormin, cormin]$ (see Fig. 2.6). The set of all such gene pairs is called \mathcal{G}_2 .

In many practical situations, the experimenter may not be comfortable in specifying a MAS or FDR criterion in advance. In this situation, it is useful to solve the inverse problem: what is the most stringent pair of criteria (α , *cormin*) that would cause a particular subset of gene pairs to be included in the screen \mathcal{G}_2 . The inverse screening procedure is displayed in Fig. 2.2.

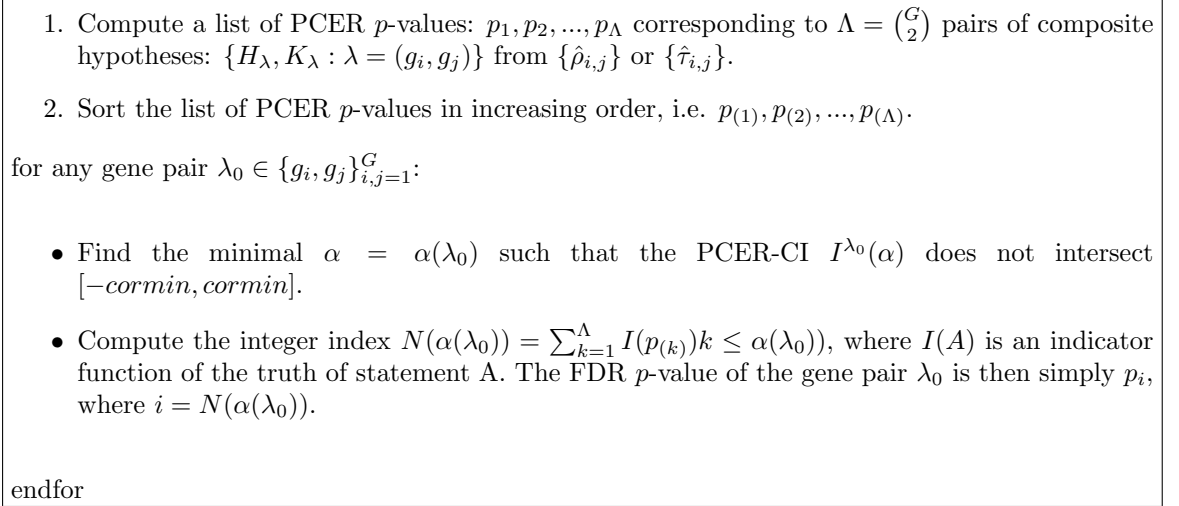


Figure 2.2: Inverse screening procedure allows the FDR p -value of a gene pair's (λ_0) strength of association to be computed.

2.2 Simulation Studies

2.2.1 Validating the Two-stage Algorithm

Validating Asymptotic Null Distribution.

We first verify that the proposed two-stage algorithm controls FDR at a specified MAS level using simulated data. Since the p -values are based on asymptotic distribution approximations (Eq. 2.5 and Eq. 2.6), we display in Fig. 2.3a the goodness of fit of the $\hat{\rho}$ sampling distribution to the Gaussian distribution using QQ plots. Note that there is good agreement to the Gaussian distribution for $N \geq 10$. Moreover, since the construction of confidence intervals requires estimation of sampling distribution variance, the accuracy of the variance approximation is vital. This can be evaluated by the mean squared approximation error (MSE) for sample size N :

$$(2.8) \quad MSE_\rho^{(N)} = \Lambda^{-1} \sum_{1 \leq i < j \leq G} (S_{\tanh^{-1}(\hat{\rho}_{i,j})}^{(N)} - (N-3)^{-1/2})^2,$$

$$(2.9) \quad MSE_\tau^{(N)} = \Lambda^{-1} \sum_{1 \leq i < j \leq G} (S_{\hat{\tau}_{i,j}}^{(N)} - (\frac{2}{N(N-1)} \frac{2(N-2)}{N(N-1)^2} \sum_{i=1}^N (C_r - \bar{C}) + 1 - \hat{\tau}))^2,$$

where $S_{\tanh^{-1}(\hat{\rho}_{i,j})}^{(N)}$ and $S_{\hat{\tau}_{i,j}}^{(N)}$ denote standard errors of $\tanh^{-1}(\hat{\rho}_{i,j})$ and $\hat{\tau}_{i,j}$ at the sample size N . The definitions of C_r and \bar{C} can be found in Appendix A.2. The $\hat{\rho}$ variance approximations are seen to be in good agreement even for small sample sizes ($N > 10$) from Fig. 2.3b.

Validating the Error Control Procedure.

In order to validate our FDR and MAS error control procedure, we simulated pairwise gene expression data based on known population covariances (Appendix A.3). The actual FDR at a MAS level is calculated as a ratio of the number of screened gene pairs whose corresponding population correlation parameters $\Gamma_{i,j}$ are less than the MAS level specified, divided by the total number of screened gene pairs. The actual MAS is the minimum true discovery of population correlation $\Gamma_{i,j}$ among the screened pairs. We specified 16 pairs of (FDR,MAS) criteria (Four FDR levels: 0.2, 0.4, 0.6, 0.8; Four MAS levels: 0.2, 0.4, 0.6, 0.8), and each is plotted as a different upper case Roman alphabet (Red) in Fig. 2.4. The 16 corresponding pairs of actual (FDR,MAS) criteria are also shown in Fig. 2.4 using the same set of lower case Roman alphabets (Blue). It can be observed that generally the actual FDR's (lower case) fall below the specified constraint (upper case) and the actual MAS's (lower case) fall above the specified constraints (upper case). Any deviations of actual FDR's and MAS's from their specified levels are due to the conservative asymptotic approximation (Eq. 2.5 and Eq. 2.6). Observe that use of Kendall correlation (Fig. 2.4b) leads to greater overestimation of error rates than the Pearson correlation (Fig. 2.4a). Overestimation of error rates will translate into a reduction of power in discovering co-expressed pairs at the specified levels.

2.2.2 Performance Comparisons

Comparisons in terms of p -values and Confidence Intervals

In Table 2.1 and Table 2.2, we compared the performance of the proposed two-stage FDR-CI screening algorithm (labeled “FDR-CI” in the tables), with two other commonly used algorithms, called thresholded FDR (labeled “FDR-only” in the tables) and thresholded MAS (labeled “MAS” in the tables). All three algorithms aim to control MAS at a level of $cormin = 0.5$. The two-stage FDR-CI and thresholded FDR algorithms aim to control FDR at a level of $\alpha = 0.05$ in addition to MAS. Both of these latter algorithms were implemented as two-stage algorithms with common Stage I, which is to select pairs of genes \mathcal{G}_1 that pass the test of association with $cormin = 0$ at a FDR level of 5%. Stage II of the two-stage FDR-CI algorithm selects \mathcal{G}_2 as a subset of \mathcal{G}_1 at the specified FDR-CI level of 5%. Stage II of the thresholded FDR algorithm simply selects a subset of \mathcal{G}_1 having a strength of association greater than 0.5. The single-stage thresholded MAS algorithm selects a subset of the original 496,506 gene pairs by thresholding Pearson correlation $\hat{\rho}_{i,j} \geq 0.5$ (Table 2.1) and Kendall coefficient $\hat{\tau}_{i,j} \geq 0.5$ (Table 2.2) without attempting to control FDR.

The number of screened and discovered gene pairs for the three algorithms is indicated in the first two columns of Table 2.1 and Table 2.2. The maximum and median of the FDR p -values of the discovered gene pairs are indicated in the third and fourth columns for each algorithm. The last column indicates the average length of the FDR-CI’s on correlation coefficients of the discovered gene pairs. We conclude from Table 2.1 and Table 2.2 that the proposed two-stage FDR-CI algorithm outperforms the other algorithms in terms of: (1) maintaining the FDR requirement that false positives not exceed 5% (column 4); (2) ensuring a substantially lower median FDR p -value than the others (column 5); (3) discovering genes that have tighter (on

Table 2.1: Performance comparison for three algorithms based on Pearson correlation coefficient for selecting gene pairs with a MAS level of 0.5. Thresholded MAS and thresholded FDR are significantly worse in terms of statistical significance (p -value) than the proposed two-stage FDR-CI algorithm (columns 4 and 5). Furthermore, the average length of the CI's on ρ 's of the discovered gene pairs are shorter for the two-stage FDR-CI algorithm than for the other algorithms (column 6).

Algorithms	# Screened	# Discovered	Max(Pv)	Median(Pv)	AvgFDRCI
MAS	496,506	174,830	2.5e-02	2.1e-03	6.5e-01
FDR-only	153,983	153,983	1.6e-02	1.4e-03	6.3e-01
FDR-CI	153,983	18,135	1.3e-05	1.3e-06	3.3e-01

Table 2.2: Performance comparison for three algorithms based on Kendall's τ statistic for selecting gene pairs with a MAS level of 0.5. Thresholded MAS and thresholded FDR are significantly worse in terms of statistical significance (p -value) than the proposed two-stage FDR-CI algorithm (columns 4 and 5). Furthermore, the average length of the CI's on τ 's of the discovered gene pairs are shorter for the two-stage FDR-CI algorithm than for the other algorithms (column 6).

Algorithm	# Screened	# Discovered	Max(Pv)	Median(Pv)	AvgFDRCI
MAS	496,506	31,151	2.0e-02	6.4e-03	6.3e-01
FDR-only	95,205	31,151	2.0e-02	6.4e-03	6.3e-01
FDR-CI	95,205	3,552	1.4e-03	4.3e-04	4.1e-01

the average) confidence intervals on biologically significant (i.e. $\Gamma \geq 0.5$) correlation coefficients (column 6).

Comparisons using Receiving Operator Characteristic (ROC) Curve

We also compared the performance of the two two-stage algorithms ("thresholded FDR" and "FDR-CI") using the Receiving Operator Characteristic (ROC) curve in which "sensitivity" is plotted against "1 - specificity". Let Λ_0 denotes the number of false hypotheses (true strength of pairwise association is smaller than or equal to the threshold *cormin*), and Λ_α denotes the number of true hypotheses (true strength of pairwise association is greater than the threshold *cormin*). We counted false positives FP (falsely rejected hypotheses) and false negatives FN (falsely accepted hypotheses). The "sensitivity" (True positive rate, pTP) can be calculated as $pTP =$

$1 - E(FN/\Lambda_\alpha)$; and the “1 - specificity” (False positive rate, pFP) can be calculated as: $pFP = E(FP/\Lambda_0)$. Same as the above, the two-stage algorithm labeled as “FDR-only” in Fig. 2.5 denotes the FDR test followed by a “hard” correlation thresholding; and that labeled as “FDR-CI” denotes the FDR test followed by a “soft” correlation thresholding (FDR-CI). In Fig. 2.5, we observe overall better performance of “FDR-CI” test than the “FDR-only” test especially at low levels of correlation thresholding. For example, at the MAS level of 0.2 and the specificity level of 0.9, the “FDR-CI” method has a three-fold higher sensitivity ($pTP \approx 0.6$) than the “FDR-only” method ($pTP \approx 0.2$).

2.3 Applications in Network Construction and Seeded Clustering

2.3.1 Constructing Relevance Networks with Controlled FDR and MAS

We demonstrate the application of our approach and compare it with the traditional approach using a yeast galactose metabolism two-color microarray data. This data represents approximately 6200 gene expression levels on two-color cDNA microarrays over 20 physiological/genetic conditions (nine mutants and one wild type strains incubated in either GAL-inducing or non-inducing media). A subset of 997 differentially expressed genes were identified by Ideker et al using a generalized likelihood ratio test procedure (Ideker *et al.* 2000). Genes having a likelihood ratio statistic $\lambda \leq 45$ were selected as differentially expressed, i.e. whose mRNA levels differed significantly from the reference under one or more treatments.

Selecting biological significance level (MAS) is a key to constructing relevance networks with controlled biological significance and statistical significance. In general, there are two ways to tackle this problem. One way is to select a small portion (e.g. 5%) of the top gene pairs ranked by the absolute magnitude of correlation (Lee *et al.* 2004). The other way is to use a cut-off value (e.g. 0.6)(Zhou *et al.* 2005, Zhou

et al. 2002, Butte and Kohane 2000). Both ways seek a good compromise between sensitivity and specificity. Zhou *et al.* advocated using the cutoff value between 0.5 and 0.7, and they considered two factors in choosing the biological significance level (Zhou *et al.* 2005):

- It should be statistically conservative to achieve high sensitivity.
- It should retain a sufficient number of gene pairs for functional analysis to achieve high specificity.

Correspondingly, they first performed a randomization test to determine that 0.6 is a statistically conservative cutoff value. They then used a series of correlation cutoff values ranging from 0.4 to 0.9 to determine the number and the ratio of the true positives (functional related gene pairs). They found that cutoff values from 0.5 to 0.7 give rise to good compromise between sensitivity and specificity (Zhou *et al.* 2005). Throughout this thesis, we select the cutoff value within this range to control biological significance.

Fig. 2.6a and Fig. 2.6b illustrate the direct implementation of the two-stage procedure to screen positively or negatively correlated gene pairs based on the Pearson correlation coefficient. The direct screening procedure is constrained by FDR level $\alpha = 0.05$ and MAS level $cormin = 0.5$. Stage I of the screen discovered $\Lambda_1 = 153,983$ out of $\Lambda = \binom{997}{2} = 496,506$ gene pairs having $FDR \leq 0.05$, leaving 153,983 correlation coefficients for which FDR-CI's must be constructed. Recall that gene pair passes the Stage II screening if the FDR-CI does not intersect the interval $[-0.5, 0.5]$. $\Lambda_2 = 18,135$ of the 153,983 gene pairs passed the Stage II screening and were declared to be both “biologically” and “statistically” significant. Similarly, using Kendall correlation coefficient, there were $\Lambda_1 = 95,205$ gene pairs that passed the Stage I screen,

and only $\Lambda_2 = 3,552$ gene pairs passed the Stage II screen constrained by the same MAS and FDR criteria as above (Table A.1).

Although for Gaussian distributed pairs the Kendall rank correlation coefficient has lower discovery power compared to the Pearson correlation coefficient, our screening procedure was nevertheless able to pull out many non-linearly correlated gene pairs that were missed by the Pearson correlation procedure. These non-linearly correlated gene pairs, just like those linearly correlated ones, may be biologically relevant too. For example, the link between gene “RPC40” and gene “YDR516C” passed both Stage I and II screening ($\alpha = 0.015$, $cormin = 0.5$) when using Kendall correlation coefficient ($\hat{\tau} = -7.5e-01$, FDR p -value = $6.2e-04$, FDR-CI = $[-9.7e-01, -5.4e-01]$), but they failed to pass even the first screening when the Pearson correlation coefficient was used ($\hat{\rho} = -6.3e-01$, FDR p -value = $1.2e-02$). From the scatter plot, we can observe an obvious non-linear correlation for this gene pair (Fig. 2.7). The poor linear fit can be verified by fitting a simple linear regression model and observing $R^2 = 0.36$. Biologically, the gene “RPC40” encodes RNA polymerase (I and III) subunit (transcription apparatus); although the specific function of gene “YDR516C” remains unclear, it is recently shown that it involves in transcriptional induction of the early meiotic-specific transcription factor IME1 (Dwight *et al.* 2002). Both genes are thus components of transcription apparatus. Applying our two-stage algorithm based on Pearson correlation coefficient alone will miss the important functional relationship. Therefore, the Kendall correlation statistic can beat the Pearson correlation statistic in some instances and hence the two correlation statistics should be used together to capture functional relationships as many as possible.

Relevance networks are implemented as a graph where n nodes (genes) are connected by m sets of edges (co-expressions). Each of the p sets of edges are discovered

using a different similarity measure (Butte *et al.* 2000, Butte and Kohane 2000). Therefore, our constructed networks are mixed networks with $m = 2$ in which edges are discovered using either Pearson correlation coefficients or Kendall correlation coefficients constrained by the same set of (FDR, MAS) criteria. In relevance networks, genes that are of considerable interest to the biologist are “hub genes” such as RPL33A and RPS4A in Fig. 2.8. Hub genes are best connected genes that dominate a large part of the network topology (Jeong *et al.* 2001, Barabási 2004). We constructed five such networks that are constrained by five pairs of constraints (FDR ≤ 0.05 , $cormin = 0.5, 0.6, 0.7, 0.8, 0.9$). Most of the “hub genes” in each discovered network fall into two categories: “RPL” and “RPS”. The former encodes “Ribosome Protein Large (60S) subunit,” and the latter encodes “Ribosome Protein Small (40S) subunit”. Both of these categories are structural components of the ribosome that is responsible for protein biosynthesis. Protein biosynthesis plays the central role in galactose metabolism because galactose is not a primary carbon source for yeast, when switching from primary carbon sources (glucose) to secondary carbon source (e.g. galactose), many different types of proteins including transporters, enzymes, and regulators have to be synthesized to be able to degrade the secondary carbon source (Wieczorke *et al.* 1999). We ranked the “hub genes” by calculating and sorting average rank of each “hub gene” over five networks (Table 2.3). The list of “hub genes” (Table 2.3) are presumably indispensable for galactose metabolism (Jeong *et al.* 2001).

Fig. 2.8 presents the discovered network topology with a FDR level of 5% (5% discovered edges are expected to be false positive) at the MAS level of $cormin = 0.9$. The network is composed of 89 connected vertices and 132 edges. Similar to some other biological networks, the network marginal degree distributions appear to be

of the form of a power-law. This was tested by verifying goodness of fit to the log-transformed power-law model ($R^2 = 0.95$) i.e., $\log P(D_i) = -\gamma \log D_i + \log \eta + \varepsilon_i$ (Barabási 2004). Here γ and η are shape and intercept parameters, i is the index of a gene in the network, ε_i is a residual fitting error, D_i is the number of edges (degree) of i th gene and $P(D_i)$ is the corresponding probability.

2.3.2 Seeded Clustering

Inspired by the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.* 1990), and based on the “guilt-by-association” assumption (Eisen *et al.* 1998), we applied the two-stage screening procedure to cluster co-expressed genes with controlled FDR and MAS. We sought to demo its application in metabolic pathway discovery by “rediscovering” the extensively studied galactose metabolic pathway, which consists of at least three types of genes including transporter genes (GAL2, HXTs etc), enzyme genes (GAL1, GAL7, GAL10 etc) and transcription factor genes (GAL4, GAL80, GAL3 etc). Some other genes are also involved in galactose metabolism but their roles are not entirely clear (Rohde *et al.* 2000, Ideker *et al.* 2001). Therefore, our aims are not only to validate our procedure by rediscovering known co-expressed genes pairs, but also to discover some unknown genes in the pathway.

We selected gene “GAL10” as the “seed gene” which encodes the UDP-glucose-4-epimerase (EC 5.1.3.3) (Fig. 2.9). We set a relatively stringent criterion ($\alpha = 0.05$, $cor_{min} = 0.6$), and $cor_{min} = 0.6$ is widely used in the literature (e.g. Zhou *et al.* 2002, Farkas *et al.* 2003). We discovered six genes (GAL10, GAL7, GCY1, GAL1, GAL2 and YOR121C) (Table A.2). Five of six genes are known to be lying in the pathway as shown in shaded squares in Fig. 2.9, which leads to a specificity of at least 83%. The sixth gene “YOR121C” is a hypothetical ORF for which no functional annotation is currently available. Our results provide strong motivation

Table 2.3: Top twenty “hub genes” from the two-stage algorithm applied to galactose metabolism data (Ideker *et al.* 2000). The rank of each gene is the average rank over five different networks. Each of five networks is constrained by a different pair of (FDR,MAS) criteria. The highest ranked gene is the most connected and stable gene under varying constraints of (FDR,MAS).

Gene Name	Average Rank	GO Annotation
RPL42B	4.2	protein biosynthesis[GO:0006412]
RPS16B	6.2	protein biosynthesis[GO:0006412]
RPL14A	7.4	protein biosynthesis[GO:0006412]
RPS3	7.4	protein biosynthesis[GO:0006412]
GTT2	8.0	glutathione metabolism[GO:0006749]
RPS4A	9.8	protein biosynthesis[GO:0006412]
RPL33A	11.6	protein biosynthesis[GO:0006412]
RPL23B	15.4	protein biosynthesis[GO:0006412]
RPS7A	15.8	protein biosynthesis[GO:0006412]
RPS4B	17.2	protein biosynthesis[GO:0006412]
RPL27A	17.8	protein biosynthesis[GO:0006412]
RPS18A	19	protein biosynthesis[GO:0006412]
RPL26B	19.8	protein biosynthesis[GO:0006412]
RPS9A	20	protein biosynthesis[GO:0006412]
RPL33B	20.6	protein biosynthesis[GO:0006412]
RPL21A	22.2	protein biosynthesis[GO:0006412]
RPL23A	22.2	protein biosynthesis[GO:0006412]
RPL9B	22.2	protein biosynthesis[GO:0006412]
RPL11B	23.8	protein biosynthesis[GO:0006412]
RPL20B	24.2	protein biosynthesis[GO:0006412]

to experimentally characterize this gene’s biological function. Known transcription factor genes (GAL4 and GAL80) were not discoverable from this microarray experiment as the GAL4 and GAL80 expressions are time shifted and only one time sample was included. The pathways discovered using other “seed genes” in the pathway such as GAL1 and GAL7 gave similar results (Table A.3, Table A.4, Table A.5).

2.4 Discussion

In this chapter, we presented a two-stage procedure for screening co-expressed gene pairs that controls both biological and statistical significance of the discovered strength of association, and hence the gene co-expression network. For the discovered co-expressions, our method also provides an “accuracy” assessment of the strength of association by constructing confidence intervals for the strength of each edge.

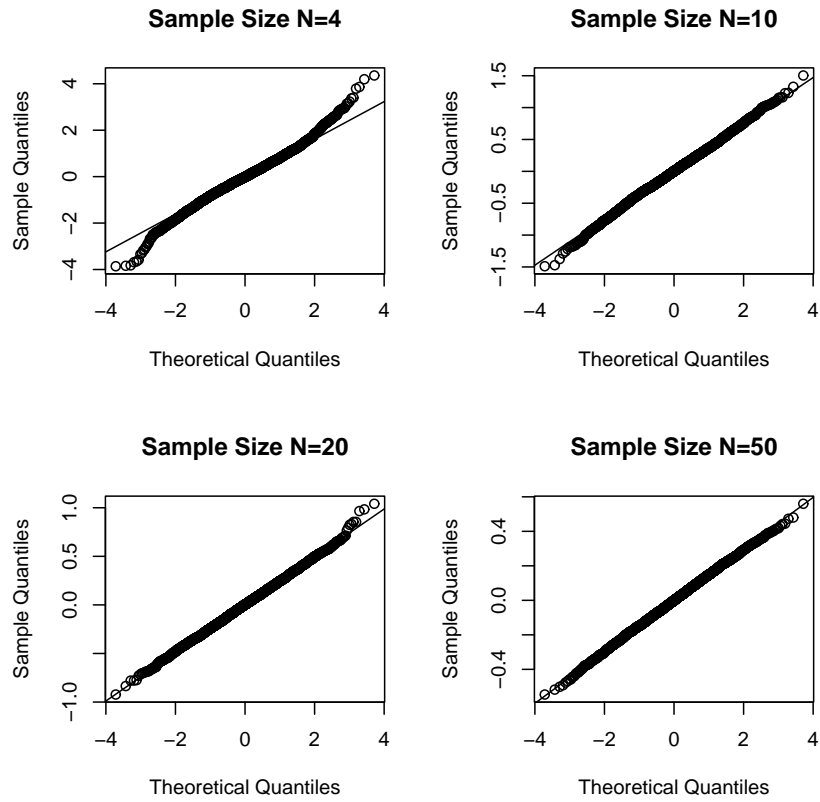
Indeed, for the typically small sample size microarray data, a simultaneous confidence interval is useful to characterize reliability of the reported strength of association. Correlation thresholding is becoming standard practice in gene co-expression analyses (e.g. Butte and Kohane 2000, Butte *et al.* 2000, Zhou *et al.* 2002, Farkas *et al.* 2003, Lee *et al.* 2004), yet “hard” thresholding lowers the discriminative power of the FDR based test (Fig. 2.5). Our “soft thresholding ” procedure is able to control error rate and maintain discriminative power (Fig. 2.4). The method requires a tight confidence interval on correlation, which may be difficult to obtain for small sample sizes. However, we have shown that our algorithm provides error rate control at a biologically relevant level with relatively large sample size (20 samples for Fig. 2.3b, Fig. 2.4). In the Chapter III, we present a complementary approach that is more suitable for small sample size data.

The algorithm is sufficiently general to be applied to many different correlation measures, e.g. Spearman’s or Hotelling’s dependency statistics. The algorithm can also be extended to different frameworks such as Gaussian Graphical Models (GGM) in which partial correlation coefficients are used as the dependency measures (Whittaker 1990). Different groups have developed approaches to infer GGM from small sample size microarray data (Wang *et al.* 2003, Schafer and Strimmer 2005a, Dobra *et al.* 2004). Schafer and Strimmer recently presented a procedure that is based on the bootstrap estimator of the partial correlation coefficient (Schafer and Strimmer 2005a). Most of the pairwise partial correlations discovered by their procedure are very close to zero. On one hand, these ultra weak correlations screened by the FDR based inference procedure are “true correlation” from a pure statistical point of view. On the other hand, the “true correlation” may be caused by a variety of factors other than functional relationship, such as positional and spatial artifacts of

gene co-expression along chromosomes (Kluger *et al.* 2003). Thus it seems necessary to combine such statistical testing with a “soft” thresholding to achieve high sensitivity and specificity (Fig. 2.5). This chapter has presented such a method to simultaneously minimize the discovered proportion of the functionally irrelevant “true correlations” and maximize that of functionally relevant ones. Our two-stage algorithm has been extended to the GGM framework and implementations are included in our R package “GeneNT” (available from <http://cran.r-project.org/>). Running the partial correlation based two-stage algorithm and combining the results with those of marginal correlation based two-stage algorithm may allow discovering additional functional links that would have been missed by running the latter algorithm alone.

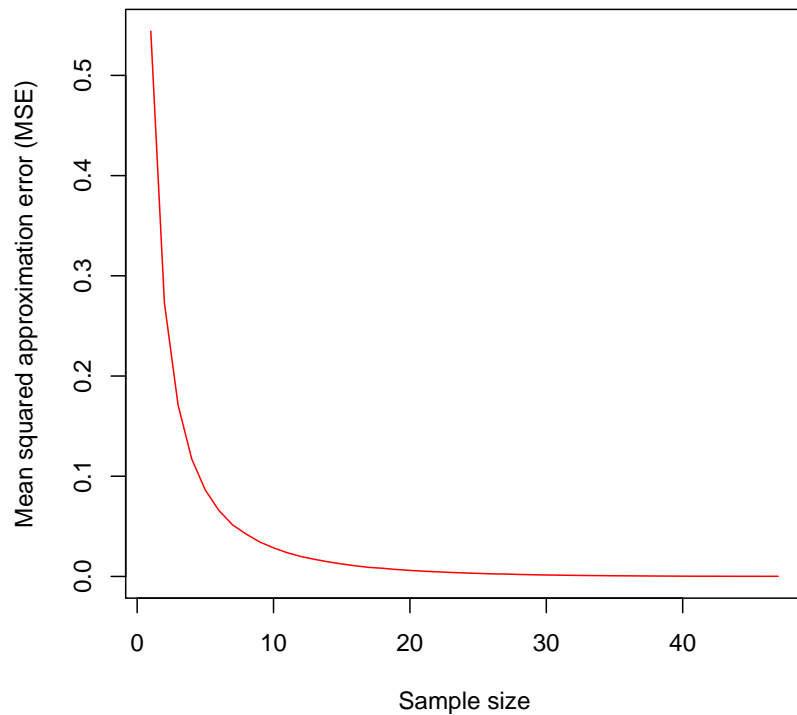
The scope of application of our statistical analysis is explicitly that of randomly sampled, complete observational data (Dobra *et al.* 2004). In this thesis, we are not concerned with developing models of causal gene networks (Dobra *et al.* 2004). This would require a different experimentation and intervention approach to understand directional influences, rather than the simple observational random sampling paradigm adopted here (Dobra *et al.* 2004).

