BAYESIAN ANALYSIS IN CANCER PATHWAY STUDIES AND PROBABILISTIC PATHWAY ANNOTATION

by

Haige Shen

Program in Computational Biology & Bioinformatics
Duke University

Date: _____

Approved:

_____
Dr. Mike West, Co-Supervisor

_____
Dr. Jen-Tsan Ashley Chi, Co-Supervisor

_____
Dr. Edwin S. Iversen, Jr.

_____
Dr. Sayan Mukherjee

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Program in Computational Biology & Bioinformatics
in the Graduate School of
Duke University

2008

ABSTRACT

BAYESIAN ANALYSIS IN CANCER PATHWAY STUDIES AND PROBABILISTIC PATHWAY ANNOTATION

by

Haige Shen

Program in Computational Biology & Bioinformatics
Duke University

Date: _____

Approved:

_____
Dr. Mike West, Co-Supervisor

_____
Dr. Jen-Tsan Ashley Chi, Co-Supervisor

_____
Dr. Edwin S. Iversen, Jr.

_____
Dr. Sayan Mukherjee

An abstract of a dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Program in Computational Biology & Bioinformatics
in the Graduate School of
Duke University

2008

Abstract

Improving the understanding of the complexity of molecular pathways underlying cancer phenotypes is essential to uncovering the dynamic processes of cancer development. As part of this, linking quantified, experimentally defined gene expression signatures with known biological pathway gene sets is a key challenge. This dissertation presents a novel Bayesian statistical approach to this pathway annotation problem.

In my approach, a formal probabilistic model delivers probabilities over pathways for an experimental signature, thus allowing a quantitative assessment and ranking of pathways putatively linked to the experimental phenotype. The fundamental advantage of this approach is formal modeling of the uncertainty in the pathway analysis. Biological understanding of the data and knowledge are incorporated in the model. In addition, coherent inference on uncertainties about gene pathway membership highlights a key benefit of this model-based approach.

Technically, this research involves advanced statistical modeling and high-dimensional computation. Analysis of the models uses Markov chain Monte Carlo techniques and variational methods for statistical computation. To evaluate model evidence, a critical component of pathway analysis, I propose an innovative Monte Carlo variational method that provides optimal upper and lower bounds on model evidence. This method, motivated and developed by genomic pathway analysis, is in fact general and represents an advance in statistical model-based computation of much broader utility.

The effectiveness and robustness of my approach are tested through simulation studies as well as analyses of real data sets, including "proof-of-principle" pathway annotation for breast tumor estrogen-receptor and ErbB2 phenotypes. A study of path-

way activities underlying the cellular response to lactic acidosis micro-environment in breast tumors involves the analyses of both *in vitro* and *in vivo* data, and demonstrates the application of the method in decomposing the complexity of gene expression-based predictions about interacting pathway activation in this cancer context.

In conclusion, this dissertation generates innovation in statistical methodology as well as in cancer genomics applications. Current and future research plans and directions include broad opportunities for application and evaluation in cancer genomics studies, as well as in other areas of genomics, and follow-on efficient computer implementations for use of the method by the research community.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to acknowledge all those people who have helped me over the last few years. Because of them, my graduate experience at Duke has been one that I will cherish forever.

My deepest gratitude is to my advisor, Dr. Mike West. I have been amazingly fortunate to have an advisor who is always generous and supportive of his students. He encouraged me to develop independent scientific thinking, guided me with his constructive ideas and insightful comments, and helped me sort out the technical details of my work. He always responded to my requests more quickly than I could have hoped. He has created all kinds opportunities for the development of my life and career. His continual patience and support helped me overcome many crisis situations and finish this dissertation.

I am deeply grateful to my co-advisor, Dr. Jen-Tsan Chi, for his encouragement and practical advice. The discussions with him enriched my ideas, and his comments on my views greatly helped me understand my work from the biological perspective. Working with him was a very valuable and happy experience in my graduate study.

I am also very grateful for having an exceptional doctoral committee and wish to thank Dr. Edwin Iverson and Dr. Sayan Mukherjee. They have always been there to listen and give me advice. Sayan, my good friend, provided much extraordinarily perceptive advice and helped me develop my work and career. The conversations with him were thought-provoking, and became one of the most enjoyable parts of my academic life. I wish to thank Dr. Philip Febbo. Phil assisted me with the scientific writing of my thesis proposal and provided many insightful and stylistic suggestions.

I would like to thank my friends and student-colleagues who helped me finish my dissertation. My collaboration with Chunlin Ji has led to interesting research results

in statistical computation methodology. His passion and dedication to academic research greatly inspired me. Many thanks to Craig Reeson, Daniel Mace and Matthew Heaton for their generous time and patience in proof-reading my dissertation. Special thanks to Parawee (Nicky) Lekprasert for sharing my feelings during the writing of this dissertation. I loved her cookies and Thai food.

I am particularly thankful to my family, who have been a constant source of love, concern, support and strength all these years. I'm especially grateful to my husband and best friend, Yanyu, for his patience and for helping me keep my life in proper perspective and balance. I have been extremely fortunate to have the best parents-in-law. This work would not have been possible without their support. This dissertation is dedicated to my son Yuangeng, who was born during this graduate study, and is the highlight and the greatest achievement in my life. Finally, I would like to express my heartfelt gratitude to my dear mom, dad and sister: Your unconditional love is my most cherished wealth. And I promise you, this is my last PhD.

Chapter 1

Introduction

Substantial evidence indicates that most cancers are derived from sequential genetic alterations that deregulate cellular growth through specific cellular signaling pathways. Although all cancer cells share common features of malignant growth associated with some common biological pathway activities, different categories and subcategories of cancers have their own specific mechanisms. Improving the understanding of the complexity of molecular pathways underlying cancer phenotypes is essential to uncovering the dynamic processes of cancer development and identifying cancer prognostic factors and therapeutic targets.

In cancer research, genome-wide gene expression profiles are generated for specific cancer-related phenotypes in experiments either on normal tissues or tumors, such as cancer micro-environments, oncogene perturbation and cancer subtypes. Identifying pathway activities associated with certain cancer-related phenotypes is the basic goal, addressed by capturing the characteristics of cancer cell behavior on the transcriptional level. Moreover, due to aggregated efforts in cancer genomic research, much genome-wide gene expression data on tumor samples from different sources is available and of great value for studies of tumor heterogeneity. Many statistical methods have been developed to analyze such cancer profiling data. Of key interest here is Bayesian factor regression modeling (Carvalho *et al.* 2007; Lucas *et al.* 2006) that provides a comprehensive tool for cancer signature identification and molecular phenotype dissection using expression data.

Pathway databases have collected information on cancer signaling pathways derived from biological molecular interaction studies. Besides these, many cancer-

oriented gene expression profiling experiments, such as oncogene or tumor suppressor gene alteration and growth factor stimulation, have been conducted on cell lines and animal model organisms. The analyses of these gene expression profiles have provided gene expression signatures characterizing or predictive of certain biological phenotypes. Presumably a set of signature genes contains genes that are likely to be participants in underlying biological processes – in other words, members of underlying biological pathways. Such existing pathway gene sets and signatures can be used as references for pathway annotation in new contexts.

Pathway annotation involves borrowing knowledge from the signaling pathway databases to investigate pathway activities involved in a current experimental context. Here, linking quantified, experimentally defined gene expression signatures with known biological pathway gene sets is a key challenge. A necessary step following annotation is to reveal the specificity of the identified pathways in the current biological context. This is the key to imputing the connectivity among pathway modules as well as to identifying biomarkers and therapeutic targets. None of the existing pathway annotation methods were developed based on modeling of the quantified association between gene expression and biological phenotypes. As a result, biological understanding of the data and pathways are really not incorporated in the methods; nor is the context-based pathway specificity problem typically addressed. The existing methods are reviewed in Chapter 2 following the introduction of pathway-oriented cancer mechanism studies and Bayesian factor regression modeling.

Motivated by these requirements in practical research, this thesis develops a Bayesian statistical approach, called probabilistic pathway annotation (PROPA), for the general problem of genome-wide expression-based pathway annotation. This involves a model-based approach to matching experimental signatures of structure or outcomes in gene expression to multiple biological pathway gene sets from curated

databases. This formal probabilistic model delivers probabilities over pathways for each experimental signature, thus allowing for a quantitative assessment and ranking of pathways putatively linked to the experimental phenotype. The fundamental advantage of this approach is formal modeling of the uncertainty in the pathway analysis. For example, accuracy and relevance of genes in *á priori* defined sets can easily be incorporated, and inference on uncertainty about gene set membership is transparent. In Chapter 3, the statistical model for this approach is introduced and described in detail.

This approach involves Bayesian inference and model comparisons. A pathway gene set is refined through inference on pathway membership probabilities; multiple pathways are compared in terms of their association with the studied phenotype through model comparisons. Both tasks rely on posterior simulation for solutions. Chapter 4 discusses the Markov chain Monte Carlo method for the PROPA model, and demonstrates it in a simulation study. Inference on gene pathway membership is discussed in the simulation example.

Multiple pathway comparisons, formulated as a Bayesian model comparison problem in PROPA, involve the evaluation of marginal likelihoods or model evidence – this generates difficult integration problems as is encountered in other areas. The high-dimensionality of the genome-wide gene expression data causes intractability in both the analytical and numerical evaluation of the integrals. This core computational biology and statistical problem is addressed in Chapter 5 with a study of the characteristics of joint probability density functions in the PROPA model. Numerical integration methods for marginal likelihood approximation, including quadrature, Laplace approximation, importance sampling and posterior sampling integration, are reviewed and evaluated. These methods either fail or are very hard to use in this situation due to the difficulty caused by the high dimension. Some of these methods

are tested in a simple simulation study that aims to show PROPA analysis as effective for pathway comparisons.

To solve the marginal likelihood evaluation problem, I propose a novel Monte Carlo variational approximation method, which involves innovation in statistical methodology generally as well as an effective method for PROPA. Traditional mean-field variational methods, which use conditionally optimal lower bounds to approximate the marginal likelihoods, have desirable mathematical forms that avoid the numerical problems caused by high-dimensional data. Inspired by this idea, I derive a general optimization method based on Monte Carlo simulation and stochastic approximation to achieve a conditionally optimal lower bound on the marginal likelihood. More importantly, I propose a novel method that combines posterior simulation and mean-field variational approximation, providing a conditionally optimal upper bound to the marginal likelihood. This lower and upper bounding strategy successfully solves the statistical computation problem in PROPA. In addition, this generates advances in the computation of marginal probabilities generally, a key problem in Bayesian statistical inference, by implementing double-sided bounding. In Chapter 6, this method is discussed in detail and demonstrated in simulation studies.

In Chapter 7, two examples on real data sets are presented to demonstrate the effectiveness and robustness of PROPA. In the first example, pathway activities related to breast tumor estrogen-receptor status are analyzed using PROPA. The comparison of this analysis with that given by gene set enrichment analysis (GSEA) (a pathway annotation method reviewed in Chapter 2) shows the effectiveness, and some of the unique features and benefits, of PROPA. The robustness of the PROPA model is studied through observing the change of annotation results caused by variation in model hyper-parameters and data distributions. The effectiveness of PROPA is then also demonstrated in the analysis of pathway activities underlying breast tumor ErbB2

status, another important phenotype in breast cancer. This example also highlights inference on gene pathway membership probabilities, a further key feature and benefit of the new model framework.

Chapter 8 presents an application of PROPA in studies of cancer developmental mechanisms. A main area of applied interest is in the study of pathway activities associated with lactic acidosis, a key micro-environmental factor in solid tumors. Pathway annotation analyses of lactic acidosis status in human mammary epithelial cells and breast tumor cells identify pathways elucidating the nature of lactic acidosis as a potential signal for cells as well as linkages to key risk signatures in breast cancers. Through this study, I demonstrate that PROPA combined with the gene signature dissection and enhancement methodology, shows considerable promise and utility in cancer genomic studies.

Current and potential future research directions are discussed in Chapter 9.

The research work in this dissertation is implemented in MATLAB and generates a software package named PROPA. The major functional components includes gene set curation, gene-phenotype data pre-processing, pathway annotation, gene set refinement and result presentation. This software package will be available at http://www.stat.duke.edu/.

Chapter 2

Pathway Annotation in Cancer Biology

Cancer begins with genetic mutations, which cause signal cascades leading to the stimulation of cell proliferation or inhibition of cell-cycle arrest, and suppression of apoptosis (Evan and Vousden 2001; Kufe *et al.* 2006; Hanahan and Weinberg 2000; Jones and Baylin 2002; Esteller and Herman 2002). During cancer development, sequential genetic alterations occur as both results and driving forces of cancerous signaling and cell evolution pressured by the extracellular environment. Therefore, the key to understanding the mechanisms of cancer development lies in the investigation of cancer signaling pathways as the interconnected modules in cancer cellular circuitry. In this chapter, I first justify the pathway approach as reasonable and feasible to cancer mechanism investigation. An introduction is given to the Bayesian factor regression modeling methodology, a powerful tool for cancer phenotype dissection based on genome-wide gene expression profiles that forms a major context for development of my pathway annotation analysis method. Then, I motivate model-based approaches through a review of existing pathway annotation methods.

2.1 Understanding Cancer Through Pathways

Cancer is the complex result of multiple sequential genetic mutations. Very few cancer types are the results of single gene mutation, a typical example of which is the inherited eye tumor caused by a loss of function of retinoblastoma gene (Rb)(Knudson 1971; Friend *et al.* 1986; Cavenee *et al.* 1983). Development of the vast majority of cancer types involves more than one genetic mutation and presents as multi-step events. Vogelstein and Kinzler (1993) proposed a progression model for colorectal

neoplasia. In this model, colon cancer development was hypothesized as a step-wise process, including abnormal initiation in colonic epithelium, propagation and local aggregation of the abnormality to form adenomatous polyps, and increase of malignancy as infiltrating adenocarcinoma. Each step involves different contributing genetic mutations. This multi-step and multi-mutant nature of cancer was also demonstrated by Hahn *et al.* (1999) in their work on tumorigenic conversion of normal human epithelial fibroblast cells.

Effects of genetic mutations differ in different stages of cancer progression, making the assessment of the effects possibly incomplete. Generally, the common mechanism under cancer progression is that the activation of oncogenes, inactivation of tumor suppressor genes and malfunction of stability genes result in the deregulation in cellular signaling pathways controlling cell proliferation and apoptosis (Vogelstein and Kinzler 2004). Despite the fact that tumors seem to be associated with the expression of oncogenes (Giuriato et al.,2004), some oncogene downstream effectors are more important than the oncogenes themselves for maintaining tumor growth once the tumors are established (realistically, tumors are usually observed in this stage). As an example, Lim and Counter (2005) demonstrated that the PI3K/ATK pathway, one of the residual activities of the upstream Ras oncogene, is sufficient for tumor maintenance in human cells.

The number of tumorigenic gene mutations appears to be much larger than what is currently known; meanwhile, mutant gene sets of different tumor types have few overlaps, and even the mutations within a single histological tumor type are highly heterogeneous. Sjöblom *et al.* (2006) presented the evidence for this in their recent study of coding sequences consensus in human breast and colorectal cancers. The diversity and heterogeneity of tumorigenic mutations make it difficult to predict the behavior, prognosis, or therapy response of tumors based on a single set of gene

mutations.

It has been indicated that different genetic mutations may have similar functions and play equivalent roles in cancer cell signaling pathways (Vogelstein and Kinzler 2004). Sjöblom *et al.* (2006) suggested that mutant genes might possibly be grouped into a limited number of pathways. Hence, exploring signaling pathway activities under cancer phenotypes to decompose the relationship between genetic mutations and cancer development may be a way to probe the complex circuitry of cancer cells and advance the understanding of cancer development.

High-throughput gene expression technology makes comprehensive genomic analyses of human cancers possible (Schena *et al.* 1995). Through the observation of genome-scale gene expression profiles, genes are linked in a shared phenotype which reflects specific underlying biological processes or pathway modules. Genes expressed at higher or lower levels in cancerous cells with a certain phenotype are involved in the biological processes, and are components of the signaling pathways contributing to this phenotype. Many studies have shown that such pathway gene sets themselves, in absence of interaction information, are able to distinguish cancer phenotypes (West *et al.* 2001; van't Veer *et al.* 2002; Bild *et al.* 2006). This provides fundamental support for pathway annotation approaches to cancer research using genome-wide gene expression data. Many statistical methods have been developed to identify differentially expressed genes. I focus on Bayesian methodology that has the ability to identify genes associated with explicit phenotypes and dissect complex molecular phenotypes in heterogeneous tumor samples to facilitate the studies of cancer development.

2.2 Bayesian Factor Regression Modeling

Bayesian factor regression modeling (BFRM), a methodology developed for analysis of high-dimensional data, has been applied in cancer biology studies based on genome-wide gene expression profiles (West 2003; Lucas *et al.* 2006; Carvalho *et al.* 2007; Lucas *et al.* 2007; Chen *et al.* 2007). This methodology considers sparsity, a key concept in practical high-dimensional data analysis, in a generalized multivariate regression framework, and combines it with ANOVA and latent factor modeling to correct experimental artifacts, dissect transcriptional responses to biological perturbations and explore the underlying gene expression patterns predictive of cancer phenotypes. Such multiple tasks are either fulfilled separately or combined in one analysis according to the goals of biological studies. Without covering the thorough details and complete framework, I briefly introduce BFRM with a focus on the form directly linked to the topic of pathway annotation for tumorigenesis studies.

The fundamental framework of BFRM is modeling gene expression as a linear combination of a number of factors with biological meaning, either explicit or yet to be identified. For each of n tissue samples, observations are made on the expression levels of p genes. The expression of gene g in sample i is denoted by $x_{g,i}$ and modeled as a conditionally Gaussian random variable

$$x_{g,i} = \mu_g + \sum_{j=1}^{r} \beta_{g,j} h_{j,i} + \sum_{j=1}^{k} \alpha_{g,j} \lambda_{j,i} + \nu_{g,i}, \tag{2.1}$$

where μ_g is the expression baseline of gene g across all samples, and $\nu_{g,i}$ is the Gaussian error term representing the intrinsic variation or measurement error. The $\{h_{j,i}\}_{j=1:r}$ and $\{\lambda_{j,i}\}_{j=1:k}$ combined by corresponding loadings $\{\beta_{g,j}\}_{j=1:r}$ and $\{\alpha_{g,j}\}_{j=1:k}$ breakdown and interpret the variation of $x_{g,i}$ across samples. The $\{h_{j,i}\}_{j=1:r}$ are known design factors for sample i. They can be any known characteristics of the sam-

ples obtained through predefinition or measurement, such as experimental groups or treatment variables, sample clinical features, expression level of a certain protein or experiment artifacts. The $\{\lambda_{j,i}\}_{j=1:k}$ are unknown latent factors for sample i to be inferred through model fitting. These latent factors are regarded as representatives of underlying biological processes or sub-phenotypic structures that impact on the expression of multiple genes in potentially complex, interacting ways.

Sparsity is an important concept in modeling and variable selection problem in high-dimensional problems. Based on the fact that most biological phenotypes only involve the transcriptional response of a relatively small number of genes in a genome, this concept has been introduced in genomics and successfully applied in microarray gene expression analysis (West 2003; Seo *et al.* 2007). The regression model coupled with a novel sparsity modeling idea is implemented in a Bayesian framework by BFRM, and becomes remarkably effective in identifying bio-markers and pathway gene signatures from the data contaminated by experimental artifacts and non-biological biases (Lucas *et al.* 2006). The sparsity concept is embedded in the model by imposing "variable selection" prior distributions on regression parameters $\{\beta_{g,j}\}_{g=1:p,j=1:r}$, namely,

$$p(\beta_{g,j}) = (1 - \pi_{g,j})\delta_0(\beta_{g,j}) + \pi_{g,j}\mathrm{N}(\beta_{g,j}; 0, \tau_j), \tag{2.2}$$

which means that $\beta_{g,j}$ either is 0 or has a normal prior with variance τ_j. $\pi_{g,j}$, the probability that $\beta_{g,j}$ has a normal distribution, is gene-specific and assigned an hierarchical shrinkage prior that heavily favors 0 probability. This sparsity modeling, benefitting from such hierarchical shrinkage priors, effectively reduces the false discovery rates and enhances the ability to isolate significant gene-variable effects (Lucas *et al.* 2006).

Bayesian analysis leads to an estimate of the posterior probability that $\beta_{g,j}$ is

non-zero. I use the notation $\pi^*_{g,j}$ for this estimate, a probabilistic assessment of the level of expression of gene g being associated with factor j. Here I refer to $\pi^*_{g,j}$ as the *association probability* of gene g with phenotype j. Typically, only a relative small number of genes will have large association probabilities, while the vast majority will have very small or 0 probabilities. Similar sparsity ideas apply to the latent factor loadings $\{\alpha_{g,j}\}_{g=1:p,j=1:k}$. As one of the products of the BFRM analysis, the association probability matrix

$$\left[\pi^*_{g,j}\right]_{p\times(r+k)} = \begin{pmatrix} \pi^*_{11} & \cdots & \pi^*_{1(r+k)} \\ \vdots & \ddots & \vdots \\ \pi^*_{p1} & \cdots & \pi^*_{p(r+k)} \end{pmatrix} \tag{2.3}$$

contains key summary information of the gene-phenotype associations implied by the gene expression profiling data set. Each column of the matrix is a probabilistically indexed gene list.

Without considering latent structure of gene expression profiles, this model becomes a multivariate regression model

$$x_{g,i} = \mu_g + \sum_{j=1}^{r} \beta_{g,j} h_{j,i} + \nu_{g,i}. \tag{2.4}$$

For all $g = 1:p$, statistical inference on regression parameters $\{\beta_{g,j}\}_{j=1:r}$ conveys the expression predictability of each gene on the regressor variables. Instead of predicting gene expression, such an analysis often aims to identify genes significantly related to design factors, and the probabilities $\pi^*_{g,j}$ are key to this goal.

Latent factor modeling added to this regression aims to identify underlying patterns having biological significance and contributing to the gene expression variation not explained by design factors (Carvalho *et al.* 2007). The association between each gene $g \in \{1:p\}$ and each of the biological phenotypes represented by these latent fac-

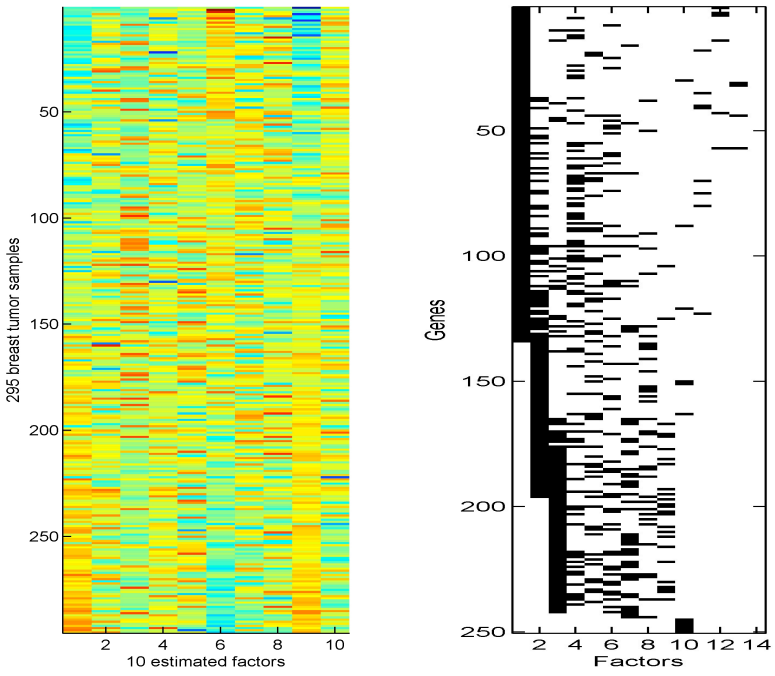
tors are indicated through the inference on the loadings $\{\alpha_{g,j}\}_{j=1:k}$. One application of this latent factor model is to decompose the expression patterns linked to a phenotype or intervention. In contrast to conventional gene clustering analysis, rather than segregating genes into individual groups this analysis emphasizes connections among the factors reflected in those genes shared by different factors, and infers the structure by estimating the factors over samples as well as the gene-factor loadings. If each factor is assumed to represent the transcriptional read-out of one or more "modules" in a gene regulatory network, the intersecting gene subsets can be viewed as the nodes connecting the modules. This cross-talk among the factors potentially provides interaction information on signaling pathways. Lucas *et al.* (2006) gave an example on this type of application, in which the pathway activities related to breast tumor estrogen-receptor and ErbB2 expression variation are dissected.