Finally we note that the two-stage procedures can be applied under the assumption of independency/positive dependency or under more general dependency assumptions (Benjamini and Hochberg 1995, Benjamini and Yekutieli 2001). The implementation of the general dependency procedure ($\nu = \sum_{\lambda=1}^{\Lambda} \lambda^{-1}$) causes loss of discovery power. The assumption of independence may not be critical in the discovery of relevance networks since biological networks are typically very sparse (Yeung *et al.* 2002).



(a)

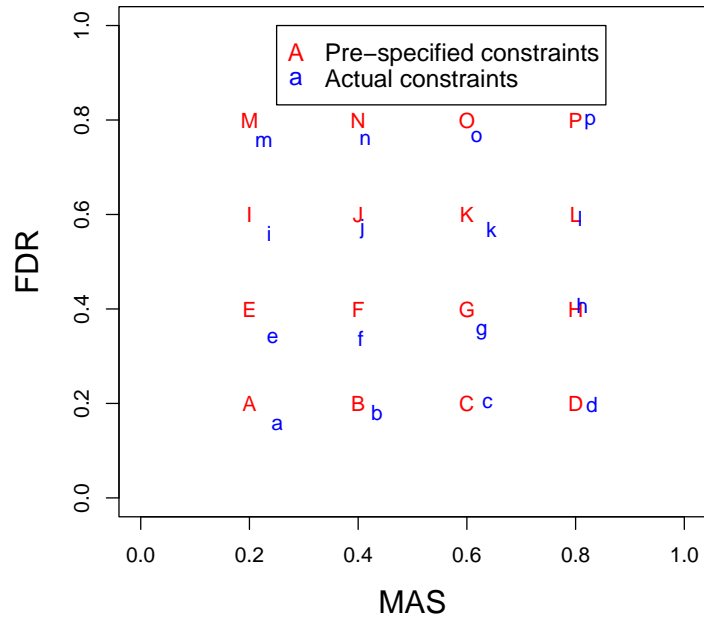
Accuracy of variance approximation



(b)

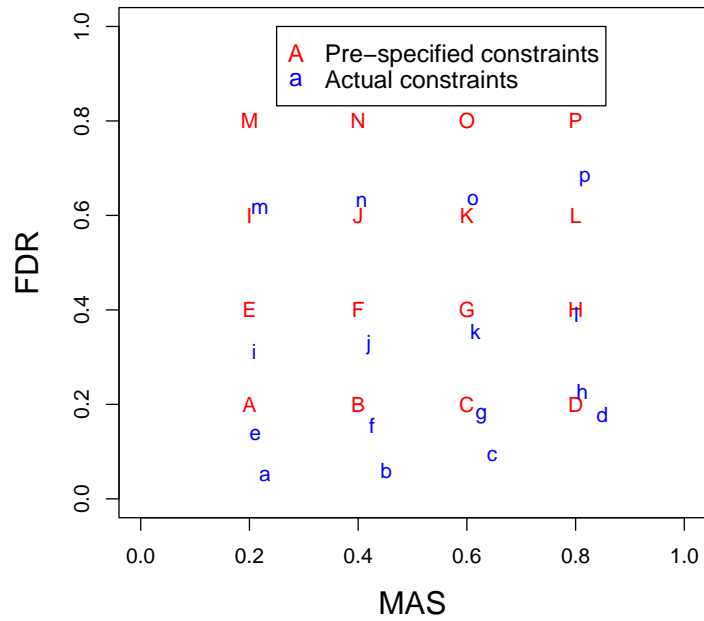
Figure 2.3: Verification of Gaussian null sampling distribution and variance approximation for Pearson correlation coefficient. (a) QQ plot of transformed sampling distribution of Pearson correlation coefficient $\hat{\rho}$ versus Gaussian distribution. (b) Mean squared approximation errors (MSE) of the variances of transformed sample Pearson correlation coefficients $\hat{\rho}$.

Pearson correlation coefficient



(a)

Kendall correlation coefficient



(b)

Figure 2.4: Verification of two-stage error control procedure based on Pearson correlation coefficient (a) and Kendall correlation coefficient (b). Sample size $N = 20$.

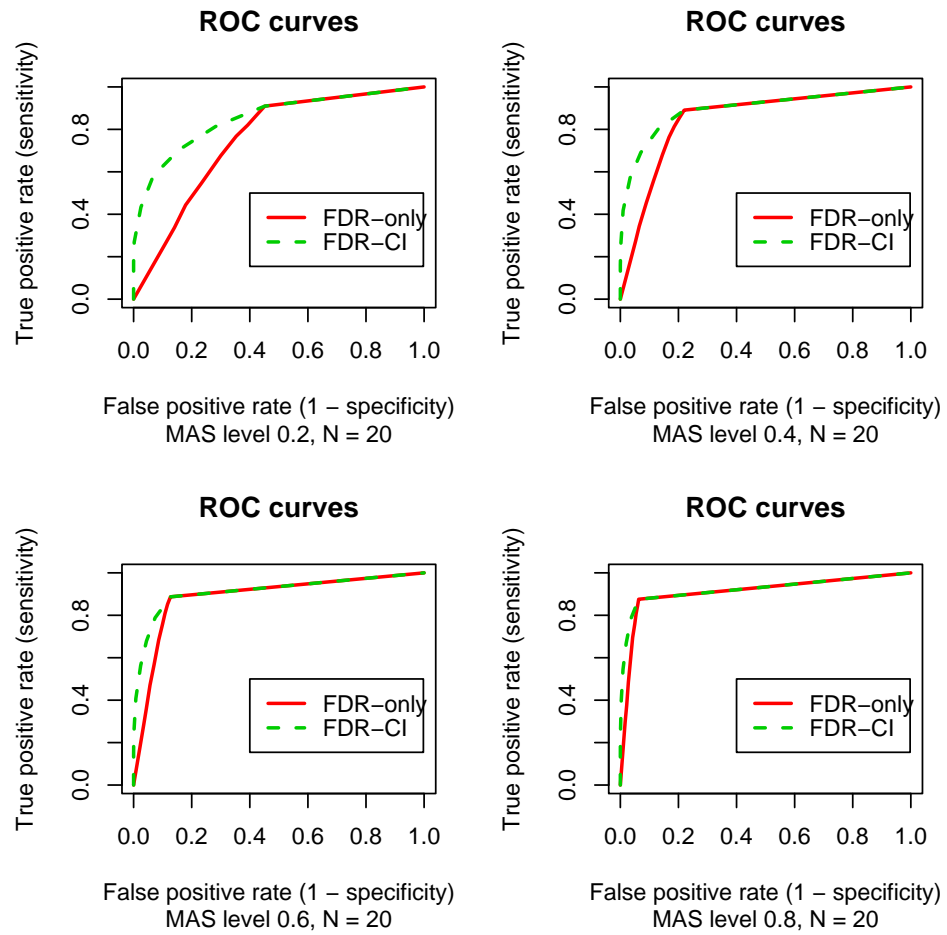


Figure 2.5: ROC curves of “FDR-CI” test procedure and “FDR-only” test procedure based on Pearson correlation statistic

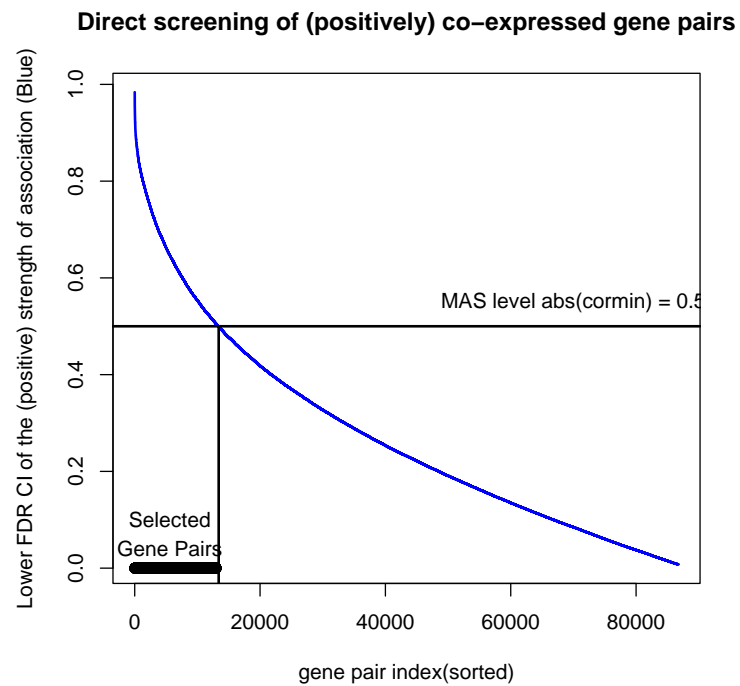
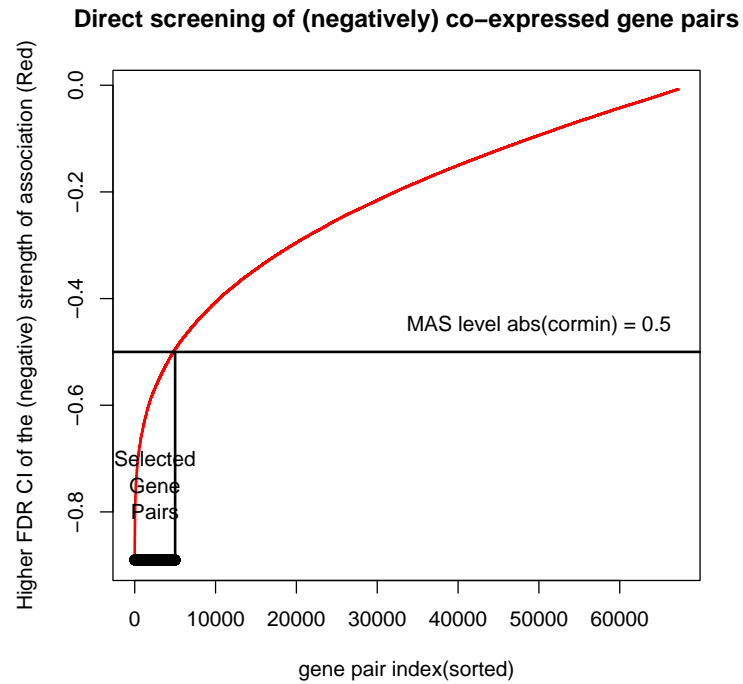
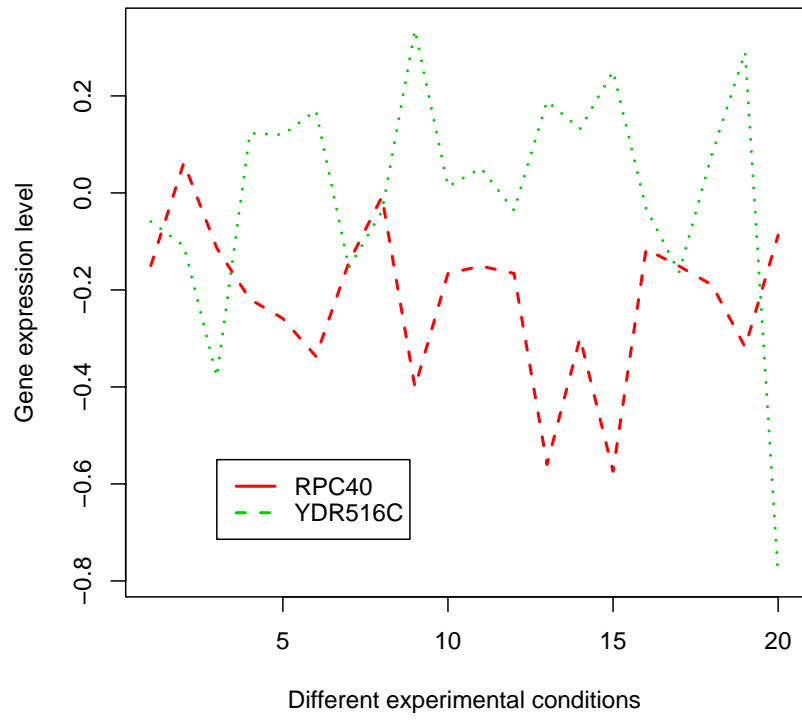
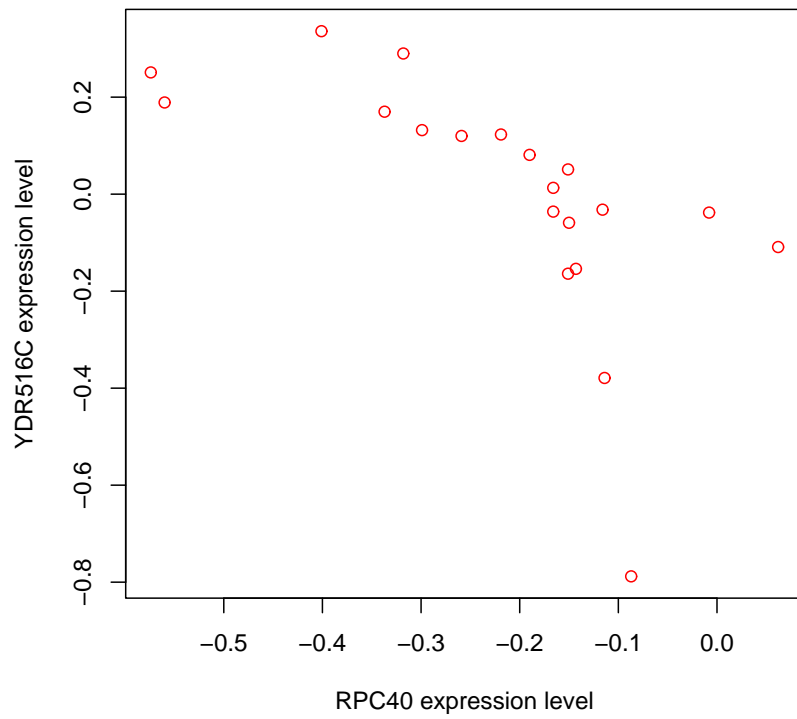


Figure 2.6: Curves specify lower endpoints (a) and upper endpoints (b) of the 5% FDR-CI's on the positive Pearson correlation coefficients (a) and negative Pearson correlation coefficients (b) for the galactose metabolism study. Only those gene pairs whose FDR-CI's do not intersect $[-cormin, cormin]$ are selected by the second stage of screening. When the MAS strength of association criterion is $cormin = 0.5$, these gene pairs are obtained by thresholding the curves as indicated.

Expression profiles of gene RPC40 and gene YDR516C

(a)

Scatterplot of RPC40 vs. YDR516C

(b)

Figure 2.7: A pair of non-linearly correlated genes.

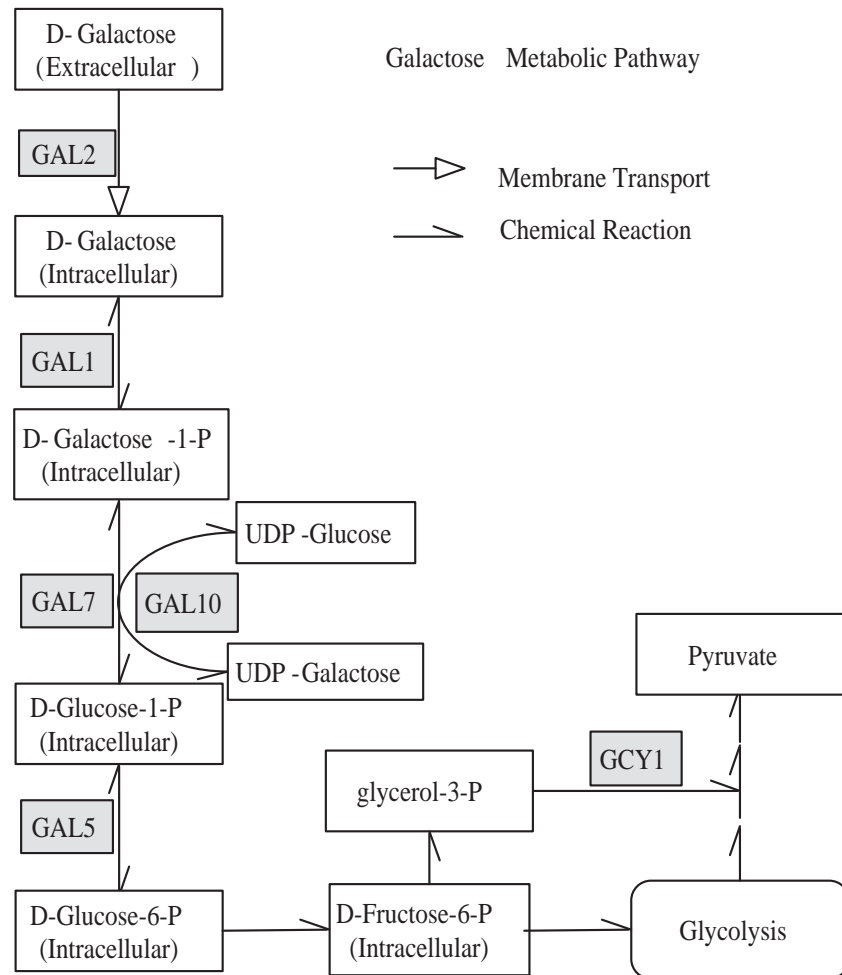


Figure 2.9: Diagram of the structural module of the galactose metabolic pathway. The shaded boxes denote the five out of six genes whose gene products lie in the galactose metabolic pathway “rediscovered” by our algorithm.

CHAPTER III

Co-expression Networks Construction - Bayesian Approach

In Chapter II, we demonstrated that the frequentist approach is able to simultaneously control statistical significance and biological significance. However, for small number N of samples and large number p of genes the correlation estimates have poor accuracy due to overfitting (Ledoit and Wolf 2004, Schafer and Strimmer 2005b). Introducing some form of strong dependency among correlation parameters can lead to improved accuracy in this small sample situation. Many approaches to introducing dependency can be adopted. For example, the full order partial correlation estimation approach, also called Gaussian Graphical Modeling (GGM), introduces a Bayes model from which all correlations are estimated using an Empirical Bayes method (Schafer and Strimmer 2005a). Bayesian hierarchical models accomplish this introduction of dependency in a simple but effective manner. We take this approach in this chapter.

3.1 The Bayesian Hierarchical Model

The framework of Bayesian hierarchical models is a powerful technique that allows for high complexity without a large number of parameters (Gelman *et al.* 2004). We assume the correlation parameters are *exchangeable* meaning that their joint distribution is invariant to permutations of their indexes. Biologically, this represents

a kind of topological invariance that imposes prior assumptions on the location of high correlations in the network. We then regularize variances of the marginal correlation densities by specifying a parent Gaussian distribution from which marginal correlation parameters are sampled. Using a prior population distribution we are able to introduce dependency into the parameters that tends to avoid problems of overfitting. Using quantiles of posterior distributions of the correlation parameters provide a seamless combination of correlation estimation and strength thresholding that can be used as an alternative to FDR-CI methods for small samples.

We use ρ to denote the true correlation coefficient between a pair of gene expression profiles (Bickel and Doksum 2000). Specifically, let $X_{g_j(n)}$ be the n -th condition index of the i -th gene profile and let $S_{X_{g_i}, X_{g_i}}$, $S_{X_{g_j}, X_{g_j}}$, and $S_{X_{g_i}, X_{g_j}}$ are sample variances and covariance as in Eq. 2.2. The true correlation coefficient is defined as

$$(3.1) \quad \rho = \frac{\mathbb{E}[S_{X_{g_i}, X_{g_j}}]}{\sqrt{\mathbb{E}[S_{X_{g_i}, X_{g_i}}] \mathbb{E}[S_{X_{g_j}, X_{g_j}}]}}$$

where $\mathbb{E}[\cdot]$ is statistical expectation. For G gene expression profiles in a gene microarray sequence, there are $\Lambda = \binom{G}{2}$ of these correlation parameters ρ that need to be estimated, denoted as $\rho_\lambda, \lambda = 1, \dots, \Lambda$. We define $\hat{\rho}_\lambda$ as the λ th sample correlation coefficient, and $\hat{\Gamma}_\lambda$ as the hyperbolic arc-tangent transformation of $\hat{\rho}_\lambda$. Then the transformed sample correlation coefficients $\hat{\Gamma}_\lambda = \text{atanh}(\hat{\rho}_\lambda)$ are asymptotically Gaussian distributed with means of ρ_λ and stabilized variance approximations of $\sigma_\lambda^2 = 1/(N - 3)$ (Fisher 1923). As in Chapter II, N is the sample size. We define $\Gamma_\lambda = \text{atanh}(\rho_\lambda)$ as the corresponding transformed true correlation coefficients.

Simulation studies show that this variance approximation works reasonably well even at a relatively small sample size, i.e. $N < 10$ (Fig. 2.3b). In this sequel we assume known variance to reduce computational complexity. In case of unknown

variances, the conditional posterior distribution can not generally be written in closed form, for this reason, Markov Chain Monte Carlo (MCMC) techniques might be applied but at high cost.

From our assumption that the $\{\rho_\lambda\}_{\lambda=1}^\Lambda$ are *exchangeable* we model $\{\rho_\lambda\}_{\lambda=1}^\Lambda$ as random variables drawn from a Gaussian distribution with unknown hyperparameters (α, β^2) (Fig. 3.1).

$$(3.2) \quad p(\Gamma_1, \dots, \Gamma_\Lambda | \alpha, \beta^2) = \prod_{\lambda=1}^{\Lambda} P(\Gamma_\lambda | \alpha, \beta^2),$$

where $P(\Gamma_\lambda | \alpha, \beta^2)$ is a Gaussian distribution with mean α and variance β^2 .

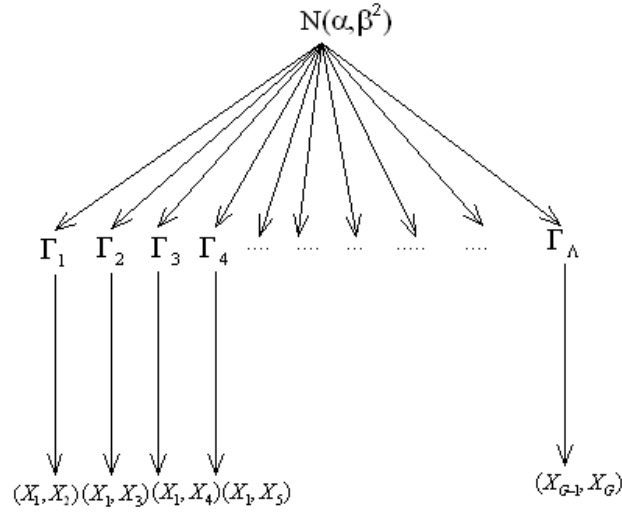


Figure 3.1: Bayesian hierarchical model structure (Gelman *et al.* 2004, Chapter V).

In order to generate conditional posterior distributions $p(\Gamma_\lambda | \alpha, \beta, y)$ for each parameter $\Gamma_\lambda, \lambda = 1, \dots, \Lambda$, we performed simulation steps as follows: (Gelman *et al.* 2004, Chapter V) (refer to Appendix A.5 for details):

- Assign prior distribution for β , e.g. uniform prior distribution $p(\beta) \propto 1$. Note, the choice of uniform prior yields a *proper* posterior density while other *noninformative* prior distributions such as, $p(\beta) \propto \beta^{-1}$ do not. (refer to Appendix

A.4 for mathematical proof.)

- Draw β from posterior distribution $p(\beta|y)$.

$$(3.3) \quad p(\beta|y) \propto \frac{p(\beta) \prod_{\lambda=1}^{\Lambda} N(\hat{\Gamma}_{\lambda}|\hat{\alpha}, \sigma_{\lambda}^2 + \beta^2)}{N(\hat{\alpha}|\hat{\alpha}, V_{\alpha})}$$

$$(3.4) \quad \propto p(\beta) V_{\alpha}^{1/2} \prod_{\lambda=1}^{\Lambda} (\sigma_{\lambda}^2 + \beta^2)^{-1/2} \exp\left(-\frac{(\hat{\Gamma}_{\lambda} - \hat{\alpha})^2}{2(\sigma_{\lambda}^2 + \beta^2)}\right),$$

where $\hat{\alpha}$ and V_{α} are defined as:

$$(3.5) \quad \hat{\alpha} = \frac{\sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_{\lambda}^2 + \beta^2} \hat{\Gamma}_{\lambda}}{\sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_{\lambda}^2 + \beta^2}},$$

and

$$(3.6) \quad V_{\alpha}^{-1} = \sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_{\lambda}^2 + \beta^2}.$$

See Appendix A.5 for detailed derivation of $p(\beta|y)$.

- Draw α from $p(\alpha|\beta, y)$. Combining the data with the uniform prior density $p(\alpha|\beta)$ yields,

$$(3.7) \quad p(\alpha|\beta, y) \sim N(\hat{\alpha}, V_{\alpha}).$$

where $\hat{\alpha}$ is a precision-weighted average of the $\hat{\Gamma}$'s and V_{α} is the total precision.

Note, we define precision as inverse of variance.

- Draw Γ_{λ} from $p(\Gamma_{\lambda}|\alpha, \beta, y)$

$$(3.8) \quad p(\Gamma_{\lambda}|\alpha, \beta, y) \sim N(\hat{\Theta}_{\lambda}, V_{\lambda}),$$

where $\hat{\Theta}_{\lambda}, V_{\lambda}$ are defined as:

$$(3.9) \quad \hat{\Theta}_{\lambda} = \frac{\frac{1}{\sigma_{\lambda}^2} \hat{\Gamma}_{\lambda} + \frac{1}{\beta^2} \alpha}{\frac{1}{\sigma_{\lambda}^2} + \frac{1}{\beta^2}},$$

and

$$(3.10) \quad V_\lambda = \frac{1}{\frac{1}{\sigma_\lambda^2} + \frac{1}{\beta^2}}.$$

The atanh-transformed posterior mean correlation coefficient $\hat{\Theta}_\lambda$ is a precision-weighted average of the prior population mean α and the λ th sample mean $\hat{\Gamma}_\lambda$.