A strength of this latent factor modeling approach lies in its ability to borrow information from existing studies to explore the pathway activities in different and often more complicated contexts, called signature dissection and enhancement (Lucas *et al.* 2007). An analysis of this type of application usually starts with a group of genes predefined as a signature of a certain interesting biological context, for example, Rb/E2F pathway genes, genes differentially expressed in gastric cancer cell lines resistant to doxorubicin, or human orthologous genes mapped from certain mouse oncogenic pathway genes. Through an evolutionary stochastic model search method, more genes and latent factors are iteratively added to the model. The added genes strengthen the evolution of the latent factors, while the added factors help to improve the explanation of the expression variation in the genes. Such a model fitting a process not only decomposes the biological event projected from the initial set of signature genes, but also extends the focus to other pathway activities potentially linked to it. This cross-study strategy based on BFRM has been applied to analyze lactic acidosis

response in breast tumors (Lucas *et al.* 2007; Chen *et al.* 2007), among other recent and current studies.

Generally, a gene expression profile analysis in BFRM provides an information-rich output including the measures of the association between genes and biological phenotypes, sub-phenotypes represented by latent factors, as well as the interactions between the phenotypes. Here is an example of a typical BFRM analysis. The microarray readings are collected from 295 breast cancer tumors exhibiting diverse clinical phenotypes. The factor analysis of the data by BFRM includes four clinical phenotypes as covariates (estrogen-recptor status, lymph-node status, progestoron-receptor status and ErbB2 status), and outputs 10 latent factors whose profiles are shown as a heatmap in Figure 2.1(a). The skeleton plot in Figure 2.1(b) illustrates the association between each factor – including the four clinical and 10 latent factors – and 250 genes, each of which is differentially expressed with at least one of the 14 factors in the sense that the association probability $\pi^*_{g,j}$ is greater then 0.75. The other genes not shown in the figure have 0 or small association probabilities with all factors. As can be seen, the factors are distinguished from each other in terms of the associated gene sets. Meanwhile, each factor is linked to one or more other factors through cross-talk introduced by common sets of associated genes.

In cancer research, estimated factor "phenotypes" and the association probabilities provide a basis for developing hypotheses in tumorigenesis studies, and are further utilized to predict cancer phenotypes such as cancer sub-types, development stage, survival and therapy response (Carvalho *et al.* 2007; Lucas *et al.* 2007). An essential work that builds a bridge between the statistical analysis results and the applications is the identification of potential or putative biological processes or gene pathway activities underlying the factors as phenotypes. For each factor phenotype, the underlying pathway activity information is embedded in the relation between



(a) (b)

Figure 2.1: Example of BFRM analysis output. (a) The profiles of the 10 latent factors in the breast tumor gene expressions. (b) Binarized association probability matrix of 250 genes with four clinical phenotypes (first four columns) and 10 latent factors. Here black if $\pi^*_{g,j} > 0.75$, white otherwise.

genes and the factor phenotype. This highlights the central importance of the corresponding columns of the posterior association probability matrix in (2.3), exemplified in Figure 2.1 (b). The pathway annotation methodology described in the following chapters of this thesis is developed for (but not limited to) this setting of studies; I focus on pathway annotation and identification based on gene-phenotype association probabilities projected from such an analysis.

2.3 Pathway Annotation Methods

2.3.1 Limitations of Biomarkers

A goal of analyzing gene expression profiles is to reveal the biological pathways linked to a sample phenotype or response to a biological perturbation. Genes with transcriptional level changes are the participants and carriers of the signals and can give valuable insight into the underlying biological processes. A variety of statistical methods have been developed to detect such signature genes or biomarkers based on the correlation between single gene expression variation and the phenotype or perturbation. However, identification of biological pathways depending on observing the expression of individual bio-markers often fails because of instability in the measurement technology as well as experimental bias. Biological variation in single gene expression can be damped by noise and irrelevant factors and become undetectable. Many studies have shown that, in the same biological context, the lists of differentially expressed genes obtained from independent experimental data sets, different platform, or even simply by different statistical methods have poor overlaps. On the other hand, it has been observed that a large portion of genes, which are truly associated with a phenotype or respond to a perturbation, are excluded from consideration due to either subtle expression variation in the genome-wide transcriptional profiling experiment or the identifying statistical method (van't Veer *et al.* 2002; Fan *et al.*

2006; Kim and Volsky 2005; Manoli *et al.* 2006; Cheadle *et al.* 2007). Additionally, setting a cutoff of statistical significance of association between a single gene and a phenotype can be arbitrary. As a result, the interpretations of the underlying biological processes based on the significant genes in isolation are also often arbitrary (Subramanian *et al.* 2005).

2.3.2 Gene-Set Pathway Annotation

To overcome the limitations of biomarker-based pathway annotation, gene-set approaches, also called gene-class testing (Allison *et al.* 2006), have been proposed on the basis of examining the group effect of pathway gene sets. The fundamental idea is to measure aggregate association between a phenotype and a set of genes in a predefined gene set. Statistical analysis then often tests the significance of this group association through comparisons with random gene sets or its association with randomized phenotypes. In these approaches, the phenotype association of individual genes are integrated in certain ways, hence even subtle expression changes of individual genes, which have relatively low correlation with the phenotype and could be ignored otherwise, contribute to the pathway association with the phenotype and may make it significant. Besides interpretability and sensitivity, such gene-set pathway annotation approaches also have advantages in assessing results conducted in independent experiments or cross-platform (Kim and Volsky 2005; Manoli *et al.* 2006; Cheadle *et al.* 2007). Here I introduce the general framework and summarize existing methods of gene-set pathway annotation.

Suppose the pathway activities underlying a phenotype is investigated through a genome-wide gene expression profiling experiment that involves p genes and n samples. Denote the full gene list by $\mathcal{G}$. The phenotype and gene expression of each sample $i \in \{1:n\}$ are observed, and denoted by y_i and $\{x_{g,i}\}_{g=1:p}$, respectively. The

value of y_i can be binary, categorical or continuous values, depending the characteristic of the phenotype. Given an association metric W, each gene $g \in \mathcal{G}$ receives a phenotype association value $w_g = W(\boldsymbol{x}_g, \boldsymbol{y})$, where $\boldsymbol{x}_g = \{x_{g,i}\}_{i=1:n}$ is the expression profile of gene g, and $\boldsymbol{y} = \{y_i\}_{i=1:n}$ is the phenotype profile. Common choices of W are Pearson's correlation, signal-to-noise ratio, fold change and t-statistics. Then all the genes in the full list are indexed or weighted by their association values. Now introduce a gene set, A, which has been predefined either based on theoretical reasoning or by prior experiments. The gene set A represents a certain biological pathway or process. Then the variation of values $\{w_g\}_{g\in A}$ in the full set $\{w_g\}_{g\in\mathcal{G}}$ provides information on the pathway-phenotype association. Intuitively, the more genes in A that have large association values, the more likely the pathway that A represents is linked to the phenotype. A gene set A that has highly associated genes is said to be enriched in the phenotype. A gene-set pathway annotation method, in the conventional framework, defines a scoring metric S on the weights to measure this gene set enrichment. Traditional methods then test the significance of $S(\{w_g\}_{g\in\mathcal{G}}, A)$ given the null distribution either in gene set space or phenotype space. To deal with multiple gene sets, certain correction methods, such as Family-Wise Error Rate (FWER) (Westfall and Young 1993) and False Positive Rate (FDR) controlling (Benjamini and Hochberg 1995), are often used to correct multiple tests.

2.3.3 Methods Based on Binary Weights

The simplest class of gene-set methods are developed using binary metrics of gene-phenotype association and Fisher's exact test. The association values $\{w_g\}_{g\in\mathcal{G}}$ are transformed to 0 or 1 representing whether or not they are considered as being differentially expressed with the phenotype. Given an association cutoff w_0, the association

weight of gene g is

$$u_g = \begin{cases} 1, & w_g \geq w_0, \\ 0, & w_g < w_0. \end{cases} . \tag{2.5}$$

A natural enrichment score in this case is the number of genes in gene set A with u equal 1. Under an assumed null hypothesis of random gene assignment to A, this count has a hypergeometric distribution, and Fisher's exact test, or a χ^2 test when A has a large number of genes, has been used to test the significance of the gene set enrichment. This method is widely implemented in gene set function annotation software packages, where the gene sets are from Gene Ontology functional categories (Dahlquist *et al.* 2002; Zhong *et al.* 2003; Zeeberg *et al.* 2003; Draghici *et al.* 2003; Berriz *et al.* 2003). A related method based on binary weights is developed in the Bayesian parametric framework and implemented in the software package GATHER (Chang and Nevins 2006). In this method, the number of genes labeled with 1 in A is modeled as a binomial variate, as is that in the complementary gene set. Bayes factors are then computed to compare the evidence for or against the hypothesis that these two distributions are different, the null hypothesis being that they are same.

Although this type of method has advantage in terms of simplicity and low computational costs, the shortcomings are obvious. The gene weighting process needs to choose a cutoff for gene-phenotype association, which highly depends on the statistical method being employed and sometimes appears to be arbitrary. Further, binary weighting loses considerable information by completely ignoring the differences among genes in their association values. The annotation is then highly influenced by false positive and false negative rates in determining associated genes. Another common problem is that genes are treated as independent, that is, intrinsic biological interactions among genes are not considered.

2.3.4 Methods Based on Full Weights

To make better use of the gene-phenotype association information $\{w_g\}_{g\in\mathcal{G}}$, another type of gene-set pathway annotation method developed to take into account *continuous* gene-phenotype association weights. Typically, such methods compute a statistic that summarizes the weight information in a gene set and represents gene set-phenotype association, and again compares the value of such a statistic to the corresponding null distribution followed by the adjustment for multiple testing. Here I review some representative methods with a focus on the definition of enrichment summary statistics and null distribution generation.

One category of methods perform non-parametric statistical tests on gene set-phenotype association. Genes in $\mathcal{G}$ are ranked with respect to the gene-phenotype association weights w_g. A certain enrichment summary statistic is defined on the ranks.

Virtaneva *et al.* (2001) first implemented this methodology and applied it in functional annotation of gene expression signatures. They use the Wilcoxon rank-sum statistic

$$S = \sum_{g\in A} R(w_g) - \frac{|A|(|A|+1)}{2}$$

as the enrichment summary statistic. R is the rank function, $|A|$ is the size of gene set A. Under certain assumptions, on the hypothesis of random rankings, S is approximately normally distributed when $|A|$ is relatively large. In this method, it is assumed that there are no transcriptional interconnections among genes.

Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.* 2005; Mootha *et al.* 2003) is a more widely used non-parametric method. The enrichment summary statistic for gene set A is defined as the maximum deviation from zero during the

weighted random walk on the ranking list road-marked by gene set A. That is,

$$S = \max_{1 \le i \le |\mathcal{G}|} \left(\frac{\sum_{g \in A} |w_g| I(w_g < w_i)}{\sum_{g \in A} |w_g|} - \frac{\sum_{g \notin A} I(w_g < w_i)}{|\mathcal{G}| - |A|} \right),$$

where I is the indicator function, and $|\mathcal{G}|$ is the size of full gene list $\mathcal{G}$. Named the enrichment score (ES), S corresponds to a weighted version of Kolmogorov-Smirnov statistic. GSEA advocates sample permutation to generate a null distribution of ES to maintain the gene-wise association and test the association between a gene set and the phenotype of interest. The null distribution is then adjusted for variation in gene set size, i.e. the number of genes in the gene set, by dividing each empirical value of ES by their mean. The empirical nominal p-value is computed for the adjusted/normalized ES (NES) to show the significance of the association between the gene set and the phenotype.

Zahn *et al.* (2006) use Van der Waerden's test as an alternative to the GSEA Kolmogorov-Smirnov test. The ranks of genes are converted to standard normal quantiles to attain a normal score

$$u_g = \Phi^{-1} \left[\frac{R(w_g)}{(|\mathcal{G}| + 1)} \right]$$

for each gene $g \in \mathcal{G}$, where Φ is standard normal cumulative density function (c.d.f.). The mean statistic

$$S = \frac{1}{|A|} \sum_{g \in A} u_g,$$

is applied to summarize the gene set enrichment. Since the w_g are obtained from a multivariate regression analysis that has the phenotype under study as one of the covariates, the null hypothesis is generated using bootstrap resampling instead

of sample permutation to preserve relationship between genes and covariates. An hypothesis test is then performed on the normal distribution standardized by dividing each S^* (corresponding to each resampling) by the estimated variance.

The other category of methods use parametric tests with an initiative of modeling the gene-phenotype association. Efron and Tibshirani (2006) propose a maxmean statistic that seems to have superior power characteristics compared to the Kolmogorov-Smirnov statistic, and demonstrate it with t-statistic inferred from an expression analysis with respect to a binary variable, say, control vs. a phenotype. Each weight is normalized as

$$u_g = \Phi^{-1}[F(w_g)], \quad g \in \mathcal{G},$$

where Φ is standard normal c.d.f., and F is the c.d.f. of t-distribution with $n-2$ degrees of freedom (n is sample size). The maxmean statistic is defined as

$$S = \max(S^+, S^-)$$

with S^+ and S^- the absolute means of $\{u_g : g \in A \text{ and } u_g > 0\}$ and $\{u_g : g \in A \text{ and } u_g < 0\}$, respectively. A null distribution is attained by computing the maxmeans of the gene set under sample permutation or gene randomization followed by standardization.

Kim and Volsky (2005) directly use gene expression fold change between two experimental groups for association weights, and empirically show the approximate normality of the mean statistic

$$S = \sum_{g \in \mathcal{G}} w_g \tag{2.6}$$

when a random gene set contains more than 10 genes. Newton *et al.* (2007) develop a random-set method based on a very similar idea but with rigorous theoretical justification. To improve the normality of the summary statistic, they suggest using more

regular w_g such as log transformed p-value or fold change. Assuming nonexistence of interconnections among genes, they show the mean statistic S (as expressed in (2.6)) can be standardized as

$$Z = (S - \mu)/\sigma,$$

where the overall mean μ and variance σ^2 have analytical forms

$$\mu = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} w_g \quad \text{and} \quad \sigma^2 = \frac{1}{|A|}\left(\frac{|\mathcal{G}| - |A|}{|\mathcal{G}| - 1}\right)\left[\left(\frac{\sum_{g\in\mathcal{G}} w_g^2}{|\mathcal{G}|}\right) - \mu^2\right].$$

According to central limit theory, Z is approximately standard normal on the null hypothesis when the gene set size is large enough. This test is equivalent to Fisher's exact test when w_g is replaced with a binary quantity u_g as shown in (2.5), and to Wilcoxon test when w_g is replaced with its rank. Notably, this standardization does not involve the estimation of moments based on random set simulation, hence is more efficient compared to the other methods based on sample permutation or gene randomization.

In summary, the gene-set pathway annotation methods that use continuous association weight information, in terms of the tests being used, have two types: nonparametric and parametric. Generally, both types at least implicitly depend on sample permutation or gene set randomization either to generate the null distribution or to estimate the sufficient statistics of the null distribution. As addressed by Tian *et al.* (2005), sample permutation formulates a null hypothesis that the tested gene set does not show stronger association with the phenotype of interest than with other randomized phenotypes; gene randomization formulates a null hypothesis that the phenotype-association pattern of the tested gene set can not be distinguished from those of random gene sets. Both hypotheses stand on reasonable interpretations of the concept of gene set enrichment, and show advantages and disadvantages when

compared to each other. Sample permutation requires the sample size to be large enough to allow the test to gain the power of rejecting the null hypothesis. Usually, for bi-categorical samples, at least eight are needed for each group. This limits its application in some experimental data sets with very few samples. Gene randomization upon genome-wide data does not have this problem. Another advantage of gene randomization lies in the computational efficiency when using mean statistic (Newton *et al.* 2007; Virtaneva *et al.* 2001). However, sample permutation is more appealing in terms of its consideration and maintenance of the interactions among genes. Gene randomization, on the contrary, assumes that all genes are expressed independently, and the transcriptional interconnections among genes are out of consideration when comparing a gene set with random sets.

2.4 Motivation for Formal Probabilistic Modeling

Analysis of many gene expression studies has generated a tremendous number of signature gene sets that can further benefit the studies of cancer signaling pathways. There are also many gene sets derived from experiments that generate other types of data, such as differential expression of a set of proteins under a certain experimental perturbation. Then a set of genes encoding these proteins, to some degree, contains information on consequent pathway activation. Meanwhile, efforts have been made to dissect the complexity of signaling pathways under cancer phenotypes by using appropriate computational methodology based on genome-scale expression profile analysis. Making use of the knowledge carried by existing signature gene sets to identify biological pathways is an important component of the whole strategy for the studies of cancer mechanisms. Gene-set pathway annotation methodology has been developed in this context.

Although the fundamental idea is appealing and well accepted, the existing meth-

ods have some common problems regardless of their individual defects. None of these methods attempts to model the gene-set pathway annotation problem based on a clear biological interpretation. What is the relationship between a pathway and a curated gene set? When the biological context under study is different from that in which the gene set was originally defined, how should we consider this intrinsic biological discrepancy in the pathway annotation? Is that a pathway is involved in the phenotype equivalent to that a gene set is enriched in the phenotype? How should we interpret the annotation results? These questions arising from the fundamental idea are not addressed by any of existing methods. In fact, without considering these questions, these methods literally are gene set enrichment analysis rather than pathway annotation methods. To incorporate rigorous biological thinking in this gene set pathway annotation problem one needs to develop a model-based approach.

Moreover, pathway annotation analysis borrows knowledge from the signaling pathway studies in other experimental contexts. A necessary step following annotation is to examine the specificity of these pathways in the current biological context, i.e. to compare and map the reference pathway to the one involved in the currently investigated context. This problem arises from these facts: first, a reference pathway gene set defined through experiments other than gene expression analysis may contain genes whose mRNA levels are apparently unrelated to the phenotype; second, a reference gene set may be incomplete or include irrelevant genes due to the noise in the predefining process; more generally, the genes comprising the reference pathway may not be equally relevant in the pathway under study due to the difference of biological contexts. Hence, a complete pathway analysis must involve a process of refining pathway gene sets. This is the key to imputing the connectivity among pathway modules as well as to identifying biomarkers and therapeutic targets. GSEA includes a operation to find a group of genes, called leading edge subset. However,

there is no theoretical support for this operation; hence, the meaning of this subset of genes is unclear. To address this problem, we also need a model-based approach.

Chapter 3

Probabilistic Pathway Annotation

In this chapter I propose a model-based approach to matching experimental gene expression signatures of a phenotype to pathway gene sets. This approach, named as probabilistic pathway annotation (PROPA), is developed in the Bayesian framework and assesses the uncertainties of association between the phenotype and multiple pathways from the database. I first introduce a Bayesian perspective on the gene set-based pathway annotation problem, then develop the statistical models, discuss multiple pathway comparisons and gene pathway membership inference followed by a summary of the practical benefits of this approach.

3.1 Bayesian Foundation

In a genome-wide gene expression profiling data set, the total number of genes under consideration is p, typically several thousands or several tens of thousands. Assume the genes are labeled by indices $g = 1 : p$ and denote the full gene list by

$$\mathcal{G} = \{1 : p\}.$$

To begin, consider just one specific *biological pathway gene set* $\mathcal{A} \subset \mathcal{G}$; this is a set of genes explicitly assumed to be involved in a defined, specific biological context. However, the exact information on the genes in $\mathcal{A}$ is unavailable. The only knowledge is that $\mathcal{A}$ includes a set of genes that have been experimentally defined as playing roles in the pathway. Denote this known *reference gene set* by A. The number of genes in A is $q < p$, and A is an incomplete and typically error-prone observation on the true pathway gene set $\mathcal{A}$.

The key here is to understand that there is a true but unknown pathway gene set $\mathcal{A}$ and to distinguish this from the current known reference gene set A. If the experiment and analysis by which A has been defined can be trusted, A should be a reasonably good representation of $\mathcal{A}$ so that the reliability of the pathway annotation analysis can be guaranteed. Some degrees of discordance would be expected between A and $\mathcal{A}$. The intrinsic noise of biological experiments and statistical analysis cause some genes to be falsely included or excluded when A is generated, i.e., there are false positives and false negatives. Apart from the experimental noise, an experiment usually only takes a snapshot of one of the several sides of a theoretical pathway rather than capturing all the information. In other word, the definition of a theoretical pathway usually is based on the key experimental context and ignores the non-essential experimental conditions; this is exactly the foundation of the knowledge-based pathway annotation approach. The impacts of those ignored experimental conditions are reflected in the discrepancy between A and $\mathcal{A}$. Moreover, the pathway gene sets are not necessarily defined by gene expression differentiation. Indirect methods can be used to define pathway gene sets; for example, the genes whose corresponding proteins differentially express with the perturbation, or the human homologs of genes involved in a biological pathway of some other organism. For these reasons, A cannot be a perfect representation of $\mathcal{A}$ in general.

A statistical analysis of a gene expression microarray data set assigns a value to each gene that represents the level of its association with a biological phenotype (or intervention). Thus, a full set of association weights are generated as

$$\Pi = \{\pi_1, \ldots, \pi_p\}.$$

Although the methodology proposed here is applicable to *any* measure of the association between genes and the phenotype, I exclusively focus in this dissertation on

the use of association probabilities generated in the framework of BFRM introduced in Section 2.2. Then Π is a full list of probabilities corresponding to a column in the association probability matrix exhibited in (2.3) from a particular analysis. In this context, the factor the column represents is the phenotype of interest.

Now the question is, given knowledge of A, how strongly does the gene-phenotype association support the claim that the phenotype involves the biological pathway $\mathcal{A}$? To phrase this problem in statistical language, I represent by $\mathcal{F}$ the gene set of the hypothetical pathway involved in the *factor phenotype*, $\mathcal{F} \subset \mathcal{G}$. The pathway annotation then is a process of assessing the concordance between $\mathcal{F}$ and $\mathcal{A}$ given the information in Π and A. In the Bayesian framework, this is quantified through the evaluation of the posterior probability

$$Pr(\mathcal{F} = \mathcal{A}|\Pi, A).$$

In a practical context that involves all the pathways in a entire database, the problem is extended to multiple assessments and a model comparison problem. There are a set of pathways $\mathcal{A}_1, \ldots, \mathcal{A}_m$, and the corresponding reference gene sets $A_1, \ldots, A_m$ available; the question is which of these pathways is more likely to be associated with the phenotype compared to the others.

In Bayesian framework, this is a model comparison problem, in which a model corresponds to a hypothesis $\mathcal{F} = \mathcal{A}_j$ $(j = 1, \ldots, m)$ and the partition of the whole gene list (specified by A_j). The formal solution of the problem is provided by the evaluation of the posterior probability of each model in a model space constructed by all the pathways to be compared and their reference gene sets. Explicitly, one needs to compute the posterior probabilities

$$\begin{aligned} &Pr(\mathcal{F} = \mathcal{A}_j|\Pi, A_1, \ldots, A_m) \\ &\quad \propto Pr(\mathcal{F} = \mathcal{A}_j|A_1, \cdots, A_m)p(\Pi|A_j, \mathcal{F} = \mathcal{A}_j), \quad j = 1:m \end{aligned} \tag{3.1}$$

with $Pr(\mathcal{F} = \mathcal{A}_j|A_1, \ldots, A_m)$ some assumed prior probabilities. The marginal likelihoods, $p(\Pi|A_j, \mathcal{F} = \mathcal{A}_j)$, as j moves across all the pathways, are the probability densities of the association data Π conditional on the hypotheses that the factor phenotype $\mathcal{F}$ is pathway $\mathcal{A}_j$.

3.2 Summary of Notation

For clarification, I summarize the key quantities and notation as following:

- $\mathcal{G} = \{1 : p\}$ is the full gene list.
- $\mathcal{A}$ is the true biological pathway gene set, and $\mathcal{A} \subset \mathcal{G}$.
- A is the known reference gene set of $\mathcal{A}$, and $A \subset \mathcal{G}$.
- $\mathcal{F}$ is the hypothetical pathway gene set involved in the factor phenotype being studied, and $\mathcal{F} \subset \mathcal{G}$.
- $\Pi = \{\pi_g\}_{g=1:p}$ is the full list of the gene-factor phenotype association probabilities.
- Goal is to compute the posterior probabilities $Pr(\mathcal{F} = \mathcal{A}|\Pi, A)$.

3.3 Statistical Models

I begin with the development of the model for one biological pathway and then extend it to the comparison of multiple pathways.