The posterior distribution (Eq. 3.8) contains all the current information about the atanh-transformed parameter ρ_λ . In particular, the *posterior mean* and *posterior interval* are derived as the following:

$$(3.11) \quad \begin{aligned} E[\Gamma_\lambda] &= E[\operatorname{atanh}(\rho_\lambda)] \\ &= \operatorname{atanh}(E[\rho_\lambda]) = \hat{\Theta}_\lambda. \end{aligned}$$

Applying function \tanh to both sides of the Eq. 3.11, we have,

$$(3.12) \quad E[\rho_\lambda] = \tanh(\hat{\Theta}_\lambda).$$

For deriving the posterior interval of the ρ_λ , we used the fact that the cumulative density function (cdf) of $\Gamma_\lambda' = \frac{\Gamma_\lambda - \hat{\Theta}_\lambda}{\sqrt{V_\lambda}}$ is Φ , the cdf of standard Gaussian random variable. Hence, we define its quantile function as Φ^{-1} , and write down the $(1 - q) \times 100\%$ posterior interval of the parameter Γ_λ' :

$$(3.13) \quad I^{\Gamma_\lambda'}(q) : [\Phi^{-1}(q/2), \Phi^{-1}(1 - q/2)].$$

After some algebraic derivation and based on the fact that \tanh is a monotonically increasing function, we have a $(1 - q) \times 100\%$ posterior interval for the parameter ρ_λ :

$$(3.14) \quad I^{\rho_\lambda}(q) : [\tanh(\sqrt{V_\lambda}(\Phi^{-1}(q/2)) + \hat{\Theta}_\lambda), \tanh(\sqrt{V_\lambda}(\Phi^{-1}(1 - q/2)) + \hat{\Theta}_\lambda)].$$

3.2 Simulation Studies

3.2.1 Comparisons in terms of Confidence Interval, Mean Squared Error, and Variance

We evaluated the performance of full Bayesian hierarchical model estimation of correlations and compared with the frequentist method of last chapter. We define the frequentist CI as the following: If L and U are statistics (i.e., observable random variables) whose probability distribution depends on some unobservable parameter θ , and

$$Pr(L \leq \theta \leq U) = q, q \in (0, 1),$$

then the random interval $[L, U]$ is a $(1 - q) \times 100\%$ *confidence interval* for θ . A frequentist interval may strictly be interpreted only in relation to a sequence of similar inferences that might be made in repeated trials, while a *Bayesian (confidence) interval* for an unknown quantity of interest can be directly regarded as having a high probability of containing the unknown quantity. Therefore, Bayesian approach where reliable prior is available, facilitates a common-sense interpretation of statistical conclusions (Gelman *et al.* 2004).

We first compared two point estimators of correlations in terms of the average width of the individual frequentist (Pearson) CI's for the correlation parameters versus that of the posterior CI's for the same set of correlation parameters at the corresponding significance levels. Obviously, more concentrated (narrower) CI's, at the given significance level, are superior to less concentrated CI's. It is clear from Fig. 3.2 and Fig. 3.3 that the average Bayesian posterior CI's are uniformly narrower than the average frequentist CI's in both small ($N = 4$) and larger sample data ($N = 20$). This dramatic contrast indicates the advantages of Bayesian approach for small sample size problems (Fig. 3.3). From Eqs. 3.4 and 3.5, the posterior

distributions of the mean $p(\alpha|\beta, y)$ and of the variance $p(\beta|y)$ are decreasing functions of Λ , i.e., the number of correlation parameter Γ 's. Therefore, narrower posterior CI's are expected for larger Λ . On the other hand, wider CI's are expected when transforming individual frequentist CI's into simultaneous FDR-CI's.

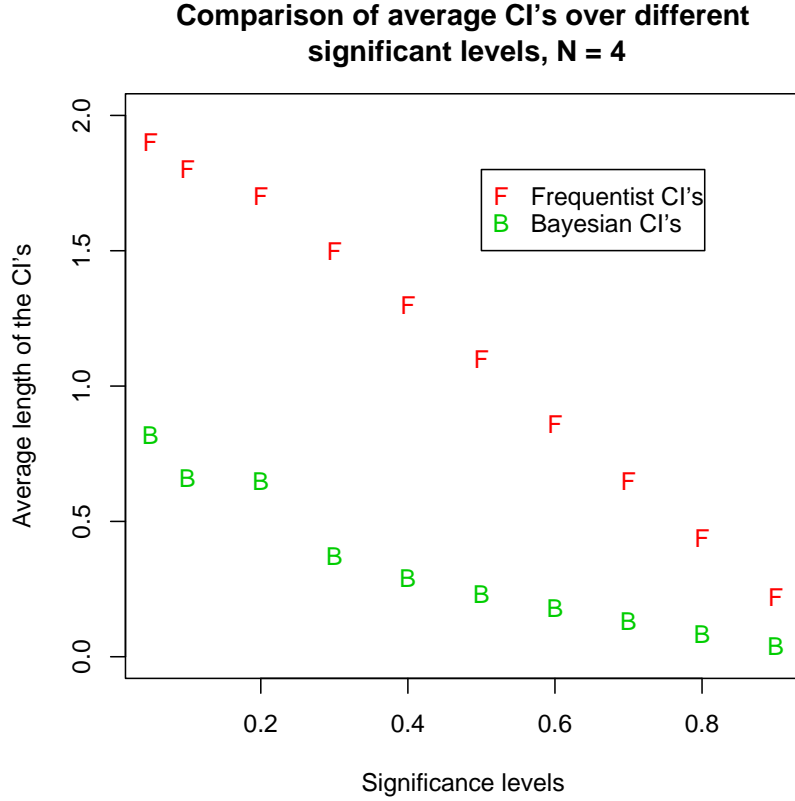


Figure 3.2: Comparison of average posterior CI's versus average individual frequentist CI's over a wide range of significance levels at a small sample size ($N = 4$).

We also compared these two correlation estimators in terms of Mean Squared Error (MSE) and variance criteria. Similar to Chapter II, the MSE is defined as:

$$(3.15) \quad MSE = \frac{1}{\Lambda} \sum_{\lambda=1}^{\Lambda} (\hat{\rho}_{\lambda} - \rho_{\lambda})^2,$$

where ρ_{λ} is the true population correlation, and $\hat{\rho}_{\lambda}$ is the sample correlation estimator, λ is the parameter index, and Λ is the total number of parameters.

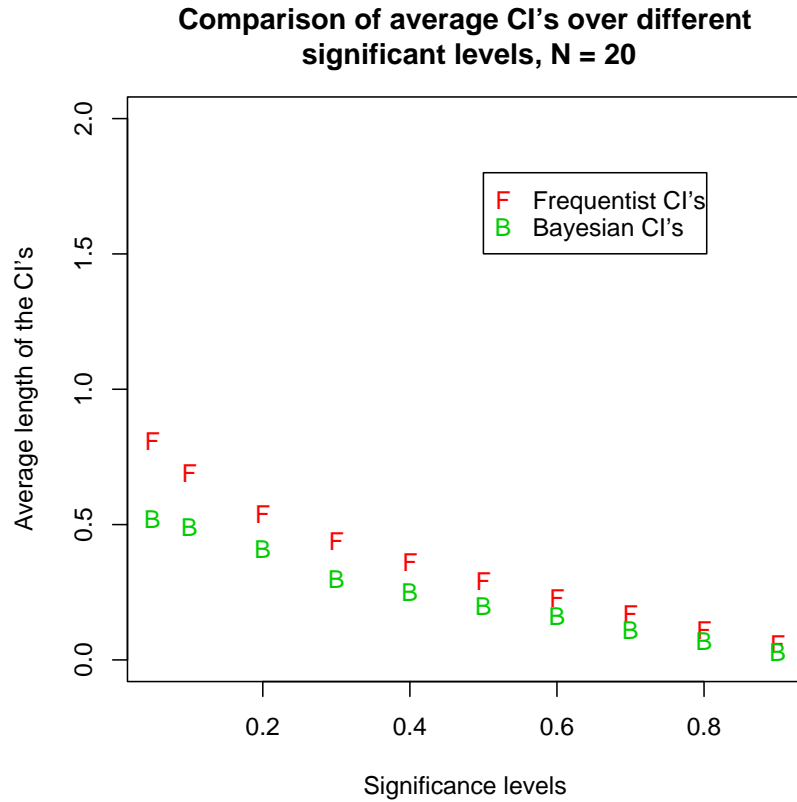


Figure 3.3: Comparison of average posterior CI's versus average individual frequentist CI's over a wide range of significance levels at a larger sample size ($N = 20$).

The simulation steps proceed as follows:

- Draw Λ population correlations from a normal distribution with known mean (α) and variance (β) (hyperparameters) as defined in Eq. 3.2.
- Re-estimate the Λ parameters either separately using the frequentist (Pearson) correlation estimator or using Bayesian hierarchical model. For the Bayesian approach, the correlation estimator is the posterior mean (Eq. 3.11).
- Compare the two estimators in terms of both MSE and variance. An estimator with low MSE and variance are considered to be superior.

Fig. 3.4 plots MSE's (upper panel) and variances (lower panels) of Bayesian corre-

lation estimators and frequentist (Pearson) correlation estimators at a small sample size (e.g. $N = 4$) and a larger sample size (e.g. $N = 20$) over 500 runs of simulations. It is evident in upper panel of the Fig. 3.4 that the MSE of Bayesian estimators is about three-fold smaller than the frequentist estimators for larger sample size. Similarly to the CI's comparisons, this indicates the advantages of the Bayesian correlation estimator for the small sample size problems (Fig. 3.4). The lower panel of the Fig. 3.4 plots variances of the Bayesian correlation estimator and the frequentist correlation estimator. Again, the comparison of variances follow the same trend as that of the MSE's (Fig. 3.4).

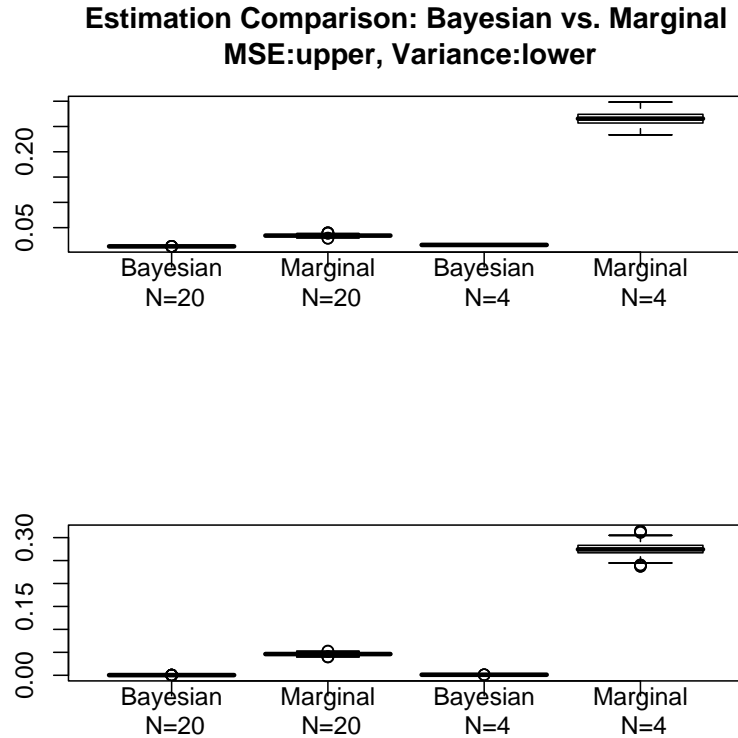


Figure 3.4: Mean Squared Errors (MSE's) and Variances of the Bayesian estimations versus the simple estimations over 500 runs of simulations.

It is worth mentioning that the above simulations were biased towards the assumptions of Bayesian hierarchical model. In order to test robustness of our algorithm to model mismatch, we also generated data using the uniform distribution but implemented with Pearson CI's and Bayesian CI's that assume mismatched Gaussian and hierarchical models, respectively. In Fig. 3.5, we compared the average width of individual Pearson CI's with that of individual Bayesian intervals. The superior performance of hierarchical Bayesian estimator (Fig. 3.2, Fig. 3.3) is clearly offset by the invalid model assumption in that average Bayesian CI's are uniformly wider than average frequentist CI's (Fig. 3.5). This simulation results highlight the importance of Fisher transformation in the section 3.1.

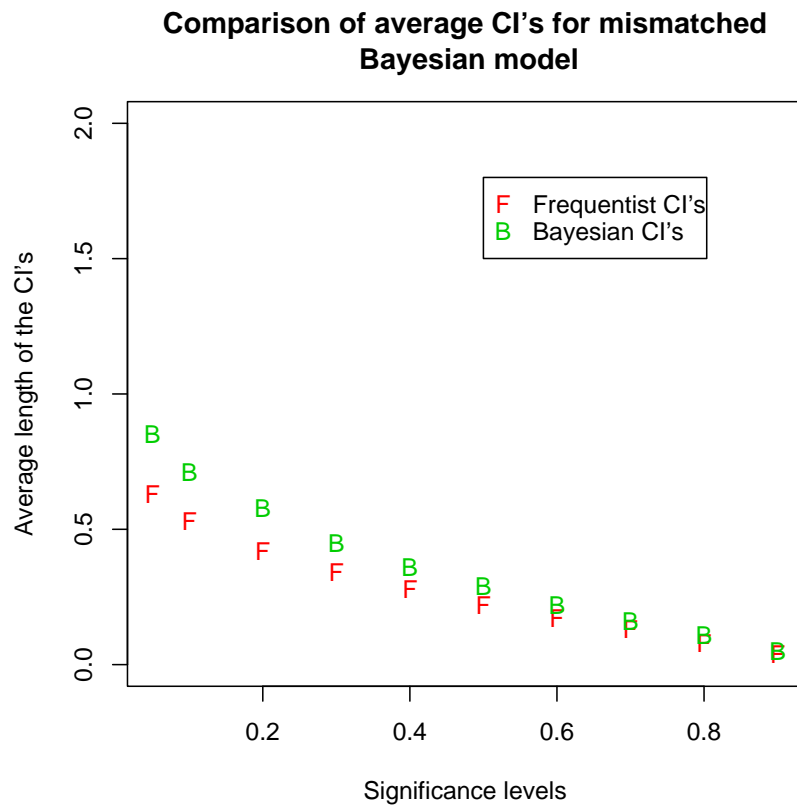


Figure 3.5: Comparison of average CI's when the Bayesian model is unsustainable.

3.2.2 Posterior Predictive Model Checking

After fitting a Bayesian model, one common practice is to check whether the model is consistent with the data. In order to examine the goodness of fit, we generated posterior predictive distributions of the following statistics: the largest observed correlation (max), the smallest observed correlation (min), the mean of observed correlations (mean), and the standard deviation of the observed correlations (sd). We approximated the posterior predictive distribution of each test statistic by the histogram of values obtained from simulations of the parameters and generated data samples, and we compared each distribution to the observed value of the test statistic. The results were displayed in Fig. 3.6 in which four boxplots represent empirical null distributions of four test statistics. Define T_0 is the test statistic calculated from observations, and H_0 is the null hypothesis, then the p -value of a test statistic is defined as:

$$(3.16) \quad p = Pr(|T| > T_0 | H_0).$$

The posterior predictive model checking results reveal the remarkably high accuracy and low variance of Bayesian estimation.

3.2.3 Evaluation of the Bayesian Hierarchical Model

In order to evaluate our Bayesian approach in terms of error control and compare with the frequentist counterpart, we simulated pairwise gene expression data based on known population covariances (Appendix A.3), and then simulated Bayesian intervals for each parameter from the hierarchical model. The actual False Positive (FP) at a given MAS level is calculated as a ratio of the number of screened gene pairs whose corresponding population correlation parameters $\rho_{i,j}$ are less than the MAS level specified, divided by the total number of gene pairs. The actual MAS is the

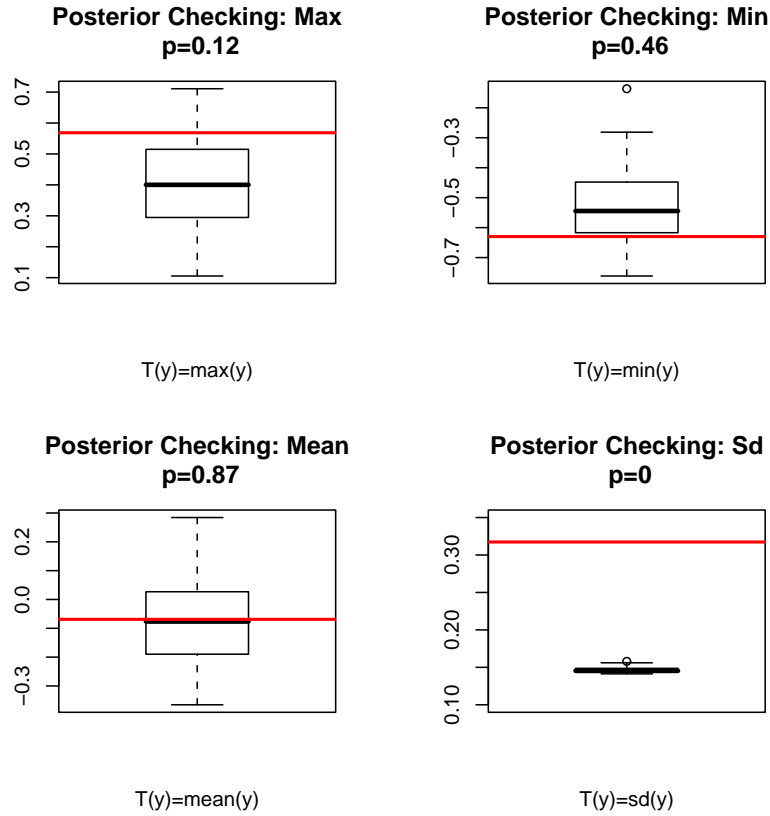


Figure 3.6: Posterior predictive distribution, observed results (red line), and p -value for each of four test statistics.

minimum true discovery of population correlation $\rho_{i,j}$ among the screened pairs. We specified 16 pairs of (FP,MAS) criteria (Four FP levels: 0.2, 0.4, 0.6, 0.8; Four MAS levels: 0.2, 0.4, 0.6, 0.8), and each is plotted as a different upper case Roman alphabet (Red) in Fig. 3.7. The 16 corresponding pairs of actual (FP,MAS) criteria are also shown in Fig. 3.7 using the same set of lower case Roman alphabets (Blue). It can be observed that generally the actual FP's (lower case) fall further below the specified constraint (upper case) than those did in Chapter II (Fig. 3.7, Fig. 2.4), and the actual MAS's (lower case) fall above the specified constraints (upper case). The more dramatic deviations of actual FP's from their specified levels are due

to multiple factors, such as, lack of multiplicity adjustment and the conservative asymptotic approximation. Simulations using some other combinations of N and Λ , as compared with the FDR-CI approach, give rise to the similar results. We conclude that Bayesian hierarchical model yields better correlations estimates. However, the false positive rate is overestimated by the Bayesian procedure and this leads to overly stringent error control.

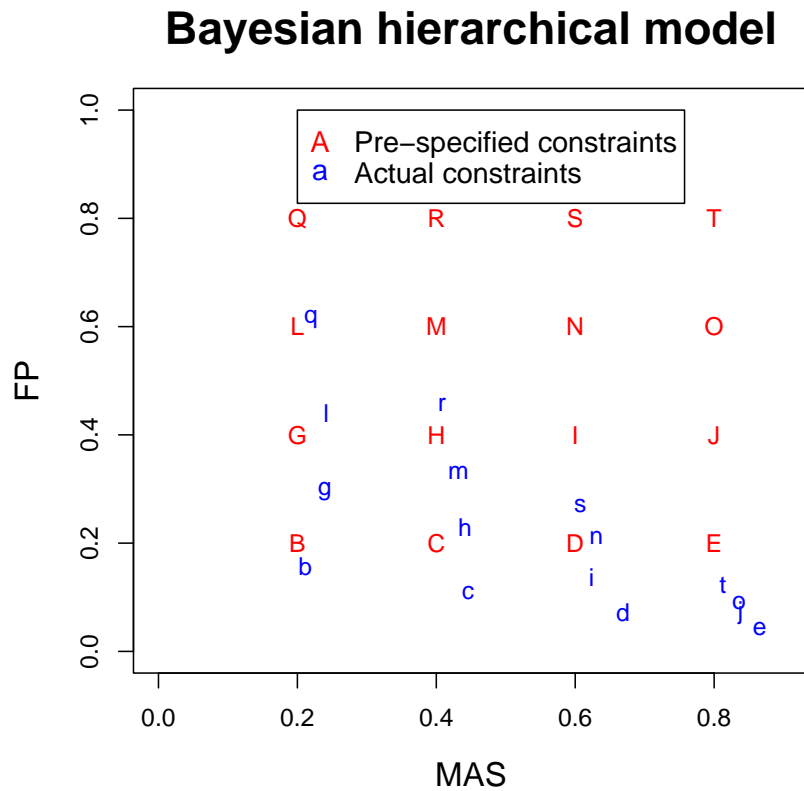


Figure 3.7: Evaluation of error control of the Bayesian hierarchical model. Sample size $N = 20$, and $\Lambda = 1000$ correlation coefficients were simulated. Simulations using smaller sample size data yield more stringent error control.

3.3 Applications to Network Construction and Seeded Clustering

3.3.1 Constructing Relevance Networks

We demonstrate the application of the Bayesian hierarchical model to high throughput data and compare it with the frequentist approach using the same subset of yeast galactose metabolism two-color microarray data that was described in Chapter II. The data contains 997 gene expression profiles across 20 genetic/physiological conditions that was identified by Ideker et al using the generalized likelihood ratio test (Ideker *et al.* 2000). This is the same data used for Chapter II.

Following the procedure described in section 3.1, we simulated the empirical posterior distribution for each of the $\binom{997}{2} = 496,506$ correlation parameters ρ_λ . The $(1 - q) \times 100\%$ *posterior interval* for each ‘parameter’ was obtained by thresholding $q/2 \times 100\%$ and $(1 - q/2) \times 100\%$ of it’s quantile function (Eq. 3.14). Analogous to the FDRCI screening procedure described in the Chapter II, a network edge is declared to be present at the significance level q and the MAS level *cormin* if it’s posterior CI does not intersect with $[-cormin, cormin]$. We sought to compare the two approaches in terms of network topological properties that are interesting to the biologists. In particular, we compared the biological functional annotations of the top hub genes of the two networks. In Chapter II, we controlled FDR at 5%, and constructed networks at five MAS levels, i.e. 0.5, 0.6, 0.7, 0.8, 0.9. Correspondingly, 18135, 9337, 4151, 1346, 133 edges were declared to be present using Pearson correlation statistic alone. Controlling the significance level at 5%, we screened the same set of numbers of edges using Bayesian hierarchal model to construct the five networks that are more comparable to those in Chapter II. A list of stable hub genes were obtained by calculating and sorting the average rank of each vertex (gene) degree over five networks (Table 3.1).

Comparing the Table 3.1 with the Table 2.3, note that the GO biological process annotation “protein biosynthesis[GO:0006412]” and/or its children annotations “hypusine biosynthesis[GO:0046515]”, “branched chain family amino acid biosynthesis[GO:0009082]”, and “tryptophan biosynthesis[GO:0000162]” are significantly enriched in both tables. This is consistent with the established fact that protein biosynthesis plays a key role in galactose metabolism (Berg *et al.* 2006). The underlying biological mechanism is that many types of proteins need to be synthesized upon switching from primary carbon source (glucose) to secondary carbon source (galactose)(Wieczorke *et al.* 1999).

A salient feature in Table 3.1 that is not possessed in Table 2.3 is that it includes several transporters and regulators such as GAP1[GO:0006865], YBR043C[GO:0006855], and ASC1[GO:0006417] etc. These proteins are essential for a smooth transition from glucose to galactose (Berg *et al.* 2006, Wieczorke *et al.* 1999). In addition, Table 3.1 also includes several biologically unknown genes that are hypothesized to be important for galactose metabolism. In general, the Bayesian data analysis results not only conform to the previous frequentist data analysis results, but also provide additional justification for the biological mechanism and motivation for illustrating new gene functions.

3.3.2 Seeded Clustering

In parallel with the application of the two-stage algorithm to rediscover the galactose metabolic pathway reported in Chapter II, we also applied the Bayesian hierarchical model to perform the seeded clustering. Performance was evaluated according to the relative ranks of a handful biologically known genes lying in the galactose metabolic pathway. The gene ranks were used instead of p -values due to substantial differences of the two statistical frameworks.

Table 3.1: Top twenty “hub genes” from Bayesian hierarchical model applied to the galactose metabolism data (Ideker *et al.* 2000). The rank of each gene is the average rank over five different networks with the same set of edge numbers as in Table 2.3. The highest ranked gene is the most connected and stable gene under varying constraints of (FP,MAS).

Gene Name	Average Rank	GO Annotation
YJR070C	4	hypusine biosynthesis[GO:0046515]
YBR043C	4.4	multidrug transport[GO:0006855]
AGA2	4.4	agglutination[GO:0000771]
RPP0	4.6	protein biosynthesis[GO:0006412]
RPL26A	4.6	protein biosynthesis[GO:0006412]
YOR263C	5	biological process unknown
TRP2	5.4	tryptophan biosynthesis[GO:0000162]
ASC1	5.6	regulation of protein biosynthesis[GO:0006417]
YIL064W	5.6	biological process unknown
BOP2	5.6	biological process unknown
GAP1	5.8	amino acid transport[GO:0006865]
RPS2	6	protein biosynthesis[GO:0006412]
RPL11A	6.2	protein biosynthesis[GO:0006412]
SSF2	6.2	ribosomal subunit assembly[GO:0042257]
ILV5	6.2	branched chain family amino acid biosynthesis[GO:0009082]
YPL185W	6.2	biological process unknown
PCK1	6.4	hexose biosynthesis[GO:0019319]
YDR100W	6.4	biological process unknown
YMR291W	6.6	biological process unknown
ATC1	6.6	bipolar bud site selection[GO:0007121]

We selected gene “GAL10” as the “seed gene” in order to compare the results with those reported in Chapter II. The comparison was made at a large sample size $N = 20$ and a smaller sample size $N = 4$ respectively aiming to examine the performance of the two methods as a function of the sample size. In the former, we used all the 20 genetic/physiological conditions under which gene expression levels were measured (Table A.6); In the later, we sampled a small subset (e.g. $N = 4$) of these 20 conditions each time without replication and repeated a number of times to obtain a “bagged” (stable) estimation of gene ranks in the seeded clusters (Table 3.2).

When all the 20 observations were used, the two approaches give rise to very similar seeded clusters indicating that the Bayesian hierarchical model approach is as powerful as the frequentist approach for relatively large sample size problems. As shown in Table 3.2, all of the top 20 seeded gene pairs have the identical rank across two methods. When multiple random subset data were used, many genes have dissimilar average ranks across the two approaches. Among the top five genes (GAL10, GAL7, GCY1 GAL1, GAL2) screened by the seeded clustering using “GAL10” as the seed gene (see Chapter II and Table A.2), 4 out of 5 (GAL10, GAL7, GAL1, GAL2) genes rank higher in Bayesian estimation than those in marginal estimation, and the remaining “GCY1” gene receives tie ranks. In addition, our results provide strong experimental motivation for examining the genes that received higher ranks in the Bayesian analysis, for example, gene YEL057C. The evaluation using “GAL7” as the “seed gene” gave the similar results.