The starting point is to view Π as data. Then statistical thinking focuses on models of the distribution of the data Π conditional on the hypothesis that $\mathcal{A}$ is the underlying biological pathway of factor phenotype, i.e. $\mathcal{F} = \mathcal{A}$. If $\mathcal{F}$ is indeed $\mathcal{A}$, the observed association probability π_g would be expected to have higher value for

gene $g \in \mathcal{A}$ than for gene $g \notin \mathcal{A}$. Assuming association probabilities are generated independently, and no other information about the genes can distinguish them except the pathway membership, one has models

$$(\pi_g | g \in \mathcal{A}, \mathcal{F} = \mathcal{A}) \sim f_1(\pi_g) \quad \text{and} \quad (\pi_g | g \notin \mathcal{A}, \mathcal{F} = \mathcal{A}) \sim f_0(\pi_g) \tag{3.2}$$

independently over all g. Here, f_1 and f_0 are two probability density functions over the unit interval; f_1 favors higher values of π_g while f_0 favors lower values.

The natural choice of f_1 and f_0 is beta density functions. Specifically, I choose $f_1(\pi) = \text{Be}(\alpha_1, 1)$ and $f_0(\pi) = \text{Be}(1, \alpha_0)$; explicitly,

$$f_1(\pi) = \alpha_1 \pi^{\alpha_1 - 1} \quad \text{and} \quad f_0(\pi) = \alpha_0 (1 - \pi)^{\alpha_0 - 1} \tag{3.3}$$

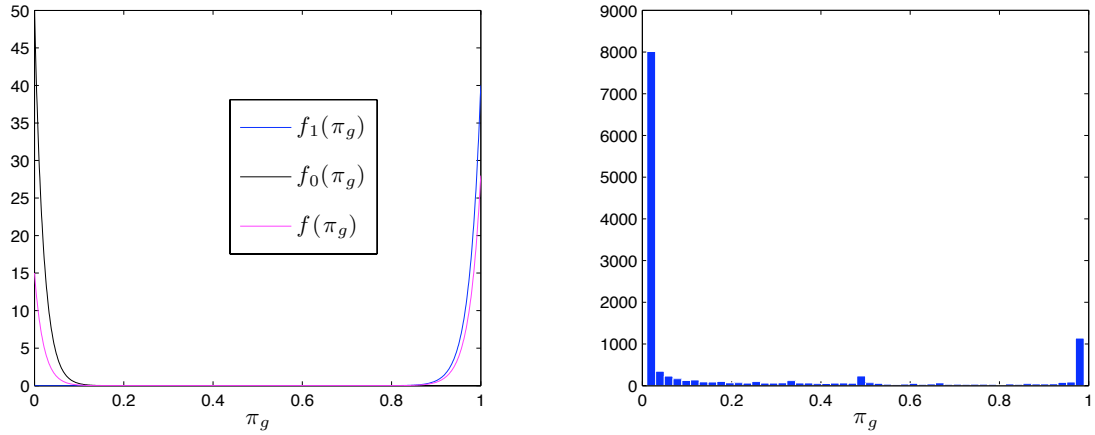
for $0 \leq \pi \leq 1$, with $\alpha_0, \alpha_1 > 1$. Clearly, from 0 to 1, f_1 is monotonically increasing, while f_0 is monotonically decreasing, demonstrated as the blue and black curves, respectively, in Figure 3.1(a). Such a specification is based on the observation of the real data generated with sparsity modeling as introduced in Section 2.2. Figure 3.1(b) shows the histogram of the association probabilities Π from the analysis of a real data set. To allow flexibility in these distributions, I give α_0 and α_1 reference priors

$$p(\alpha_0) \propto \alpha_0^{-1} \quad \text{and} \quad p(\alpha_1) \propto \alpha_1^{-1} \tag{3.4}$$

with constraint $1 < \alpha_0 < a$ and $1 < \alpha_1 < a$. Here a is a large number, say, 100, serving as the upper limit of the values that α_0 and α_1 can take. Setting an upper bound for α_0 and α_1 is needed for marginal likelihood computation, which will be discussed in Chapter 5 and 6, rather than as a requirement of modeling.

Introduce a set of indicators $\{z_g\}_{g \in \mathcal{G}}$, where

$$z_g = \begin{cases} 1, & g \in \mathcal{A}, \\ 0, & g \notin \mathcal{A}, \end{cases}$$



(a) (b)

Figure 3.1: (a) $f_1(\pi_g)$ and $f_0(\pi_g)$ are Beta density functions specified in (3.3); model the density function $f(\pi_g)$ as a mixture of them. (b) Histogram of association probabilities Π generated in a real expression data analysis.

given $\mathcal{F} = \mathcal{A}$. The z_g are the unknown pathway membership indicators. So the p.d.f. of π_g conditional on α_0, α_1 and z_g is

$$p(\pi_g|\alpha_0, \alpha_1, z_g, \mathcal{F} = \mathcal{A}) = [f_1(\pi_g; \alpha_1)]^{z_g} [f_0(\pi_g; \alpha_0)]^{1-z_g} . \tag{3.5}$$

As mentioned at the beginning, $\mathcal{A}$ is unknown, and so is the pathway membership of each gene g, i.e. the value of z_g. Model this uncertainty by giving z_g a Bernoulli distribution

$$p(z_g|\beta_g) = \mathrm{B}(z_g; \beta_g) = \beta_g^{z_g}(1 - \beta_g)^{1-z_g}, \tag{3.6}$$

where β_g is the expectation of z_g, representing the prior pathway membership probability of gene g. Marginalizing the distribution of π_g in (3.5) with respect to z_g leads to the p.d.f. of π_g conditional on β_g,

$$p(\pi_g|\alpha_0, \alpha_1, \beta_g, \mathcal{F} = \mathcal{A}) = \beta_g f_1(\pi_g; \alpha_1) + (1 - \beta_g) f_0(\pi_g; \alpha_0),$$

which is a mixture of two beta densities $f_1(\pi_g)$ and $f_0(\pi_g)$ weighted by β_g and $1 - \beta_g$.

This conditional p.d.f. is illustrated in Figure 3.1(a) as the pink curve.

I further model β_g as a parameter whose distribution is conditional on whether $g \in A$. This modeling is reflecting the natural idea that the determination of the true pathway membership for a gene should consider its reference set membership. For convenience, denote the set of genes not in A as B, i.e. $B = \mathcal{G} \backslash A$. Given A and B, β_g is modeled via prior distributions

$$\begin{aligned} (\beta_g | g \in A, \mathcal{F} = \mathcal{A}) &\sim \text{Be}(\phi_A r_A, \phi_A(1 - r_A)), \\ (\beta_g | g \in B, \mathcal{F} = \mathcal{A}) &\sim \text{Be}(\phi_B r_B, \phi_B(1 - r_B)), \end{aligned} \tag{3.7}$$

with specified $r_A, r_B \in (0, 1)$ and $\phi_A, \phi_B > 0$. The r_A and r_B are prior means of β_g given $g \in A$ and $g \in B$, respectively,

$$r_A = E(\beta_g | g \in A, \mathcal{F} = \mathcal{A}) \quad \text{and} \quad r_B = E(\beta_g | g \in B, \mathcal{F} = \mathcal{A}). \tag{3.8}$$

The values of r_A and r_B are subject to one's expectation of the true positive rate and false negative rate: r_A is the *á priori* probability that gene g in A is a true member of $\mathcal{A}$, and r_B is the probability that gene g in B is actually a true member of $\mathcal{A}$. Given the assumption that the gene set A is a fairly good representation of the true pathway gene signature $\mathcal{A}$, r_A should be relatively large. The value of r_B depends on an assessment of how many genes in B are likely to be associated with the factor phenotype $\mathcal{F}$. The number of genes in A, typically tens to a few hundreds (shown in Figure 3.2) is small and ignorable compared to the full gene list $\mathcal{G}$, which typically has thousands to tens of thousands genes. Therefore, a reasonable value of r_B should be approximately equal to the ratio of the number of signature genes to the total number of genes. This ratio usually is small, for example, 0.005. The ϕ_A and ϕ_B are the precision parameters in the density functions in (3.7) and constrain the variation range of β_g. The fundamental idea of this statistical modeling is illustrated in the

diagram in Figure 3.3.

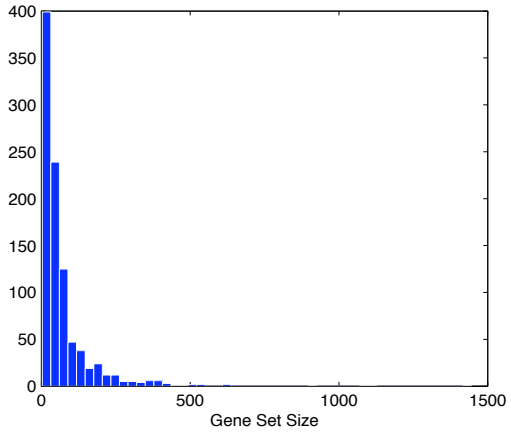


Figure 3.2: Histogram of the number of genes in 956 gene sets from a human biological pathway gene set database. The number of genes in a pathway gene set varies from a few to over a thousand. Nevertheless, a typical range for this number is tens to a few hundreds.

Under this model specification the likelihood function is

$$\begin{aligned}\mathcal{L}(\alpha_0, \alpha_1, \beta_1, \ldots, \beta_p) &= \prod_{g=1}^{p} p(\pi_g | \alpha_0, \alpha_1, \beta_g, A, \mathcal{F} = \mathcal{A}) \\ &= \prod_{g=1}^{p} \left[\beta_g f_1(\pi_g; \alpha_1) + (1 - \beta_g) f_0(\pi_g; \alpha_0)\right], \end{aligned} \tag{3.9}$$

with $f_1(\pi_g; \alpha_1)$ and $f_0(\pi_g; \alpha_0)$ specified in (3.3), and the likelihood can be rewritten as

$$\mathcal{L}(\alpha_0, \alpha_1, z_1, \ldots, z_p) = \prod_{g=1}^{p} \left[f_1(\pi_g; \alpha_1)\right]^{z_g} \left[f_0(\pi_g; \alpha_0)\right]^{1-z_g}. \tag{3.10}$$

3.4 Data Independence

An assumption in the statistical models discussed above is that the distribution of the association probability π_g, conditional on hypothesis $\mathcal{F} = \mathcal{A}$ and reference gene set A,

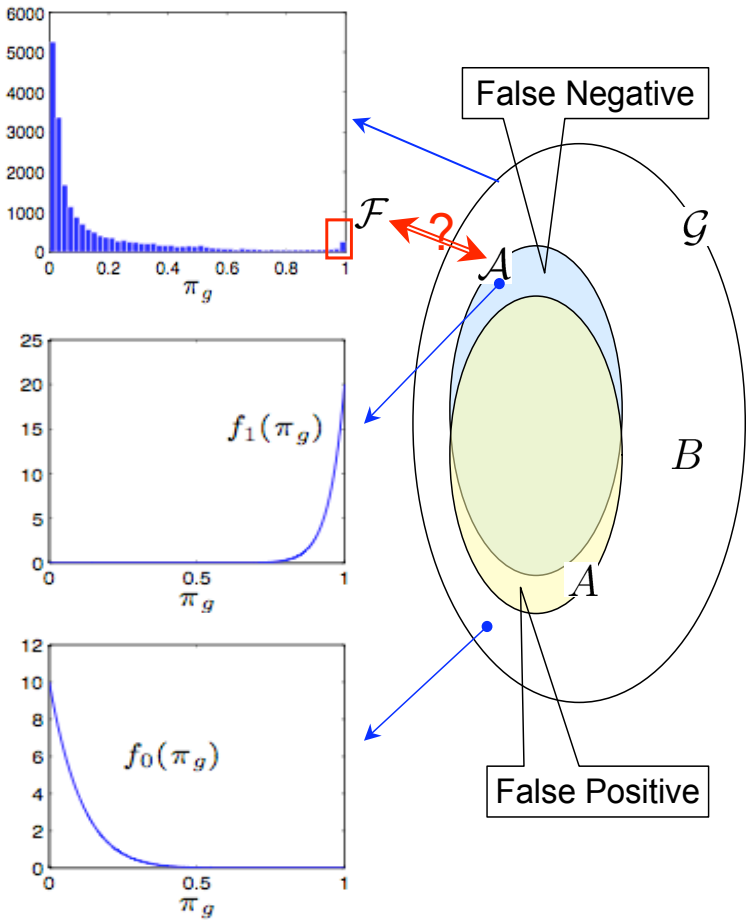


Figure 3.3: Diagram of the PROPA models. The unknown true pathway gene set $\mathcal{A}$ and the corresponding predefined reference gene set A have reasonably good overlap and some discrepancy. The light blue area represents the false negative genes. The light yellow area represents the false positive genes. The left top histogram of Π shows only a small set of genes in $\mathcal{G}$ are associated with the factor phenotype. This gene signature $\mathcal{F}$ should overlap with $\mathcal{A}$ if the hypothetical pathway under this phenotype is indeed the one specified by $\mathcal{A}$. If $\mathcal{F} = \mathcal{A}$, true pathway genes π_g have distribution f_1 while non-pathway genes π_g have distribution f_0.

is independent of the other association probabilities $\{\pi_k\}_{k\neq g}$. Note this is assuming the observations π_g to be independent, not at all that there is no interaction or coregulation among genes. Here I justify the conditional independence assumption for the π_g data distribution.

Suppose genes g_1 and g_2 are known to be coregulated in the biological context under study. One can think about an example that g_1 and g_2 are HER2 and ErbB2, respectively. They are the same gene named differently and correspond to two probe sets on a microarray. Ideally, these "two genes" should have same expression profile. Given that the statistical analysis of the gene expression data is correct, it is the experimental noise that causes the discrepancy of their expression profiles, and as consequence the association probabilities with the factor phenotype, π_{g_1} and π_{g_2}, are different. Suppose the value of the datum π_{g_1} is observed. Then *any* form of dependence would imply that one or more of the components of the distribution for π_{g_2}, as described in equations (3.5)-(3.7) at $g = g_2$, would now depend on the value of π_{g_1}. By looking at this in detail, it is argued that there really should be no such involvement of π_{g_1}, i.e., that the conditional independence assumption (conditional on all model parameters and model structure) is relevant.

Since g_1 and g_2 are really two noisy versions of the same gene, it is expected that *either* $g_1, g_2 \in \mathcal{A}$ *or* $g_1, g_2 \notin \mathcal{A}$. This is key in highlighting the role of the indicators z_g in driving the assumption of conditional independence of the π_g, as follows.

Conditional independence of π_{g_1} and π_{g_2} given z_{g_2}:

Condition on $z_{g_2} = 1$ so that $g_2 \in \mathcal{A}$ *for sure*. Under the assumed independence, this conditioning information together with other model parameters implies that $\pi_{g_2} \sim f_1(\cdot; \alpha_1)$, a density generally concentrated on high values. Does knowing π_{g_1} change this? No, for the following reasons:

- If π_{g_1} is high, this simply suggests $g_1 \in \mathcal{A}$ too, but the relevance of this additional information to the belief about π_{g_2} is overridden by already knowing that $g_2 \in \mathcal{A}$. That is, $f_1(\pi_{g_2}; \alpha_1)$ already favors high values, and the news that π_{g_1} is high simply confirms the view that π_{g_2} is likely to be high.

- Conversely, if π_{g_1} is low, that suggests $g_1 \notin \mathcal{A}$; since it is expected, on biological grounds, that either $g_1, g_2 \in \mathcal{A}$ or $g_1, g_2 \notin \mathcal{A}$, and since one is conditioning at this point on $g_2 \in \mathcal{A}$ for sure, the only rationale for a low π_{g_1} is as a false negative – due to noise in the experimental data. Hence one would reject that information as relevant to $p(\pi_{g_2}|\cdot)$ and maintain the conditional independence.

A parallel argument applies to the case with condition $z_{g_2} = 0$.

Conditional independence of z_{g_2} and π_{g_1}:

The above discussion indicates the relevance of the assumption of conditional independence of the π_g in the specific sense that

$$p(\pi_{g_2}|\pi_{g_1}, z_{g_2}, \mathcal{F} = \mathcal{A}) = p(\pi_{g_2}|z_{g_2}, \mathcal{F} = \mathcal{A}) \tag{3.11}$$

The complete independence assumption then relies on the assumed lack of dependence of z_{g_2} on the observed value of π_{g_1}. The model specifies $(z_{g_2}|\beta_{g_2}) \sim \mathrm{B}(\beta_{g_2})$ and β_{g_2} has the mixture prior of equation (3.7), so that marginally with respect to β_{g_2},

$$(z_{g_2}|g_2 \in A, \mathcal{F} = \mathcal{A}) \sim \mathrm{B}(r_A) \quad \text{and} \quad (z_{g_2}|g_2 \notin A, \mathcal{F} = \mathcal{A}) \sim \mathrm{B}(r_B),$$

with, generally, r_B near zero and r_A near 1. Conditional on $g_2 \in A$, does learning π_{g_1} change the thinking about $z_{g_2} = 1$? No, for the following reasons:

- Observing a high value of π_{g_1} suggests that $g_1 \in \mathcal{A}$. Since it is known that $g_1, g_2 \in \mathcal{A}$ or $g_1, g_2 \notin \mathcal{A}$, this suggests $z_{g_2} = 1$. Conditioning upon $g_2 \in A$, one

is already favoring $z_{g_2} = 1$, and the news that π_{g_1} is high simply confirms the view that z_{g_2} is likely to be 1.

- Conversely, observing a low value of π_{g_1} suggests $g_1 \notin \mathcal{A}$ and $z_{g_2} = 0$, but, again, conditioning upon $g_2 \in A$, one is favoring $z_{g_2} = 1$. Again this apparent conflict can only be interpreted as arising from a false negative in the sense of a low value of π_{g_1} due to the experimental noise. One should reject that information as relevant to $p(z_{g_2}|\cdot)$.

A parallel argument applies to the case with condition $g_2 \notin A$.

The above then supports the lack of dependence of z_{g_2} on any observed datum π_{g_1}, so that

$$p(z_{g_2}|\pi_{g_1}, A, \mathcal{F} = \mathcal{A}) = p(z_{g_2}|A, \mathcal{F} = \mathcal{A}). \tag{3.12}$$

Equations (3.11) and (3.12) combined imply and therefore support the independence assumptions, both conditional on z_g and unconditional, and hence the treatment of the π_g as randomly sampled from the hierarchical mixture model described in equations (3.5)-(3.7).

3.5 Model Comparison

One core goal in PROPA is to compare multiple pathways in terms of how strongly they are associated with the factor phenotype based on the information provided by Π. As discussed in Section 3.1, the Bayesian solution to this problem is based on evaluation of the posterior probabilities shown in (3.1). The required marginal likelihood is the expectation of the likelihood with respect to the prior distribution of model parameters, namely

$$p(\Pi|A_j, \mathcal{F} = \mathcal{A}_j) = \int_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}) p(\boldsymbol{\theta}|A_j, \mathcal{F} = \mathcal{A}_j) \mathrm{d}\boldsymbol{\theta} \tag{3.13}$$

with the prior $p(\boldsymbol{\theta}|A_j, \mathcal{F} = \mathcal{A}_j)$ specified in (3.6), (3.4) and (3.7), and likelihood $\mathcal{L}(\boldsymbol{\theta})$ in (3.9) or (3.10), depending on the configuration of parameters $\boldsymbol{\theta}$.

In this model comparison problem, all the models $j = 1 : m$ have the same complexity. They are only differentiated by their reference gene sets A_j. This simplifies the specification of model priors and allows us to only think about incorporating biological knowledge into the priors. In the absence of such knowledge, I put non-informative prior on models and only focus on the evaluation of the evidence, in terms of the marginal likelihood of the data Π, for each model. Hence, computing marginal likelihood is the central problem in PROPA and will be discussed in Chapter 5 and Chapter 6.

The Bayes factor, or weight of evidence, is another quantity to look at, especially when evaluating multiple models. This is defined as the ratio of the posterior odds to the prior odds of the model to be tested. In PROPA, for each pathway $\mathcal{A}_j$, $j \in 1 : m$,

$$BF_{\mathcal{F}=\mathcal{A}_j} = \frac{Pr(\mathcal{F} = \mathcal{A}_j|\Pi, A_1, \ldots, A_m)/[1 - Pr(\mathcal{F} = \mathcal{A}_j|\Pi, A_1, \ldots, A_m)]}{Pr(\mathcal{F} = \mathcal{A}_j|A_1, \cdots, A_m)/[1 - Pr(\mathcal{F} = \mathcal{A}_j|A_1, \cdots, A_m)]}. \quad (3.14)$$

In most cases, a single biological phenotype or perturbation involves complex pathway activities and may be associated with several pathways in the database. Sometimes, the association between the phenotype and one or two of these pathways are strong and appear to dominate the model selection. These dominant pathways have such large posterior probabilities and Bayes Factors that the other pathways that are actually associated with the phenotype have posterior probabilities and Bayes factors that are too small to be considered significant. In such cases, simply making conclusions based on posterior probabilities and Bayes factors of pathways would result in missing important information. Combining these quantities with pathway ranking based on these quantities should be a more reasonable and reliable way to summarize the annotation result.

Moreover, the overriding purpose of PROPA is to generate biological hypotheses. This analysis provides quantified evidence (posterior probability, Bayes factor and rank) of association between each pathway and the factor phenotype, bringing out potentially interesting pathways and serving as references for biologists to develop biological hypotheses.

3.6 Pathway Membership Probability

As has been mentioned in Section 2.4, another question of concern in pathway analysis is gene pathway membership. This question is addressed via calculation of pathway membership probabilities. In the PROPA model, the posterior pathway membership probability of gene g is

$$\beta_g^* = Pr(g \in \mathcal{A}|\Pi, A, \mathcal{F} = \mathcal{A}) = Pr(z_g = 1|\Pi, A, \mathcal{F} = \mathcal{A}).$$

The corresponding Bayes factor is

$$\begin{aligned} BF_{g\in\mathcal{A}} &= \frac{p(\Pi|g \in \mathcal{A}, A, \mathcal{F} = \mathcal{A})}{p(\Pi|g \notin \mathcal{A}, A, \mathcal{F} = \mathcal{A})} \\ &= \frac{Pr(z_g = 1|\Pi, A, \mathcal{F} = \mathcal{A})/Pr(z_g = 0|\Pi, A, \mathcal{F} = \mathcal{A})}{Pr(z_g = 1|A, \mathcal{F} = \mathcal{A})/Pr(z_g = 0|A, \mathcal{F} = \mathcal{A})}. \end{aligned}$$

Refer to (3.6), (3.8) and (3.7). This Bayes factor in explicit form is

$$BF_{g\in\mathcal{A}}|g \in A = \frac{\beta_g^*/(1-\beta_g^*)}{r_A/(1-r_A)},$$

and

$$BF_{g\in\mathcal{A}}|g \in B = \frac{\beta_g^*/(1-\beta_g^*)}{r_B/(1-r_B)}.$$

This Bayes factor measures the evidence given by data, Π, for gene $g \in \mathcal{A}$ versus $g \notin \mathcal{A}$. If a gene in reference gene set A has a Bayes factor much less than one, it

means the data give evidence that this gene is not a true member of $\mathcal{A}$. Such genes are false positives as shown in Figure 3.3. A gene in B with a large value of the Bayes factor may be a member of $\mathcal{A}$ missed by A and correspond to a false negative, again as in Figure 3.3.

3.7 Summary

PROPA is a formal model-based framework for matching experimental signatures of structure or outcomes in gene expression – represented in terms of weighted gene lists – to multiple biological pathway gene sets from curated databases. In the canonical setting here, the gene weights are explicit – gene-factor phenotype association probabilities. The formal probabilistic model delivers probabilities over pathways for each factor phenotype, allowing for a quantitative assessment and ranking of pathways putatively linked to the phenotype. The fundamental advantage of this approach is formal modeling of the uncertainty in the pathway analysis supported by clear biological interpretation. For example, *á priori* information on the accuracy and relevance of genes in reference gene sets are incorporated, and inference on both pathway-phenotype association and gene pathway membership is coherent and transparent.

Compared with existing gene-set pathway annotation approaches, PROPA is a fully probabilistically coherent model, and provides for posterior inferences regardless of the size of reference gene set or origin of gene-phenotype association probabilities. The analysis does not, and should not, involve sample permutation, and is not limited by the sample size of the data set. It also does not, and should not, involve gene randomization. The gene interaction or coregulation information is maintained by the reference gene set. Instead of comparing a pathway gene set with random sets of genes, PROPA finds significant pathways by comparing the pathway gene sets to

each other in a coherent way.

Additionally, although I develop the model using the association probability generated by BFRM, the fundamental modeling idea and framework are applicable to other metrics of gene-phenotype association, for example, t-tests and Pearson correlation coefficients. For such data, certain adjustment may need to be performed to reduce the false discovery rate. I suggest the use of the association probability because it is canonical in my perspective.