3.4 Discussion

Numerous previous studies have demonstrated the suitability of using gene co-expression networks for functional discoveries (e.g. Butte and Kohane 2000, Zhou

Table 3.2: Comparison of Bayesian estimations versus Marginal estimations using “seeded” clustering at a small and a larger sample sizes. In the former, the ranks were averaged over 100 estimations, in each of which a subset data of sample size $N = 4$ was randomly sampled from the whole data of sample size $N = 20$. In the later, the ranks were obtained using the whole data of sample size $N = 20$.

		$N = 4$				$N = 20$	
Gene1	Gene2	Bayesian	Frequentist	Gene1	Gene2	Bayesian	Frequentist
GAL10	GAL1	5.25	5.35	GAL10	GAL7	1	1
GAL10	GAL2	6.65	7.4	GAL10	GCY1	2	2
GAL10	GAL7	6.7	6.85	GAL10	GAL1	3	3
GAL10	GCY1	7.7	7.7	GAL10	GAL2	4	4
GAL10	YOR121C	8.05	7.8	GAL10	YOR121C	5	5
GAL10	YEL057C	8.55	10.6	GAL10	YEL057C	6	6
GAL10	SSU1	8.6	7.65	GAL10	YDR010C	7	7
GAL10	FKS1	8.75	8.25	GAL10	SSU1	8	8
GAL10	PCL10	9.95	7.85	GAL10	PCL10	9	9
GAL10	YJL212C	11	8.85	GAL10	YJL212C	10	10
GAL10	MET14	11.1	10.4	GAL10	FKS1	11	11
GAL10	YDR010C	11.3	10.9	GAL10	MET14	12	12
GAL10	MCM1	11.35	12.3	GAL10	MCM1	13	13
GAL10	EXG1	11.85	13.1	GAL10	EXG1	14	14
GAL10	CRH1	12.05	12.95	GAL10	ARG1	15	15
GAL10	ARG7	12.8	12.3	GAL10	CRH1	16	16
GAL10	YPR157W	13.2	15.35	GAL10	PRY2	17	17
GAL10	PRY2	14.4	13.3	GAL10	YPR157W	18	18
GAL10	YKR012C	14.6	16.25	GAL10	YKR012C	19	19
GAL10	CPA2	16.15	14.85	GAL10	CPA2	20	20

et al. 2002). The differences lie in the co-expression network construction methodology, i.e. the correlation statistic and the error control procedure used. For the error control procedure, the main difference from existing approaches is that we test whether the magnitude correlation is different from 0 or a non-zero positive number. We noted that for small sample size the frequentist (Pearson) test declares many small but statistically significant correlations to be biologically relevant. However, these may be caused by non-biological effects such as spatial and positional effects of genes along the chromosome (Kluger *et al.* 2003).

Fuente *et al.* proposed a limited order partial correlation estimation that is dependent on a fixed number of neighboring nodes (order) in the network (Fuente *et al.* 2004). This approach is statistically sound, and it accounts for biological knowledge that the functional relationship between a gene pair is typically regulated by only a small number of surrounding genes in the gene regulation network. However, the application is limited by a number of empirical difficulties, such as: estimating the true order for tens of thousands of partial correlation parameters, and estimating degree of freedom for null distribution. Schafer and Strimmer's full order partial correlation approach estimates the correlation between a gene pair conditioning on all the rest of genes in the network (Schafer and Strimmer 2005a). The Gaussian Graphical Model (GGM) approach, while effective in variance reduction, may be an overly conservative way of correlation estimation. The implicit biological assumption that the functional association of a gene pair is dependent on all the other genes in the gene regulation network does not seem to have adequate biological support.

As discussed in the previous chapter, one should seek a good combination of level of significance and correlation strength. The Bayesian approach prescribed here imposes a model of the parameters as random variables sampled from a parental

population distribution. This model structure allows the regularization of variances by introducing dependency between the parameters. Using simulations, we have shown the superior performance of Bayesian hierarchical model approach to marginal estimation approach, in terms of width of the CI's, MSE and variance, especially for small sample size. The posterior distribution provides a natural way of correlation thresholding that bridges between statistical correlation and biological relevancy.

In deriving the posterior distributions of the correlation 'parameters', the conjugate prior and likelihood (i.e. Gaussian parental distribution) were assumed in order to keep the posterior distributions in a closed form. The computational load is thus greatly reduced and we avoided MCMC techniques, making the application to the larger networks become more feasible.

CHAPTER IV

Network Constrained Clustering

In the Chapters II and III, we presented a pair of complementary approaches to infer gene interaction network topology. A pair of genes in the network can be either directly associated or indirectly associated through one or more intermediate genes. The partially connected network structure can be viewed as a network constraint that can be used as side information to improve gene clustering performance. Gene clusters group genes according to similar function. We focus on imposing network constraints in gene clustering in this chapter, and we will describe how we might impose network constraints in ordering pathway components in the Chapter V. In network constrained clustering, we use the shortest-path distance as the estimate of distance between non-adjacent genes in the network. Network constrained clustering proceeds in two consecutive steps (Zhu *et al.* 2005c, Zhu and Hero 2005d, Zhu and Hero 2005e). First we extract a giant connected component. Then we calculate a “network constrained pairwise distance matrix” from which clustering is accomplished.

4.1 Network Constrained Clustering - Method

4.1.1 Extract the Giant Connected Component

Only gene pairs that are in the same Connected Component (CC) of the relevance network have finite distances and can be clustered. The CC is defined as the cluster of nodes within which any pair of nodes is mutually reachable from each other. The largest connected component, usually of importance to both biological function and network topology (Ma *et al.* 2004, Ma *et al.* 2004, Zhu and Qin 2005), is called the Giant Connected Component (GCC) (see Fig. 1.3c, genes A, B, C, D, E, F form a GCC). The GCC of an undirected graph $G = (V, E)$, where V is the set of all vertices and E is the set of all edges, is the maximal set of vertices $U \subset V$ such that every pair of vertices u and v in U are reachable from each other. Our network constrained clustering method is applied to the GCC. Analogously to previous studies, we assume that almost all important genes are included in the GCC. The standard depth first search (DFS) algorithm (Cormen *et al.* 1990) was used to extract the GCC from the gene microarray data.

4.1.2 Compute “Network Constrained Distance Matrix”

Let $\hat{\Gamma}_{ij}$ be the sample correlation coefficient between gene i and j , e.g. estimated from a gene microarray sequence by Pearson or Kendall correlation statistic. Let w_{ij} be the weight of the edge between gene i and gene j . Similar to Zhou et al (Zhou *et al.* 2002), the w_{ij} is defined as:

$$(4.1) \quad w_{ij} = (1 - \text{abs}(\hat{\Gamma}_{ij}))^p$$

The integer p is an exponential tuning parameter used to enhance the differences between low and high correlation. We define the matrix $W = [w_{ij}]$ as the “Traditional distance matrix” (e.g. Fig. 1.3b).

We use the standard Floyd-Warshall algorithm to search among all-pairs for the shortest-paths within the GCC. Let $d_{ij}^{(k)}$ be the weight of a shortest-path from vertex i to vertex j such that all intermediate vertices on the path (if any) are in set $\{1, 2, \dots, k\}$. When $k = 0$, there is no intermediate vertex between vertices i and j , and we define $d_{ij}^{(0)} = w_{ij}$. A recursive definition of $d_{ij}^{(k)}$ is given by (Cormen *et al.* 1990):

$$(4.2) \quad d_{ij}^{(k)} = \begin{cases} w_{ij} & \text{if } k = 0, \\ \min(d_{ij}^{(k-1)}, d_{ik}^{(k-1)} + d_{kj}^{(k-1)}) & \text{if } k \geq 1, \end{cases}$$

where $d_{ij}^{(k-1)}$ is the length of shortest-path when k is not a vertex on the path, and $d_{ik}^{(k-1)} + d_{kj}^{(k-1)}$ is that k is a vertex on the path. We define the matrix $D = [d_{ij}]$ as the “Network constrained distance matrix” (e.g. Fig. 1d). It can be used as input to many distance matrix based clustering software packages such as: hierarchical clustering (Eisen *et al.* 1998) and K -medoids (Hartigan and Wong 1979). The calculation of matrix D can be easily extended to higher Eukaryote since the algorithm runs in polynomial time, i.e. $O(V^3 + V + E)$.

The above algorithm uses the shortest-path between a gene pair to approximate the corresponding geodesic distance. The geodesic approximation is motivated by the fact that locally a smooth manifold is well “approximated” by a linear hyperplane, and so, geodesic distance is estimated by summing the sequence of such local approximations over the shortest-path through the GCC (Costa *et al.* 2004, Silva *et al.* 2002). Interested readers should refer to Costa *et al.* 2004 for mathematical proof.

We note that the network constrained clustering can be also performed on the whole large-scale network that is composed of many other smaller connected com-

ponents. The above algorithm for computing pairwise distance remains unchanged if the pair of genes lie in the same connected component. Otherwise, we set the distance finitely large compared with within-component distances.

4.2 Network Constrained Clustering - Results

4.2.1 Sensitivity Analysis

The FDR, MAS, and exponential tuning parameter p are three parameters involved in calculating the network constrained distance matrix. It would be interesting to investigate the sensitivity of the results to variance in these parameters. The biological significance level $MAS = 0.6$ has been widely adopted as a correlation cut-off in the literature, e.g. Zhou *et al.* 2002, Zhou *et al.* 2005. The selection of FDR statistical significance level is intimately associated with the sample size, and the underlying biological mechanism. Our selection of $FDR = 5\%$ imposes the stringent statistical criterion that on the average only 5% of the genes discovered and included in the network will be false positives.

The parameter p in Eq. (4.1) is an exponential tuning factor used to enhance the differences between expression similarity and dissimilarity. As pointed out by Zhou *et al.* (Zhou *et al.* 2002), for a fixed correlation threshold, as p is increased more transitive genes will be revealed at the expense of higher false discovery rate. In Fig. 4.1 we present results of an empirical study of the influence of p on clustering performance for the yeast galactose metabolism data set (Ideker *et al.* 2000).

A subset of 205 gene expression profiles whose Gene Ontology (GO) annotation (Ashburner *et al.* 2000) falls into four functional classes were used (Yeung *et al.* 2003). We investigate the effect of p by examining how closely the clusters reproduce these functional classes as p varies. We used both the RAND index (Rand 1971) and the adjusted RAND index (Hubert *et al.* 1985) as measures of consistency between the

clustering results and GO annotations. Fig. 4.1 shows that the network constrained clustering best conforms to the GO annotations when $p = 6$. Note that Zhou et al. also suggested using $p = 6$ to define the edge weight in their analysis (Zhou *et al.* 2002).

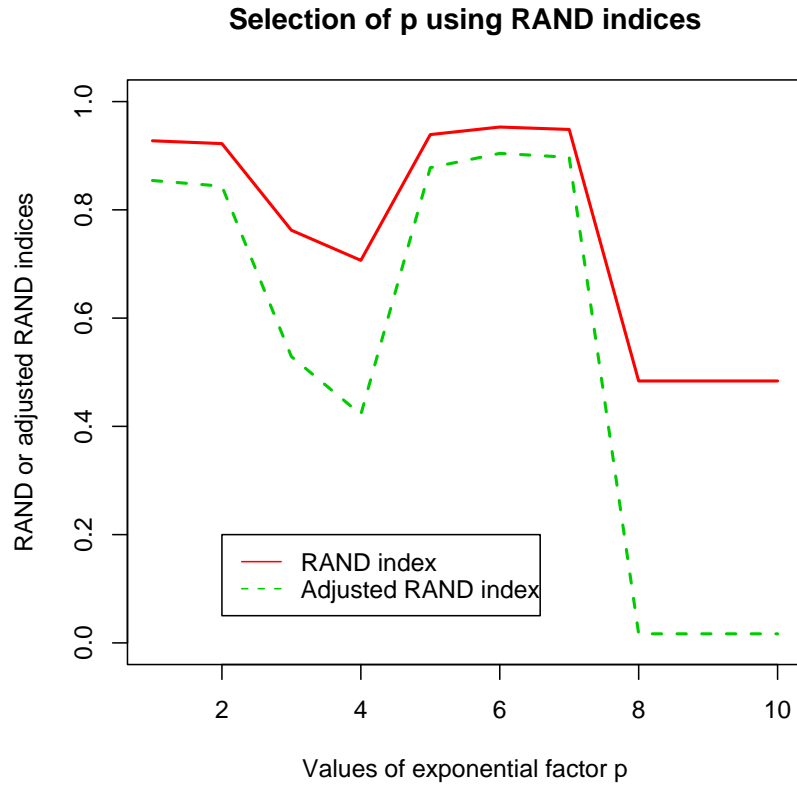


Figure 4.1: Selection of p using RAND indices.

4.2.2 Yeast Galactose Metabolism Data

Data Processing and Network Construction

We empirically evaluated the performance of the proposed clustering approach by applying it to a relatively well-known yeast galactose metabolism signaling pathway and comparing it with the traditional clustering approaches. We used the subset of 997 genes that were identified by Ideker et al using a standard generalized likelihood

ratio testing procedure (Ideker *et al.* 2000). Genes having a likelihood statistic $\lambda \leq 45$ were selected as differentially expressed and whose mRNA levels differed significantly from reference under one or more perturbations.

By measuring the pairwise gene correlations using both Pearson and Kendall correlation coefficients, we applied a two-stage algorithm to screen gene pairs with $\text{FDR} \leq 5\%$ and $\text{MAS} = 0.6$ (Zhu *et al.* 2005a). The resulting network is a mixed network within which edges are discovered with Pearson and Kendall correlation statistics. Our network construction algorithm and the screening criteria ensures false discovery of no more than 5% of the edges having strength of association greater than 0.6 (Zhu *et al.* 2005a).

Network Constrained Clustering

We extracted the GCC from the co-expression network using a DFS type algorithm (see Methods). The GCC contains 772 genes within which almost all known structural genes in the pathway are included. This confirms the notion that GCC of the network has not only structural but also functional significance (Ma *et al.* 2004, Ma and Zeng 2003, Zhu *et al.* 2005a). The network constrained distance matrix for GCC was computed according to Eq. 4.1 and Eq. 4.2 using GCC selected genes (see Methods) while the distance matrix for the traditional clustering method was computed according to Eq. 4.1 only.

As mentioned in Chapter II, the yeast galactose metabolism pathway consists of at least three types of genes including transporter genes (GAL2, HXT1-10, the roles of other HXT genes are not entirely clear), enzyme genes (GAL1, GAL7, GAL10 etc) and transcription factor genes (GAL3, GAL4, GAL80 etc) (Wieczorke *et al.* 1999). Transcription factor genes are not discoverable from this microarray experiment as their expressions are typically time shifted and only one time sample was included.

Since the pathway has been relatively well studied, we sought to compare our network constrained clustering approach with the traditional clustering approach through rediscovering the 14 important genes in the structural module (GAL2, HXT1-10, and enzyme genes: GAL1, GAL7, GAL10) of the yeast galactose metabolism pathway.

For comparison to a widespread clustering algorithm we used agglomerative hierarchical clustering (implemented in R function `hclust()`). We expect that other traditional clustering methods such as K -means or K -medoids would give similar results. For calculating distance between clusters, we implemented a “complete” method in which the longest geodesics between genes in the two clusters are used as distance between clusters. As empirically demonstrated in (Speed 2003), the “complete” method gives rise to better cluster separation.

Fig. 4.2 shows the traditional clustering approach using all 997 genes and Fig. 4.3 shows the traditional clustering approach using the 772 genes in the GCC. In both cases, the 14 structural genes are separated into three subclusters (Fig. 4.2 and Fig. 4.3). In Fig. 4.2, all GAL genes are nicely grouped in a cluster, but not the HXT genes, while in Fig. 4.3, all HXT genes are grouped into a single cluster, but the algorithm failed to combine GAL gene clusters with HXT gene clusters. Fig. 4.2 and Fig. 4.3 show that the GCC gene selection process has some desirable effects on clustering by removing a few unrelated genes (Tseng and Wong 2005) that are not relevant to the biological pathway. However, using the GCC gene selection procedure alone does not significantly improve clustering performance.

We think that this undesirable separation of genes in the pathway is due to the presence of gene expression dissimilarity **between** subclusters and gene expression similarity **within** subclusters. To test this hypothesis, we plotted the correlation matrix of 14 genes in the structural module and did hierarchical clustering (Fig. 4.4).

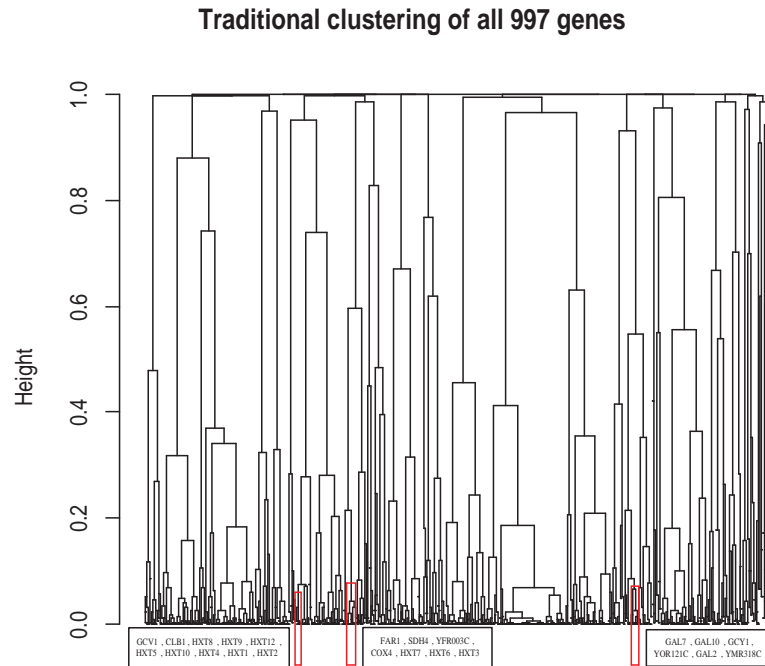


Figure 4.2: Traditional clustering: agglomerative hierarchical clustering using all 997 differentially expressed genes. The 14 structural genes are separated into three clusters (red rectangular).

The color intensities in Fig. 4.4 correspond to the levels of correlations (increasing correlations are represented from yellow to red). It is evident from Fig. 4.4 that expression correlations within GAL genes and HXT genes are much higher than the correlations between the two groups. This explains the separation of these two gene clusters in the associated clustering dendrogram (Fig. 4.2 and Fig. 4.3). Among the HXT gene clusters, HXT3, HXT6 and HXT7 are highly correlated (red (dark) zone in Fig. 4.4). It explains the actual separation of these three genes from the remaining HXT genes shown in the clustering dendrogram (Fig. 4.2). Fig. 4.2, Fig. 4.3 and Fig. 4.4 showed that traditional clustering methods failed to group functionally

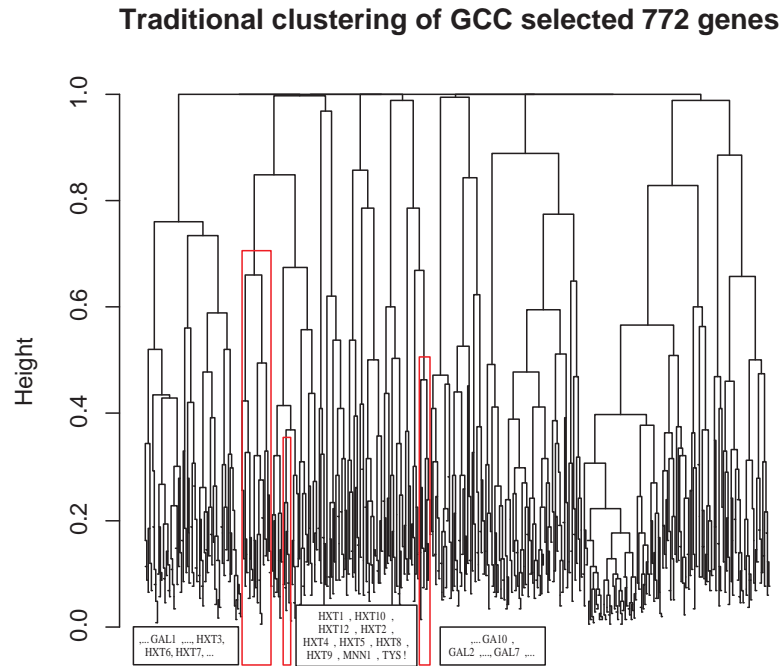


Figure 4.3: Traditional clustering: agglomerative hierarchical clustering using the GCC selected 772 genes. The 14 structural genes are separated into three clusters (red rectangular). Dots indicate incomplete clusters are shown due to space limitation.

related genes with dissimilar expression profiles (low correlations) into one cluster.

Fig. 4.5 presents results of applying our network constrained clustering algorithm to the 772 genes selected by GCC extraction. Note that all 14 structural genes that failed to be clustered together by the traditional approach (Fig. 4.2) are grouped into a single tight cluster by the network constrained clustering approach. As has been shown, the GCC selection process contributes only moderately to the apparent success. This demonstrates that employment of the network constrained distance matrix can lead to significant improvement in clustering performance.

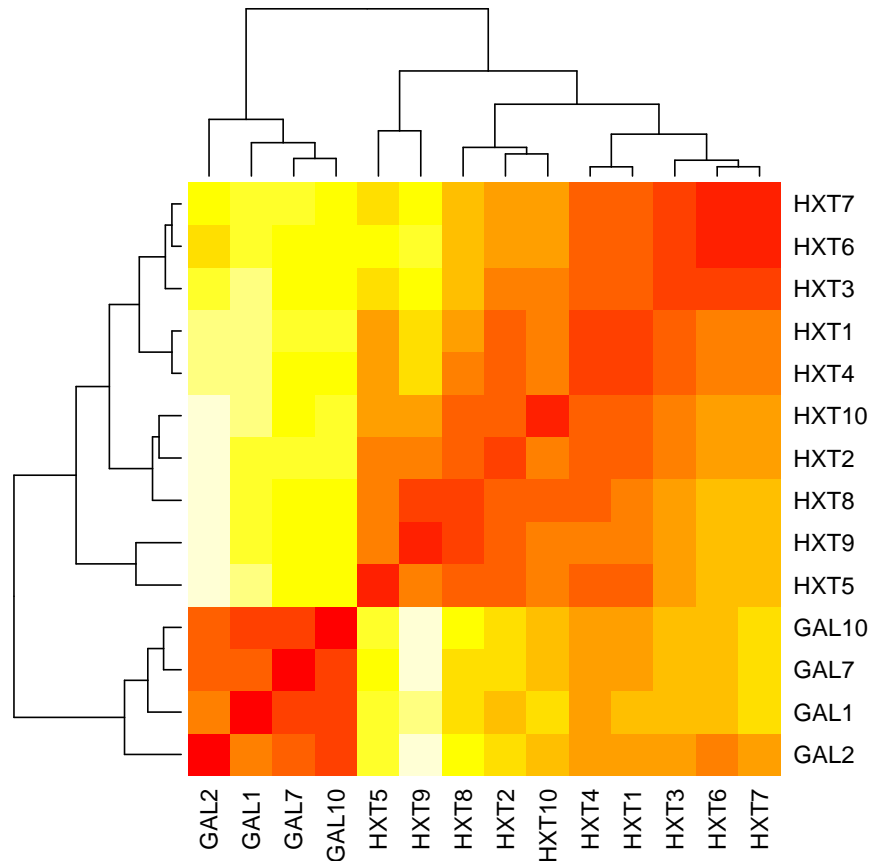


Figure 4.4: Correlation matrix of 14 structural genes with clustering dendrogram. White to grey corresponds to the low correlations to high correlations.

4.2.3 Retinal Gene Expression Data

The aim of the retinal gene expression experiment is to investigate the gene pathway of photoreception differentiation during retinal development and to discover unknown genes related to this pathway. The retinal data represents a total of 45,101 gene expression profiles over five time points measured in both wide type and Nrl (Swaroop *et al.* 1992)(the Maf-family transcription factor, key regulator of photoreceptor differentiation in mammals) knockout mice (Akimoto *et al.*, 2005). The data is available from the NCBI Gene Expression Omnibus (GEO) with accession ID: GSE4051.

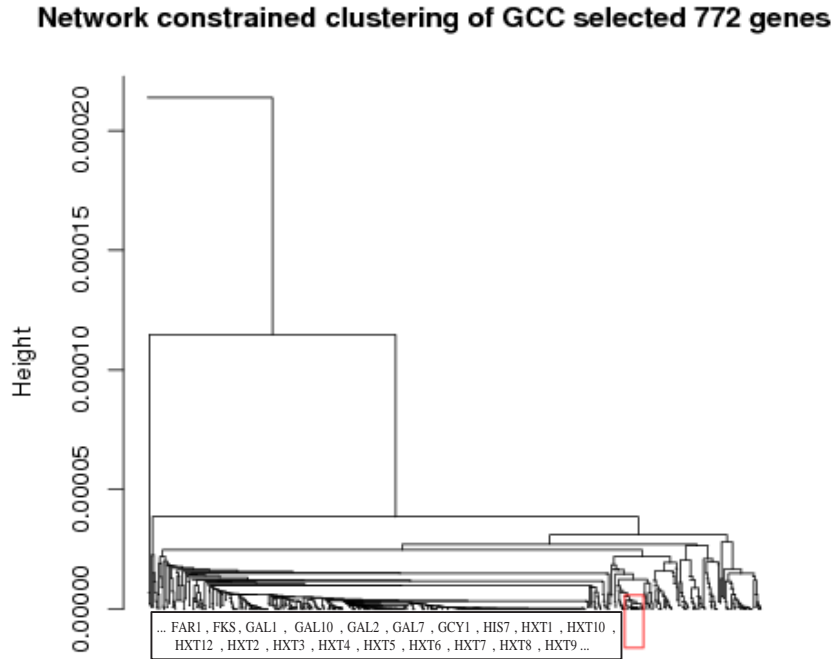


Figure 4.5: Network constrained clustering: agglomerative hierarchical clustering using network constrained distance matrix calculated from relevance network (Eq. 4.2).

The data was preprocessed using the “rma” method (Bolstad *et al.*, 2003), and it was subjected to an initial screening using the two-stage screening method proposed by Hero *et al.* (Hero *et al.*, 2004) in which the top 1000 genes ranked by FDR and Fold Change are kept for further analysis. We constructed a co-expression network similarly to the yeast analysis ($FDR \leq 5\%$ and $MAS = 0.6$) in the last subsection. A GCC of size 790 genes was extracted. These 790 genes were used in our NC clustering algorithm according to Eq. 4.1 and Eq. 4.2 while the total 1000 genes were used in the traditional hierarchical clustering algorithm according to Eq. 4.1 only.

As above we used GO annotation as the objective criteria to compare the two

clustering approaches. GO is a set of standard hierarchical vocabularies to describe the *biological process*, *molecular function* and *cellular component* of genes. It is conveniently represented as a graph where nodes represents standard vocabularies and edges represent the relationship (either “is-a” or “part of”) between vocabularies. A child node is the more specific vocabulary than its parent node(s). A list of probe sets obtained from any clustering method can be mapped to a GO graph (e.g. *biological process* graph), the appearance counts of all nodes of the GO graph can be calculated as well as their p -values of chi-square statistics. The most significant node(s)(corresponding to the smallest p -value(s)) usually describe(s) the biological functions of the probe set list. Specifically, all genes that having GO annotation “visual perception [GO:0007601]” are expected to belong to photoreceptor differentiation pathway.

We thoroughly compared the two clustering results with respect to three criteria (appearance counts, separation and p -values of the GO category: visual perception) at each cluster number ranging from 1 to 20. Only the largest 20 clusters were investigated as the remaining clusters contained fewer than 5 genes. The first two clustering criteria measure stability of the “visual perception” cluster as a function of cluster numbers, and the third criterion measures the enrichment of the interested GO vocabulary as a function of cluster numbers. Fig. 4.6 and Fig. 4.7 demonstrate the “visual perception” cluster acquired by NC clustering is quite stable over different cluster numbers but not that acquired by traditional clustering. Fig. 4.8 demonstrates that the interested GO vocabulary “visual perception” is much more enriched by NC clustering over different cluster numbers. In Fig. 4.6, the initial (cluster number=1) count difference (28 vs. 30) is due to the GCC gene selection criterion.

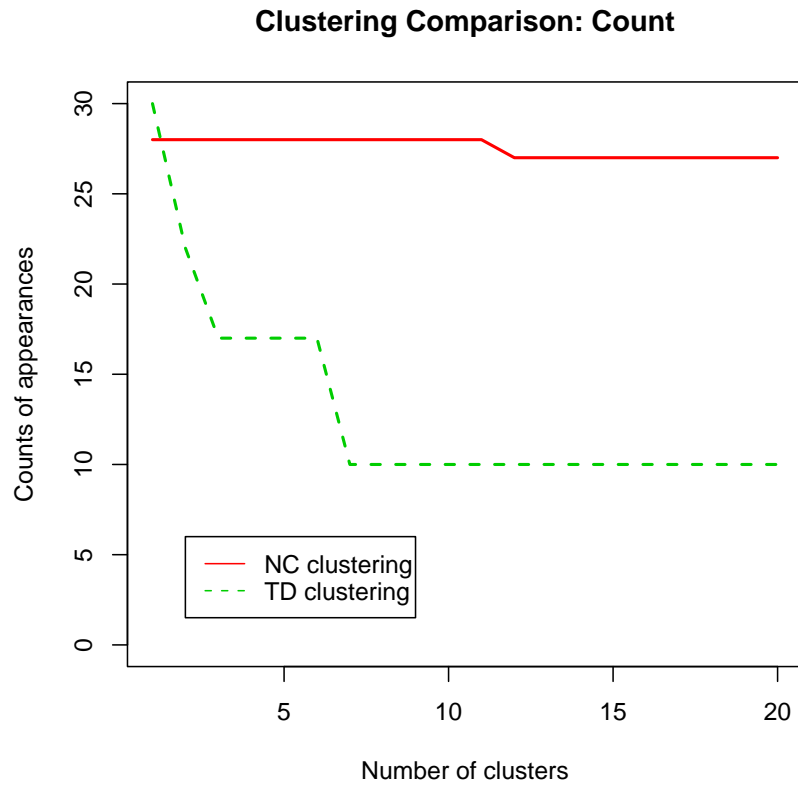


Figure 4.6: Clustering comparison - GO vocabulary “visual perception” counts.

4.3 Software Availability

The proposed network construction method and network constrained clustering method have been implemented in a R package “GeneNT” that is freely available from <http://cran.r-project.org/> with detailed documentation and examples. To promote the accessibility of the methods described in this thesis to the more general users, in collaboration with Ritu Khanna, the programmer and analyst in Swaroop lab, we further implemented the methods in an open source C based clustering software with GUI (Fig. 4.9). The software implemented the generalized network constrained clustering algorithm that is applicable to the whole network. The pro-

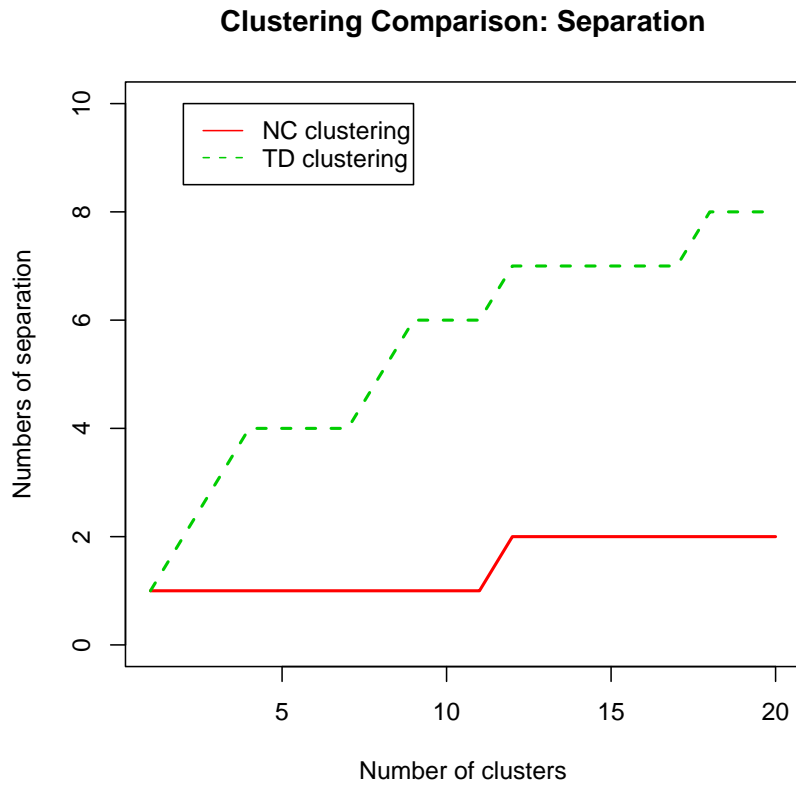


Figure 4.7: Clustering comparison - GO vocabulary “visual perception” separation.

prototype of the open source software was provided by M. J. L. de Hoon et al (de Hoon *et al.* 2004) from Human Genome Center, Institute of Medical Science, University of Tokyo. For more information, and software download, please visit http://www-personal.umich.edu/~zhud/cluster_31.htm.

4.4 Discussion

Inferring signaling pathway components from gene expression data is one of most active research areas in microarray data analysis. In this chapter, we proposed a new clustering approach to solve the first sub-problem of the signaling pathway reconstruction problem, i.e. discovery of pathway components. Our approach is based on

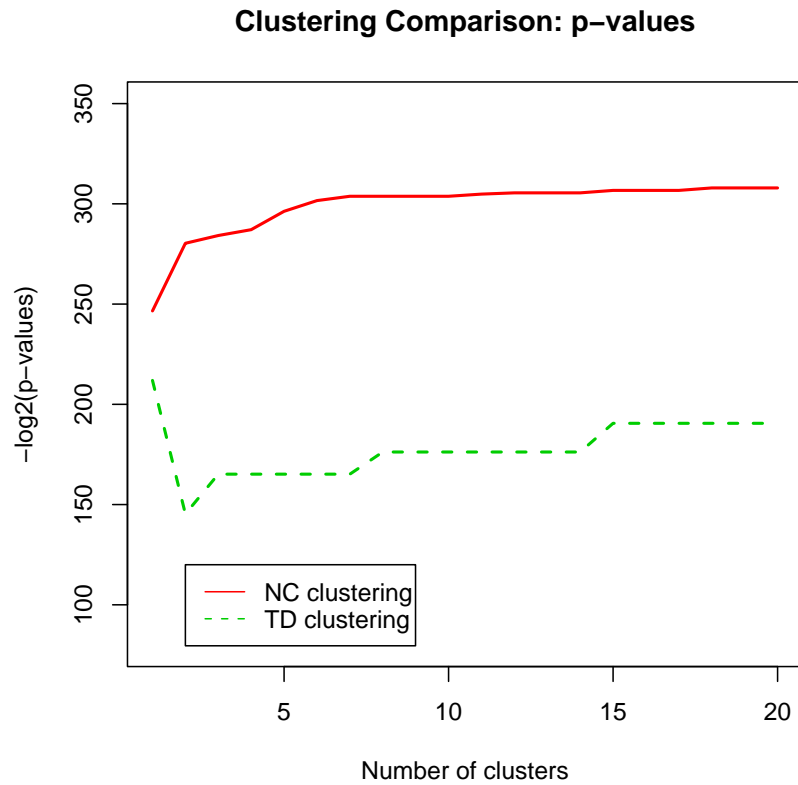


Figure 4.8: Clustering comparison - GO vocabulary “visual perception” p -values.

co-expression analysis that remains to be one of the most popular approaches. While at this stage many functional predictions made through co-expression analysis are based on the assumption of “Guilt-by-Association,” there are still few methods for functional predictions from dissimilar expression profiles. Transitive co-expression analysis (Zhou *et al.* 2002) is a systematic method to accomplish functional prediction from dissimilar gene expression profiles (Zhou and Gibson 2004, Zhou *et al.* 2005).

Systematic network analysis approaches have been widely applied to many biological networks such as metabolic networks, e.g. (Gagneur *et al.* 2003). Many theoretical approaches have been implemented to analyze metabolic networks including

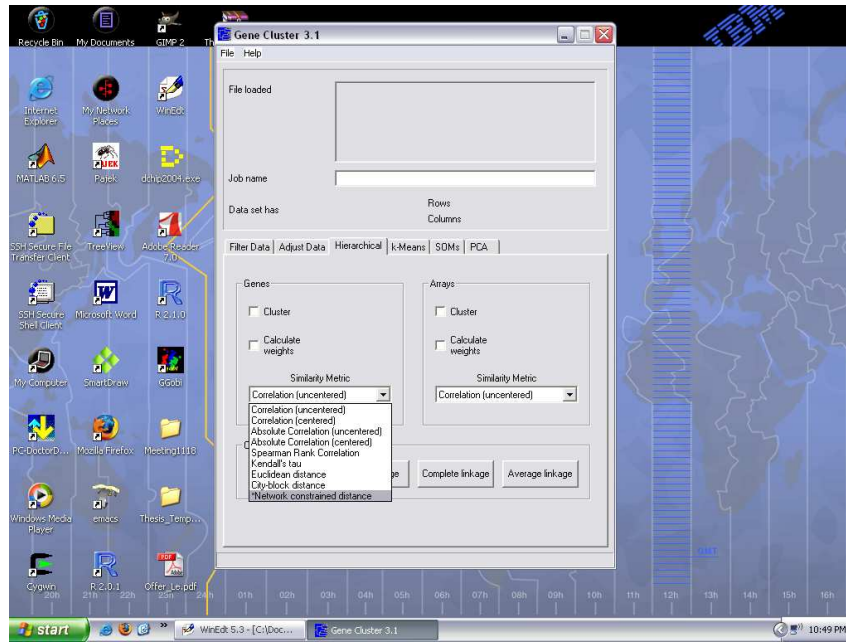


Figure 4.9: A significant update to the open source clustering software (joint work with Ritu Khanna).

network decomposition and isomorphism methods. For example, Ma et al. (Ma *et al.* 2004) presented a network decomposition approach to analyze metabolic pathways, by considering the global network structure rather than local marginal connectivity. They showed that chemical reactions in the same cluster are indeed functionally related. Our approach extends this to gene co-expression networks extracted from microarray data. Our network constrained clustering differs significantly from the traditional clustering approach in at least two aspects: 1) it uses GCC selected genes instead of all differentially expressed genes for clustering; 2) it uses a hybrid distance matrix that is composed of both direct distances and shortest-path distances for clustering instead of the traditional distance matrix that is composed of only direct distance matrix. The latter has been shown to lead to clustering improvements.

There are, however, biological function constrained clustering approaches that have previously been shown to possess clear advantages over the traditional clus-

tering approaches. One early attempt to introduce constraints into gene clustering was to account for the functional constraint revealed from well-studied metabolic pathways (Hanisch *et al.* 2002). Two more recent approaches have been to shrink the expression correlation based distance towards zero if the corresponding pair is functionally related as defined by Gene Ontology (Cheng *et al.* 2004, Huang *et al.* 2006). These approaches are successful implementations of the constraint. Our network constrained clustering approach extracts co-expression network information directly from the expression data. Other sources of functional association can easily be incorporated into our framework.

Gene co-expression networks differ from metabolic networks and protein-protein interaction networks in that the edges are inferred from hypothetical rather than physical interactions. Statistical methods are more useful in dealing with the inherent uncertainties. The method we adopted constructs the co-expression network by simultaneously controlling biological and statistical significance. Our network constrained clustering method has the following features: 1) it tends to group functional related genes into a tight cluster disregarding whether these genes have similar expression profiles; 2) it is sufficiently flexible because the calculated network constrained distance matrix can be fitted in many popular distance-based clustering software packages; 3) the algorithm runs in polynomial time.

CHAPTER V

de Novo Signaling Pathway Reconstruction

5.1 Introduction

In this chapter, we focus on estimation of the order of genes along a pathway assuming the unordered terminal and intermediate pathway components are known. Signaling pathways are the primary means of regulating cell growth, metabolism, differentiation, and apoptosis. The sensing and processing of extracellular stimuli are mediated by *signal transduction cascades*, that molecular circuits seek to detect, amplify, and integrate to generate responses such as changes in enzyme activity, activation/deactivation of transcription factors, gene expression, or ion-channel activity (Berg *et al.* 2006). Biochemically, the extracellular signal is transmitted through a series of molecular modifications (e.g. phosphorylation, dephosphorylation, acetylation, methylation) and interactions (e.g. protein-protein interaction, protein-DNA interaction).

As discussed before, recent bioinformatics research efforts have been shifted from the single gene analysis to signaling pathway analysis and network. With the evolution of signaling pathway research methods, the definition of such pathways also evolves. In earlier decades when genetic epistatic experiments were the predominant approach to reconstruct the signaling pathways, signaling pathway was defined as:

“The cascade of processes by which an extracellular signal (typically a hormone or neurotransmitter) interacts with a receptor at the cell surface, causing a change in the level of a second messenger for example calcium or cyclic AMP and ultimately effects a change in the cells functioning” (Berg *et al.* 2006). In the post-genomic era, simultaneous quantifying the abundance levels of thousands of biomolecules enables “high throughput” signaling pathway reconstruction. Lu *et al.* defined a signaling pathway as a specified group of genes that have coordinated association with a phenotype of interest (Lu *et al.* 2005). Subramanian *et al.* gave a more general definition: the groups of genes that share common biological function, chromosomal location, or regulation (Subramanian *et al.* 2005). Subramanian’s approach looks at a hypothetical set of genes and detects significant enrichment toward the top of a rank-ordered list. Both of these studies give the analyst great power toward solving the first sub-problem in signaling pathway reconstruction, i.e. discovery of pathway components. However, in the past the epistatic relationships among pathway components have been ignored, and these relationships are the key to understanding the underlying biological mechanism. Our application of Network Inference from Co-Occurrences (NICO) method can be used to solve this second sub-problem, i.e. ordering the pathway components (Rabbat *et al.* 2006). We propose a new definition of signaling pathway as: “a series of gene interaction that leads to an endpoint biological function from a membrane receptor.”

There are abundant biological and/or computational approaches to discovering signaling pathway components. Biological approaches include the traditional low throughput protein-protein interaction analysis such as immunoprecipitation, western blot and pull-down assay and high throughput protein-protein interaction analysis such as yeast two-hybrid assay. Computational approaches are mainly focused

on clustering genes according to function. Examples include network constrained clustering introduced in Chapter IV and other methods (Eisen *et al.* 1998, Hartigan and Wong 1979, Yeung *et al.* 2001, Schliep *et al.* 2003, Zhu *et al.* 2005c). These analyses have led to discovery of many signaling pathway components. The ultimate goal of pathway reconstruction analysis is to decipher the order through which the signal is transmitted. However, despite its importance, there has only been limited research on ordering pathway components.

The classical approach to pathway discovery is called genetic epistasis analysis, in which a pair of genes are mutated in the same strain and the phenotype of the double mutant was compared with those of the corresponding single mutants. The predominant phenotype defines the epistatic relationship between genes (Avery and Wasserman 1992). The success of this approach is contingent on the measured phenotype, therefore, the analysis of different pathways requires a variety of experiments. For example, satisfactory answers to the following questions are prerequisites to effective epistasis analysis: what kind of phenotype to measure, how to quantify this phenotype, e.g. morphology. In addition, as pointed out by Van Driessche *et al.* (Van Driessche *et al.* 2005), “the rules of epistasis cannot be applied consistently if the experimental procedures are not identical for all pairs of genes in a certain pathway.”

Recently, Van Driessche *et al.* (Van Driessche *et al.* 2005) proposed a new epistasis analysis using microarray gene expression profiles as a more objective phenotype. Their approach greatly relaxed the stringent requirement of experimental expertise in doing the classical epistasis analysis because the knowledge of relationship between gene function and phenotype is not essential. They reconstructed part of the Protein Kinase A Pathway by making ten combinations of single or double mutations in six

genes. The approach is limited to reconstructing very small size pathways due to the combinatorial explosion of the number of mutations needed. More mutations are either prohibited by the cost or by possibly lethal effects. In addition, the approach implicitly requires that the mutations have marked gene expression variation so that the epistatic relationship can be determined using a computational method without requiring replicated experiments.

In the last decade we have witnessed a rapid accumulation of high throughput genomics data, however reliable knowledge extraction from this data lags far behind. Instead of acquiring new data, Liu and Zhao proposed a pure computational approach to reconstruct the order of the pathway components from existing genomics and proteomics data (Liu and Zhao 2004). Assuming all terminal and intermediate components (unordered) are known, each permutation of the pathway components was scored using a score function, which was defined as un-weighted sum of score functions for gene expression data alone and for protein-protein interaction data alone. The score function of gene expression data was derived, based on the hypergeometric distribution, by testing whether the correlation between adjacent gene pairs is significantly higher than the random gene pairs in the pathway (Liu and Zhao 2004). The score function of protein-protein interaction data was derived based on the binomial distribution, in which the parameter (false negative rate) was estimated from protein-protein interactions in the DIP (Database of Interacting Proteins) database, and the binomial random variable corresponds to the observation whether the adjacent proteins interact or not. Using the simplified Mitogen Activated Protein Kinase (MAPK) pathway as an example, they reported that the “known” MAPK pathway was scored the second highest among all pathway permutations, which is much better than that obtainable using genomics data or proteomics data alone.

Being probably the first pure computational approach of its kind, the advantage of this approach is that it exploits existing data. The approach of Liu and Zhao also provides compelling evidence of the advantages of integration of multiple data sources. However, the approach also has a number of limitations:

- It heavily relies on the availability of high throughput data.
- It integrates only numeric data sources. Many kinds of non-numerical meta information, e.g. published literature and biologist’s expert knowledge, are difficult to include in the current probability model.
- Similar to the classical epistasis analysis, the approach is limited to reconstructing nonlethal signaling pathways.
- The approach is also limited to ordering short pathways due to the computational complexity introduced by the permutation step.

Here we review and apply a new maximum likelihood approach that exploits information about which genes are in each pathway to reconstruct a “gene regulation network topology” in the form of a first-order Markov chain transition matrix (denoted as NICO method throughout this chapter). The NICO method was originally developed by Rabbat et al (Rabbat *et al.* 2006) for tomographic reconstruction of telecommunications networks. Information on the genes composing a pathway can be integrated from multiple data sources (solid curves in Fig. 5.1). Non-zero transition probabilities correspond to directed edges in the network. We use this probability transition matrix to determine the maximum likelihood order of genes in each pathway. The applied technique naturally combines pathway information (both composition information and epistasis information) that are derived from multiple data sources.

provide
the full
expansion
of this
term!

To summarize, the features of this proposed techniques are:

- As shown in Fig. 5.1, the unordered pathway composition information can be either integrated from high throughput experiments or from meta-information.
- The prior information on pathway epistasis can be easily integrated into the first order Markov model in the format of prior on the transition matrix. For example, kinase and phosphatase appear in front of their substrate in the pathways. The corresponding entries of the prior transition matrix can be inflated to larger transition probabilities as compared to other entries in the same row of the matrix.
- ^{NICO} The approach is able to order relatively large pathways using Monte Carlo importance sampling.

It is often the case that the available pathway composition information and prior epistatic information are not sufficient to resolve the ambiguous epistasis relationship among a subset of genes. Using the NICO method, we can provide confidence coefficients on the ordering of genes and these can be used to suggest future experiments to the biologist to resolve the ambiguity. More specifically, more than one pathway order may have the same confidence as measured by the likelihood score. Comparing these “equally likely” candidate pathways may allow biologists to identify the non-redundant set of genetic experiments to resolve the ambiguity (see dotted curves in Fig. 5.1). In this sense, applying the technique may be incorporated into a sequential design of experiments context, resulting in significant savings in experimental effort.

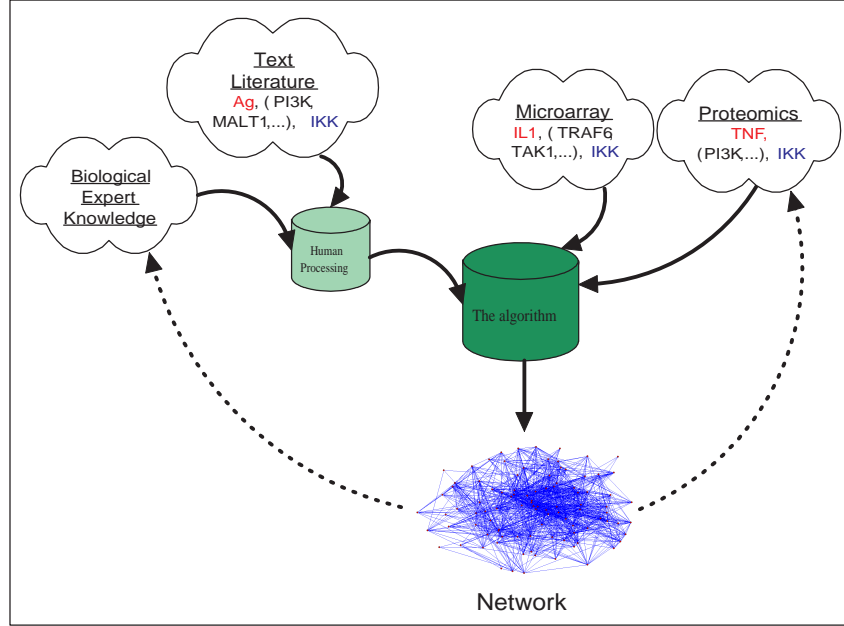



Figure 5.1: The schematic representation of the signaling pathway reconstruction algorithm. The starting pathway component is in red (left), and the ending pathway component is in blue (right). Pathway components in the parenthesis are intermediate and unordered. The solid lines represent the inputs to the algorithm (different sources of pathway information). The dotted lines represent the outputs from the algorithm (the maximum likelihood pathway(s)).

5.2 Methods

5.2.1 Mathematical Formulation of the Problem

We assume a biologically known signaling pathway is an ordered path $\mathbf{z} = (z_1, z_2, \dots, z_N)$ that is sampled from a discrete-time first-order Markov chain where the states of the chain, z_i , are genes or proteins in the pathway. In reality, the pathways derived from many data analysis schemes are unordered, defined as a gene co-occurrence observation, i.e. a string of genes or proteins \mathbf{x} . One can interpret \mathbf{x} as having been obtained from \mathbf{z} after subjecting \mathbf{z} to a random permutation, τ . A biologically known signaling network consisting of an ensemble of signaling pathways can be viewed as a collection of T independent permuted Markov processes, $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)}\}$. The problem that we need to solve is to recover the signaling network topology given

a set of co-occurrence strings \mathcal{X} that are obtained from multiple sources such as: cluster analysis of high throughput data; text literature mining or biological expert knowledge. Treating the unobserved permutations, τ^1, \dots, τ^T , as hidden variables. For the sake of completeness, here we review an expectation-maximization (EM) algorithm for computing the maximum likelihood estimates of the Markov chain parameters: the initial state distribution $\boldsymbol{\pi}$ (starting genes in the signaling pathways) and transition matrix \mathbf{A} (signaling network topology). This EM algorithm was originally derived in the NICO framework to solve a network tomography problem in telecommunication networks (Rabbat *et al.* 2006). In section 5.2.2, we review the standard approach to estimating parameters of a Markov chain when fully ordered pathways are available. In section 5.2.3, we review ^{the implemented} an EM algorithm for estimating Markov chain parameters from unordered pathways. For relatively large pathways, we review a Monte Carlo E-step that approximates E-step computation (section 5.2.4). Finally, we review ^{the extension of} how to extend NICO into a fully Bayesian framework that facilitates incorporating prior pathway information (section 5.2.5). 

5.2.2 Estimating a Markov Chain from Direct Observations

The sections 5.2.2, 5.2.3, 5.2.4, 5.2.5 were in a large part summarized from descriptions of the NICO methodology by our collaborators (Rabbat *et al.* 2006). Each independent Markov process is fully defined by the parameters \mathbf{A} and $\boldsymbol{\pi}$,

$$(5.1) \quad P[Z_t = j | Z_{t-1} = i] = A_{i,j},$$

for $i, j \in S$, where S is the number of Markov states (distinct set of pathway components), and $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_{|S|})$ is the vector of marginal state probabilities, $\pi_k = P(Z_t = k)$. The biological signal has to be initially emitted from one of the S pathway components, and the signal emitted from pathway component i , if it is not

the terminal component, has to be received by one of the $|S|$ pathway components, indexed by j . Mathematically, the former corresponds to the constraint in Eq. 5.2 and the later corresponds to the constraint in Eq. 5.3:

$$(5.2) \quad \sum_{i=1}^{|S|} \pi_i = 1,$$

$$(5.3) \quad \sum_{j=1}^{|S|} A_{i,j} = 1.$$

The probability of a length- N signaling pathway (z_1, z_2, \dots, z_N) being generated by the chain $(S, \mathbf{A}, \boldsymbol{\pi})$ is

$$(5.4) \quad P[Z_1 = z_1, Z_2 = z_2, \dots, Z_N = z_N | \mathbf{A}, \boldsymbol{\pi}] = \pi_{z_1} \prod_{t=2}^N A_{z_{t-1}, z_t}.$$

Now, suppose that instead of one pathway $\mathbf{w} = (w_1, w_2, \dots, w_N)$, we have a set \mathcal{W} of T distinct pathways which are assumed to have been generated independently by this Markov process. The log-likelihood for this set of pathways is simply

$$(5.5) \quad \log P[\mathcal{W} | \mathbf{A}, \boldsymbol{\pi}] = \sum_{m=1}^T \log P[\mathbf{w}^{(m)} | \mathbf{A}, \boldsymbol{\pi}].$$

Maximum likelihood (ML) estimates of $\boldsymbol{\pi}$ and \mathbf{A} are obtained by maximizing $\log P[\mathcal{W} | \mathbf{A}, \boldsymbol{\pi}]$ under the constraints in Eqs. 5.2 and 5.3, i.e.

$$(5.6) \quad \hat{A}_{i,j} = \frac{\sum_{m=1}^T \sum_{t=2}^{N_m} w_{t-1,i}^{(m)} w_{t,j}^{(m)}}{\sum_{j=1}^{|S|} \sum_{m=1}^T \sum_{t=2}^{N_m} w_{t-1,i}^{(m)} w_{t,j}^{(m)}}$$

$$(5.7) \quad \hat{\pi}_i = \frac{1}{T} \sum_{m=1}^T w_{1,i}^{(m)},$$

where $m = 1, \dots, T$ is the pathway index, $t = 2, \dots, N_m$ is the pathway component index, and $(w_{t,i} = 1) \Leftrightarrow (z_t = i)$.