Having introduced the conceptual and technical details of the PROPA framework, I now turn to the core methodological issues – evaluation of the determining marginal likelihoods (or measure of evidence) of equation (3.13), and the accompanying questions of computation of posterior distribution for the full set of model parameters and uncertain variables $\boldsymbol{\theta}$. It turns out that the evidence computations are most effectively addressed following posterior computations using Bayesian simulation methods, so I begin with such methods in the next chapter.

Chapter 4

Posterior Simulation

Markov chain Monte Carlo (MCMC) methods are powerful computational methods widely used in Bayesian statistical analysis. As has been addressed in the previous chapter, PROPA is able to refine a pathway gene set based on the belief that this pathway is associated with the phenotype in the current context of study. Such a refinement is implemented through the estimation of posterior pathway membership probabilities for genes. I address these computations, and the broader questions of posterior analysis, using MCMC methods. Moreover, the central problem in PROPA is model comparison based on computing marginal likelihood, an integration problem that appears to be intractable for conventional integration methods due to the high dimensionality of data. In Chapter 6, I introduce a new Monte Carlo integration method for PROPA marginal likelihood computation, and that method builds on posterior simulation. In this chapter, I discuss the core MCMC method introduced for PROPA, and demonstrate the performance with a simulation study.

4.1 Gibbs Sampling

With the parameterization and priors specified in Chapter 3, the joint probability density function of the PROPA model is

$$
\begin{aligned}
& p(\Pi, \alpha_0, \alpha_1, \boldsymbol{\beta}, \boldsymbol{z} | A, \mathcal{F} = \mathcal{A}) \\
= \quad & \alpha_0^{-1} \alpha_1^{-1} \prod_{g=1}^{p} \{ [\beta_g \alpha_1 \pi_g^{\alpha_1 - 1}]^{z_g} [(1-\beta_g) \alpha_0 (1-\pi_g)^{\alpha_0 - 1}]^{1-z_g} \\
& [\mathrm{Be}(\beta_g; \phi_A r_A, \phi_A (1-r_A))]^{I(g \in A)} [\mathrm{Be}(\beta_g; \phi_B r_B, \phi_B (1-r_B))]^{I(g \notin A)} \}, \qquad (4.1)
\end{aligned}
$$

where $\boldsymbol{\beta} = \{\beta_g\}_{1:p}$ and $\boldsymbol{z} = \{z_g\}_{1:p}$, and I is an indicator function. Benefiting from these conjugate priors, the full conditional distribution of each parameter is easy to derive, and these are summarized as follows:

- Given $\boldsymbol{z}$, α_0 and α_1 are independent of the other parameters and have gamma density functions truncated below 1 and above a. When a takes a large value, at which both density functions decay to 0, they can be simplified to one-sided truncation (truncation below 1):

$$p(\alpha_0|\boldsymbol{z}, \Pi, A, \mathcal{F} = \mathcal{A}) = \mathrm{Ga}_{>1}\left(\alpha_0; \sum_{g=1}^{p}(1 - z_g), -\sum_{g=1}^{p}(1 - z_g)\log(1 - \pi_g)\right), \tag{4.2}$$

$$p(\alpha_1|\boldsymbol{z}, \Pi, A, \mathcal{F} = \mathcal{A}) = \mathrm{Ga}_{>1}\left(\alpha_1; \sum_{g=1}^{p} z_g, -\sum_{g=1}^{p} z_g \log \pi_g\right). \tag{4.3}$$

- Given z_g, β_g is also independent of the other model parameters and has a Beta distribution:

$$\begin{aligned} p(\beta_g|z_g, g \in A, \mathcal{F} = \mathcal{A}) &= \mathrm{Be}\left(\beta_g; z_g + \phi_A r_A, (1 - z_g) + \phi_A(1 - r_A)\right), \\ p(\beta_g|z_g, g \in B, \mathcal{F} = \mathcal{A}) &= \mathrm{Be}\left(\beta_g; z_g + \phi_B r_B, (1 - z_g) + \phi_B(1 - r_B)\right). \end{aligned} \tag{4.4}$$

- z_g has a Bernoulli complete conditional distribution independent of $\{\beta_k : k \in \mathcal{G}, k \neq g\}$ and $\{z_k : k \in \mathcal{G}, k \neq g\}$:

$$p(z_g|\alpha_0, \alpha_1, \beta_g, A, \mathcal{F} = \mathcal{A}) = \mathrm{B}\left(z_g; \rho_g\right), \tag{4.5}$$

with

$$\rho_g = \frac{\beta_g \alpha_1 \pi_g^{\alpha_1 - 1}}{\beta_g \alpha_1 \pi_g^{\alpha_1 - 1} + (1 - \beta_g)\alpha_0(1 - \pi_g)^{\alpha_0 - 1}}. \tag{4.6}$$

Here ρ_g is the conditional probability of $g \in \mathcal{A}$.

Gibbs sampling based on these full conditional distributions includes the following steps:

Step 1 : Set starting value for each parameter;

Step 2 : Based on the current value of $\boldsymbol{z}$, sample α_0 and α_1 from the density functions in (4.2) and (4.3), respectively, and update α_0 and α_1 with these samples;

Step 3 : Based on the current value of $\boldsymbol{z}$, sample $\beta_1, \ldots, \beta_p$ in parallel from the density functions in (4.4) and update $\boldsymbol{\beta}$ with these samples;

Step 4 : Based on the current values of α_0, α_1 and $\boldsymbol{\beta}$, sample $z_1, \ldots, z_g$ in parallel from the distribution shown in (4.5) and update $\boldsymbol{z}$ with these samples.

Repeat **Step 1** to **Step 4** $(T + M)$ times and discard the first T samples of each parameters (suppose sample Markov chains are stationary after T iterations). Then model parameters can be estimated using the M posterior samples.

The MCMC procedure is generally fast mixing, and rapid, clean convergence has been confirmed in experiences across many examples.

4.2 Simulation

The simulated data set concerns $p = 18$ synthetic genes with association probabilities

$$\Pi = [0.9698,\ 0.9335,\ 0.9182,\ 0.9369,\ 0.7260,\ 0.0832,\ 0.5776,\ 0.4869,\ 0.3831,$$
$$0.0094,\ 0.0563,\ 0.0529,\ 0.6118,\ 0.0918,\ 0.1603,\ 0.0872,\ 0.1548,\ 0.2257].$$

Here the first five genes with relatively high association probabilities are likely the members of the hypothetical pathway $\mathcal{F}$, while the genes with probabilities lower than 0.3 are not likely to be the members of $\mathcal{F}$. The pathway memberships of the other four genes with probabilities 0.5776, 0.4869, 0.3831 and 0.6118 are less certain.

Clearly, this data set is far from the real case in terms of the number of genes in $\mathcal{G}$ as well as the ratio of the number of pathway genes to the number of non-member genes. The choice of such a small size of data is for the need of testing the marginal likelihood approximation methods, which will be discussed in Chapter 5 and 6, as well as for its utility in clearly illustrating the central aspects and ideas. Further, the small member $p = 18$ is a tolerable size of gene list $\mathcal{G}$ for analytically computing the exact values of the marginal likelihoods, so that numerical approximations of various kinds can be compared to the exact values. The simulation also illuminates how the Gibbs sampler performs under the PROPA model. Focus on one pathway $\mathcal{A}$ as an example here, corresponding to the reference gene set A that consists of exactly the first 8 genes. The data set Π and this gene set A are illustrated in Figure 4.1.

In this example, let $r_A = 0.8$, $r_B = 0.1$ and $\phi_A = \phi_B = 8$. Then the prior and conditional densities of β_g given $g \in A$ and $g \in B$ are shown in Figure 4.2, respectively. As can be seen, such settings of these hyper-parameters allow β_g to change in only relatively small ranges. The prior mean of β_g, r_A or r_B, is the key parameter determining the posterior values of β_g, while the value of z_g makes β_g fluctuate around its prior mean. The amplitude of this fluctuation is constrained by the precision ϕ_A or ϕ_B.

Figure 4.3 demonstrates the trajectories and histograms of the MCMC samples of parameters α_0 and α_1. The red portions in the trace plots represent the burn-in period, 200 samples here. This figure only shows 3000 MCMC samples. The chains generated by this Gibbs sampler appear to mix rapidly. Meanwhile, the histograms show that the posterior samples of α_0 and α_1 have gamma-like distributions.

Figure 4.4 shows the estimated posterior means of each z_g, $g \in \mathcal{G}$. As can be seen, the posterior mean of z_g, i.e. the posterior pathway membership probability of gene g, is largely driven by the data, π_g. The genes in A that have low association

probabilities and genes in B that have high association probabilities are more likely to be false positives and false negatives, respectively, in terms of being members of $\mathcal{A}$. Through this inference, the incorporated prior information, r_A, r_B, ϕ_A and ϕ_B then provide a quantified standard to identify these falsely labeled genes. Here, for each gene the Bayes factor of pathway membership is estimated by

$$BF_{g\in\mathcal{A}}|g \in A \approx \frac{\hat{\beta}_g^{\,*}/(1-\hat{\beta}_g^{*})}{r_A/(1-r_A)}, \tag{4.7}$$

and

$$BF_{g\in\mathcal{A}}|g \in B \approx \frac{\hat{\beta}_g^{\,*}/(1-\hat{\beta}_g^{*})}{r_B/(1-r_B)}, \tag{4.8}$$

where $\hat{\beta}_g^{*}$ is the estimate of pathway membership probability defined in (3.6); this can be obtained by taking the posterior Monte Carlo sample mean of z_g or, alternatively, the mean of ρ_g as shown in (4.6) after the chain reaches equilibrium. The pathway membership evidence in decibans (dB), defined as $10\log_{10} BF_{g\in\mathcal{A}}$, of each gene is plotted in Figure 4.5. Gene 6, a member of gene set A, has membership evidence close to -20dB, strongly suggesting it is not a member of the true pathway signature $\mathcal{A}$ (false positive). Gene 13 is not a member of A, but it has membership evidence greater than 10dB, which is substantial evidence of gene 13 being a member of $\mathcal{A}$ (false negative).

Although the pathway membership of a gene is largely driven by its association probability, this relationship is pathway specific. Another two gene sets are constructed assumedly corresponding to two different pathways. Gene set A_4 contains genes $\{1:4\}$ and represents pathway $\mathcal{A}_4$. Gene set A_6 contains genes $\{1:6\}$ and represents pathway $\mathcal{A}_6$. The pathway membership evidence for each gene in these two pathways are presented in Figure 4.6 (a) and (b), respectively. In the first case, there is no strong evidence for any false positives or false negatives, which means the

predefined gene set A_4 is likely to be an accurate representative of pathway signature $\mathcal{A}_4$. For $\mathcal{A}_6$, the pathway membership inference result gives decisive evidence that gene 6 is not a member of the theoretical pathway $\mathcal{A}_6$.

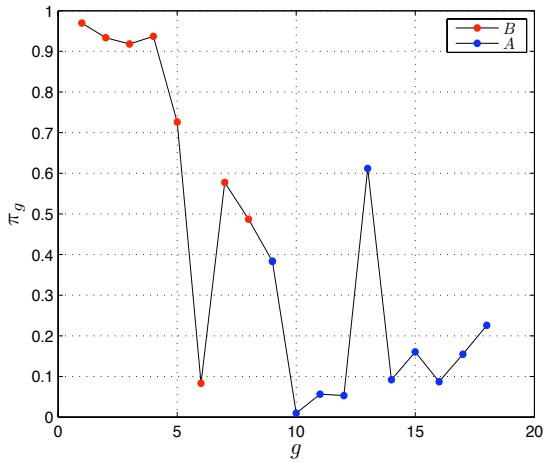


Figure 4.1: The association probabilities in the simulated data set. The red dots correspond to the genes in pathway reference gene set $A = \{1, 2, 3, 4, 5, 6, 7, 8\}$, and the blue dots correspond to those not in A (i.e. in B).

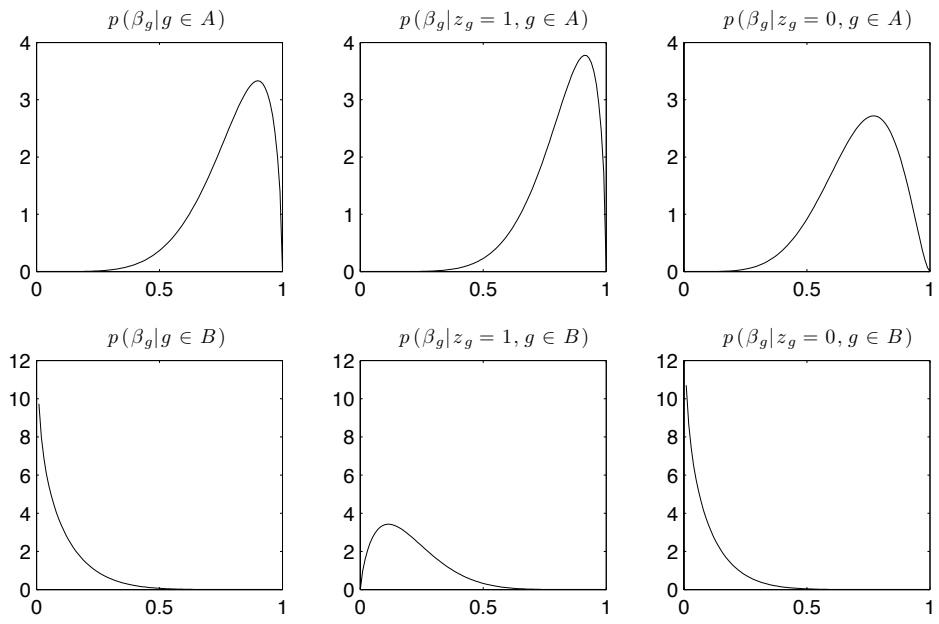


Figure 4.2: The prior and conditional p.d.f.s of membership probability β_g when $r_A = 0.8$, $r_B = 0.1$, $\phi_A = 8$ and $\phi_B = 8$.

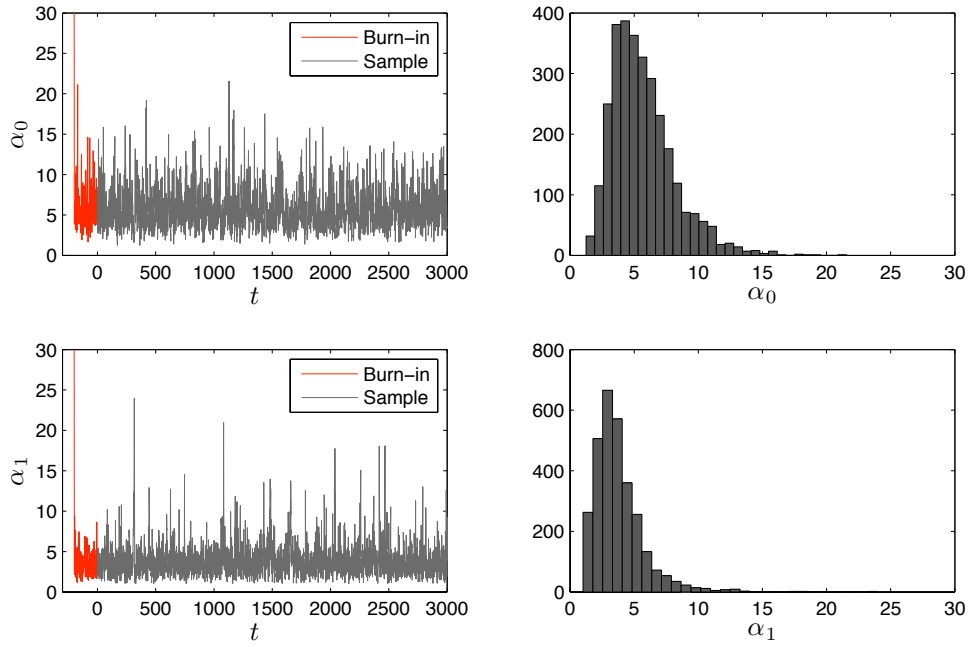


Figure 4.3: The MCMC sample trajectories and histograms of α_0 and α_1. Burn-in = 200, samples = 3000.

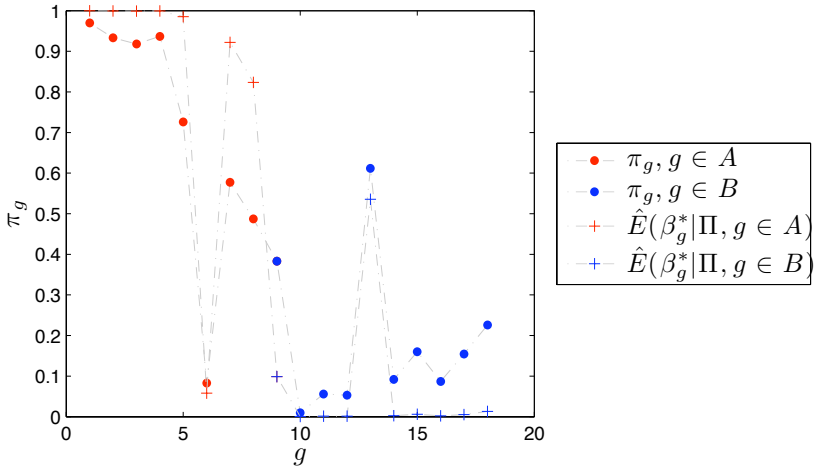


Figure 4.4: The estimated posterior means of β_g^* and z_g for each $g \in \mathcal{G}$.

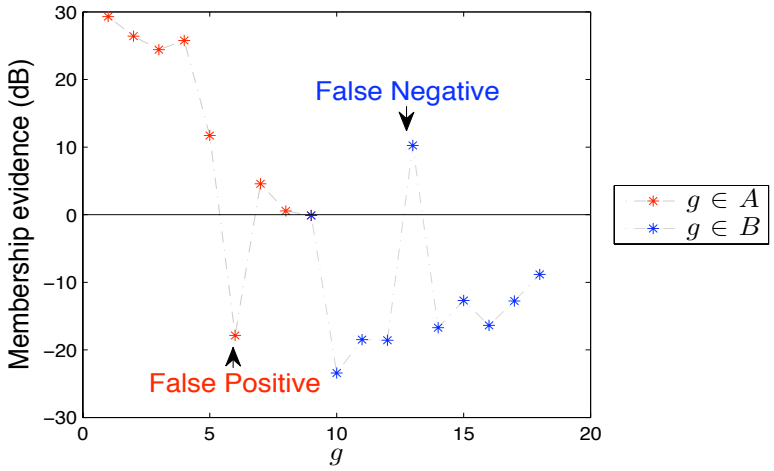
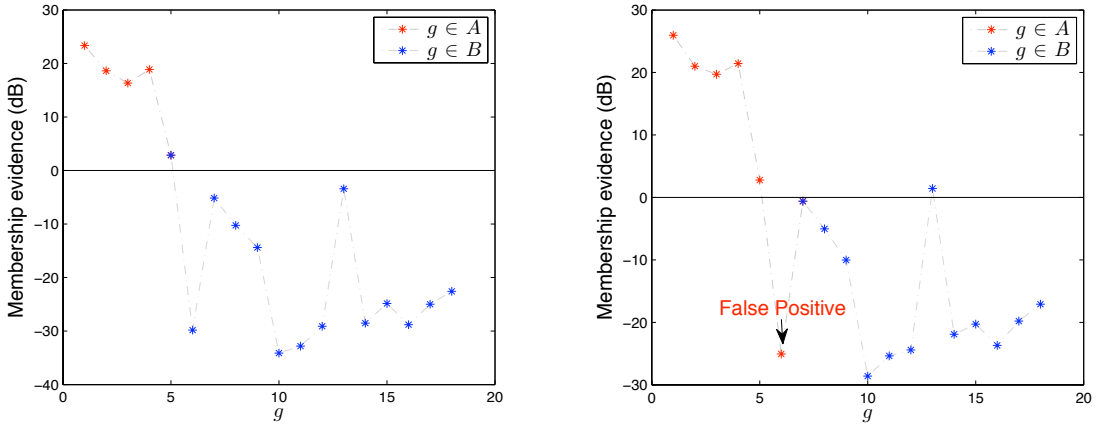


Figure 4.5: Pathway membership evidence for each gene $g \in \mathcal{G}$. The red star below the black line represents the gene that has been predefined as a member gene of pathway $\mathcal{A}$ but should not be as strongly suggested by PROPA; the blue star above the black line represents the gene that has been predefined as not a member gene of pathway $\mathcal{A}$ but may be as suggested by PROPA.



(a) (b)

Figure 4.6: Pathway membership evidence for each gene $g \in \mathcal{G}$ with respect to two different pathway gene sets. (a) The pathway gene set includes the first four genes (A_4); no evidence of any false positives or false negatives. (b) The pathway gene set includes the first six genes (A_6); PROPA strongly suggests the sixth red gene in the gene set is not a member of the pathway.

Chapter 5

Numerical Integration and Approximation

Marginal likelihood is a key quantity for Bayesian model evaluation, comparison and selection. Otherwise known as "evidence", the marginal likelihood is the data probability averaged with respect to all the model parameters with respect to their prior probability distribution. Computing marginal likelihood is an integration problem that, however, appears to be difficult in most realistic applications because of the intractability of the likelihood function. Sometimes the analytical form of antiderivative is not available. It may also be the case that the antiderivative is given as the sum or product of an enormously large number of terms. In such cases, numerical integration and approximation is needed.

In this chapter, the characteristics of the joint density function under PROPA model are studied. I briefly review some existing numerical integration and approximation methods, including quadrature, Laplace approximation, and Monte Carlo integration. Difficulties in using these methods to estimate the marginal likelihood in PROPA are addressed. A simulation study demonstrates the effectiveness of PROPA in assessing pathway-phenotype association and the marginal likelihood approximation with quadrature and Laplace approximation.

5.1 Curse of Dimensionality

Refer to the marginal likelihood computation in PROPA, that is, the expression of (3.13) with the likelihood function $p(\Pi|\boldsymbol{\theta}, A, \mathcal{F} = \mathcal{A})$ shown in (3.9) and the prior

specified in (3.4) and (3.7). Simple algebra leads to the form

$$
\begin{aligned}
& p(\Pi|A, \mathcal{F} = \mathcal{A}) \\
&= \int_1^a \int_1^a \alpha_1^{-1}\alpha_0^{-1} \prod_{g=1}^{p} \left[r_A \alpha_1 \pi_g^{\alpha_1 - 1} + (1 - r_A)\alpha_0 (1 - \pi_g)^{\alpha_0 - 1}\right]^{I(g \in A)} \\
&\qquad \left[r_B \alpha_1 \pi_g^{\alpha_1 - 1} + (1 - r_B)\alpha_0 (1 - \pi_g)^{\alpha_0 - 1}\right]^{I(g \notin A)} \mathrm{d}\alpha_0 \mathrm{d}\alpha_1 \qquad (5.1) \\
&= \int_1^a \int_1^a \left[C_1 \alpha_0^{-1} \mathrm{Ga}_{1\cdot 1}(\alpha_1) + C_2 \alpha_1^{-1} \mathrm{Ga}_{2\cdot 0}(\alpha_0) + \sum_{i=3}^{2^p} C_i \mathrm{Ga}_{i\cdot 1}(\alpha_1)\mathrm{Ga}_{i\cdot 0}(\alpha_0)\right] \mathrm{d}\alpha_0 \mathrm{d}\alpha_1. \qquad (5.2)
\end{aligned}
$$

where, for $i = 1, \ldots, 2^p$, C_i is a constant, and $\mathrm{Ga}_{i\cdot 1}(\alpha_1)$ and $\mathrm{Ga}_{i\cdot 0}(\alpha_0)$ are gamma density functions for α_1 and α_0 specified by certain parameters that can be easily derived. As can be seen, the integrand is a weighted sum of 2^p terms, each of which has an antiderivative in an analytical form, so the exact value of this integral should be available. However, the computatioal complexity in this evaluation is $\mathcal{O}(2^p)$, meaning the computational cost increases exponentially in p. It is known that genome-wide expression data usually contains thousands or even tens of thousands of genes. The computatioal expense in such cases makes it impractical to obtain the exact value of this integral. Therefore, numerical integration methods are applied to approximate this marginal likelihood.