5.2.3 Estimating a Markov Chain from Shuffled Observations via the EM Algorithm

We are now ready to proceed to solve the general pathway reconstruction problem in which we only observe the pathway components, but not their orders. In order to cast this into the familiar framework of EM algorithm, we suppose that the observed T unordered pathways are partial data, denoted as $\mathcal{X} = \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)}$ and their orders are missing data, modeled as a set of shuffling matrices $\mathcal{R} = \mathbf{r}^{(1)}, \mathbf{r}^{(2)}, \dots, \mathbf{r}^{(T)}$ so that $(r_{t,t'}^{(m)} = 1) \Leftrightarrow (\mathbf{x}_{t'}^{(m)} = \mathbf{w}_t^{(m)})$. Given both $\mathbf{r}^{(m)}$ and $\mathbf{x}^{(m)}$, we could recover the unshuffled sequence $\mathbf{w}^{(m)}$ by applying (Rabbat *et al.* 2006)

$$(5.8) \quad w_{t,i}^{(m)} = \prod_{t'=1}^{N_m} (x_{t',i}^{(m)})^{r_{t,t'}^{(m)}},$$

adopting the convention $0^0 = 1$.

We are now ready to write the complete log-likelihood. Starting by observing that

$$(5.9) \quad \log P[\mathcal{X}, \mathcal{R} | \mathbf{A}, \boldsymbol{\pi}] = \log P[\mathcal{X} | \mathcal{R}, \mathbf{A}, \boldsymbol{\pi}] + \log p[\mathcal{R}],$$

and that $p[\mathcal{R}]$ is just a constant (assuming uniform distribution over the set of all possible permutations), we have

$$(5.10) \quad \log P[\mathcal{X}, \mathcal{R} | \mathbf{A}, \boldsymbol{\pi}] \propto \log P[\mathcal{X} | \mathcal{R}, \mathbf{A}, \boldsymbol{\pi}] = \sum_{m=1}^T \log P[\mathbf{x}^{(m)} | \mathbf{r}^{(m)}, \mathbf{A}, \boldsymbol{\pi}]$$

The EM algorithm proceeds by computing the expected value of the complete log-likelihood $\log P[\mathcal{X}, \mathcal{R} | \mathbf{A}, \boldsymbol{\pi}]$ with respect to the missing data, conditioned on the observations and on the current estimate of the model parameters $\hat{\mathbf{A}}$ and $\hat{\boldsymbol{\pi}}$,

$$(5.11) \quad Q(\mathbf{A}, \boldsymbol{\pi}; \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}) = \mathbb{E}[\log P[\mathcal{X}, \mathcal{R} | \mathbf{A}, \boldsymbol{\pi}] | \mathcal{X}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}].$$

A key observation which facilitates the derivation of the E-step is that the complete log-likelihood is linear with respect to simple functions of the missing variables:

- the first row of each matrix $\mathbf{r}^{(m)}$, that is, for $m = 1, \dots, T$ and $t' = 1, \dots, N_m$;

- sums of products of pairs of variables: $\alpha_{t',t''}^{(m)} \equiv \sum_{t=2}^{N_m} r_{t,t'}^{(m)} r_{t-1,t''}^{(m)}$, for $m = 1, \dots, T$, and $t', t'' = 1, 2, \dots, N_m$.

Since the conditional expectation of a linear function of a random variable is simply that linear function computed at the expected value of the random variable, in the E-step we just have to compute the conditional expectations of $r_{t,t'}^{(m)}$ and $\alpha_{t',t''}^{(m)}$ and plug them into the complete log-likelihood function. After some algebraic derivations, the conditional expectation function $Q(\mathbf{A}, \boldsymbol{\pi}; \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}})$ is (Rabbat *et al.* 2006)

$$(5.12) \quad Q(\mathbf{A}, \boldsymbol{\pi}; \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}) = \sum_{m=1}^T \sum_{t',t''=1}^{N_m} \sum_{i,j=1}^{|S|} \bar{\alpha}_{t',t''}^{(m)} x_{t'',i}^{(m)} x_{t',j}^{(m)} \log A_{i,j} + \sum_{m=1}^T \sum_{t'=1}^{N_m} \sum_{i=1}^{|S|} \bar{r}_{1,t'}^{(m)} x_{t',i}^{(m)} \log \pi_i,$$

where the $\bar{\alpha}_{t',t''}^{(m)}$ and $\bar{r}_{1,t'}^{(m)}$ are defined as:

$$(5.13) \quad \bar{\alpha}_{t',t''}^{(m)} \equiv E[\alpha_{t',t''}^{(m)} | \mathcal{X}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}] = P[\alpha_{t',t''}^{(m)} = 1 | \mathcal{X}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}],$$

and

$$(5.14) \quad \bar{r}_{1,t'}^{(m)} \equiv E[r_{1,t'}^{(m)} | \mathcal{X}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}] = P[r_{1,t'}^{(m)} = 1 | \mathcal{X}, \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}].$$

The model parameter estimates are then updated according to

$$(5.15) \quad (\hat{\mathbf{A}}_{\text{new}}, \hat{\boldsymbol{\pi}}_{\text{new}}) = \operatorname{argmax}_{\mathbf{A}, \boldsymbol{\pi}} Q(\mathbf{A}, \boldsymbol{\pi}; \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}),$$

and the process is repeated cyclically until some convergence criterion is met. Eq. 5.12 is the E-step, and Eq. 5.15 is the M-step. Maximization under the constraints in 5.2 and 5.3 leads to the following simple update equations (Rabbat *et al.* 2006):

- Transition matrix:

$$(5.16) \quad (\hat{\mathbf{A}}_{i,j})_{\text{new}} = \frac{\sum_{m=1}^T \sum_{t',t''=1}^{N_m} \bar{\alpha}_{t',t''}^{(m)} x_{t'',i}^{(m)} x_{t',j}^{(m)}}{\sum_{j=1}^{|S|} \sum_{m=1}^T \sum_{t',t''=1}^{N_m} \bar{\alpha}_{t',t''}^{(m)} x_{t'',i}^{(m)} x_{t',j}^{(m)}}.$$

- Initial probabilities:

$$(5.17) \quad (\hat{\pi}_i)_{\text{new}} = \frac{\sum_{m=1}^T \sum_{t'=1}^{N_m} \bar{r}_{1,t'}^{(m)} x_{t',i}^{(m)}}{\sum_{i=1}^{|S|} \sum_{m=1}^T \sum_{t'=1}^{N_m} \bar{r}_{1,t'}^{(m)} x_{t',i}^{(m)}}.$$

The EM algorithm can easily be modified to handle the special case that the starting and ending genes of each pathway are known and only the intermediate pathway components are unordered. The knowledge of the endpoints of each pathway imposes the constraints

$$(5.18) \quad r_{1,1}^{(m)} = 1,$$

and

$$(5.19) \quad r_{N_m, N_m}^{(m)} = 1.$$

Under the first constraint, estimates of the initial state probabilities are simply given by

$$(5.20) \quad \hat{\pi}_i = \frac{1}{T} \sum_{m=1}^T x_{1,i}^{(m)}.$$

Thus, only the transition matrix has to be estimated using the EM algorithm. Let

$$(5.21) \quad \Psi_N = \{r \in \Psi_N : r_{1,1} = 1, r_{N,N} = 1\},$$

denote the collection of permutations of N pathway components with fixed endpoints. The M-step (update for $\hat{\mathbf{A}}$) remains exactly same. The E-step can be computed using summary statistics (Rabbat *et al.* 2006):

$$(5.22) \quad \tilde{\gamma}^{(m)} = \sum_{r \in \tilde{\Psi}_{N_m}} P[\mathbf{x}^{(m)} | \mathbf{r}, \hat{\mathbf{A}}, \hat{\pi}]$$

$$(5.23) \quad \tilde{\gamma}_{t',t''}^{(m)} = \sum_{r \in \tilde{\Psi}_{N_m}} P[\mathbf{x}^{(m)} | \mathbf{r}, \hat{\mathbf{A}}, \hat{\pi}] \sum_{t=2}^{N_m} r_{t,t'} r_{t-1,t''},$$

for $t', t'' = 1, \dots, N_m$, and setting $\bar{\alpha}_{t,t',t''}^{(m)} = \frac{\tilde{\gamma}_{t,t',t''}^{(m)}}{\tilde{\gamma}}$.

5.2.4 Monte Carlo E-Step by Important Sampling

For a large pathway, the combinatorial nature of the equations (5.13) and (5.14), that is, the need to sum over all permutations of the pathway, may render exact computation intractable. We review a sampling-based approximation version of the E-step, which avoids the combinatorial nature of its exact version (Rabbat *et al.* 2006). Without loss of generality, we focus on a particular length- N pathway $\mathbf{x} = x_1, x_2, \dots, x_N$ to lighten the notation. We also drop the hats from $(\hat{\mathbf{A}}, \hat{\boldsymbol{\pi}})$ and use simply $(\mathbf{A}, \boldsymbol{\pi})$ to denote the current Markov chain parameter estimates in the EM algorithm (Rabbat *et al.* 2006).

An intuitive Monte Carlo approximation to the sums in Eqs. 5.13 and 5.14 would be based on random permutations, sampled from the uniform distribution over Ψ_N (the collections of all permutations of N components). For a large Ψ_N , only a small fraction of these random permutations will have non-negligible posterior probability, $P[\mathbf{r}|\mathbf{x}, \mathbf{A}, \boldsymbol{\pi}]$, and so a very large number of uniform samples is needed to obtain a good approximation to $\bar{r}_{1,t'}$ and $\bar{\alpha}_{t',t''}$. Ideally, one could sample permutations directly from the posterior distribution $P[\mathbf{r}|\mathbf{x}, \mathbf{A}, \boldsymbol{\pi}]$; however, sampling from this distribution would require determining its value from all $N!$ permutations in Ψ_N . Instead, *importance sampling* (IS) was employed (Rabbat *et al.* 2006): the step is that one sample L permutations, $\mathbf{r}^1, \dots, \mathbf{r}^L$, from a distribution $R[\mathbf{r}]$, from which it is easier to sample than $P[\mathbf{r}|\mathbf{x}, \mathbf{A}, \boldsymbol{\pi}]$, and then apply a corrective re-weighting to obtain approximations to $\bar{r}_{1,t'}$ and $\bar{\alpha}_{t',t''}$. The importance sampling estimates are given by (Rabbat *et al.* 2006)

$$(5.24) \quad \bar{r}_{1,t'} \simeq \frac{\sum_{i=1}^L z_i r_{1,t'}^i}{\sum_{i=1}^L z_i},$$

$$(5.25) \quad \bar{\alpha}_{t,t',t''} \simeq \frac{\sum_{i=1}^L z_i \sum_{t=2}^{N_m} r_{t,t'}^i r_{t-1,t''}^i}{\sum_{i=1}^L z_i},$$

where z_i is the correction factor (or weight) for sample r_i , given by

$$(5.26) \quad z_i = \frac{P[\mathbf{r}^i | \mathbf{x}, \mathbf{A}, \boldsymbol{\pi}]}{R[\mathbf{r}^i]},$$

the ratio between the desired distribution and the sampling distribution employed.

5.2.5 Incorporating Prior Information

Prior information about the Markov chain parameters \mathbf{A} and $\boldsymbol{\pi}$ can easily be incorporated into the algorithm by applying independent Dirichlet priors to each row of the transition matrix and to the initial state distribution (Rabbat *et al.* 2006). The Dirichlet distribution that was used to incorporate prior knowledge about the Markov chain parameters is exactly *a priori*. It is fixed before performing the inference via the EM algorithm and it does not change from iteration to iteration. Hence, we have

$$(5.27) \quad P[\boldsymbol{\pi} | \mathbf{u}] \propto \prod_{i=1}^{|\mathcal{S}|} \pi_i^{u_i-1}$$

$$(5.28) \quad p[\mathbf{A} | \mathbf{v}] \propto \prod_{i=1}^{|\mathcal{S}|} \prod_{j=1}^{|\mathcal{S}|} A_{i,j}^{v_{i,j}-1},$$

where the parameter u_i and $v_{i,j}$ should be non-negative in order to have proper priors. The larger that u_i is relative to the other $u_{i'}$, $i' \neq i$, the greater our prior belief that pathway component i is a starting component of the pathway rather than the others. In the case that the pathway terminal components were known, we set $u_i = 1$. Similarly, the larger $v_{i,j}$ relative to other $v_{i,j'}$ for $j' \neq j$, the more likely we expect that, *a priori*, the signal is transmitted from pathway component i to pathway component j relative to the transmissions from i to the other pathway components.

Plugging Eqs. 5.27 and 5.28 into the complete log-likelihood (Eq. 5.12), it is found that incorporating priors into the EM algorithm only results in a change to the M-step (Rabbat *et al.* 2006). In particular, instead of the ML estimator recursions of Eq. 5.17, we have recursions for the maximum a posteriori (MAP) estimates,

$$(5.29) \quad (\hat{\pi}_i)_{\text{new}} = \frac{u_i + \sum_{m=1}^T \sum_{t'=1}^{N_m} \bar{r}_{1,t'}^{(m)} x_{t',i}^{(m)}}{\sum_{i=1}^{|S|} \left(u_i + \sum_{m=1}^T \sum_{t'=1}^{N_m} \bar{r}_{1,t'}^{(m)} x_{t',i}^{(m)} \right)},$$

and instead of Eq. 5.16, we have

$$(5.30) \quad (\hat{\mathbf{A}}_{i,j})_{\text{new}} = \frac{v_{i,j} + \sum_{m=1}^T \sum_{t',t''=1}^{N_m} \bar{\alpha}_{t',t''}^{(m)} x_{t',i}^{(m)} x_{t'',j}^{(m)}}{\sum_{j=1}^{|S|} \left(v_{i,j} + \sum_{m=1}^T \sum_{t',t''=1}^{N_m} \bar{\alpha}_{t',t''}^{(m)} x_{t',i}^{(m)} x_{t'',j}^{(m)} \right)}.$$

After convergence the corresponding a posteriori log-likelihood can be approximated using Eq. 5.12 and Eqs. 5.27 and 5.28,

$$(5.31) \quad \log P(\mathcal{X} | \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}) P(\hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}) \simeq Q(\mathbf{A}, \boldsymbol{\pi}; \hat{\mathbf{A}}, \hat{\boldsymbol{\pi}}) + \log P(\hat{\boldsymbol{\pi}} | \mathbf{u}) + \log p(\hat{\mathbf{A}} | \mathbf{v}).$$

5.3 Results

Using three representative signaling pathways, we intend to show three useful properties of the NICO method: reconstruction of the order of genes in the pathway assuming the intermediate and terminal components are known; ease of incorporating prior knowledge in the form of a prior on the transition matrix; identifying the most important missing information that prevents high confidence path order reconstruction. The latter will be useful for specifying the most informative future experiment if needed (Fig. 5.1).

5.3.1 Protein Kinase A Pathway

The protein kinase A (PKA) pathway is an essential signaling pathway for development. The central component cyclic AMP (cAMP)-dependent protein kinase A is able to phosphorylate a variety of proteins and thereby affect their activity.

Malfunction of this pathway lead to developmental arrest or attenuation, precocious development and aberrant sporulation and germination (Loomis *et al.* 1998, Van Driessche *et al.* 2005). Van Driessche et al used this pathway to demonstrate a microarray based epistasis approach (Van Driessche *et al.* 2005). They reconstructed an incomplete pathway by making ten combinations of single or double mutations in six genes. The relationships between several pairs of genes could not be determined from their analysis. For example, the level of interaction between *acaA* and *pkaR* was not tested because the corresponding mutations were not analyzed or difficult to make. Despite this missing information, our approach is able to reconstruct the reported pathway based only on the information about terminal components and the unordered intermediate components in each pathway (Fig. 5.2, Fig. 5.3). This suggests that our techniques ^{approach} may enable biologists to reconstruct pathways without having to perform exhaustive experiments on all pairwise interactions.

PKA Pathway

acaA, (*pkaC*, *pkaR*), Development

regA, (*pkaR*, *pkaC*), Development

yakA, (*pufA*, *pkaC*), Development

Figure 5.2: The (unordered) protein kinase A signaling pathway. Membrane receptors are in red (left), and transcription factors are in blue (right). Activation or inhibition information between pathway components are omitted. The pathway is mainly adapted from Van Driessche et al (Van Driessche *et al.* 2005).

Since the protein kinase A pathway is a relatively small pathway, it is perhaps not surprising that we are able to reconstruct it in a straightforward manner. For larger pathways, without prior epistatic knowledge available pathway composition information often only allows the pathway be reconstructed to a “certain low resolution”, i.e. up to certain ambiguities in relative ordering within the pathway. Incorporat-

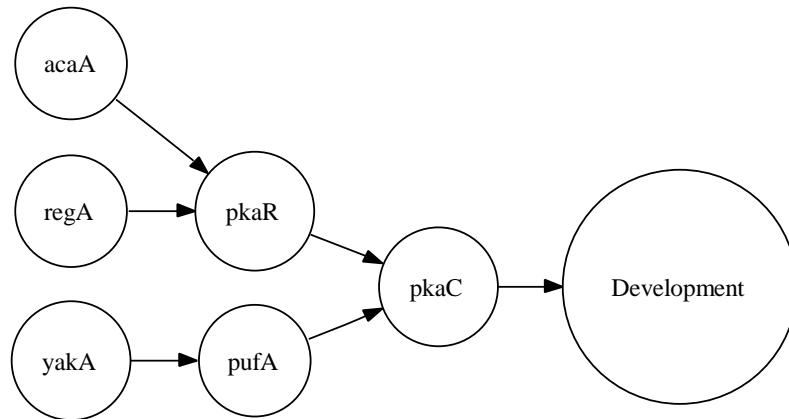


Figure 5.3: The reconstructed protein kinase A signaling network topology from unordered pathway composition data (Fig. 5.2).

ing prior knowledge can often help to reveal the order of the whole pathway, or an ensemble of pathways, i.e. a signaling network. In the next subsection we illustrate the NICO method on the more complicated SAPK/JNK pathway.

5.3.2 SAPK/JNK Pathway

Stress-activated protein kinases (SAPK)/Jun N-terminal kinases (JNK) are members of the MAPK family and are activated by a variety of environmental stresses, inflammatory cytokines, growth factors and GPCR agonists. Stress signals are delivered to this cascade by small GTPases of the Rho family (Rac, Rho, Cdc42) (Weston *et al.* 2002). Similar to our study of the protein kinase A pathway, we attempt to reconstruct pathway order based only on the terminal components and on unordered list of intermediate pathway components (Fig. 5.4).

In the framework of first-order Markov chains, epistasis relationships of the pathway components are fully defined by the probability transition matrix \mathbf{A} . For the observed unordered SAPK/JNK pathway, multiple pathway orders as defined by the corresponding probability transition matrixes may have the same likelihood score.

$$(5.33) \quad \hat{A}' = \begin{pmatrix} 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.333 & 0 & 0.667 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.875 & 0 & 0.125 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

SAPK/JNK Pathway

GF, (HPK, MEKK, MKK), **JNK**
GF, (EKK, HPK, MKK), **JNK**
GF, (RAC, RAS, MEKK, MKK), **JNK**
GF, (RAS, CDC42, RAC, MKK, MEKK), **JNK**
GF, (RAS, RAC), **RHO**
CS1, (RAC, MEKK, MKK, CDC42), **JNK**
CS2, (MEKK, MKK, RAC), **JNK**
FASL, (GCKs, MKK, MEKK), **JNK**
OS, (ASK1, MEKK, MKK), **JNK**

Figure 5.4: The (unordered) SAPK/JNK signaling pathway. Membrane receptors are in red (left), and transcription factors are in blue (right). Activation or inhibition information between pathway components are omitted. “GF” stands for Growth Factor, “CS” stands for Cellular Stress, “FASL stands for Fas Ligand”, “OS” stands for Oxidation Stress. The pathway is adapted from <http://www.cellsignal.com/>.

Often prior epistasis information or pathway composition information may not suffice to resolve all ordering ambiguity in the pathway. In such cases it would be useful to predict the crucial pieces of information necessary to resolve remaining ambiguity. We next show how the NICO method can be applied to perform such a prediction for the Nuclear Factor κ B (NF κ B) pathway.

5.3.3 NF κ B Pathway

NF κ B proteins function as dimeric transcription factors that control genes regulating a broad range of biological processes including innate and adaptive immunity, inflammation, apoptosis, stress responses, B cell development and lymphoid organogenesis (Pomerantz and Baltimore 2002). NF κ B pathways mediate the signal transduction from extracellular stimuli to these transcription factors including controlled cytoplasmic-nuclear shuttling and modulation of transcriptional activity (Ghosh *et al.* 2002).

We specified the terminal components of different stimuli receptors (start) and NF κ B (end), and pathway components corresponding to each stimuli (Fig. 5.6). The latter can often be derived from a combination of computational approaches (e.g. clustering) and the biologist's expert knowledge. The biological expert knowledge is acquired gradually over ^{many} years from multiple sources such as literature, science seminar, and experimental results. We also incorporated several pieces of prior biological information including the epistasis relationships between PI(3)K and PLC γ 2 (Humphries *et al.* 2004), between PLC γ 2 and PKC (Humphries *et al.* 2004), between PKC and MALT1 (Che *et al.* 2004), between MALT1 and TRAF6 (TNF-receptor-associated factor 6) (Sun *et al.* 2004), between TRAF6 and TAK1 (TGF β -activated kinase 1) (Morlon *et al.* 2005), between TAK1 and IKK (Sun *et al.* 2004), between PI(3)K and Akt/Cot complex (Kane *et al.* 2002), and between JNK and β TrCP (β Transducin Repeat-Containing Protein) (Spiegelman *et al.* 2001). The biology background is as follows: Upon PI(3)K activation the Akt/Cot complex is likely recruited to the membrane through the Akt PH domain, which binds the phospholipid PIP3 (Kane *et al.* 2002). JNK induces β TrCP to activate NF κ B pathway (Spiegelman *et al.* 2001). Tyrosine phosphorylation of phospholipase PLC γ 2 is a crucial

activation switch that initiates, and maintains, intracellular calcium mobilization in response to extracellular stimuli (Humphries *et al.* 2004). PKC was reported to be able to activate MALT1 upon receiving extracellular stimuli (Che *et al.* 2004). MALT1 binds and activates TRAF6 (Sun *et al.* 2004). TRAF6 activates TAK1 through the adaptor protein TAB2 (Morlon *et al.* 2005) and TAK1 activates IKK (Sun *et al.* 2004).

Our application of the NICO method successfully reconstructed most of the pathway component orders after 7 iterations and used approximate search for path longer than 8 components. The sole ambiguity is between $\text{NF}\kappa\text{B}$ complex1 and complex2 (Fig. 5.7). Indeed in this case the ambiguity can be detected by investigating the relative maxima of the likelihood function (Eq. 5.31). A relative maximum that is approximately equal to the global maximum indicates an ambiguity that is localized by the positions of the relative maxima over the space of transition matrices \mathbf{A} (Eqs. 5.32, 5.33, Appendix. A.6). To resolve this ambiguity, our analysis indicates that biologists should focus on investigating the epistatic relationship between these two complexes.

5.3.4 Assembling Signaling Pathways into Signaling Networks

Biological signaling pathways tend to share a fair amount of common signal components, and we often define these as signaling networks. The latter provides a more complete view of cellular regulatory mechanisms. Fig. 5.8 presents a signaling network assembled from SNK/JNK and $\text{NF}\kappa\text{B}$ pathways.

5.4 Software Availability

The NICO method ^{has} have been implemented in a set of Matlab codes by Mike Rabbat. Dongxiao Zhu wrote wrapper functions to apply the method to reconstructing

the

signaling pathways. The set of codes will be soon available for download from authors' website.

5.5 Discussion

In this chapter, we reviewed and applied a model based approach to reconstruct the order of an unordered list of pathway components along with terminal genes (Rabbat *et al.* 2006). Compared to previous genetic and computational approaches, the approach does not directly depend on the numeric format of the data, thus it enjoys the features of versatility, flexibility and a high level of data abstraction. The knowledge of pathway intermediate components and terminal components can be derived either from numeric data using computational/statistical methods or from meta-data using biological expertise, e.g., terminal genes of a pathway are often specified as membrane receptor (start) and transcription factor (end). In this sense, the approach represents progress in data integration for gene pathway discovery. Moreover, the adapted Bayesian framework permits seamless incorporation of prior epistatic knowledge in the form of a prior on the transition matrix. When ambiguities do exist our algorithm can identify them and provide information on the most fruitful set of future experiments to resolve the ambiguities.

Many researchers have found the topology of networks of signaling pathways to be scale-free and sparse. In such topologies a small number of nodes (hub nodes) are highly connected while the remaining nodes are not. The hub nodes may form interaction motifs (functional modules) that are often shared by multiple pathways. Our pathway ordering approach may be used to exploit the scale-free property by better defining these multiple pathways. One limitation of our approach shared by previous approaches, is that our method assumes a linear pathway model without

any feedback loops. Many signaling pathways have been found to be interconnected and regulated via positive/negative feedback loops. Examples are the *p53* signaling pathways that correspond to a variety of intrinsic and extrinsic stress signals that impacts upon cellular homeostatic mechanisms (Vogelstein *et al.* 2000). These pathways consist of multiple positive/negative feedback loops, e.g. between *p53* and MDM2. The linear pathway model assumption may result in suboptimal pathway reconstruction. In future work, this limitation might be overcome by integrating more sophisticated graphical models into the methodology.