In order to choose appropriate numerical integration methods, one needs to observe the behavior of the integrand (joint density function). Denote the integrand here by $h(\alpha_0, \alpha_1)$ and consider the form shown in (5.1),

$$
\begin{aligned}
h(\alpha_0, \alpha_1) =& \alpha_1^{-1}\alpha_0^{-1} \prod_{g=1}^{p} \{\left[r_A \alpha_1 \pi_g^{\alpha_1 - 1} + (1 - r_A)\alpha_0 (1 - \pi_g)^{\alpha_0 - 1}\right]^{I(g \in A)} \\
& \left[r_B \alpha_1 \pi_g^{\alpha_1 - 1} + (1 - r_B)\alpha_0 (1 - \pi_g)^{\alpha_0 - 1}\right]^{I(g \notin A)}\}. \qquad (5.3)
\end{aligned}
$$

Figure 5.1 presents some examples of $h(\alpha_0, \alpha_1)$ for various example choices of the data

set Π, gene set A and hyper-parameters r_A and r_B. The three figures correspond to three simulated data sets with $p = 18$, 100, and 1000, respectively. For each data set, $h(\alpha_0, \alpha_1)$ is plotted with respect to three different gene sets A_1, A_2 and A_3. The heated areas represent the integration domains where the integrand has high values. Each heated spot corresponds to one distinct gene set.

As can be seen, $h(\alpha_0, \alpha_1)$ appears (in these example cases) to be a unimodal function, though this is not guaranteed for a function which is a mixture of gamma densities (as 5.2). In these examples, $h(\alpha_0, \alpha_1)$ is usually concentrated in a relatively small area whose location varies with the data and gene set. Even so, some adaptive techniques probably can still be found to facilitate standard numerical methods for this integral.

Many numerical integration methods, such as some of those introduced in the following section, inevitably involve the evaluation of the integrand at certain locations in the integration domain, especially where the integrand has high values. As indicated in (5.3), the integrand $h(\alpha_0, \alpha_1)$ is a product of $p + 2$ terms. When p is large, the multiplication can easily cause a computer floating-point overflow problem. This is demonstrated in Table 5.1, where the values of $\log h(\alpha_0, \alpha_1)$ and $h_g(\alpha_0, \alpha_1)$ are exemplified with respect to different p. Even for a moderate value of p, say, 10^3, $h(\alpha_0, \alpha_1)$ has exceeded the largest floating-point number representable on the computer. When such overflows happen, the computer evaluation of the quantities lose precision. So the overflow problem in large p cases is non-trivial in this marginal likelihood computation - the curse of dimensionality.

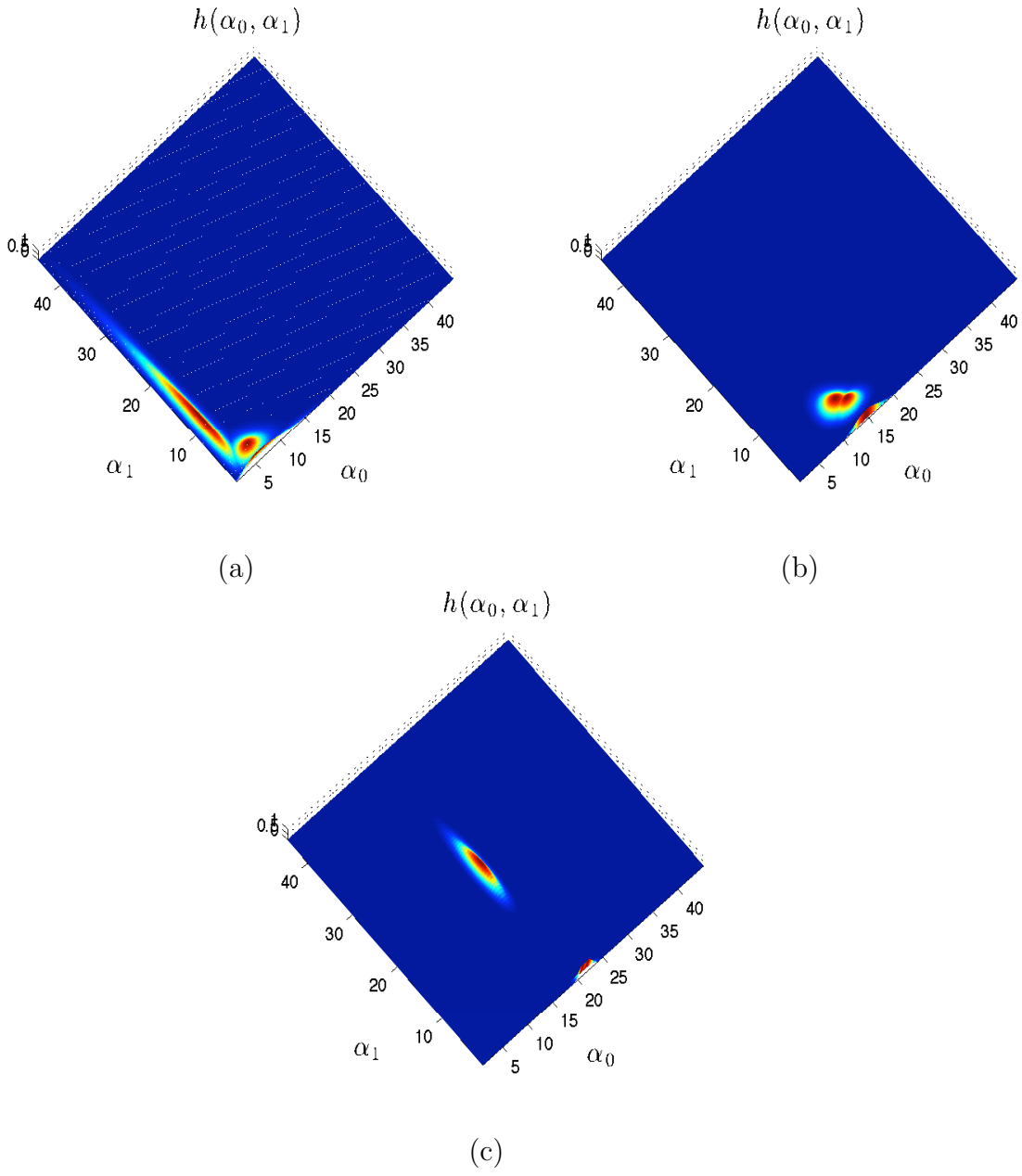


Figure 5.1: Examples of integrand $h(\alpha_0, \alpha_1)$. (a) $p = 18$; (b) $p = 100$; (c) $p = 1000$. In each of the three cases, the logarithm of integrand is evaluated for three different gene set A's. The heated areas represent the places where $h(\alpha_0, \alpha_1)$ has high values corresponding to one gene set.

Table 5.1: Examples of values of $\log h(\alpha_0, \alpha_1)$ and $h(\alpha_0, \alpha_1)$ with respect to different p. $h(\alpha_0, \alpha_1)$ generally increases with p and overflows when p reaches 10^3. In such cases, $\log h(\alpha_0, \alpha_1)$ is a projection of the real value of $h(\alpha_0, \alpha_1)$.

p	$\log h_g(\alpha_0, \alpha_1)$	$h_g(\alpha_0, \alpha_1)$
18	6.6849	800.2307
100	131.6183	1.4491e+57
1000	1.8950e+03	Inf
10000	1.7184e+03	Inf

5.2 Numerical Integration Methods

5.2.1 Quadrature

Quadrature is the standard, most straightforward numerical integration approach (Ueberhuber 1997; Davis and Rabinowitz 1984). This approach evaluates the integrand $h(\theta)$ at a number of pre-determined locations $\theta^{(1)}, \ldots, \theta^{(N)}$, which usually are regularly spaced throughout the whole integration interval, then uses the weighted sum of these evaluation to approximate the integral, namely,

$$\int_{\Theta} h(\theta) \mathrm{d}\theta \approx \sum_{i=1}^{N} w_i h(\theta_{(i)}).$$

There are many rules for assigning weights to the evaluated results at the locations, such as Trapezoidal rule, Simpson's rule and Boole's rule corresponding to the employment of 2-, 3- and 5-point Newton-Cotes formulas derived from Lagrange interpolation. The simplest one is the Trapezoidal rule, by which the evaluation results of integrand are equally weighted by the spacing interval (except for the first and last locations). This rule is widely used for its simplicity and accuracy.

A more complicated method is Gaussian quadrature, which optimizes the numerical integration by selecting specific evaluation locations based on certain integration interval-dependent rules rather than equally spacing the whole integration interval.

The complication of selecting evaluation locations lowers the desirability of Gaussian quadrature. Simply decreasing the space between evaluation locations in quadrature can equally increase the accuracy, while the increased computational cost is not much more than Guaussian quadrature.

Quadrature is more frequently used for approximating one-dimensional integrals for its ease and good accuracy in most cases (it is also called cubature in a multiple integral). However, the number of evaluations increases exponentially with the dimension of integral, which leads to tremendous computational cost and a decrease in approximation accuracy. Thus quadrature is limited in multi-dimensional integrals, especially when the integrand is not smooth or well diffused. Some more sophisticated methods, for example, sparse grids (Gerstner and Griebel 1998), have been developed based on certain techniques of combining one-dimensional quadrature results to solve multi-dimensional integration problems.

5.2.2 Laplace Approximation

Laplace's method approximates the marginal likelihood by fitting a normal density function at the maximum of the joint density $h(\theta) = p(\theta|D)p(\theta)$, where D is data, and computing the volume under the Gaussian curve within the domain of θ (Tierney and Kadane 1986). The marginal likelihood estimator is then

$$\hat{p}(D) = \int_{\Theta} h(\theta)\mathrm{d}\theta \approx (2\pi)^{d/2}|\hat{\Psi}|^{1/2}h(\hat{\theta})$$

where d is the dimension of θ and $\hat{\theta} = \arg\max_{\theta} \log h(\theta)$, corresponding to the MAP estimate of θ, and $\hat{\Psi} = (-\hat{H})^{-1}$ where $\hat{H}$ is the Hessian matrix of $h(\theta)$ evaluated at $\hat{\theta}$. Newton-Raphson's method is frequently applied to obtain $\hat{\theta}$. The fundamental idea of Laplace approximation is using a normal density $N(\hat{\theta}, \hat{\Psi})$ to approximate the

posterior density $p(\theta|D)$, and estimating the marginal likelihood at $\hat{\theta}$. This method is most relevant when the posterior distribution is unimodal.

5.2.3 Monte Carlo Integration

Generally, Monte Carlo integration methods involve the evaluation and averaging of certain objective functions, which depend on the integrand and specific approaches, at locations randomly sampled in the domain of integration. Usually a large number of samples of integration variables are required to attain good accuracy. The computational cost for sampling makes the Monte Carlo integration methods uncompetitive with quadrature in one or two dimensional integration. However the convergence rate is independent of the number of dimensions. Hence, Monte Carlo integration methods are more often used for the approximation of higher dimensional integration. In this section, I briefly summarize importance sampling and Metropolis-based numerical estimators without discussion of the specific integral estimation problem in the PROPA models.

Importance Sampling

The importance sampling method approximates the integral with a weighted sum of integrand values evaluated at the locations sampled from an importance distribution $g(\theta)$, which has an explicit mathematical form, i.e.

$$\hat{p}(D) = \frac{1}{N} \sum_{i=1}^{N} \frac{h(\theta^{(i)})}{g(\theta^{(i)})},$$

where $\theta^{(1)}, \ldots, \theta^{(N)} \sim g(\theta)$. The p.d.f. of the importance distribution $g(\theta)$ must have a support larger than that of the integrand and be easy to sample. There are no rigorous criteria for choosing a good $g(\theta)$ except that the integration is more efficient

if $g(\theta)$ has a shape similar to $h(\theta)$ and $g(\theta)$ is relatively better diffused than $h(\theta)$ (Robert and Casella 2004).

Quantile integration is an extension of importance sampling (Johnson 1992). Given the importance distribution, the samples of θ are quantile interval means of $g(\theta)$, i.e.

$$\hat{p}(D) = \frac{1}{N}\sum_{i=1}^{N}\frac{h(w^{(i)})}{g(w^{(i)})},$$

where $w^{(i)}$ satisfies $\int_{\Theta^{(i)}}(\theta - w^{(i)})g(\theta)\mathrm{d}\theta = 0$ and $\int_{\Theta^{(i)}} g(\theta)\mathrm{d}\theta = 1/N$. Compared with importance sampling, this approach is theoretically more efficient, but much harder to implement.

The simplest case is sampling from the prior distribution $p(\theta)$, thus,

$$\hat{p}(D) = \frac{1}{N}\sum_{i=1}^{N} p(D|\theta^{(i)}).$$

This is generally inadequate because poor similarity between the likelihood, $p(D|\theta)$, and the prior leads to high variability in the estimate.

Posterior Sampling Estimator

In the importance sampling method for marginal likelihood estimation, the ideal importance distribution would be the posterior distribution $p(\theta|D)$, because the joint distribution $h(\theta)$ as a function of θ is $p(\theta|D)$ times a normalizing constant. However, the explicit form of $p(\theta|D)$ is not available, because the normalizing constant is the marginal likelihood itself. Importance sampling in such a case can be extended to a large class of Monte Carlo integration approaches using marginal posterior samples of each integration variable provided by Metropolis-Hestings sampling or its specific case, Gibbs sampling when the Markov chains reaches equilibrium. Gamerman and

Lopes (2006) summarized the existing Monte Carlo marginal likelihood estimators based on posterior samples. The key approaches are:

- Harmonic mean estimator (Newton and Raftery 1994)

$$\hat{p}(D) = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{p(D|\theta^{(i)})} \right)^{-1},$$

  where $\theta^{(1)}, \ldots, \theta^{(N)} \sim p(\theta|D)$. This estimator is extremely sensitive to small likelihood values.

- Generalized harmonic mean estimator (Gelfand and Dey 1994)

$$\hat{p}(D) = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{g(\theta^{(i)})}{h(\theta^{(i)})} \right)^{-1},$$

  where $\theta^{(1)}, \ldots, \theta^{(N)} \sim p(\theta|D)$, and $g(\theta)$ can be any density function on the same support of $h(\theta)$, but has to be chosen carefully to obtain good accuracy in the estimation. Clearly, $\hat{p}(D) = p(D)$, when $g(\theta) = p(\theta|D)$.

- Newton and Raftery's estimator (Newton and Raftery 1994)

  This estimator uses an iterative scheme,

$$\hat{p}^{(t)}(D) = \frac{\sum_{i=1}^{N} p(D|\theta^{(i)}) \left[\delta \hat{p}^{(t-1)}(D) + (1-\delta) p(D|\theta^{(i)}) \right]^{-1}}{\sum_{i=1}^{N} \left[\delta \hat{p}^{(t-1)}(D) + (1-\delta) p(D|\theta^{(i)}) \right]^{-1}},$$

  where $\theta^{(1)}, \ldots, \theta^{(N)} \sim \delta p(\theta) + (1-\delta) p(\theta|D)$ with $0 < \delta < 1$.

- Laplace-Metropolis estimator (Lewis and Raftery 1997)

$$\hat{p}(D) = (2\pi)^{d/2} |\tilde{\Psi}|^{1/2} h(\tilde{\theta}).$$

d is the dimension of θ. This estimator is the same as traditional Laplace approximation except that $\tilde{\theta}$ is from the posterior samples of θ that maximizes $h(\theta)$, and $\tilde{\Psi}$ is the MCMC variance of posterior samples.

- Bridge sampling estimator (Meng and Wong 1996)

  The marginal likelihood is estimated iteratively as

$$\hat{p}^{(t)}(D) = \frac{\frac{1}{N_2}\sum_{i=1}^{N_2}\tilde{\omega}^{(i)}\left[\frac{N_1}{N_1+N_2}\tilde{\omega}^{(i)} + \frac{N_2}{N_1+N_2}\hat{p}^{(t-1)}(D)\right]^{-1}}{\frac{1}{N_1}\sum_{i=1}^{N_1}\left[\frac{N_1}{N_1+N_2}\omega^{(i)} + \frac{N2}{N_1+N_2}\hat{p}^{(t-1)}(D)\right]^{-1}},$$

  where $\omega^{(i)} = h(\theta^{(i)})/g(\theta^{(i)})$ for $i = 1, \ldots, N_1$, and $\tilde{\omega}^{(i)} = h(\tilde{\theta}^{(i)})/g(\tilde{\theta}^{(i)})$ for $i = 1, \ldots, N_2$. $\theta^{(1)}, \ldots, \theta^{(N_1)} \sim p(\theta|D)$, and $\tilde{\theta}^{(1)}, \ldots, \tilde{\theta}^{(N_2)} \sim g(\tilde{\theta})$. The accuracy of this estimator depends on the distance between the proposal density g and the posterior density.

- Candidate's estimator (Chib 1995)

  A simple and intuitive estimator is

$$\hat{p}(D) = \frac{h(\theta)}{\hat{p}(\theta|D)},$$

  where $\hat{p}(\theta|D)$ is an approximation density function of the posterior density. Precise marginal likelihood in principle can be obtained by taking any value of θ within the support of $h(\theta)$ if $\hat{p}(\theta|D) = p(\theta|D)$. However, $\hat{p}(\theta|D)$ is an approximation. Hence, the value of θ need to be chosen so that the distance between $\hat{p}(\theta|D)$ and $p(\theta|D)$ can be minimized. Usually, θ is chosen to be the posterior sample mean or mode, or any value around the mean or mode. The Laplace-Metropolis estimator can be viewed as a special case of Candidate's estimator, where $\hat{p}(\theta|D)$ is a normal density whose mean and variance are estimated using posterior samples of θ.

5.3 Simulation

In this section, I present a simulation study of marginal likelihood computation for the PROPA model. The purpose is to demonstrate the effectiveness of PROPA and the performance of the applicable numerical integration methods, specifically, quadrature and Laplace approximation. Quadrature requires the evaluation of the integrand, hence it is limited when p is large. However, I choose this method for the simulation study because of its simplicity and generally good accuracy. Laplace approximation is also easy to implement in PROPA. The greatest advantage of this method is that this estimate has a form allowing log transformation of the integrand. The main difficulty of importance sampling lies in the choice of the importance density function, and this is hard when the integrand is a mixture of a large number of density functions. Figure 5.1 showed that $h(\alpha_0, \alpha_1)$ is not well diffused in the integration domain and varies widely depending on the data and model. Finding an appropriate importance density is a hard problem. The same difficulty exists for the generalized harmonic mean estimator. Some posterior sampling estimators are limited by the inevitable evaluation of either joint likelihood or likelihood function because of their summation forms, such as the harmonic mean estimator, Newton and Raftery's estimator and bridge sampling estimator. The Laplace-Metropolis estimator is a substitution of Laplace approximation when analytically fitting the normal density is difficult. To the integration problem in the PROPA models, it is not necessary. Candidate's estimator also has an appealing form that allows log transform of joint density during the marginal likelihood approximation. However, determining the posterior approximation density $\hat{p}(\theta|D)$ appears to be hard in the integration problem in the PROPA models. The usual procedure for determining $\hat{p}(\theta|D)$ is to choose a parameterized density kernel that looks like the empirical posterior distribution of θ, then use the posterior samples to estimate these parameters to get $\hat{p}(\theta)$.

Since the posterior distribution of parameters α_0 and α_1 is truncated, it is difficult to approximate using the posterior samples.

In this simulation study, I use the data set described in Section 4.2. The number of genes in this simulated data set is $p = 18$, a small value that allows for exact evaluation of the joint density function and the marginal likelihood. Seventeen gene sets are generated and denoted by $\{A_{s_A}\}_{s_A=1:17}$, representing pathways $\{\mathcal{A}_{s_A}\}_{s_A=1:17}$. Each gene set A_{s_A} is composed of the first s_A genes. Figure 5.2 shows the association probabilities of all the genes and illustrates how the 17 gene sets are constructed. Again, the prior means of gene pathway membership $r_A = 0.8$ and $r_B = 0.1$. The joint density function conditional on each gene set is exhibited as a contour plot in Figure 5.3. For different gene sets, the joint density is concentrated in different domains of α_0 and α_1. Generally, the concentration field shifts from the large α_1 and small α_0 area to the small α_1 and large α_0 areas with the increase of s_A. This is decided by the arrangement of π_g in this data set, which is generally in a decreasing order.

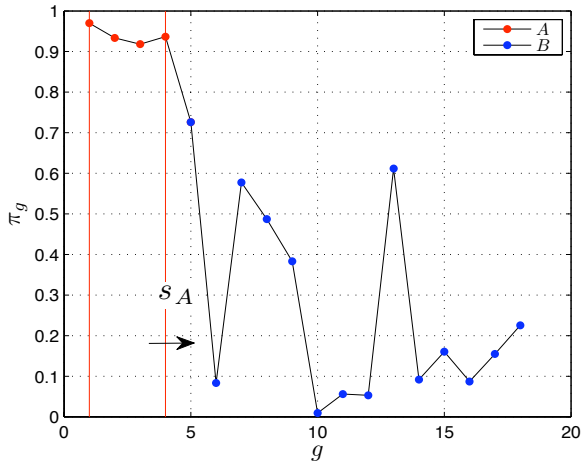


Figure 5.2: Association probabilities in the simulated data set. The red dots correspond to the genes in pathway gene set A, and the blue dots correspond to those not in A (i.e. in B). A_{s_A} is the pathway gene set that includes the first s_A genes of the 18. By increasing s_A from 1 to 17, 17 gene sets are generated.

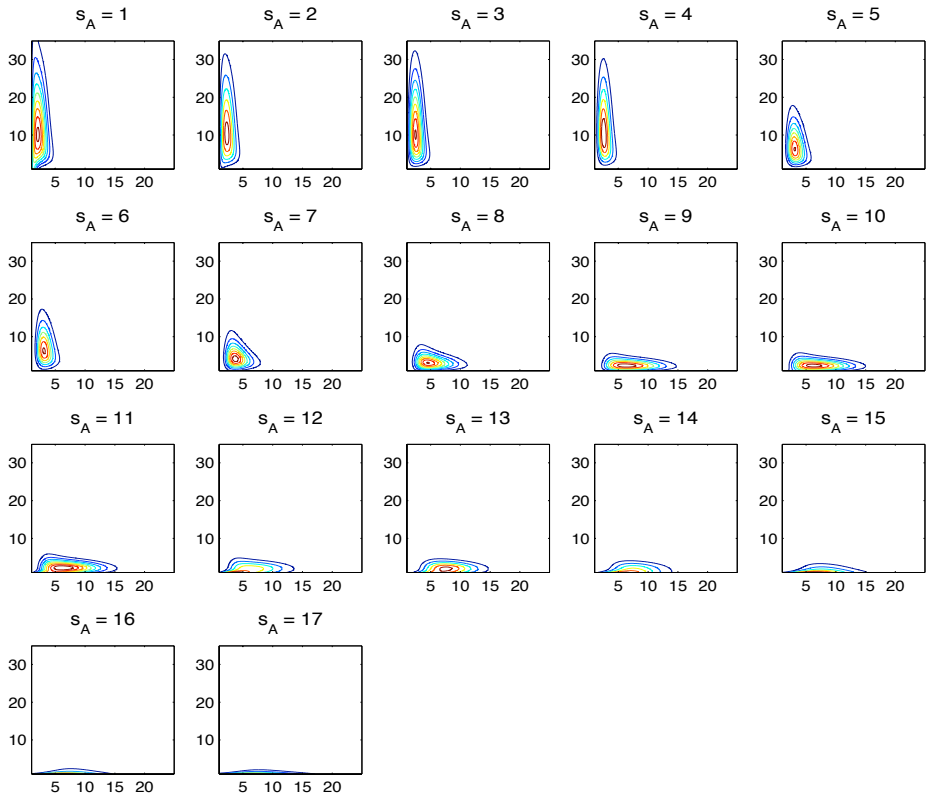


Figure 5.3: Contour plots of $h(\alpha_0, \alpha_1)$ corresponding to 17 pathway gene sets. In each plot, x-axis is α_0; y-axis is α_1.

Since $p = 18$ the exact values of the marginal likelihood given each gene set can be easily computed by using the analytical form derived in (5.2). The standardized log marginal likelihood (scaled to $[0, 1]$ by dividing the maximum), as shown in Figure 5.4, increases when $s_A = 1 : 5$ and reaches the peak at $s_A = 4$ and 5, giving evidence that the gene sets containing the first 4 or 5 genes are most likely to be associated with the phenotype. This is consistent with the original simulation design in that the first few genes are the signature genes of the hypothetical pathway $\mathcal{F}$. The variation of log marginal likelihood across the rest of the gene sets can easily be interpreted by the fluctuation of the π_g across genes.