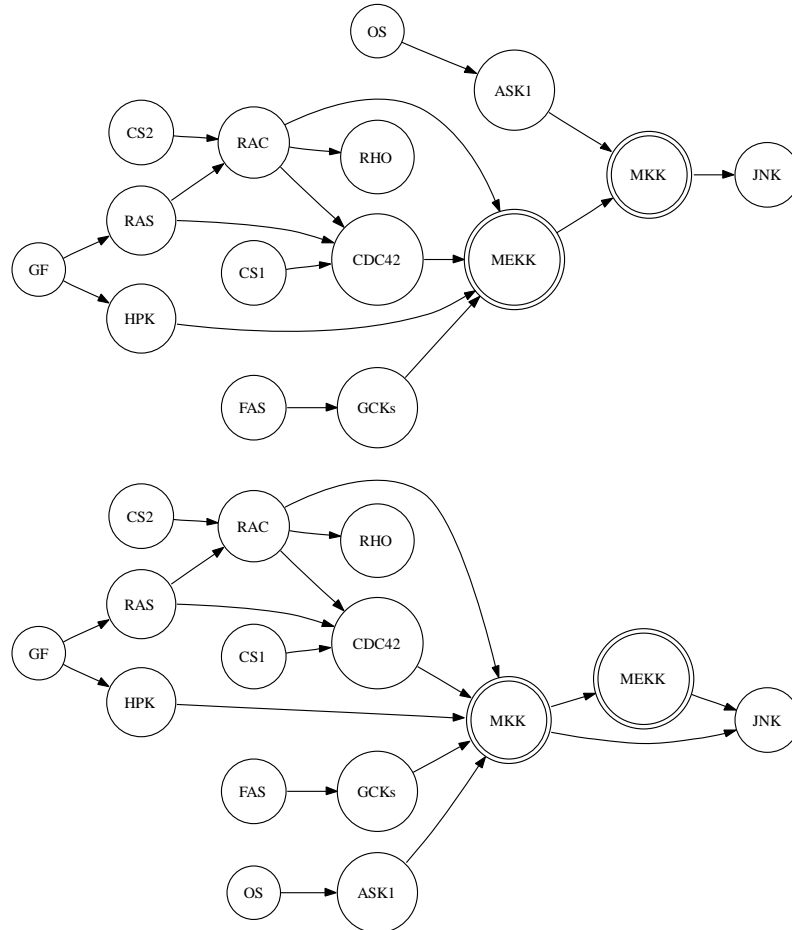


Figure 5.5: Upper panel: The correct SAPK/JNK signaling network topology defined by the probability transition matrix Eq. 5.32 estimated from unordered pathway composition data (Fig. 5.4) improved by incorporating a prior information on gene-gene interactions, in particular the interactions between the two double-circled components. Lower panel: The incorrect SAPK/JNK signaling network topology defined by the probability transition matrix Eq. 5.33 estimated from unordered pathway composition data without incorporating prior information.

NF κ B Pathway

Ag, (PKC, PI(3)K, PLC γ 2, MALT1, TAK, TAB1/2, IKK, TRAF6, NF κ BC2, NF κ BC1), **NF κ B**
Ag-MHC, (TRAF6, PLC γ 2, MALT1, TAK, TAB1/2, PKC, IKK, NF κ BC1, NF κ BC2), **NF κ B**
IL-1, (IKK, TRAF6, TAK, TAB1/2, NF κ BC1, NF κ BC2), **NF κ B**
dsRNA, (NF κ BC1, PKR, IKK, NF κ BC2), **NF κ B**
TNF, (IKK, MEKK, NF κ BC1, NF κ BC2), **NF κ B**
GF, (AKT.COT, IKK, PI(3)K, NF κ BC2, NF κ BC1), **NF κ B**
LT, (PI(3)K, IKK, NF κ BC1, NF κ BC2, AKT.COT), **NF κ B**
LT, (IKK, NIK, NF κ BC1, NF κ BC2), **NF κ B**
UV, (bTrCP, NF κ BC1, NF κ BC2, JNK), **NF κ B**

Figure 5.6: The (unordered) NF κ B signaling pathways. Membrane receptors are in red (left), and transcription factors are in blue (right). Activation or inhibition information between pathway components are omitted. “Ag” stands for Antigen, “Ag-MHC” stands for Major Histocompatibility Complex (MHC) Antigen, “IL-1” stands for Interleukemia-1, “dsRNA” stands for double stranded RNA, TNF stands for Tumor Necrosis Factor, “GF” stands for Growth Factor, “LT” stands for heat-labile enterotoxin. “NF κ BC1” and “NF κ BC2” stand for NF κ B complexes 1 and 2. The pathway is adapted from <http://www.cellsignal.com/>.

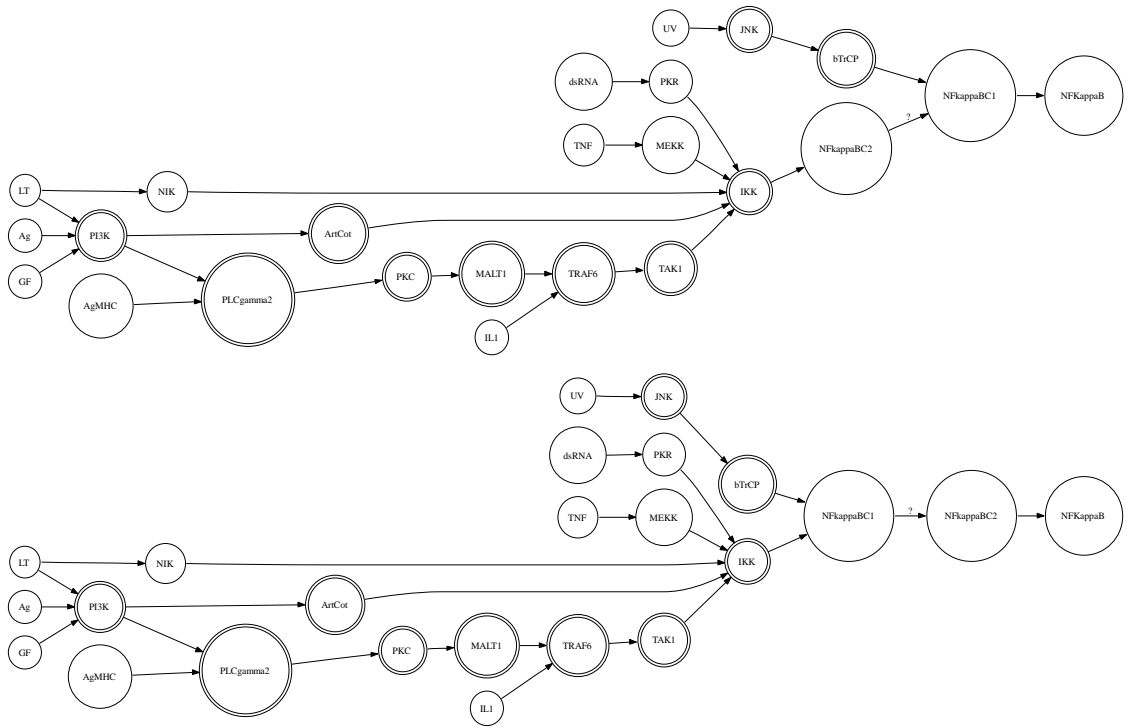


Figure 5.7: The two possible $\text{NF}\kappa\text{B}$ signaling network topologies defined by the probability transition matrix Eq.A.16 and Eq.A.17 estimated from unordered pathway composition data (Fig. 5.6) after incorporating prior information. The relationships between the two double-circled components are disambiguated from prior information. The epistasis relationship labeled with “?” remains ambiguous and deserves further investigation.

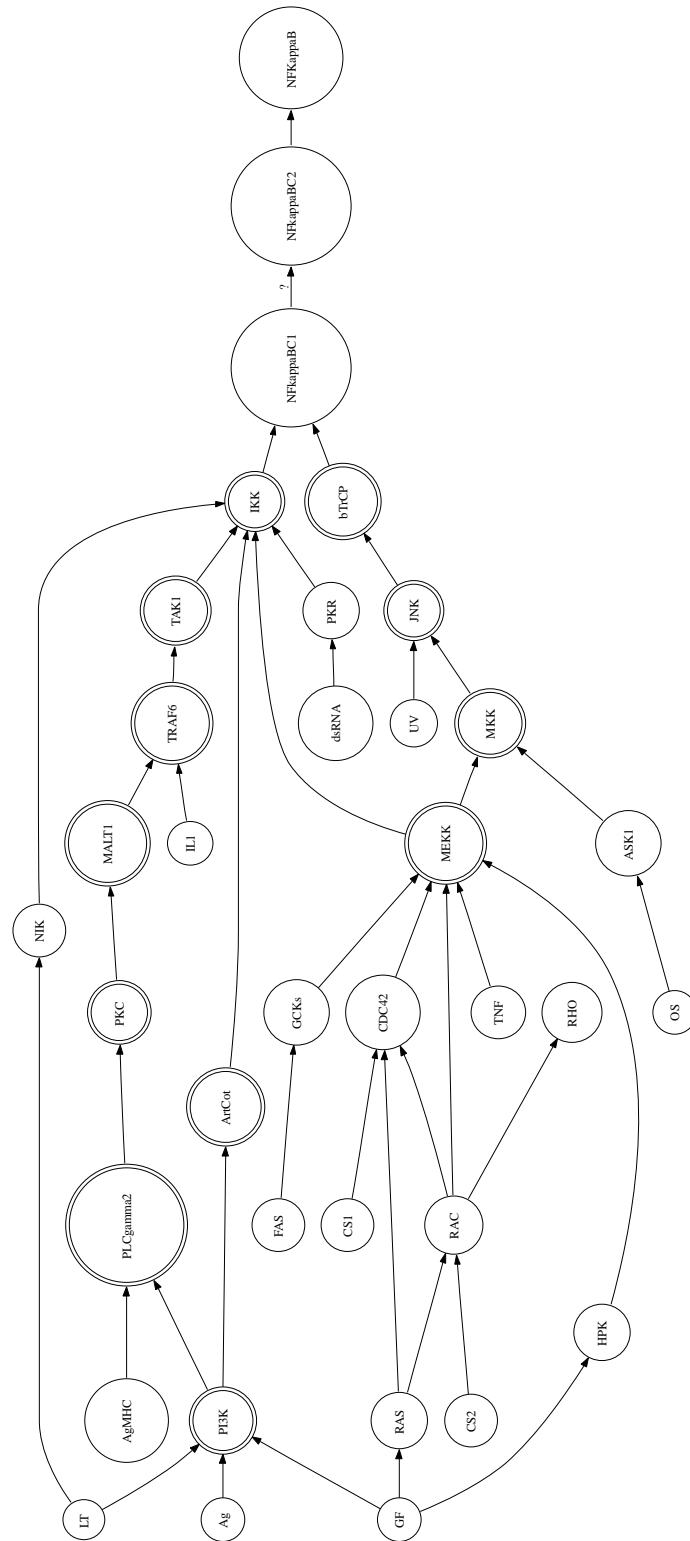


Figure 5.8: The signaling networks assembled from SNK/JNK and NFκB pathways.

CHAPTER VI

Conclusion, Discussion and Future Works

In this thesis, we have addressed the problem of reconstructing gene interaction networks and signaling pathways. We provided a series of logically coherent approaches to attack the problem, including network construction from high throughput data, clustering genes according to similar function while discounting expression dissimilarity, and pathway reconstruction from multiple data sources. We presented a full statistical formulation of the network construction problem, and solved it using a combination of frequentist and Bayesian approaches. By taking into account the underlying network constraint, we then proposed an improved gene clustering approach that is able to group the whole pathway into a single cluster. Given partially known pathway components, e.g. inferred from clustering and/or biologist expert knowledge, we employed a first-order Markov model to reconstruct the order of the entire pathway.

Bio-molecules, including genes, proteins and metabolites etc, exist in a complicated network of tight regulation and interaction. There is considerable interest in inferring the network topology from high throughput data, which is the key to systematic biological discovery. Under the framework of a statistical hypothesis test, the null network topology model may be fully connected, meaning that all pairs

of bio-molecules have direct relationships, e.g. co-regulation, interaction, chemical modification etc. However, the null network model does not reflect biological reality and does not conform to the rules of parsimony in life. In the real world, many biological networks are found to be only partially connected and very sparse. For example, in the metabolic networks of the selected single cell organisms, the “concentration” (defined as the ratio of the total number of network edges over the maximal allowable number of edges) of the edges is estimated to be less than 1% (Zhu and Qin 2005).

Unfortunately, many current data analysis schemes implicitly assume an unconstrained model, e.g., the null network model introduced in Chapter I. More familiar examples are the ‘one-gene-at-a-time’ approaches reviewed in Chapter I, and the traditional clustering approaches reviewed in Chapter IV. One extension of our work could be employment of a complexity constrained model, e.g., implemented by a shrinkage method, in analyzing high throughput biological data. A statistical motivation for such a method lies in the “small n , large p paradigm” and the complexity reducing dependency structure among response variables. A biological motivation is the existence of only a few well connected hub genes or proteins among biomolecules. We briefly discuss some well-known examples of complexity constraints here.

For identifying differentially expressed genes, shrinkage methods have received much recent attention. Examples include Significant Analysis of Microarrays (SAM, Tusher *et al.* 2001), Empirical Bayes (EB, Efron *et al.* 2001) and Penalized Linear Regression Model (Wu *et al.* 2005a, Wu *et al.* 2005b). In SAM and EB methods, the idea of penalizing for complexity of the model was implemented in the framework of variance shrinkage that adds a constant ‘fudge factor’ to the denominator of the ordinary t -test statistic. The fudge factor, estimated from a large number of genes,

penalizes the ranks of those differentially expressed genes with very small variances. Wu et al cast the differential expression detection problem in the familiar framework of linear regression (Wu *et al.* 2005a, Wu *et al.* 2005b). By using the alternative penalized regression model, a penalized t/F-statistic for screening differentially expressed gene was developed. Compared with the former *ad hoc* shrinkage methods such as SAM (Tusher *et al.* 2001), the latter provides a more rigorous and unified statistical framework.

Recently, network constraints have been imposed to identify differentially expressed genes. For example, Morrison et al. adjusted the gene rank obtained from the regular statistical tests using the network structure inferred from gene annotations (gene ontology) or expression profile correlations (Morrison *et al.* 2005). Thus the original gene rank was altered by the corresponding network connectivity that can be treated as a network constraint. This approach is able to reveal additional functionally important genes having weak differential expression. We define the single gene approach as “network constrained screening of differentially expressed genes”. However, there are relatively few studies on imposing multi-gene network constraints to analyze high throughput data analysis. In this thesis, we proposed a generalized multi-gene network constraint using clustering and signaling pathway reconstruction. We think that our success might open an avenue for future research on network constrained high throughput data analysis.

The possibility of future implementations of complexity constraints are certainly warranted. We propose two possible future directions as the closure of this thesis.

One possible future direction is to adapt sophisticated shrinkage methods to the network construction problem. Shrinkage methods developed in diverse statistical areas can be readily be adapted to cope with the small n , large p challenge in inferring

bio-molecule networks from high throughput data. Some of more promising methods are: large-scale multiple test, penalized discriminant analysis, penalized regression, support vector machine (SVM), supervised and unsupervised principal component analysis (Hastie *et al.* 2001). A key point is how to incorporate the network construction problem into the sparsity constrained statistical framework. In particular, the core of network construction problem is to reliably declare the presence and absence of network edges from noisy data with complicated dependency structure. This is highly similar to a number of statistical problems such as large-scale multiple testing (including this work), Bayesian hierarchical model (including this work), logistic regression, SVM and discriminant analysis. Therefore, the recent developments of shrinkage methods for these classical problems can be readily applied to network constructions. More generally, in stead of declaring network edges in a binary manner, we can also view it as multinomial outcomes, in which possible network edges are classified into multiple classes based on levels of confidence. More methods might be adapted, such as decision tree-based methods and their extensions (Hastie *et al.* 2001), e.g. random forest and neural networks.

Another possible future direction is network constrained discovery. Graphical models and network optimization have already been applied to many areas of contemporary bioinformatics. Some of more successful applications are: network flow algorithms applied to protein domain decomposition (Xu *et al.* 2000), protein function prediction (Nabieva *et al.* 2005) and subgraph searching algorithms applied to mining coherent dense subgraphs (Hu *et al.* 2005). The recent developments in network reconstruction techniques with error control provide new opportunities. Network constrained high throughput data analysis remains a very promising area of research.

APPENDICES

APPENDIX A

Technical Details and Supplemental Tables

A.1 Construct PCER-CI for ρ

Here we present the details of constructing asymptotic PCER-CI for ρ as described in section 2.1.2.

Based on the fact that z is the $z = \tanh^{-1}(\hat{\rho})$ monotonic function of $\hat{\rho}$, the asymptotic PCER $(1 - \alpha) \times 100\%$ Confidence Interval: $I^\lambda(\alpha)$ on each true Pearson correlation coefficient ρ of the set \mathcal{G}_1 is: $\tanh(z - \frac{z_{\alpha/2}}{(N-3)^{1/2}}) \leq \rho \leq \tanh(z + \frac{z_{\alpha/2}}{(N-3)^{1/2}})$, where $P(N(0, 1) > z_{\alpha/2}) = \alpha/2$.

A.2 Construct PCER-CI for τ

Here we present the details of constructing asymptotic PCER-CI for τ as described in section 2.1.2.

The asymptotic PCER $(1 - \alpha) \times 100\%$ Confidence Interval: $I^\lambda(\alpha)$ on each true Kendall correlation coefficient τ of the set \mathcal{G}_1 is constructed as follows:

- Compute $C_r = \sum_{\substack{t=1 \\ t \neq r}}^N Q((X_r, Y_r), (X_t, Y_t))$, for $r = 1, 2, \dots, N.$, where $Q((a, b), (c, d))$

is given by:

$$(A.1) \quad Q((a, b), (c, d)) = \begin{cases} 1 & \text{if } (d - b)(c - a) > 0, \\ 0 & \text{if } (d - b)(c - a) = 0, \\ -1 & \text{if } (d - b)(c - a) < 0. \end{cases}$$

- Let $\bar{C} = \frac{1}{N} \sum_{r=1}^N C_r$ and define $\hat{\sigma}_\tau = \frac{2}{N(N-1)} \frac{2(N-2)}{N(N-1)} \sum_{i=1}^N [(C_r - \bar{C})^2 + 1 - \hat{\tau}^2]$
- $I^\lambda(\alpha) : \hat{\tau} - z_{\alpha/2} \hat{\sigma}_\tau \leq \tau \leq \hat{\tau} + z_{\alpha/2} \hat{\sigma}_\tau$.

A.3 Simulating Bivariate Data Based on Pre-specified Population Covariances

Here we present the steps to simulate bivariate data based on pre-specified population covariances as described in section 2.2.1.

Pearson correlation coefficient ρ

- Specify a covariance matrix \mathbf{V} and a mean vector μ .
- Form the Cholesky decomposition of \mathbf{V} , i.e. find the lower triangular matrix L such that $\mathbf{V} = LL^T$.
- Simulate a vector \mathbf{z} with independent $N(0, 1)$ elements.
- A vector simulated from the required multivariate normal distribution is then given by $\mu + L\mathbf{z}$.

Kendall's τ

- Specify a value for τ .
- Simulate an $N \times N$ indicator matrix M given τ as follows:

$$(A.2) \quad M[n, m]_{1 \leq n < m \leq N} = \begin{cases} 1 & \text{if Bernulli}(\frac{1+\tau}{2}) \text{ is TRUE,} \\ -1 & \text{if Otherwise.} \end{cases}$$

- Simulate i.i.d pairs (X_r, Y_r) ($r = 1, 2, \dots, N$) according to M matrix and definition

$$(A.3) \quad Q((a, b), (c, d)) = \begin{cases} 1 & \text{if } (d - b)(c - a) > 0, \\ -1 & \text{if } (d - b)(c - a) < 0. \end{cases}$$

No tied observations are generated. Alternatively, $\hat{\tau}$ can be directly calculated from the indicator matrix M without generating the i.i.d pairs (Eq. 2.3).

A.4 Selecting Prior Distribution

Here we present the mathematical details of choosing a prior as described in section 3.1. They were adapted from the solution to exercises 2.8 in Gelman *et al.* 2004.

We need to show the joint posterior density $p(\Gamma, \alpha, \beta|y)$ is improper if we select the hyperprior distribution $p(\beta) \propto \beta^{-1}$, while $p(\Gamma, \alpha, \beta|y)$ is proper if we select the hyperprior distribution $p(\beta) \propto 1$.

We first factor the joint posterior distribution $p(\Gamma, \alpha, \beta|y) \propto p(\beta|y)p(\alpha|\beta, y)p(\Gamma|\alpha, \beta, y)$. Note that $p(\alpha|\beta, y)$ and $p(\Gamma|\alpha, \beta, y)$ have proper densities. The joint posterior density $p(\Gamma, \alpha, \beta|y)$ is proper if and only if the marginal density $p(\beta|y)$ is proper, i.e. has a finite integral for β from 0 to ∞ .

In Eq. 3.3, as β approaches 0, everything multiplying $p(\beta)$ approaches a nonzero constant limit $C(y)$. Thus the behavior of $p(\beta|y)$ near 0 is determined by the prior density $p(\beta)$. It is easy to show that the function $p(\beta) \propto 1/\beta$ is not integrable for any small interval around 0, and so it leads to a nonintegrable posterior density.

If prior density $p(\beta) \propto 1$, then the posterior density is integrable near zero. We need to examine the behavior as $\beta \rightarrow \infty$ and find an upper bound that is integrable. The exponential term is clearly less than or equal to 1. We can rewrite the remaining

terms as $(\sum_{j=1}^J [\prod_{k \neq j} (\sigma_k^2 + \beta^2)])^{-1/2}$. For $\beta > 1$ we make this quantity bigger by dropping all of the σ^2 to yield $(J\beta^{2(J-1)})^{-1/2}$. An upper bound on $p(\beta|y)$ for β large is $p(\beta)J^{-1/2}/\beta^{J-1}$. When $p(\beta) \propto 1$, this upper bound is integrable if $J > 2$, and so $p(\beta|y)$ is integrable if $J > 2$.

A.5 Deriving Posterior Distribution $p(\beta|y)$

Here we present the mathematical details of deriving posterior distribution $p(\beta|y)$ as described in section 3.1. They were adapted from Chapter V of Gelman *et al.* 2004.

We factor the marginal posterior density of the hyperparameters as follows:

$$(A.4) \quad p(\alpha, \beta|y) = p(\alpha|\beta, y)p(\beta|y),$$

which is equivalent to:

$$(A.5) \quad p(\beta|y) = \frac{p(\alpha, \beta|y)}{p(\alpha|\beta, y)}.$$

We then derive $p(\alpha, \beta|y)$ and $p(\alpha|\beta, y)$ respectively as the following. For hierarchical model, we can simply consider the information supplied by data about the hyperparameters directly:

$$(A.6) \quad p(\alpha, \beta|y) \propto p(\alpha, \beta)p(y|\alpha, \beta).$$

For many problems, decomposition in Eq. A.6 is of no help since $p(y|\alpha, \beta)$ cannot generally be written in closed form. For the Gaussian distribution, the marginal likelihood has a particularly simple form. The marginal distributions of the sample correlation $\widehat{\Gamma}_\lambda$ are independent (but not identically distributed) Gaussian:

$$(A.7) \quad p(\widehat{\Gamma}_\lambda|\alpha, \beta) \propto N(\alpha, \sigma_\lambda^2 + \beta^2).$$

Thus we can write the marginal posterior density as

$$(A.8) \quad p(\alpha, \beta|y) \propto p(\alpha, \beta) \prod_{\lambda=1}^{\Lambda} N(\hat{\Gamma}_\lambda|\alpha, \sigma_\lambda^2 + \beta^2).$$

From inspection of Eq. A.8 with β assumed known, and with a uniform conditional prior density $p(\alpha|\beta)$, where $p(\alpha|\beta, y)$ is also Gaussian, i.e.

$$(A.9) \quad p(\alpha|\beta, y) \propto N(\hat{\alpha}, V_\alpha),$$

where

$$(A.10) \quad \hat{\alpha} = \frac{\sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_\lambda^2 + \beta^2} \hat{\Gamma}_\lambda}{\sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_\lambda^2 + \beta^2}},$$

and

$$(A.11) \quad V_\alpha^{-1} = \sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_\lambda^2 + \beta^2}.$$

$\hat{\alpha}$ is a precision-weighted average of Γ 's and V_α is the total precision. We define precision as inverse of variance. From Eqs. A.5, A.8 and A.9,

$$(A.12) \quad p(\beta|y) = \frac{p(\alpha, \beta|y)}{p(\alpha|\beta, y)}$$

$$(A.13) \quad \propto \frac{p(\beta) \prod_{\lambda=1}^{\Lambda} N(\Gamma_\lambda|\alpha, \sigma_\lambda^2 + \beta^2)}{N(\alpha|\hat{\alpha}, V_\alpha)}$$

This identity holds for any value of α , in particular, it holds if we set α to $\hat{\alpha}$, which makes evaluation of the expression quite simple.

$$(A.14) \quad p(\beta|y) \propto \frac{p(\beta) \prod_{\lambda=1}^{\Lambda} N(\hat{\Gamma}_\lambda|\hat{\alpha}, \sigma_\lambda^2 + \beta^2)}{N(\hat{\alpha}|\hat{\alpha}, V_\alpha)}$$

$$(A.15) \quad \propto p(\beta) V_\alpha^{1/2} \prod_{\lambda=1}^{\Lambda} (\sigma_\lambda^2 + \beta^2)^{-1/2} \exp\left(-\frac{(\hat{\Gamma}_\lambda - \hat{\alpha})^2}{2(\sigma_\lambda^2 + \beta^2)}\right),$$

where $\hat{\alpha}$ and V_α are defined in Eqs. A.10 and A.11. Both expressions are functions of β , which means that $p(\beta|y)$ is a complicated function of β .

Table A.1: Sample output of screening co-expressed gene pairs based on Kendall correlation coefficient. It was described in section 2.3.1.

index1	index2	gene1	gene2	corr	p-value	q-value	lower	higher
971	972	HXT7	HXT6	0.965703	2.63E-09	0.000277	0.893359	1
266	356	RPL11B	GTT2	0.947368	5.22E-09	0.000277	0.834336	1
445	446	ERR1	ERR2	0.947368	5.22E-09	0.000277	0.84075	1
260	261	RPL9B	RPL9A	0.936842	7.69E-09	0.000277	0.821361	1
268	269	RPS23B	RPS23A	0.936842	7.69E-09	0.000277	0.827631	1
254	266	RPL24A	RPL11B	0.93404	8.52E-09	0.000277	0.829735	1
230	356	RPS6B	GTT2	0.926316	1.13E-08	0.000277	0.822449	1
239	301	RPS16B	YPL142C	0.926316	1.13E-08	0.000277	0.822449	1
247	334	RPS18A	ENT4	0.926316	1.13E-08	0.000277	0.755724	1
254	356	RPL24A	GTT2	0.923486	1.25E-08	0.000277	0.794477	1
275	348	YLL044W	SEC65	0.923486	1.25E-08	0.000277	0.797236	1
277	334	RPL42A	ENT4	0.923486	1.25E-08	0.000277	0.81526	1
230	266	RPS6B	RPL11B	0.91579	1.65E-08	0.000277	0.793336	1
233	313	RPL21A	RPS3	0.91579	1.65E-08	0.000277	0.812017	1
253	266	RPL24B	RPL11B	0.91579	1.65E-08	0.000277	0.799229	1
267	356	RPL11A	GTT2	0.91579	1.65E-08	0.000277	0.805438	1
294	295	RPL20B	RPL20A	0.91579	1.65E-08	0.000277	0.772159	1
300	302	RPL33B	RPL33A	0.91579	1.65E-08	0.000277	0.777149	1
249	250	RPL27A	RPL27B	0.912932	1.83E-08	0.000277	0.802223	1

Table A.2: Clustering co-expressed genes with controlled FDR (5%) at a MAS level of 0.6 using “GAL10” as the “seed gene”. Known genes in the pathway are in bold face. Pearson correlation coefficient was used as metric. It was described in section 2.3.2.

index1	index2	gene1	gene2	corr	p-value	q-value	lower	higher
2	2	GAL10	GAL10	1	0.00E+00	0.00E+00	1	
2	1	GAL10	GAL7	0.925103	5.35E-09	2.67E-06	0.727108	0.981023
2	4	GAL10	GCY1	0.91733	1.27E-08	4.20E-06	0.701969	0.97899
2	3	GAL10	GAL1	0.905611	3.99E-08	9.95E-06	0.665053	0.975901
2	59	GAL10	GAL2	0.893609	1.12E-07	2.23E-05	0.628426	0.972709
2	5	GAL10	YOR121C	0.891345	1.34E-07	2.23E-05	0.621649	0.972104

Table A.3: Clustering co-expressed genes with controlled FDR (5%) at a MAS level of 0.5 using “GAL7” as the “seed gene”. Known genes in the pathway are in bold face. (a) Pearson correlation coefficient as metric. It was described in section 2.3.2.

index1	index2	gene1	gene2	corr	p-value	q-value	lower	higher
1	1	GAL7	GAL7	1	0.00E+00	0.00E+00	1	1
1	2	GAL7	GAL10	0.925103	5.35E-09	2.67E-06	0.737186	0.980188
1	62	GAL7	YMR318C	0.892639	1.21E-07	4.03E-05	0.638563	0.971244
1	68	GAL7	YBR042C	0.882089	2.71E-07	5.84E-05	0.608213	0.968289
1	3	GAL7	GAL1	0.880999	2.93E-07	5.84E-05	0.605123	0.967982
1	70	GAL7	FAR1	0.864743	8.72E-07	1.45E-04	0.559998	0.963377
1	59	GAL7	GAL2	0.851884	1.88E-06	2.68E-04	0.525538	0.959693

Table A.4: Clustering co-expressed genes with controlled FDR (5%) at a MAS level of 0.5 using “GAL7” as the “seed gene”. Known genes in the pathway are in bold face. (b) Kendall correlation coefficient as metric. It was described in section 2.3.2.

index1	index2	gene1	gene2	corr	p-value	q-value	lower	higher
1	1	GAL7	GAL7	1	7.07E-10	7.05E-07	1	1
1	3	GAL7	GAL1	0.705263	1.38E-05	6.86E-03	0.442609	0.967917
1	2	GAL7	GAL10	0.652632	5.74E-05	1.88E-02	0.355487	0.949776

Table A.5: Clustering co-expressed genes with controlled FDR (5%) at a MAS level of 0.5 using “GAL1” as the “seed gene”. Known genes in the pathway are in bold face. Pearson correlation coefficient as metric. It was described in section 2.3.2.

index1	index2	gene1	gene2	corr	p-value	q-value	lower	higher
3	3	GAL1	GAL1	1	0.00E+00	0.00E+00	1	1
3	2	GAL1	GAL10	0.905611	3.99E-08	1.99E-05	0.660385	0.976295
3	10	GAL1	FKS1	0.89891	7.22E-08	2.40E-05	0.639567	0.974545
3	1	GAL1	GAL7	0.880999	2.93E-07	7.30E-05	0.585731	0.969822

Table A.6: Clustering co-expressed genes with Bayesian hierarchical model at the significance level 5% using “GAL10” as the “seed gene”. Known genes in the pathway are in bold face ($N = 20$). It was described in section 3.3.2.

Gene1	Gene2	2.5%	50%	97.5%
GAL10	GAL7	0.699967273	0.843269806	0.919377659
GAL10	GCY1	0.695895931	0.83904824	0.917448689
GAL10	GAL1	0.685628575	0.824914454	0.906837751
GAL10	GAL2	0.664031223	0.817631953	0.903466008
GAL10	YOR121C	0.652511568	0.814118521	0.901500909
GAL10	YDR010C	0.574348042	0.77081336	0.875409524
GAL10	YEL057C	0.582835775	0.769743768	0.880618535
GAL10	SSU1	0.584487078	0.769335123	0.879019784
GAL10	PCL10	0.552529392	0.751817344	0.871763977
GAL10	YJL212C	0.543601479	0.747480187	0.862433646
GAL10	MET14	0.525320838	0.723128249	0.852859396
GAL10	FKS1	0.515021843	0.719874179	0.854759107
GAL10	MCM1	0.474061933	0.697313988	0.834101087
GAL10	EXG1	0.446476056	0.666889754	0.818233838
GAL10	ARG1	0.382292245	0.63708452	0.807736956
GAL10	CRH1	0.344971636	0.594425382	0.773435199
GAL10	PRY1	0.299057555	0.588919717	0.774038296
GAL10	YPR157W	0.29645952	0.576125639	0.765975044
GAL10	CPA2	0.303356019	0.571475575	0.745218878
GAL10	YKR012C	0.262900828	0.566724743	0.748081117

APPENDIX B

*

Bibliography

- Akimoto, M., Cheng, H., Zhu, D., Brzezinski, J.A., Khanna, R., Filippova, E., Oh, C.T.E., Jing, Y., Linares, J.L., Brooks, M., Zarepars, S., Mears, A., Hero, A.O., Glaser, T. and Swaroop, A. 2006. Targeting of green fluorescent protein to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors. *Proc. Natl. Acad. Sci. USA*, **103**(10), 3890-3895.
- Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.
- Altschul, S., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver,

- L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25-29.
- Avery, L. and Wasserman, S. 1992. Ordering gene function: the interaction of epistasis in regulatory hierarchies. *Trends. Genet.*, **8**, 312-316.
- Barabási, A. 2004. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101-113.
- Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R. and Califano, A. 2005. Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, **37**, 382-390.
- Batagelj, A. and Mrvar, A. 1998. Pajek - Program for large network analysis. *Connections*, **21**, 47-57.
- Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B. Met.*, **57**, 289-300.
- Benjamini, Y. and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165-1188.
- Benjamini, Y. and Yekutieli, D. 2005. False discovery rate adjusted multiple confidence intervals for selected parameters. *J. Am. Stat. Assoc.*, **100**, 71-80.
- Berg, J.M., Tymoczko, J.L. and Stryer, L. 2006. Biochemistry. W. H. Freeman, New York, USA.
- Bickel, P.J. and Doksum, K.A. 2000. Mathematical statistics: basic ideas and selected topics. 2nd Edition. Prentice Hall, Upper Saddle River, NJ, USA.

- Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185-193.
- Butte, A. and Kohane, I.S. 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, **5**, 415-426.
- Butte, A., Tamayo, P., Slonim, D., Golub, T.R. and Kohane, I.S. 2000. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA*, **97**, 12182-12186.
- Che, T., You, Y., Wang, D., Tanner, M.J., Dixit, V.M. and Lin, X. 2004. MALT1/Paracaspase Is a Signaling Component Downstream of CARMA1 and Mediates T Cell Receptor-induced NF-kappaB Activation. *J. Biol. Chem.*, **279**(16), 15870 -15876.
- Cheng, J., Sun, S., Tracy, A., Hubbell, E., Morris, J., Valmeekam, V., Kimbrough, A., Cline, M.S., Liu, G., Shigeta, R., Kulp, D. and Siani-Rose, M.A. 2004. NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis. *Bioinformatics*, **20**, 1462-1463.
- Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D. and Siani-Rose, M.A. 2004. A knowledge-based clustering algorithm driven by Gene Ontology. *J. Biopharm. Stat.*, **14**(3), 687-700.
- Cormen, T.H., Leiserson, C.E. and Rivest, R.L. 1990. Introduction to Algorithms. MIT Press, Cambridge, MA, USA.
- Costa, J.A. and Hero, A.O. 2004. Geodesic Entropic Graphs for Dimension and En-

- tropy Estimation in Manifold Learning. *IEEE Trans. on Signal Processing*, **52**(8), 2210-2221.
- Daniel, H. 1944. The relation between measures of correlation in the universe of sample permutations. *Biometrika*, **33**, 129-135.
- DeRisi, J., Iyer, V. and Brown, P.O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.
- Dobra, A., Hans, C., Nevins, R., Yao, G. and West, M. 2004. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90**, 196-212.
- Dudoit, S., Fridlyand, J. and Speed, T.P. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**(457), 77-87.
- Dwight, S.S., Harris, M.A., Dolinski, K., Ball, C.A., Binkley, G., Christie, K.R., Fisk, D.G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D. and Cherry, J.M. 2002. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acid Research*, **30**, 69-72.
- Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. 2001. Empirical Bayes Analysis of a Microarray Experiment. *J. Am. Stat. Assoc.*, **96**, 1151-1160.
- Eisen, M., Spellman, P., Brown, P.O. and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863-14868.
- Farkas, I., Jeong, H., Vicsek, T., Barabasi, A.L. and Oltvai, Z.N. 2003. The topol-

- ogy of transcription regulatory network in the yeast, *Saccharomyces cerevisiae*. *Physica. A.*, **318**, 601-612.
- Fisher, R.A. 1923. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 1-32.
- Fleury, G., Hero, A.O., Yoshida, S., Carter, T., Barlow, C. and Swaroop, A. 2002. Pareto analysis for gene filtering in microarray experiments. *Proc. XI European Signal Processing Conference*, Toulouse, France, Sept 2002.
- Friedman, N., Linial, M., Nachman, I. and Pe'er, D. 2000. Using Bayesian Networks to Analyze Expression Data. *J. Comput. Biol.*, **7**, 601-620.
- Fuente, A., Bing, N., Hoeschele, I. and Mendes, P. 2004. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, **20**, 3565-3574.
- Gagneur, J., Jackson, D.B. and Casari, G. 2003. Hierarchical analysis of dependency in metabolic networks. *Bioinformatics*, **19**, 1027-1034.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. 2004. Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton, FL, USA.
- Ghosh, S. and Karin, M. 2002. Missing pieces in the NF-kappaB puzzle. *Cell*, **109**, S81-S96.
- Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

- Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. and Kell, D.B. 2004. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends. Biotechnol.*, **22**, 245-252.
- Hanisch, D., Zien, A., Zimmer, R. and Lengauer, T. 2002. Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**, S145-S154.
- Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N., Macisaac, K.D., Danford, T.D., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E. and Young, R.A. 2004. Transcriptional Regulatory Code of a Eukaryotic Genome. *Nature*, **431**, 99-104.
- Hartigan, J.A. and Wong, M.A. 1979. A k -means clustering algorithm. *Applied Statistics*, **28**, 100-108.
- Hastie, T., Tibshirani, R. and Friedman, J. 2001. The elements of statistical learning. Springer, New York, USA.
- Hero, A.O. and Fleury, G. 2004. Pareto-Optimal Methods for Gene Ranking. *Journ. of VLSI Signal Processing, Special Issue on Genomic Signal Processing*, **38**, 259-275.
- Hero, A.O., Fleury, G., Mears, A. and Swaroop, A. 2004. Multicriteria gene screening for analysis of differential expression with DNA microarrays. *EURASIP Journal on Applied Signal Processing*, **1**, 43-52.
- Hollander, A. and Wolfe, D. 1999. Nonparametric statistical methods. Wiley-Interscience, Hoboken, NJ, USA.

- de Hoon, M.J., Imoto, S., Nolan, J. and Miyano, S. 2004. Open source clustering software. *Bioinformatics*, **20**(9), 1453-1454.
- Hu, H., Yan, X., Huang, Y., Han, J. and Zhou, X.J. 2005. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, **21**, 213-221.
- Huang, D. and Pan, W. 2006. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics*, Feb 24; [Epub ahead of print].
- Hubert, A. 1985. Comparing partitions. *J. Classif.*, **2**, 193-198.
- Humphries, L.A., Dangelmaier, C., Sommer, K., Kipp, K., Kato, R.M., Griffith, N., Bakman, I., Turk, C.W., Daniel, J.L. and Rawlings, D.J. 2004. Tec kinases mediate sustained calcium influx via site-specific tyrosine phosphorylation of the phospholipase CgammaSrc homology 2-Src homology 3 linker. *J. Biol. Chem.*, **279**, 37651-37661.
- Ideker, T., Thorsson, V., Siegel, A.F. and Hood, L.E. 2000. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805-817.
- Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. and Hood, L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929-934.
- Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf,

- U. and Speed, T.P. 2003. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, **4**, 249-264.
- Jeong, H., Mason, S., Barabasi, A.L. and Oltvai, Z.N. 2001. Lethality and centrality in protein networks. *Nature*, **411**, 41-42.
- Kane, L.P., Mollenauer, M.N., Xu, Z., Turck, C.W. and Weiss, A. 2002. Akt-Dependent Phosphorylation Specifically Regulates Cot Induction of NF- κ B-Dependent Transcription. *Mol. Cell. Biol.*, **22**(16), 5962-5974.
- Kerr, M.K., Martin, M. and Churchill, G.A. 2000. Analysis of variance fo gene expression microarray data. *J. Comput. Biol.*, **7**, 819-837.
- Kim, S.Y. and Volsky, D.J. 2005. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.
- Kluger, Y., Yu, H., Qian, J. and Gerstein., M. 2003. Relationship between gene co-expression and probe localization on microarray slides. *BMC Genomics*, **4**, 49.
- Kennedy, G.C., Matsuzaki, H., Dong, S., Liu, W.M., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., Liu, W., Yang, G., Di, X., Ryder, T., He, Z., Surti, U., Phillips, M.S., Boyce-Jacino, M.T., Fodor, S.P. and Jones, K.W. 2003. Large-scale genotyping of complex DNA. *Nat. Biotechnol.*, **21**, 1233-1237.
- LaFramboise, T., Weir, B.A., Zhao X., Beroukhim, R., Li, C., Harrington, D., Sellers, W.R. and Meyerson, M. 2005. Allele-Specific Amplification in Cancer Revealed by SNP Array Analysis. *PLoS Comput. Biol.*, **1**(6):e65.
- Ledoit, O. and Wolf, M. 2004. A well conditioned estimator for largedimensional covariance matrices. *J. Multiv. Anal.*, **88**, 365-411.

- Lee, H., Hsu, A., Sajdak, J., Qin, J. and Pavlidis, P. 2004. Coexpression analysis of human genes across many microarray data sets. *Genome. Res.*, **14**, 1085-1094.
- Lee, M-LT. 2004. Analysis of Microarray Gene Expression Data. Kluwer Academic Publishers, Boston, MA, USA.
- Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.R., Thompson, C.M., Simon I., Zeitlinger J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J., Volkert T.L., Fraenkel, E., Gifford D.K. and Young, R.A. 2002. Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799-804..
- Li, C. and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA*, **98**, 31-36.
- Liang, S., Fuhrman, S. and Somogyi, R. 1998. Reveal a general reverse engineering algorithm for inference of genetic network architecture. *Pacific Symposium on Biocomputing*, **3**, 18-29.
- Liu, Y., and Zhao, H. 2004. A computational approach for ordering signal transduction pathway components from genomics and proteomics data. *BMC Bioinformatics*, **5**:158.
- Lockhart, D., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675-1680.

- Loomis, W.F. 1998. Role of PKA in the timing of developmental events in *Dicystostelium* cells. *Microbiol. Mol. Biol. Rev.*, **62**(3), 684-694.
- Lu, T., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S. and Park, P.J. 2005. Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. USA*, **12**, 13544-13549.
- Ma, H.W., Buer, J. and Zeng, A.P. 2004. Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics*, **5**, 199.
- Ma, H.W. and Zeng, A.P. 2003. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, **19**, 1423-1430.
- Ma, H.W., Zhao, X.M., Yuan, Y.J. and Zeng, A.P. 2004. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics*, **20**, 1870-1876.
- Magwene, P.M. and Zeng, A.P. 2004. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol*, **5**(12):R100.
- McLachlan, G., Bean, R. and Peel, D. 2002. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413-422.
- Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**(3), 267-273.

- Morlon, A., Munnich, A. and Smahi A. 2005. TAB2, TRAF6 and TAK1 are involved in NF- κ B activation induced by the TNF-receptor, Edar and its adaptator Edaradd. *Human Molecular Genetics*, **14**(23), 3751-3757.
- Morrison, J.L., Breitling, R., Higham, D.J. and Gilbert, D.R. 2005. GeneRank: Using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, **6**:233.
- Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. and Singh, M. 2005. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **S1**, i302-i310.
- Nixon, T., Ronson, C. and Ausubel, F.M. 1986. Two-component regulatory systems responsive to environmental stimuli share strongly conserved domains with the nitrogen assimilation regulatory genes ntrB and ntrC. *Proc. Natl. Acad. Sci. USA*, **83**, 7850-7854.
- Patterson, S.D. and Aebersold, R.H. 2003. Proteomics: the first decade and beyond. *Nat. Genet.*, **33**, 311-323.
- Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J. and dAlchBuc, F. 2003. Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, **19**, ii138-ii148.
- Pomerantz, J.L. and Baltimore, D. 2002. Two pathways to NF- κ B. *Mol. Cell*, **10**, 693-695.
- Rabbat, M.G., Figueiredo, A.T. and Nowak, R.D. 2006. Network inference from co-occurrence. University of Wisconsin - Madison technical report ECE-06-2.

- Rand, W. 1971. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 846-850.
- Rao, A., Hero, A.O., Engel, J.D., States, D.J. and Zhu, D. 2005. Inferring Time-varying Network Topologies from Gene Expression Data. *Proc. of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS'05)*, Newport, May 2005.
- Reiner, A., Yekutieli, D. and Benjamini, Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 386-375.
- Rickman, D.S., Bobek, M.P., Misek, D.E., Kuick, R., Blaivas, M., Kurnit, D.M., Taylor, J. and Hanash, S.M. 2001. Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Research*, **61**(18), 6885-6891.
- Rohde, J., Trinh, J. and Sadowski, I. 2000. Multiple signals regulate GAL transcription in yeast. *Mol. Cell. Biol.*, **20**, 3880-3886.
- Schafer, J. and Strimmer, K. 2005. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754-764.
- Schafer, J. and Strimmer, K. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.*, **4**, 32.
- Schliep, A., Schonhuth, A. and Steinhoff, C. 2003. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*, **19**, i255-i263.
- Shedden, K., Chen, W., Kuick, R., Ghosh, D., Macdonald, J., Cho, K.R., Giordano, T.J., Gruber, S.B., Fearon, E.R., Taylor, J.M.G. and Hanash, S. 2005. Compar-

- ison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. *BMC Bioinformatics*, **6**:26.
- Silva, V. and Tenenbaum, J.B. 2002. Global versus local methods in nonlinear dimensionality reduction. *Neural Information Processing Systems 15 (NIPS)*, Vancouver, Canada, Dec. 2002.
- Speed, T. ed. 2003. Statistical analysis of gene expression microarray data. Chapman & Hall/CRC Press, Boca Raton, Fla, USA.
- Spiegelman, V.S., Stavropoulos, P., Latres, E., Pagano, M., Ronai, Z., Slaga, T.J. and Fuchs, S.Y. 2001. Induction of β -Transducin Repeat-containing Protein by JNK Signaling and Its Role in the Activation of NF κ B. *J. Biol. Chem*, **276**(29), 27152-27158.
- Stock, M., Victoria, L. and Goudreau, P.N. 2000. Two-component signal transduction. *Annual Review of Biochemistry*, **69**, 183-215.
- Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249-255
- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**(43), 15545-15550.
- Sun, L., Deng, L., Ea, C.K., Xia, Z.P. and Chen, Z.J. 2004. The TRAF6 ubiquitin ligase and TAK1 kinase mediate IKK activation by BCL10 and MALT1 in T lymphocytes. *Cell*, **14**(3), 289-301.

- Swaroop A., Xu, J.Z., Pawar. H., Jackson, A., Skolnick C. and Agarwal, N. 1992. A conserved retina-specific gene encodes a basic motif/leucine zipper domain. *Proc. Natl. Acad. Sci. USA*, **89**, 266-270.
- Szallasi, Z. and Liang, S. 1998. Modeling the normal and neoplastic cell cycle with “realistic Boolean genetic networks”: Their application for understanding carcinogenesis and assessing therapeutic strategies. *Pacific Symposium on Biocomputing*, **3**, 66-76.
- Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. and Wong, W.H. 2001. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variation and assessment of gene effects. *Nucleic Acids Research*, **29**, 2549-2557.
- Tseng, G.C. and Wong, W.H. 2005. Tight clustering: A Resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**, 10-16.
- Tusher, V., Tibshirani, R. and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116-5121.
- Van Driessche, N., Demisar, J., Booth, E.O., Hill, P., Juvan, P., Zupan, B., Kuspa, A. and Shaulsky, G. 2005. Epistasis analysis with global transcriptional phenotypes. *Nat. Genet.*, **37**(5), 471-477.
- Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science*, **270**, 484-487.
- Vogelstein, B., Lane, D. and Levine, A.J. 2000. Surfing the p53 network. *Nature*, **408**(6810), 307-310.

- Wang, J., Myklebost, O. and Hovig, E. 2003. MGraph: graphical models for microarray data analysis. *Bioinformatics*, **19**, 2210-2211.
- Weston, C.R. and Davis, R.J. 2002. The JNK signal transduction pathway. *Curr Opin. Genet. Dev.*, **12**, 14-21.
- Wieczorke, R., Krampe, S., Weierstall, T., Freidel, K., Hollenberg, C.P. and Boles, E. 1999. Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in *Saccharomyces cerevisiae*. *FEBS Lett.*, **464**, 123-128.
- Whittaker, J. 1990. Graphic models in applied multivariate statistics. Wiley, New York, USA.
- Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R.S. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**(6), 625-638.
- Wu, B. 2005a. Differential gene expression detection using penalized linear regression models: the improved SAM statistics. *Bioinformatics*, **21**(8), 1565-1571.
- Wu, B. 2005b. Differential gene expression detection and sample classification using penalized linear regression models. *Bioinformatics*, **22**(4), 472-476.
- Wu, Z., Irizarry, R.A., Gentleman, R., Murillo, F.M. and Spencer, F. 2004. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *J. Am. Stat. Assoc.*, **9**, 909-917.
- Wu, Z. and Rafael, A. 2005. Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays. *J. Comput. Biol.*, **12**(6), 882-893.
- Wuensche, A. 1998. Genomic regulation modeled as a network with basins of attraction. *Pacific Symposium on Biocomputing*, 89-102.

- Xu, Y., Xu, D. and Gabow, H.N. 2000. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, **14**, 1091-1104.
- Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. and Speed, T.P. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**(4):e15.
- Yang, Y.H., Buckley, M.J., Dudoit, S. and Speed, T.P. 2002. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, **11**(1), 108-136.
- Yang, Y.H., Xiao, Y. and Segal, M.R. 2005. Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics*, **21**(7), 1084-1093.
- Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L. 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **10**, 977-987.
- Yeung, K.Y., Medvedovic, M. and Bumgarner, E.A. 2003. Clustering gene-expression data with repeated measurements. *Genome Biology*, **4**:R34.
- Yeung, M., Tegner, J. and Collins, J.J. 2002. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA*, **99**, 6163-6168.
- Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J. and Jarvis1, E.D. 2004. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, **20**, 3594-3603.

- Zarepari,S., Hero,A.O., Zack,D.J., Williams,R. and Swaroop,A. (2004) Seeing the unseen: Microarray-based gene expression profiling in vision. *Invest Ophthalmol Vis Sci.*, **45**, 2457-2462.
- Zhang, L., Miles, M.F. and Aldape, F.D. 2003. A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818-821.
- Zhao, X., Li, C., Paez, J.G, Chin, K., Janne, P.A., Chen, T.H., Girard, L., Minna, J., Christiani, D., Leo, C., Gray, J.W., Sellers, W.R. and Meyerson, M. 2004. An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. *Cancer Research*, **64**, 3060-3071.
- Zhou, X.J. and Gibson, G. 2004. Cross-species comparison of genome-wide expression patterns. *Genome. Biol.*, **5**(7), 232.
- Zhou, X.J., Kao, M. and Wong, W.H. 2002. Transitive functional annotation by shortest path analysis of gene expression data. *Proc. Natl. Acad. Sci. USA*, **99**, 12783-12788.
- Zhou, X.J., Kao, M., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O.M., Finch, C.E., Morgan, T.E. and Wong, W.H. 2005. Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat. Biotechnol.*, **23**, 238-243.
- Zhu, D., Hero, A.O., Qin, Z.S. and Swaroop, A. 2005a. High throughput screening of co-expressed gene pairs with controlled False Discovery Rate (FDR) and Minimum Acceptable Strength (MAS). *J. Comput. Biol.*, **12**, 1029-1045.
- Zhu, D. and Hero, A.O. 2005b. Identifying differentially expressed genes from probe level intensities in longitudinal Affymetrix microarray experiments. *IEEE Interna-*

- tional Workshop on Statistical Signal Processing (SSP'05)*, Bordeaux, France, July 2005.
- Zhu, D., Hero, A.O., Cheng, H., Khanna, R. and Swaroop, A. 2005c. Network constrained clustering for gene microarray data. *Bioinformatics*, **21**(21), 4014-4021.
- Zhu, D. and Hero, A.O. 2005d. Network constraint clustering for gene microarray data. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, USA, March 2005.
- Zhu, D., Hero, A.O. and Swaroop, A. 2005e. An unsupervised posterior analysis of signaling pathways from gene microarray data. *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS'05)*, New Port, Rhode Island, USA, May 2005.
- Zhu, D. and Hero, A.O. 2005f. Gene co-expression network discovery with controlled statistical and biological significance. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, USA, March 2005.
- Zhu, D. and Hero, A.O. 2005g. Bayesian hierarchical model for estimating gene association networks from microarray data. *Proc. IEEE International Workshop on Genomics signal processing and statistics (GENSIPS'06)*, College Station, Texas, USA, May 2006.
- Zhu, D. and Qin, Z.S. 2005. Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics*, **6**:8.

ABSTRACT

Reconstructing Signaling Pathways from High Throughput Data

by

Dongxiao Zhu

Chair: Alfred O Hero

Many bioinformatics problems can be tackled from a fresh angle offered by the network perspective. Taking into account the network constraints on gene interaction, we propose a series of logically-coherent approaches to reconstruct signaling pathways from high throughput expression profiling data. These approaches proceed in three consecutive steps: co-expression network construction with controlled biological and statistical significance, network constrained clustering, and reconstruction of the order of pathway components.

The first step relies on detecting pairwise co-expression of genes. We attack the problem from both frequentist statistics and Bayesian statistics perspectives. We designed and implemented a frequentist two-stage co-expression detection algorithm that controls both statistical significance (False Discovery Rate, FDR) and biological significance (Minimum Acceptable Strength, MAS) of the discovered co-expressions. In order to regularize variances of the correlation estimation in small sample sce-

nario, we also designed and implemented a Bayesian hierarchical model, in which correlation parameters are assumed to be *exchangeable* and sampled from a parental Gaussian distribution. Using simulated data and the galactose metabolism data, we demonstrated advantages of our approaches and compared the differences among them.

The second problem considered is distance-based clustering that accounts for “network constraints” extracted from the Giant Connected Component (GCC) of the network discovered from the data. The clustering is performed using a “hybrid” distance matrix composed of direct distance between adjacent genes and “shortest-path” distance between non-adjacent genes in the network. The third problem considered is the reconstruction of the order of pathway components. We applied a first-order Markov model, originally developed and applied to a network tomography problem in telecommunication networks, to reconstruct three well-known signaling pathways from unordered pathway components. We suggest that the methods proposed here can also be applied to other high throughput data analysis problems.