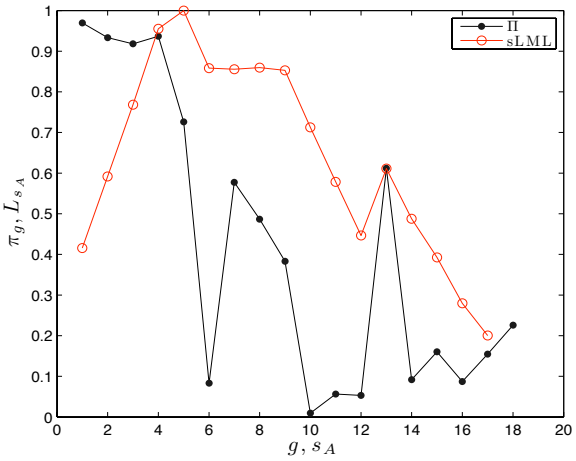


Figure 5.4: Standardized log marginal likelihood for each of the 17 pathways in the simulation.

Figure 5.5 presented the exact log marginal likelihoods of the 17 gene sets and the corresponding approximate values obtained by using quadrature integration and Laplace approximation. Among the 17 values, the maximum and minimum of quadrature approximation errors are 0.055 and 3×10^{-5}. For Laplace approximation, the minimum error is 0.002; but three approximate values, corresponding to A_2, A_{12} and A_{16}, are obviously wrong; the approximate value for A_{17} is not available due to the numerical problem. Joint density functions with A_2, A_{12} and A_{16} are exhibited in Figure 5.6. The poor accuracy of Laplace approximation may be either due to the irregularity (nonunimodality) or poor Gaussianality of $h(\alpha_0, \alpha_1)$ in the integration domain. Particularly, although $h(\alpha_0, \alpha_1)$ appears to be unimodal in the contour plots, the subtle irregularity still exists and may cause the algorithm converging to a local maximum as well as erratic/unstable values of the Hessian at the local Maximum. Setting different starting points of integration variables usually is the approach to dealing with the irregularity. However, this depends on observation of the global maximum of $\log h(\alpha_0, \alpha_1)$, which can be inaccurate and lead to poor approximation.

In summary, quadrature has good performance in approximating the evidence

when the number of genes in the data set is not large. Laplace approximation is a fast approach and potentially can deal with data with a large number of genes, though the performance is not reliable. An approach needs to be developed to work with large data sets and provide a reliable estimate of marginal likelihood.

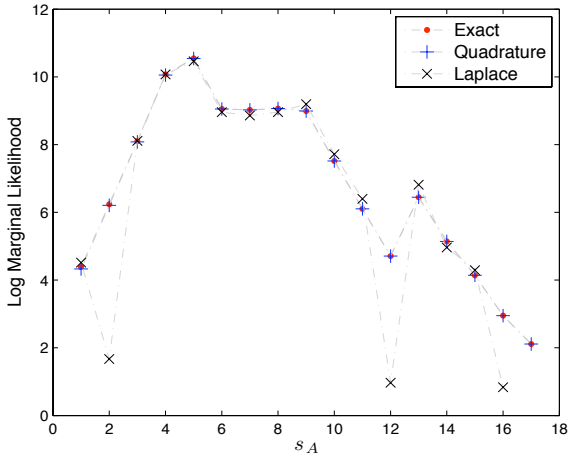
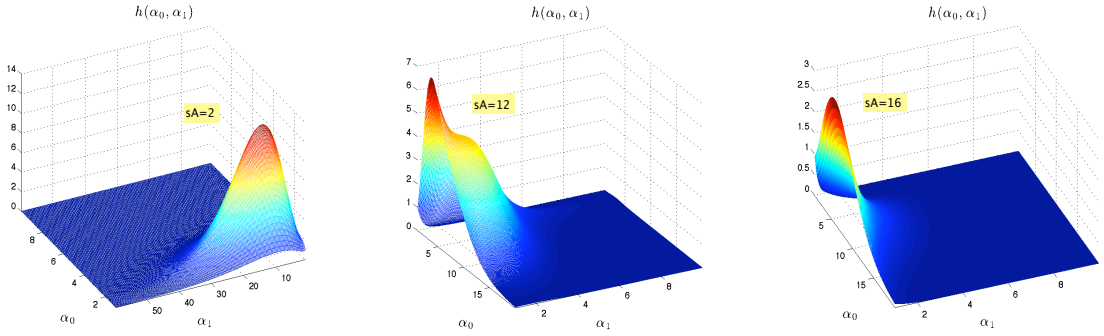


Figure 5.5: Log marginal likelihoods of 17 pathways in the simulation. Red circles are exact values; blue +'s are estimates with quadrature integration; black x's are estimates with Laplace approximation.



(a) (b) (c)

Figure 5.6: $h(\alpha_0, \alpha_1)$ when $s_A = 2$, 12 and 16. These are the cases where the Laplace approximation has the worst accuracy .

Chapter 6

Variational Methods

Variational methods provide approximate inference algorithms that yield lower bounds on marginal probabilities of interest (Jaakkola and Jordan 1997; Jordan *et al.* 1999). These type of methods have been intensively studied and applied to Bayesian inference and model comparison (or selection) problems (Corduneanu and Bishop 2001; McGrory and Titterington 2007). Specifically, for marginal likelihood computation, an existing variational approximation method, called variational Bayesian EM by Beal (2003), uses the Expectation Maximization (EM) algorithm to achieve the optimal lower bound of a log marginal likelihood under a posterior independence assumption for model parameters. Model comparisons or hypothesis testing based on the lower bounds of marginal likelihoods, however, is insufficient and not persuasive.

I propose a new method for log marginal likelihood approximation. This method is built on the same foundation as existing variational approaches, but uses Monte Carlo simulation to obtain optimum bounds – *both upper and lower* – for a log marginal likelihood. Bound optimizations on the two sides correspond to the two-way minimization of Kullback-Leibler (K-L) divergence between two variational densities and the joint posterior density of model parameters. The upper bound optimization depends on the posterior samples of parameters generated through an MCMC method, while the lower bound optimization utilizes Monte Carlo stochastic approximation approach that appears to be more general compared to variational Bayesian EM. Simulation studies are performed to show the effectiveness of this method.

6.1 Approximation with Lower Bounds

The computation of marginal likelihood in PROPA encounters a floating-point overflow problem when the number of genes is large. As has been discussed in the previous section, a desirable solution is to approximate the marginal likelihood with a certain form in which the quantities causing overflow are put in a log transform. Although not specifically developed for solving such high-dimensional problems, the Cheeseman-Stutz (Cheeseman and Stutz 1996) method and the Variational Bayesian EM (Beal 2003) meet such a requirement and provide lower bounds on the log marginal likelihood. Both methods have been developed for the settings that contain hidden variables and depend on the EM algorithm providing solutions.

Cheeseman-Stutz Approximation

Denote the marginal likelihood by $p(D|M)$, where D is the data and M is the model with parameters $\boldsymbol{\theta}$. For any hidden variables $\boldsymbol{z}$, the log marginal likelihood can be approximated as

$$\begin{aligned}\log p(D|M) \approx L_{CS} &= \log\left(p(D,\hat{\boldsymbol{z}}|M)\frac{p(D|\hat{\boldsymbol{\theta}},M)}{p(D,\hat{\boldsymbol{z}}|\hat{\boldsymbol{\theta}},M)}\right)\\ &= \log p(D,\hat{\boldsymbol{z}}|M) + \log p(D|\hat{\boldsymbol{\theta}},M) - \log p(D,\hat{\boldsymbol{z}}|\hat{\boldsymbol{\theta}},M)\end{aligned}$$

where $\hat{\boldsymbol{\theta}}$ is the a maximum likelihood (ML) or a maximum *á posteriori* (MAP) estimate of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{z}}$ is the expectation of $\boldsymbol{z}$ given $\hat{\boldsymbol{\theta}}$. $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{z}}$ are obtained via the EM algorithm. This approximation was initially proposed by Cheeseman and Stutz (1996) and later was noted by Minka (2001) as a lower bound of log marginal likelihood in the context of mixture models. Beal (2003) extended this conclusion to any model and provided a proof.

Variational Bayesian EM

The mean-field variational methods were initially developed in statistical physics and extensively studied by machine learning and Bayesian learning communities for deterministic approximation of marginal distributions (MacKay 1995; Jordan *et al.* 1999; Jaakkola and Jordan 2000; Humphreys and Titterington 2000; Corduneanu and Bishop 2001; Ueda and Ghahramani 2002; Beal and Ghahramani 2003; Jordan 2004; Wang and Titterington 2004; McGrory and Titterington 2007). Beal (2003) reviewed and examined these methods for Conjugate-Exponential models. The implementation of the variational method for the purpose of marginal likelihood approximation was called Variational Bayesian EM (VBEM).

Consider Jensen's inequality in the marginal likelihood approximation context,

$$\log p(D|M) \geq \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{Z}} q(\boldsymbol{z}, \boldsymbol{\theta}) \log \frac{p(D, \boldsymbol{z}, \boldsymbol{\theta}|M)}{q(\boldsymbol{z}, \boldsymbol{\theta})} \mathrm{d}\boldsymbol{z} \mathrm{d}\boldsymbol{\theta},$$

where $\boldsymbol{z}$ and $\boldsymbol{\theta}$, as before, are hidden variables and model parameters, respectively, and $q(\boldsymbol{z}, \boldsymbol{\theta})$ is any p.d.f. supported by $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$. The inequality sets a lower bound of log marginal likelihood as an integral in which the joint density function as part of the integrand is under a log transform. This looks appealing when one thinks about solving the numerical problem in computing marginal likelihood.

By factorizing the variational density with respect to the hidden variables and model parameters, i.e. $q(\boldsymbol{z}, \boldsymbol{\theta}) = q_{\boldsymbol{z}}(\boldsymbol{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, VBEM iteratively performs the following steps to find the optimum $q_{\boldsymbol{z}}(\boldsymbol{z})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ that maximize this lower bound:

$$q_{\boldsymbol{z}}^{(t+1)}(\boldsymbol{z}) = \frac{1}{C_{\boldsymbol{z}}} \exp \left[\int_{\boldsymbol{\Theta}} q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \log p(D, \boldsymbol{z}|\boldsymbol{\theta}, M) \mathrm{d}\boldsymbol{\theta} \right]$$

and

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) = \frac{1}{C_{\boldsymbol{\theta}}} p(\boldsymbol{\theta}|M) \exp\left[\int_{\boldsymbol{Z}} q_{\boldsymbol{z}}^{(t+1)}(\boldsymbol{z}) \log p(D, \boldsymbol{z}|\boldsymbol{\theta}, M) \mathrm{d}\boldsymbol{z}\right],$$

where $C_{\boldsymbol{z}}$ and $C_{\boldsymbol{\theta}}$ are normalizing constants. Beal (2003) has shown that this method can always obtain a tighter lower bound than the Cheeseman-Stutz approximation.

Since optimization in this functional form usually is infeasible, the factors of the mean-field variational density, $q_{\boldsymbol{z}}(\boldsymbol{z})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, need to be assumed in certain parameterized density function forms that enable computation. The lower bound optimization is then carried out with respect to the parameters in the assumed density forms. This iterative algorithm converges to the local maximum lower bound of the log marginal likelihood.

Upper Bound on Marginal Likelihood

Clearly, performing model comparisons based only on the lower bounds of marginal likelihoods can be inappropriate as the approximation error is not quantitatively limited. An ideal approximation in such cases requires an upper bound coupled with the lower bound to confine the true value of marginal likelihood. One of the quantities that could serve as an upper bound is the maximum likelihood given that the optimization process truly finds the global maximum of the likelihood. Obviously, this upper bound almost always will be very loose and of little help in restricting the estimate of marginal likelihood. The expectation of the data likelihood under the posterior distribution of parameters is also an upper bound on the marginal likelihood (Beal 2003). This is also a poor upper bound because the tightness of this bound drifts away as sample size increases. In the following section I propose a new, tight upper bound for log marginal likelihood.

6.2 Monte Carlo Variational Method

I propose a new variational method providing optimized lower bound *and* upper bound for the marginal likelihood. In this method, a Monte Carlo stochastic approximation algorithm is employed as an alternative but more general approach to the optimization of the lower bound as defined by Jensen's inequality and optimized with EM algorithm in VBEM. The upper bound is derived under the variational framework and optimized by using the posterior samples of model parameters obtained with MCMC methods.

I begin with derivation of the lower bound and upper bound of log marginal likelihood in the general framework of variational methods, and follow this up with the description of the bound optimization methods under the assumption of exponential family mean-field variational density forms. Convergence of the optimization algorithms proposed is discussed. Then this approach is applied to the marginal likelihood approximation in PROPA models, and the performance is studied in simulation studies.

6.2.1 Lower and Upper Bounding Marginal Likelihood

Denote the marginal likelihood in a general form by

$$p(D|M) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}, D|M) \mathrm{d}\boldsymbol{\theta},$$

where D is the data, M is the model, and $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\} \in \boldsymbol{\Theta}$ represents all the model parameters. In case some of parameters can be analytically integrated out, the dimension of this integral is reduced. Then the integration over the rest of the parameters fits in the discussion.

For any density function $q(\boldsymbol{\theta})$ that has the same support as the posterior distri-

bution $p(\boldsymbol{\theta}|D, M)$, by Jensen's inequality,

$$\log p(D|M) \geq \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, D|M)}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}, \tag{6.1}$$

a lower-bound for the log marginal likelihood can be set up with specific $q_L(\boldsymbol{\theta})$, i.e.

$$L = \int_{\boldsymbol{\Theta}} q_L(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}, D|M)}{q_L(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}.$$

When $q_L(\boldsymbol{\theta})$ takes an analytical form, this lower bound is simply the expectation of $\log \frac{p(\boldsymbol{\theta}, D|M)}{q_L(\boldsymbol{\theta})}$ with respect to this density $q_L(\boldsymbol{\theta})$.

Now take a step back. The inequality in (6.1) turns into equality only when the free distribution $q(\boldsymbol{\theta})$ is the posterior distribution of $\boldsymbol{\theta}$, i.e. $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|D, M)$. Consider the equality

$$\begin{aligned} \log p(D|M) &= \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D, M) \log \frac{p(\boldsymbol{\theta}, D|M)}{p(\boldsymbol{\theta}|D, M)} \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D, M) \log p(\boldsymbol{\theta}, D|M) \mathrm{d}\boldsymbol{\theta} - \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D, M) \log p(\boldsymbol{\theta}|D, M) \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{6.2}$$

The second term $-\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D, M) \log p(\boldsymbol{\theta}|D, M) \mathrm{d}\boldsymbol{\theta}$ is the mathematical entropy of $p(\boldsymbol{\theta}|D, M)$. According to Gibbs inequality, for any probability density $q_U(\boldsymbol{\theta})$ that has same support as $p(\boldsymbol{\theta}|D, M)$,

$$-\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D, M) \log p(\boldsymbol{\theta}|D, M) \mathrm{d}\boldsymbol{\theta} < -\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D, M) \log q_U(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

Plugging this into (6.2) results in an upper bound of log marginal likelihood

$$\begin{aligned} U &= \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log p(\boldsymbol{\theta},D|M)\mathrm{d}\boldsymbol{\theta} - \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log q_U(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log\frac{p(\boldsymbol{\theta},D|M)}{q_U(\boldsymbol{\theta})}\mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{6.3}$$

Clearly, given the form of $q_U(\boldsymbol{\theta})$, this upper bound is the expectation of $\log\frac{p(\boldsymbol{\theta},D|M)}{q_U(\boldsymbol{\theta})}$ with respect to the posterior distribution $p(\boldsymbol{\theta}|D,M)$. If $p(\boldsymbol{\theta}|D,M)$ does not have an explicit form (this actually is always the case because its normalizing constant $p(D|M)$ is unknown), the evaluation of this upper bound becomes a Monte Carlo integration problem that depends on the samples from $p(\boldsymbol{\theta}|D,M)$.

Now the log marginal likelihood is bounded as

$$L \leq \log p(D|M) \leq U.$$

To optimize the approximation of the log marginal likelihood, one needs to find the density functions $q_U(\boldsymbol{\theta})$ and $q_L(\boldsymbol{\theta})$ that minimize the upper bound and maximize the lower bound, respectively. Suppose both $q_U(\boldsymbol{\theta})$ and $q_L(\boldsymbol{\theta})$ take parameterized form $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ with $\boldsymbol{\gamma}\in\boldsymbol{\Gamma}$ a vector of tunable parameters. The goal is to find $\boldsymbol{\gamma}_U$ and $\boldsymbol{\gamma}_L$, so that $q(\boldsymbol{\theta};\boldsymbol{\gamma}_U)$ and $q(\boldsymbol{\theta};\boldsymbol{\gamma}_L)$ minimize the upper bound and maximize the lower bound, respectively. Then, conditional on the choice of the form of $q(\boldsymbol{\theta};\boldsymbol{\gamma})$, the log marginal likelihood is optimally bounded as

$$L_o \leq \log p(D|M) \leq U_o,$$

where

$$U_o = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log\frac{p(\boldsymbol{\theta},D|M)}{q(\boldsymbol{\theta};\boldsymbol{\gamma}_U)}\mathrm{d}\boldsymbol{\theta},$$

and

$$L_o = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta};\boldsymbol{\gamma}_L) \log \frac{p(\boldsymbol{\theta}, D|M)}{q(\boldsymbol{\theta};\boldsymbol{\gamma}_L)} \mathrm{d}\boldsymbol{\theta}. \tag{6.4}$$

This lower bound optimization is the basic idea of existing variational Bayesian methods. I name $q(\boldsymbol{\theta};\boldsymbol{\gamma}_U)$ and $q(\boldsymbol{\theta};\boldsymbol{\gamma}_L)$ the upper bound variational density function and lower bound variational density function, respectively.

Now the computation of log marginal likelihood relies on the availability of three things: 1) a good form of the variational density function kernel $q(\boldsymbol{\theta};\boldsymbol{\gamma})$, 2) an optimization strategy for $\boldsymbol{\gamma}$, and 3) the samples from posterior distribution $p(\boldsymbol{\theta}|D, M)$. The natural way (in fact the only possible way in the context) to generate $p(\boldsymbol{\theta}|D, M)$ samples is the MCMC method. I then focus on the construction and optimization of variational densities.

The choice of the kernel of $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ usually is based on the consideration of computational convenience. The tightness of the bounds are then determined by how good $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ can be as an approximation to the posterior density $p(\boldsymbol{\theta}|D, M)$. In particular, a factorized form for $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ is considered based on the mean field theory, i.e.

$$q(\boldsymbol{\theta};\boldsymbol{\gamma}) = \prod_{k=1}^{K} q_k(\theta_k;\boldsymbol{\gamma}_k). \tag{6.5}$$

For nice algebraic properties as well as generality, let $q_k(\theta_k;\boldsymbol{\gamma}_k)$ for each $k \in \{1, \ldots, K\}$ be a p.d.f. (or p.m.f.) from the exponential family. Here $\boldsymbol{\gamma}_k = \{\boldsymbol{\gamma}_{k,j}\}_{j=1:J_k}$ is a vector of the natural parameters of density $q_k(\theta_k;\boldsymbol{\gamma}_k)$. Within the exponential family, a natural choice of density kernel for $q_k(\theta_k;\boldsymbol{\gamma}_k)$ is the one that most likely captures the characteristics of the marginal posterior distribution $p(\theta_k|D, M)$. The goodness of $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ as an approximation to $p(\boldsymbol{\theta}|D, M)$ then depends on how much fluctuation the interaction between θ_k $(k = 1, \ldots, K)$ and $\{\theta_i\}_{i=1:K, i\neq k}$ causes in the mean fields. Now I discuss the optimization of the lower bound and the upper bound separately.

6.2.2 MCSA for Lower Bound Optimization

This lower bound optimization shares the same basic idea with the existing mean-field variational approximation method. Given a variational density function $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ as an approximation to the posterior density $p(\boldsymbol{\theta}|D,M)$, this lower bound maximization is equivalent to the minimization of the K-L divergence of $p(\boldsymbol{\theta}|D,M)$ from $q(\boldsymbol{\theta};\boldsymbol{\gamma})$, i.e. finding

$$\boldsymbol{\gamma}_L = \arg\min_{\boldsymbol{\gamma}} \{D_{KL}\left[q||p\right]\},$$

where

$$D_{KL}\left[q||p\right] = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta};\boldsymbol{\gamma}) \log \frac{q(\boldsymbol{\theta};\boldsymbol{\gamma})}{p(\boldsymbol{\theta}|D,M)} \mathrm{d}\boldsymbol{\theta},$$

a non-negative quantity. This is easy to show. Simply plugging the equality $p(\boldsymbol{\theta},D|M) = p(\boldsymbol{\theta}|D,M)p(D|M)$ into Jensen's inequality shown in (6.1) (with $q(\boldsymbol{\theta})$ substituted with $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ here) yields

$$\begin{aligned}
\log p(D|M) &\geq \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta};\boldsymbol{\gamma}) \log \frac{p(\boldsymbol{\theta}|D,M)p(D|M)}{q(\boldsymbol{\theta};\boldsymbol{\gamma})} \mathrm{d}\boldsymbol{\theta} \\
&= \log p(D|M) - \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta};\boldsymbol{\gamma}) \log \frac{q(\boldsymbol{\theta};\boldsymbol{\gamma})}{p(\boldsymbol{\theta}|D,M)} \mathrm{d}\boldsymbol{\theta} \\
&= \log p(D|M) - D_{KL}\left[q||p\right] \\
&= L(\boldsymbol{\gamma}),
\end{aligned}$$

so that raising $L(\boldsymbol{\gamma})$ decreases $D_{KL}\left[q||p\right]$.

The existing mean-field variational method, as introduced in Section 6.1, uses a variational density form factorized over hidden variables and model parameters (or constructs such settings by treating certain model parameters as hidden variables), and depends on the EM algorithm providing solutions to the lower bound optimization. The Monte Carlo EM (MCEM) algorithm has been proposed (in the context

of maximum likelihood estimation involving missing data) to deal with the cases that the expectation step in the EM algorithm has no analytic solutions (Celeux and Diebolt 1992; Chan and Ledolter 1995). In this algorithm, Monte Carlo method is applied to estimate the expectation. By combining with a stochastic approximation process, the convergence of this stochastic version of EM was established under mild conditions (Delyon *et al.* 1999). This work is the inspiration for a stochastic approximation version of the variational Bayesian method proposed here. The Monte Carlo method is used to calculate the expectations of the solutions, and optimize the lower bound iteratively via a stochastic approximation process. The resulting algorithm is easy to implement, and its convergence can also be guaranteed under conditions that are applicable to many practical situations.

Refer to (6.4). As a function of the parameter vector $\boldsymbol{\gamma}$, the lower bound of log marginal likelihood is written as

$$L(\boldsymbol{\gamma}) = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}; \boldsymbol{\gamma}) \log \frac{p(\boldsymbol{\theta}, D|M)}{q(\boldsymbol{\theta}; \boldsymbol{\gamma})} \mathrm{d}\boldsymbol{\theta}$$

with $q(\boldsymbol{\theta}; \boldsymbol{\gamma})$ taking a factorized exponential family form presented in (6.5). Let the first order derivative of $L(\boldsymbol{\gamma})$ with respect to $\gamma_{k,j}$ equal zero for each $k \in \{1, \ldots, K\}$

and $j \in \{1, \ldots, J_k\}$, namely,

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\gamma})}{\partial \gamma_{k,j}} &= -\frac{\partial}{\partial \gamma_{k,j}} \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}; \boldsymbol{\gamma}) \log \frac{q(\boldsymbol{\theta}; \boldsymbol{\gamma})}{p(\boldsymbol{\theta}, D|M)} \mathrm{d}\boldsymbol{\theta} \\
&= -\int_{\boldsymbol{\Theta}} \left[\frac{\partial}{\partial \gamma_{k,j}} \left(\log \frac{q(\boldsymbol{\theta}; \boldsymbol{\gamma})}{p(\boldsymbol{\theta}, D|M)} \right) q(\boldsymbol{\theta}; \boldsymbol{\gamma}) + \log \frac{q(\boldsymbol{\theta}; \boldsymbol{\gamma})}{p(\boldsymbol{\theta}, D|M)} \frac{\partial q(\boldsymbol{\theta}; \boldsymbol{\gamma})}{\partial \gamma_{k,j}} \right] \mathrm{d}\boldsymbol{\theta} \\
&= -\int_{\boldsymbol{\Theta}} \left[\frac{\partial \log q(\boldsymbol{\theta}; \boldsymbol{\gamma})}{\partial \gamma_{k,j}} q(\boldsymbol{\theta}; \boldsymbol{\gamma}) + \log \frac{q(\boldsymbol{\theta}; \boldsymbol{\gamma})}{p(\boldsymbol{\theta}, D|M)} \frac{\partial \log q(\boldsymbol{\theta}; \boldsymbol{\gamma})}{\partial \gamma_{k,j}} q(\boldsymbol{\theta}; \boldsymbol{\gamma}) \right] \mathrm{d}\boldsymbol{\theta} \\
&= -\int_{\boldsymbol{\Theta}} \left[1 + \log \frac{q(\boldsymbol{\theta}; \boldsymbol{\gamma})}{p(\boldsymbol{\theta}, D|M)} \right] \frac{\partial \log q_k(\theta_k; \boldsymbol{\gamma}_k)}{\partial \gamma_{k,j}} q(\boldsymbol{\theta}; \boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\theta} \qquad (6.6) \\
&= 0.
\end{aligned}$$

Then the solution of this system of $\sum_{k=1}^{K} J_k$ equations is $\boldsymbol{\gamma}_L$, which maximizes the lower bound of the log marginal likelihood. Here, stochastic approximation is used to solve this system of equations.

Stochastic approximation (SA) (Kushner and Yin 2003) is a class of algorithms for finding the roots of possibly non-linear equation $f(x) = 0$, in the situation where only noisy measurements of $f(x)$ are available. The Robbins-Monro algorithm (Robbins and Monro 1951), the simplest form of SA, is a recursive process

$$x^{(t+1)} = x^{(t)} + s^{(t+1)} \zeta^{(t+1)}$$

with some initial $x^{(0)}$. Here $\left\{ s^{(t)}, t \geq 1 \right\}$ is a sequence of step sizes that satisfies standard conditions: $\sum_{t=1}^{\infty} s^{(t)} = \infty$ and $\sum_{t=1}^{\infty} \left[s^{(t)} \right]^2 < \infty$. For any $t \geq 1$, $\zeta^{(t)}$ is a noisy measurement of $f(x)$, i.e.

$$\zeta^{(t)} = f(x) + \xi^{(t)},$$

where $\left\{ \xi^{(t)}, t \geq 1 \right\}$ is the so called noise sequence.

In my case, x is $\gamma_{k,j}$ and the function $f(\gamma_{k,j})$ has an integral form as shown in

(6.6). Assume Monte Carlo samples $\{\boldsymbol{\theta}^{(i)}\}_{i=1:N}$ from distribution $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ are available (these samples are easy to generate when $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ is a factorized exponential-family density). For each $k \in \{1,\ldots,K\}$ and $j \in \{1,\ldots,J_k\}$, $f(\gamma_{k,j})$ can be evaluated by its Monte Carlo estimate, namely,

$$\zeta\left(\gamma_{k,j}\right) = -\frac{1}{N}\sum_{i=1}^{N}\left\{\left[1+\log\frac{q(\boldsymbol{\theta}^{(i)};\boldsymbol{\gamma})}{p(\boldsymbol{\theta}^{(i)},D|M)}\right]\frac{\partial \log q_k(\theta_k^{(i)};\boldsymbol{\gamma}_k)}{\partial\gamma_{k,j}}\right\},$$

By the central limit theorem

$$\xi\left(\gamma_{k,j}\right) = \left[\zeta\left(\gamma_{k,j}\right) - f\left(\gamma_{k,j}\right)\right] \rightarrow \mathrm{N}\left(0,\frac{\sigma_f^2}{N}\right),$$

which means $\xi\left(\gamma_{k,j}\right)$ is Gaussian noise.

By using the Robbins-Monro algorithm, $\boldsymbol{\gamma} = \{\gamma_{k,j}\}_{k=1:K,j=1:J_k}$ can be estimated iteratively via

$$\gamma_{k,j}^{(t+1)} = \gamma_{k,j}^{(t)} + s^{(t+1)}\zeta\left(\gamma_{k,j}^{(t)}\right).$$

Then, using the estimate $\hat{\boldsymbol{\gamma}}_L$ produced through the above iterative procedure and the Monte Carlo samples $\{\boldsymbol{\theta}^{(i)}\}_{i=1:N}$ from $q(\boldsymbol{\theta};\hat{\boldsymbol{\gamma}}_L)$, one obtains the estimate of the optimal lower bound conditional on the kernel form of the variational density function,

$$\hat{L}_o = \frac{1}{N}\sum_{i=1}^{N}\log\frac{p(\boldsymbol{\theta}^{(i)},D|M)}{q(\boldsymbol{\theta}^{(i)};\hat{\boldsymbol{\gamma}}_L)}. \tag{6.7}$$

When the iterative steps in the stochastic approximation go to infinity, this estimated lower bound converges to the true maximum lower bound L_o with probability one. The proof of this conclusion is presented in Shen *et al.* (2007).

6.2.3 MCMC for Upper Bound Optimization

The upper bound optimization tunes the variational density parameters $\boldsymbol{\gamma}$ so that $U(\boldsymbol{\gamma})$ reaches minimum, i.e. finding

$$\boldsymbol{\gamma}_U = \arg\min_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \{U(\boldsymbol{\gamma})\}.$$

Minimizing $U(\boldsymbol{\gamma})$ is equivalent to minimizing $U(\boldsymbol{\gamma}) - \log p(D|M)$. Refering to (6.2) and (6.3), one has

$$\begin{aligned} U(\boldsymbol{\gamma}) - \log p(D|M) &= \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log p(\boldsymbol{\theta}|D,M)\mathrm{d}\boldsymbol{\theta} - \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log q(\boldsymbol{\theta};\boldsymbol{\gamma})\mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\log\frac{p(\boldsymbol{\theta}|D,M)}{q(\boldsymbol{\theta};\boldsymbol{\gamma})}\mathrm{d}\boldsymbol{\theta} \\ &= D_{KL}\left[p||q\right], \end{aligned}$$

the K-L divergence of variational density $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ from posterior density $p(\boldsymbol{\theta}|D,M)$. Hence, minimizing the upper bound $U(\boldsymbol{\gamma})$ is equivalent to minimizing this K-L divergence with respect to $\boldsymbol{\gamma}$, i.e. to find

$$\boldsymbol{\gamma}_U = \arg\min_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \{D_{KL}\left[p||q\right]\}.$$

Since $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ comes from the exponential family, which is well known as being log-concave, $D_{KL}\left[p||q\right]$ must be convex with respect to $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_k\}_{k=1:K}$. Consequently, the global minimum can be found by setting the partial derivatives equal to zero. For all $k \in \{1,\ldots,K\}$ and $j = 1,\ldots,J_k$, let

$$\begin{aligned} \frac{\partial}{\partial\gamma_{k,j}} D_{KL}\left[p||q\right] &= -\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|D,M)\left[\frac{\partial}{\partial\gamma_{k,j}}\log q_k(\theta_k;\boldsymbol{\gamma}_k)\right]\mathrm{d}\boldsymbol{\theta} \\ &= 0. \end{aligned}$$

For $q_k(\theta_k;\boldsymbol{\gamma}_k)$ from exponential family, these equations can be solved either ana-

lytically (when $q_k(\theta_k;\boldsymbol{\gamma}_k)$ is Bernoulli, Binomial or Gaussian) or through numerical methods when the Monte Carlo samples of $p(\boldsymbol{\theta}|D,M)$ are available. By the Strong Law of Large Numbers, the estimated solution $\hat{\boldsymbol{\gamma}}_k$ converges almost surely to the true solution if $\hat{\gamma}_{k,j}$ can be analytically expressed. Then the solutions of all these equations together form $\hat{\boldsymbol{\gamma}}_U$, which converges almost surely to $\boldsymbol{\gamma}_U$ that truly minimizes $D_{KL}[p||q]$. Given the MCMC samples of $p(\boldsymbol{\theta}|D,M)$, $\{\boldsymbol{\theta}^{(i)}\}_{i=1:N}$, the upper bound U_o can be estimated by

$$\hat{U}_o = \frac{1}{N}\sum_{i=1}^{N}\log\frac{p(\boldsymbol{\theta}^{(i)},D|M)}{q(\boldsymbol{\theta}^{(i)};\hat{\boldsymbol{\gamma}}_U)}.$$

$\hat{U}_o$ as consequence converges almost surely to the true globe minimum upper bound of the log marginal likelihood under the conditions of the variational density form.

6.3 Application to PROPA Models

Here I describe how to approximate the log marginal likelihoods in PROPA models through lower bound and upper bound optimization using MCVA outlined in the previous section. This marginal likelihood evaluation solves the central problem in PROPA and is a good example for the applications of this variational method as well.

6.3.1 Integrand and Parameters

In computation of marginal likelihood, the general principle is to integrate out as many parameters as possible to reduce the complexity and increase the accuracy of the last inevitable numerical integration step. The log marginal likelihood has been presented as a reduced form in (5.1). It is an integral over the parameters α_0 and α_1. Using the mean-field variational method to approximate this integral would involve approximating the joint posterior density of α_0 and α_1 with a variational density factorized with respect to these two parameters. According to the simulation in

Chapter 4, α_0 and α_1 are continuous variables with truncated gamma-like posterior densities, increasing the complexity of the variational p.d.f. form and the difficulty in optimization. Meanwhile, the conditional posterior distributions of α_0 and α_1 (as shown in (4.2) and (4.3)) depend on the sum of $1 - z_g$ or z_g, $g \in \mathcal{G}$, suggesting strong posterior dependency between them. One would expect that the factorized variational density is a poor approximation to the joint posterior and leads to a poor approximation of the log marginal likelihood. Therefore, I take another path leading to a simple and probably more accurate solution of this log marginal likelihood approximation problem.

Go back to the joint density function in a form with augmentation variables shown in (4.1). Integrate out $\{\beta_g\}_{g\in\mathcal{G}}$, α_0 and α_1, and leave $\{z_g\}_{g\in\mathcal{G}}$ as integration variables. Equation (4.1) contains an independent gamma density kernel of α_0 and α_1, so it is trivial to get

$$p(\Pi, \boldsymbol{z}|A, \mathcal{F} = \mathcal{A}) = \prod_{g=1}^{p} \left[\left(\frac{r_A}{\pi_g}\right)^{z_g} \left(\frac{1-r_A}{1-\pi_g}\right)^{1-z_g} \right]^{I(g\in A)} \left[\left(\frac{r_B}{\pi_g}\right)^{z_g} \left(\frac{1-r_B}{1-\pi_g}\right)^{1-z_g} \right]^{I(g\notin A)}$$

$$\frac{\Gamma(\nu_1)\Gamma(\nu_0)}{\lambda_1^{\nu_1}\lambda_0^{\nu_0}} (1 - \Phi(1; \nu_0, \lambda_0))(1 - \Phi(1; \nu_1, \lambda_1))$$

with $\nu_1 = \sum_{g=1}^{p} z_g$, $\nu_0 = \sum_{g=1}^{p} (1 - z_g)$, $\lambda_1 = -\sum_{g=1}^{p} (z_g \log \pi_g)$, $\lambda_0 = -\sum_{g=1}^{p} (1 - z_g) \log(1 - \pi_g)$, and Φ the gamma cumulative density function. It is implied here that the supports of α_0 and α_1 are upper bounded by a, which is large enough to make $\Phi(a; \nu_0, \lambda_0) = 1$ and $\Phi(a; \nu_1, \lambda_1) = 1$. Then the marginal likelihood is

$$p(\Pi|A, \mathcal{F} = \mathcal{A}) = \sum_{\boldsymbol{z}\in\boldsymbol{Z}} p(\Pi, \boldsymbol{z}|A, \mathcal{F} = \mathcal{A}).$$

The variational p.m.f. is constructed in a factorized form as

$$q(\boldsymbol{z};\boldsymbol{\gamma}) = \prod_{k=1}^{p} q_{z_k}(z_k;\boldsymbol{\gamma_k}). \tag{6.8}$$

Since z_k for each $k \in \{1,\ldots,p\}$ is a binary random variable, the density kernel of each factor must be Bernoulli, i.e.

$$q_{z_k}(z_k;\gamma_k) = \gamma_k^{z_k}(1-\gamma_k)^{1-z_k} \tag{6.9}$$

with parameter $\gamma_k = Pr(z_k = 1)$.

The reason for choosing this parameterization is that the marginal posterior distributions of z_g, $g \in \mathcal{G}$, are more likely to be close to independence in high dimensional cases. Refer to Chapter 4. In each iteration of sampling in the posterior simulation, the distributions of α_0 and α_1 are generally determined by $\sum_{g\in\mathcal{G}}(1-z_g)$ and $\sum_{g\in\mathcal{G}} z_g$, which means the fluctuation of a small number of z_g's may not have a large influence on the conditional distributions of α_0 and α_1. It is even more likely that these fluctuating z_g's are cancelling out each other's effect on the distributions of α_0 and α_1. Hence, for any z_k ($k \in \mathcal{G}$), the fluctuation of z_g ($g \in \mathcal{G}, g \neq k$), is not likely to have an impact on z_k through α_0 and α_1. Parameters $\{z_g\}_{g\in\mathcal{G}}$ may be close to being independent in the posterior. So $q_{z_k}(z_k;\gamma_k)$ should be a good approximation to the joint posterior distribution of $\{z_g : g \in \mathcal{G}\}$.

6.3.2 Bound Optimization

Refer to equation (6.6). To optimize the lower bound of the log marginal likelihood here, one needs to solve a system of non-linear equations of p variables $\{\gamma_k\}_{k=1:p}$. For

all $k \in \{1:p\}$, the equation in an explicit form is

$$f(\gamma_k) = \sum_{\boldsymbol{z}\in\boldsymbol{Z}} \left[1 + \log \frac{\prod_{g=1}^{p} \gamma_g^{z_g}(1-\gamma_g)^{1-z_g}}{p(\Pi, \boldsymbol{z}|A, \mathcal{F}=\mathcal{A})}\right](z_k - \gamma_k) = 0.$$

The numerical method described in Section 6.2.2 is used to solve these equations. At each iterative step t, samples of $z_1, \ldots, z_p$ are drawn from the mass function presented in (6.9), which is parameterized by the $\gamma_1^{(t-1)}, \ldots, \gamma_p^{(t-1)}$.

Compared with the lower bound, the upper bound optimization benefits more from the new setting of the integral. I discuss it in detail as follows.

The goal here is to find the optimum parameters $\boldsymbol{\gamma} = \{\gamma_k\}_{k=1:p}$ minimizing the K-L divergence

$$D(\boldsymbol{\gamma}) = \sum_{\boldsymbol{z}\in\boldsymbol{Z}} p(\boldsymbol{z}|\Pi, A, \mathcal{F}=\mathcal{A}) \log\left(\frac{p(\boldsymbol{z}|\Pi, A, \mathcal{F}=\mathcal{A})}{q(\boldsymbol{z};\boldsymbol{\gamma})}\right)$$

with $q(\boldsymbol{z};\boldsymbol{\gamma})$ in the form of (6.8). This corresponds to finding the solution of a set of equations $\left\{\frac{\partial D(\boldsymbol{\gamma})}{\partial \gamma_k} = 0\right\}_{k=1:p}$. Explicitly, for all $k \in \{1:p\}$,

$$\begin{aligned}
\frac{\partial D(\boldsymbol{\gamma})}{\partial \gamma_k} &= \frac{\partial}{\partial \gamma_k}\left\{-\sum_{\boldsymbol{z}\in\boldsymbol{Z}} p(\boldsymbol{z}|\Pi, A, \mathcal{F}=\mathcal{A}) \log\left(\gamma_k^{z_k}(1-\gamma_k)^{1-z_k}\right)\right\} \\
&= -\frac{1}{\gamma_k(1-\gamma_k)} \sum_{\boldsymbol{z}\in\boldsymbol{Z}} p(\boldsymbol{z}|\Pi, A, \mathcal{F}=\mathcal{A})(z_k - \gamma_k) \\
&= -\frac{1}{\gamma_k(1-\gamma_k)} \left(E(z_k|\Pi, A, \mathcal{F}=\mathcal{A}) - \gamma_k\right) = 0 \qquad (6.10)
\end{aligned}$$

Clearly, in the domain $[0, 1]$ the solution is $\gamma_k = E(z_k|\Pi, A, \mathcal{F}=\mathcal{A})$, the posterior mean of z_k. This is to say, given a factorized form, the variational p.m.f. has the minimum K-L divergence from the posterior p.m.f. of $\boldsymbol{z}$ when each factor $q_{z_k}(z_k)$ is

the marginal posterior p.m.f. of z_k, i.e.

$$q_U(\boldsymbol{z}) = \prod_{k=1}^{p} p(z_k|\Pi, A, \mathcal{F} = \mathcal{A}).$$

The MCMC method as described in Chapter 4 provides the marginal posterior samples of each model parameter when the Markov chain reaches equilibrium. So the value of γ_k that optimizes the log marginal likelihood upper bound can simply be estimated by the mean of the MCMC samples $\{z_k^{(i)}\}_{i=1:N}$, namely,

$$\hat{\gamma}_k = \hat{E}(z_k|\Pi, A, \mathcal{F} = \mathcal{A}) = \frac{1}{N}\sum_{i=1}^{N} z_k^{(i)}.$$

Then the MCMC samples $\{\boldsymbol{z}^{(i)}\}_{i=:N}$ from the joint posterior distribution $p(\boldsymbol{z}|\Pi, A, \mathcal{F} = \mathcal{A})$ are used to estimate the optimal upper bound of log marginal likelihood by

$$\hat{U}_o = \frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{p(\Pi, \boldsymbol{z}^{(i)}|A, \mathcal{F} = \mathcal{A})}{\hat{q}_U(\boldsymbol{z}^{(i)})}\right],$$

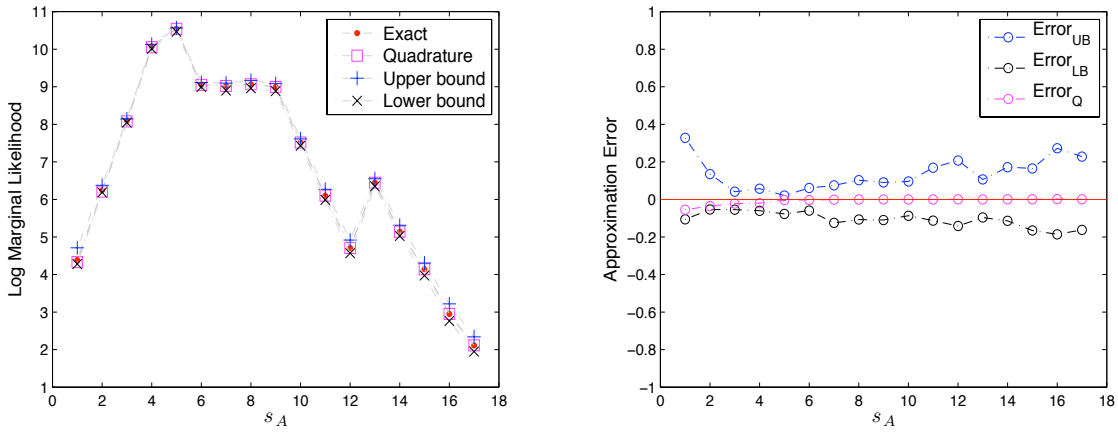
where $\hat{q}_U(\boldsymbol{z}) = \prod_{k=1}^{p} \hat{\gamma}_k^{z_k}(1-\hat{\gamma}_k)^{1-z_k}$.

6.4 Simulation Studies

To demonstrate the performance of MCVA method, I simulate two data sets and compute the log marginal likelihoods of a few simulated gene sets in the study.

The first data set has been used in Section 4.2 and 5.3 to study posterior simulation and numerical integration. It contains 18 genes, from which 17 gene sets are constructed, and uses hyper-parameters $r_A = 0.8$ and $r_B = 0.1$. For each gene set, the exact value, quadrature estimate and MCVA optimal bounds of the log marginal

likelihood are computed and displayed in Figure 6.1(a). The approximation error, defined as an estimated value minus the true value for each gene set, is presented in Figure 6.1(b). Quadrature has the best performance. The error plots show that the variational optimal bounds are indeed the upper and lower bounds of the log marginal likelihoods. Although their approximation errors are larger than quadrature, these bounds are good enough to distinguish different gene sets/models in this example.



(a) (b)

Figure 6.1: MCVA simulation study on the 18-gene data set. (a) The exact values, optimal upper bounds and optimal lower bounds of log marginal likelihoods of 17 gene sets. (b) The approximation errors of upper bounds and lower bounds.

In the second simulated data set, $p = 100$, and eleven gene sets are produced precisely as in generating the 17 gene sets for the 18-gene data set, i.e. $A_{s_A} = \{1, \ldots, s_A\}$ for each $s_A \in \{15, \ldots, 25\}$. The data and gene sets are illustrated in Figure 6.2. Let the hyper-parameters $r_A = 0.9$ and $r_B = 0.05$. The optimal bounds of log marginal likelihoods by MCVA are computed for each gene set and shown in Figure 6.3(a). It is unrealistic to compute the exact values when $p = 100$. Quadrature integration, as shown in the previous example, can produce a good approximation, and it may still work when p is less than 10^3. Hence, the values of quadrature approximation are used for reference. Figure 6.3(b) presents the difference (called

error here) between each bound and the quadrature estimate as well as the distance between the upper and lower bounds (the upper bound minus the corresponding lower bound). The optimal upper bounds have smaller difference from the quadrature estimates than the optimal lower bounds. The bound distances show the correctness of the bound estimation. In this simulation, the upper and lower bounds given by MCVA are sufficient to distinguish the log marginal likelihoods of all the models, and the resulting PROPA identification of pathways is very accurate.

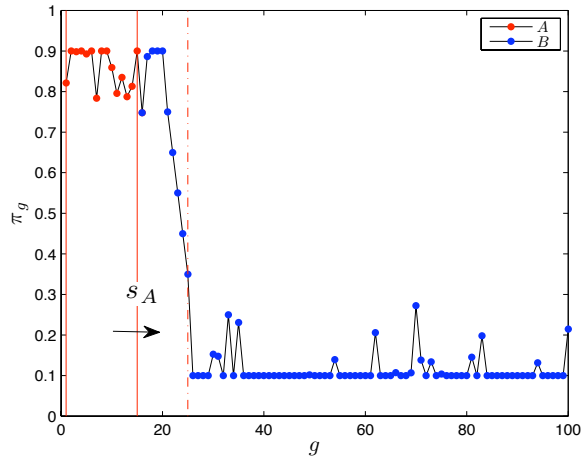
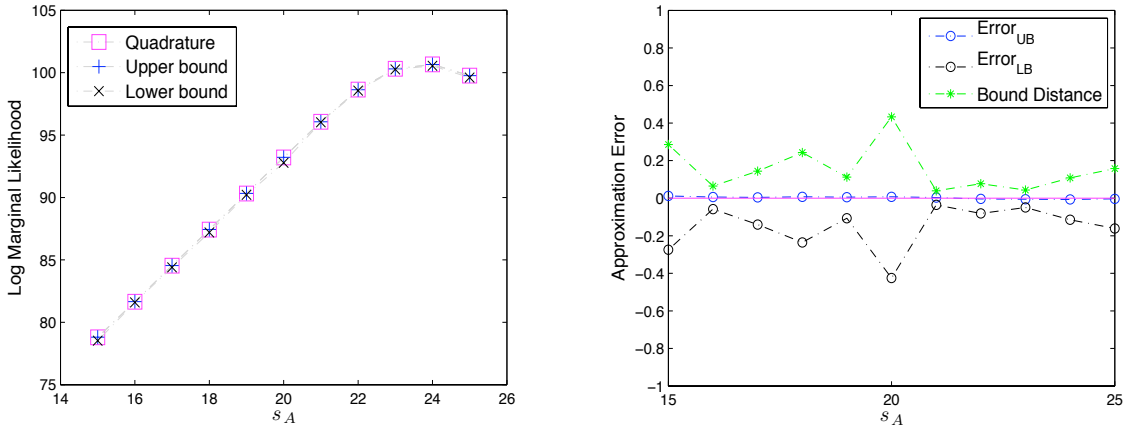


Figure 6.2: Association probabilities in the simulated data set with 100 genes. The red dots correspond to the genes in pathway gene set A, and the blue dots correspond to those not in A (i.e. in $\bar{A}$). A_{s_A} is the pathway gene set that includes the first s_A genes. By increasing s_A from 15 to 25, eleven gene sets are generated.

In these two simulation studies, both the lower and the upper bound optimization methods have good performance in terms of accuracy. This double-sided bounding provides sufficient information to facilitate model comparisons. The lower bound optimization is especially fast when the dimension of data is relatively low. This advantage of lower bound optimization is lost with the increase of dimension. When p is greater than 10^3, convergence becomes unacceptably slow. To solve the approximation problem in high-dimensional cases, I use a compromised strategy: lower bounding

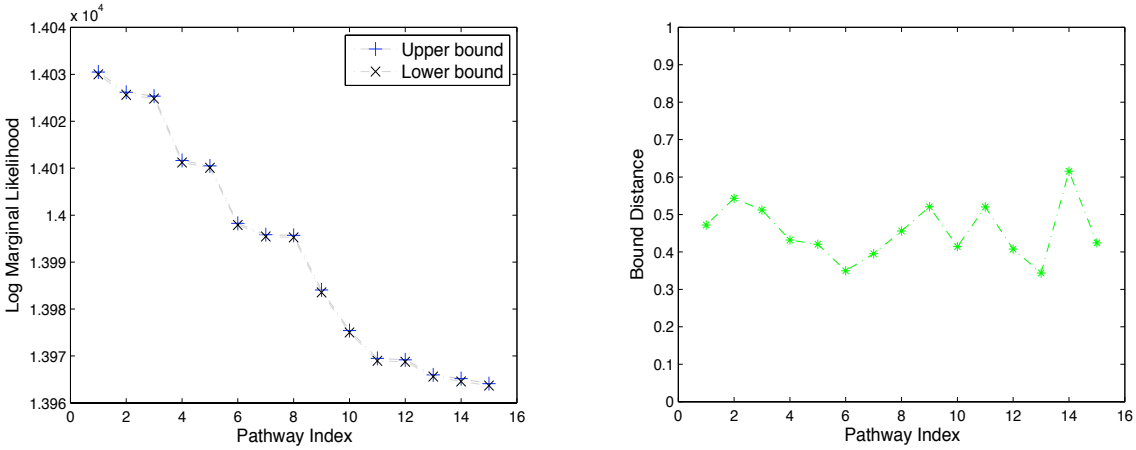


(a) (b)

Figure 6.3: MCVA simulation study on the 100-gene data set. (a) The quadrature estimates, optimal upper bounds and optimal lower bounds of log marginal likelihoods for eleven gene sets. (b) The distance between the upper bounds and lower bounds.

the log marginal likelihood with a pseudo-optimal lower bound that is obtained by using (6.7), where $q(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}_L) = q(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}_U)$. When the factorized variational density q is a good approximation of the joint posterior distribution of $\boldsymbol{\theta}$, the variational densities corresponding to optimal upper and lower bounding are likely to converge to the same function. Clearly, the value of the pseudo-optimal lower bound is always less than the lower bound obtained through MCVA lower bound optimization.

I demonstrate the approximation using this strategy with a real data set and 15 pathway gene sets from a database. The data set contains probabilities of association between 19,645 genes and the lactic acidosis (a cancer micro-enviormental factor) status of human mammary epithelial cell cultures (refer to Section 8.1). Figure 6.4(a) presents the optimal upper bounds and pseudo-optimal lower bounds of log marginal likelihoods for the 15 pathway gene sets. The distance between each pair of bounds are shown in Figure 6.4(b). As can be seen, the pseudo-optimal lower bounds are close to the optimal upper bounds, and such bounds are tight enough to discriminate the evidence for different models.



(a) (b)

Figure 6.4: MCVA study on the real data set with 19,645 genes. (a) The optimal upper bounds and pseudo-optimal lower bounds of log marginal likelihoods for 15 pathway gene sets. (b) The distance between the upper and lower bounds (upper bounds minus lower bounds).

6.5 Discussion

The proposed variational method provides both lower and upper bounds of log marginal likelihoods that are optimized under a certain variational density form. These two bounds not only facilitate more reliable model comparisons but also give a way to see the adequacy of the variational density as an approximation to the posterior density of model parameters. It is also worth noting that this variational method is generalizable in terms of the variational density form. It is not necessary to take a factorized form. This is important when the model parameters left in the integration are so dependent on each other that the factorized variational density is a poor approximation to the posterior density of these parameters.

Additionally, the co-existence of the upper and lower bounds can relax the requirement for optimization. Clearly, a single bound strongly relies on the optimization because its distance to the true value of log marginal likelihood itself is not bounded. The requirement of reducing this distance is imposed on the bound optimization al-

gorithms and makes marginal likelihood computation difficult when it involves a very large number of model parameters. In particular, the lower bound optimization that depends on EM or MCSA is more stressed by dimensionality, and convergence can be unbearably slow.

This stress can be reduced when the bounds on both sides are available. If the distances between these two bounds are small enough to distinguish different models, further optimizing them is not necessary. Furthermore, since the same variational density form $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ is used for both bounds (though this is not necessary otherwise), if one of the bounds is more easily optimized, the corresponding optimum variational density can be applied to compute the other bound, which may be good enough for the purpose of model comparisons even though not optimized. This is due to the fact that the tightness of the bounds is essentially determined by how good an approximation $q(\boldsymbol{\theta};\boldsymbol{\gamma})$ can be to $p(\boldsymbol{\theta}|D,M)$. In such a context, the upper bound shows a particular advantage because of the better convergence property of the MCMC method-based optimization algorithm compared to MCSA.

Chapter 7

Breast Cancer ER and ErbB2 Pathway Annotation

The simulation studies in the previous chapters have demonstrated the effectiveness of PROPA from the both modeling and computational aspects. In this chapter, two further "proof-of-principle" examples are presented to illustrate the use of the PROPA in real biological contexts: pathway annotation for estrogen-receptor (ER) status and human epidermal growth factor receptor 2 (ErbB2) status in breast cancers. These two phenotypes of breast cancers have been intensively studied and are clinically relevant. Although the development mechanism of these two phenotypes of cancers are not fully uncovered, many conclusions have been confirmed and well accepted, thus making them useful to evaluate the performance of PROPA. I annotate the pathway activities under these two cancer phenotypes to test the effectiveness of PROPA. The pathway annotation results are compared to those by GSEA, the most widely used gene set pathway annotation method. When solving the pathway annotation problem using the Bayesian modeling approach, one concern is the sensitivity of analysis results to the choice of priors (hyper-parameters) while allowing the model to have the flexibility to incorporate *á priori* knowledge. Another concern is the robustness of the model to real data. In the ER example, the influences of model hyper-parameters and data are addressed. Besides pathway annotation, gene pathway membership inference is exemplified by the studies in both simulated and real circumstances based on the ErbB2 over-expression signature.

7.1 Pathway Annotation Format

The pathway annotation by PROPA is based on the comparison of multiple pathways typically from a pathway gene set database. The information sufficiency of the pathway database gives PROPA the strength to identify the pathways potentially associated with the investigated biological phenotype. I assembled 956 human molecular signature gene sets from the MSigDB C2 collection provided by Broad Institute (http://www.broad.mit.edu/gsea/msigdb/) and re-curated them based on the Entrez human gene database.

In these examples and the application in the following chapter, the absence of prior knowledge on phenotype-pathway association is assumed, and all the pathways are given equal prior probabilities. In the summary of a PROPA analysis result, the approximated log marginal likelihoods of all pathways are plotted in increasing order. Histograms of the log marginal likelihood upper bounds are presented as well. These graphics aim to give some ideas of how much the levels of the association between the "top" pathways and the phenotype are distinguished from the other pathways. Some quantities representing the evidence of associations are depicted in a summary table for a number of top pathways. In this table, the top pathways are listed according to their ranks based on log marginal likelihood variational upper bound. Besides the rank and name of each pathway, the following quantities are provided as reference:

- *Post. Pr*: The posterior probability of the pathway as shown in (3.1). The variational upper bound (refer to *logML (UB)* as follows) is used to compute this posterior probability.

- *Post. Pr (sub)*: The posterior probability of the pathway after removing the dominant pathway(s). In some analysis, some pathways have such high posterior probabilities that the other pathways all appear to have zero posterior

probability, though they might also be interesting pathways in terms of the association with the phenotype. The posterior probabilities of the pathways are recomputed without including the top dominant pathways.

- *BF*: Bayes factor of the pathway as shown in (3.14).
- *BF (sub)*: Bayes factor of the pathway after removing the dominant pathway(s).
- *Size*: The number of genes in the pathway gene set from database.
- *LogML (UB)*: The optimal upper bound of log marginal likelihood for the pathway obtained by using the variational method described in Section 6.3.2;
- *LogML (LB)*: The lower bound of log marginal likelihood for the pathway obtained by using the variational density function corresponding to the optimual upper bound (refer to Section 6.4);
- *UB-LB*: The distance between the upper bound and lower bound of log marginal likelihood for the pathway. When this distance is large, the approximation of the log marginal likelihood might be poor, and the reliability of association between this pathway and the phenotype is questionable.

7.2 Breast Tumor ER Pathway Annotation

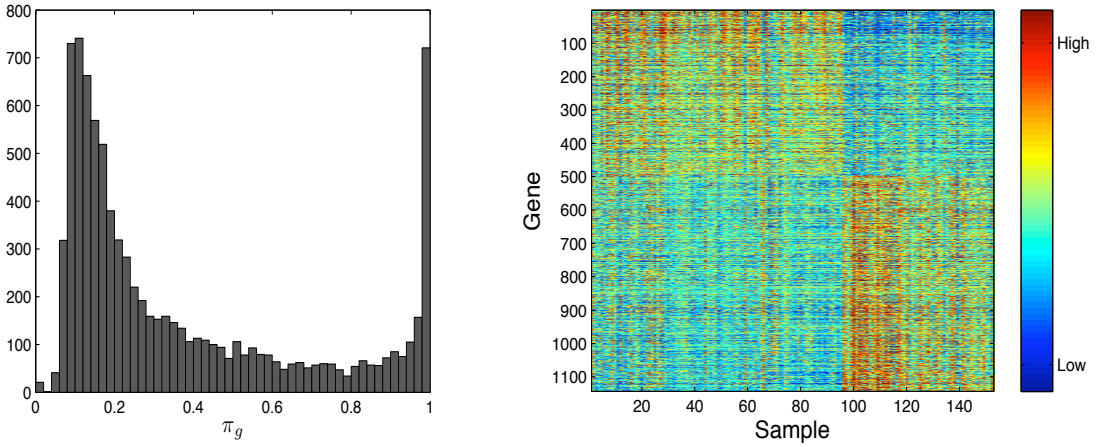
Estrogen is an important factor in the development and progression of breast carcinoma. Estrogen-receptor α (ERα), one of the two forms of estrogen-receptor, is the primary mediator of estrogenic actions in breast cancer. Upon estrogen binding, ERα is activated and becomes a transcription activator in mammary cells. Estrogen-bound ERα can bind to the estrogen response element of the target genes and activate the transcription, stimulating mammary cell proliferation through the downstream signaling pathways. It has been shown that over-expression of ERα and estrogen

binding increase the cell division and DNA synthesis, which elevate the risk for replication errors and induce breast tumorigenesis. About two-thirds of breast cancers show over-expression of ERα at the time of diagnosis. Both basic science and clinical data indicate the value of ERα level as an important predictor of breast cancer prognosis and disease-free survival. In general, ER-positive breast cancer is associated with more differentiated tumors and favorable prognosis (Deroo and Korach 2006; Moggs and Orphanieds 2001).

7.2.1 Data and Results Overview

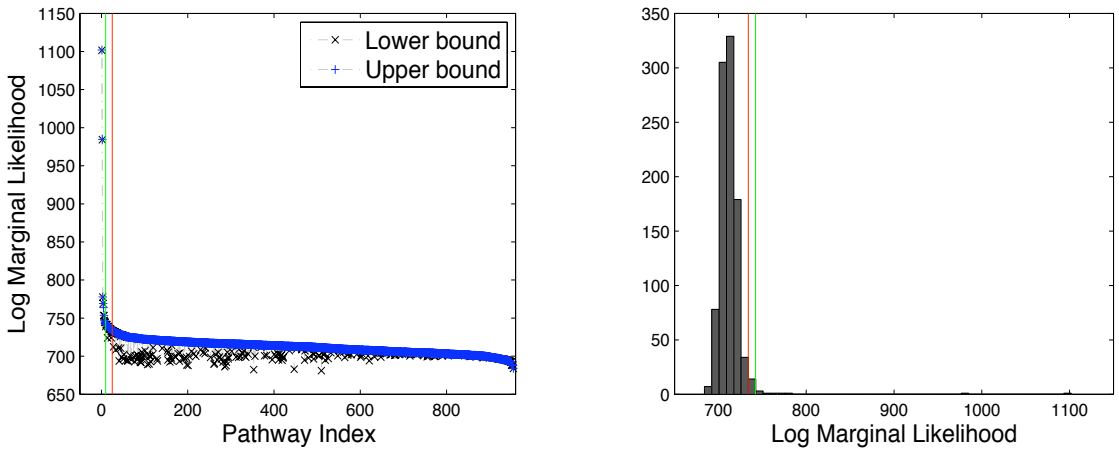
A data set including 153 primary breast tumor samples is used to investigate the pathway activities associated with ER status (Carvalho *et al.* 2007). The ER level of each tumor sample has been measured by immunohistochemical (IHC) staining. Overall, 95 tumors are ER-positive, and 58 tumors are ER-negative. For each sample, the cRNA has been derived and hybridized to Affymetrix Human genome U95Av2 microarray to yield genome-wide gene expression data. The gene expression data and ER status of these tumor samples are analyzed with a regression analysis using the BFRM software, which yields the posterior probability, as well as the sign of association (positive or negative) between the gene expression on each probe set and ER status. The 12,532 probe sets are then collapsed to 8,764 unique genes according to their corresponding Entrez gene IDs. The association probability histogram of the 8764 genes is presented in Figure 7.1(a). Figure 7.1(b) is the expression heatmap of 1,140 genes with highest association probabilities. These comprise the gene expression signature – the factor phenotype under study – of the ER status in this breast tumor set.

The optimal upper bound and non-optimal lower bound of log marginal likelihood are computed for each of the 956 pathways by using the variational method described



(a) (b)

Figure 7.1: Association probabilities and expression profiles of the genes correlated with the ER status of breast tumors. (a) Histogram of association probabilities. (b) Expression profiles of 1,140 genes whose association probabilities are greater than 0.9; the expression level change from low to high corresponds to the color change from blue to red.



(a) (b)

Figure 7.2: Log marginal likelihoods for 956 pathways in breast tumor ER status pathway annotation. (a) Pathway log marginal likelihood upper bounds (blue +) and lower bounds (black ×); pathways are sorted in a decreasing order of log marginal likelihood; pathways on the left side of the green and red lines are the top 10 and 25 pathways, respectively. (b) Histogram of the pathway log marginal likelihood upper bounds; bars on the right side of the green and red lines correspond to the top 10 and 25 pathways, respectively.

in Section 6.4. These bounds are plotted in Figure 7.2(a), where the pathways are arranged in a decreasing order of the upper bound. For some pathways, the distance between the upper bound and the lower bound is too large to give a reliable approximation for the log marginal likelihood, i.e. the mathematical quantities do not provide sufficient information for us to judge whether or not these pathways are associated with the phenotype compared to the others. However, for most of the top 25 pathways, especially those of interest, the bound distances are small enough for good estimation of their true log marginal likelihoods. Both Figure 7.2(a) and (b) show obvious drops of log marginal likelihoods within the top $\sim$20 pathways, suggesting that the top $\sim$20 pathways are likely to have significant association with ER status compared to the rest. The first several pathways have larger log marginal likelihoods and appear to be dominant in this analysis.

7.2.2 Significant Pathways

The top 25 pathways associated with beast tumor ER status are summarized in Table 7.1, and the association probability histograms of a subset of these gene sets are presented in Figure 7.3. In each of the plots, the red bars are the association probability histogram of the genes in the gene set positively correlated with the ER status, while the blue bars are the association probability histogram of the genes negatively correlated with ER status.

The first two gene sets are breast tumor ER negative and positive signatures defined by van't Veer *et al.* (2002) through the DNA microarray analysis of a set of primary breast tumors. The association probability histograms of this two gene sets are shown in Figure 7.3(a) and (b). As can be seen, almost all the genes in A_1, especially those with high probabilities, are negatively associated with ER status; almost all the genes in A_2, especially those with high probabilities, are positively associated

with ER status. This observation confirms the reliability of the annotation provided by PROPA. Besides the ER signatures, PROPA has also identified some other pathway signatures whose linkage to breast tumor ER status have been confirmed by previous research.

- Breast cancer prognosis signatures

  van't Veer *et al.* (2002) defined a breast cancer prognosis signature corresponding to four gene sets in the database. All of them show up in the top list, A_6, A_{12}, A_{18} and A_{21}. A_6 and A_{12} represent the breast cancer prognosis negative signature, containing genes whose expression are negatively correlated with breast cancer outcome. These two gene sets have different number of genes due to different curation. A_{18} and A_{21} represent the breast cancer prognosis positive signature, containing genes whose expression are positively correlated with breast cancer outcome. Clinical research has concluded that patients with ER-negative tumors generally have worse prognosis than those with ER-positive tumors (Maynard *et al.* 1978). As shown in Figure 7.3(c), (e), (h) and (i), this correlation between these prognosis signature and tumor ER-status are clearly reflected in the data set and identified by PROPA.

- Undifferentiated cancer signature

  The level of cell differentiation, technically quantified as histological grade, is an important measure in cancer evaluation. Undifferentiated (high-grade) cancers, irrespective of any tissue types, often maintain rapid and abnormal cellular proliferation and invasion, hence are associated with poor prognosis. The undifferentiated cancer transcriptional signature was defined by Rhodes *et al.* (2004) and includes the genes higher expressed in the cells of undifferentiated cancers compared to well-differentiated cancers of different tissues. Compared with ER-positive breast cancers, ER-negative cancers are more likely to be poorly

differentiated, and consequently appear to be more aggressive and associated with poor patient outcomes (Maynard *et al.* 1978; Pichon *et al.* 1996). This provides the rationale to the finding of the undifferentiated cancer signature (A_9) in the pathway annotation analysis of breast cancer ER phenotype here. Figure 7.3(d) shows that almost all the signature genes are negatively correlated with the tumor ER status, i.e. upregulated in the ER-negative tumors in this data set, supporting this relationship between the breast tumor ER phenotype and the cancer differentiation phenotype.

- Myb pathway signature

  The Myb pathway gene set A_{15} includes the genes regulated by Myb (A-Myb or c-Myb) transcription factors in MCF-7 mammary cells, primary lung epithelial cells or primary lung fibroblasts. Lei *et al.* (2004) summarized these genes and showed that the Myb-regulated gene sets in the three different cell types are nearly non-overlapping; less than half of the genes in the gene set are Myb-regulated in MCF-7 mammary cells. Importantly, previous work has shown that estrogen-induced MCF-7 mammary cell proliferation involves the expression of A-Myb, B-Myb and C-Myb proteins (Hodges *et al.* 2003), providing the evidence of association between breast tumor ER status and Myb pathway activation. As shown in Figure 7.3(g), less than half of the Myb pathway signature genes have relatively high probabilities of either positive or negative correlation with ER status, which is concordant with the summary of the gene set given by Lei *et al.* (2004). Further verification of this annotation would requires identifying of the Myb-regulated genes in MCF-7 cells from A_6 and observing the distribution of their ER-association probabilities.

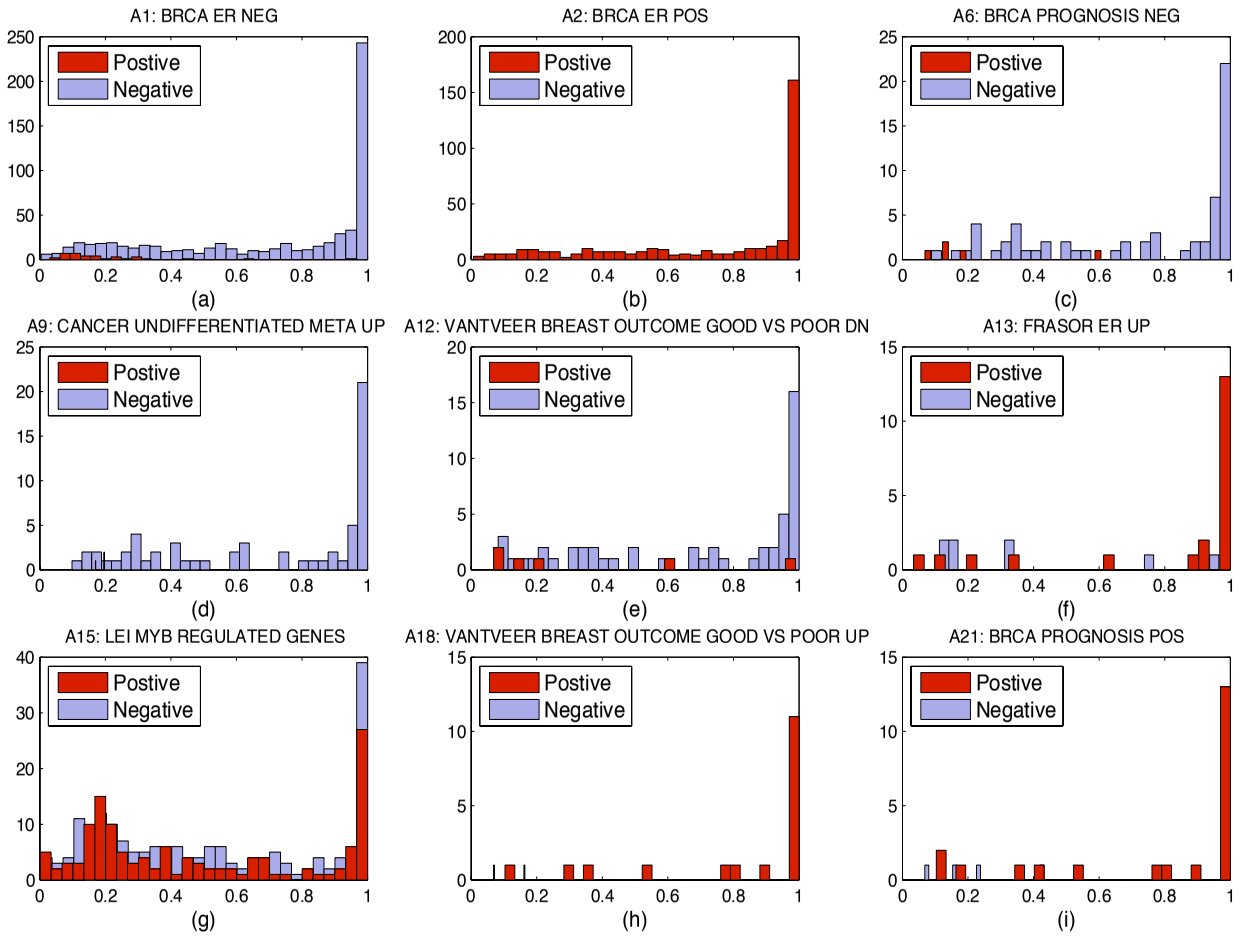


Figure 7.3: Association probability histograms of the ER-related pathway gene sets identified by PROPA. For each gene set, genes higher expressed in ER-positive tumors (positively correlated with tumor ER status) correspond to red bars; genes higher expressed in ER-negative tumors (negatively correlated with tumor ER status) correspond to blue bars.

Table 7.1: Summary of the top 25 ER-related pathways identified by PROPA

Rank	Pathway	Post. Pr	Post.Pr (sub)	BF	BF (sub)	Size	LogML (UB)	LogML (LB)	UB-LB	GSEA			
										NES Rank	p-value	FDR (q-value)	Corr
1	BRCA ER NEG	1.00E+00		Inf		692	1101.9	1101.47	0.44	1	0	0.02	-
2	BRCA ER POS	1.07E-51		0		380	984.54	984.19	0.35	1	0	0.02	+
3	LEE TCELLS2 UP	1.61E-141		0		712	777.72	777.66	0.06	47	0.008	0.21	-
4	FLECHNER KIDNEY TRANSPLANT REJECTION UP	3.32E-145		0		81	769.23	769.11	0.11	40	0.035	0.21	-
5	CARIES PULP UP	1.44E-147		0		186	763.79	753.12	10.67	67	0.045	0.21	-
6	BRCA PROGNOSIS NEG	5.73E-152		0		69	753.65	753.4	0.26	4	0	0.21	-
7	SERUM FIBROBLAST CELLCYCLE	8.57E-155		0		83	747.15	746.94	0.21	22	0.036	0.23	-
8	TARTE PLASMA BLASTIC	8.67E-156		0		295	744.86	744.71	0.15	29	0.019	0.21	-
9	CANCER UNDIFFERENTIATED META UP	5.29E-156		0		65	744.36	744.15	0.22	26	0.027	0.22	-
10	LI FETAL VS WT KIDNEY DN	6.90E-157		0		157	742.33	739.01	3.31	121	0.095	0.28	-
11	CARIES PULP HIGH UP	4.47E-157	3.59E-01	0	529	83	741.89	741.45	0.45	136	0.095	0.3	-
12	VANTVEER BREAST OUTCOME GOOD VS POOR DN	3.67E-157	2.95E-01	0	394.59	58	741.7	741.48	0.21	3	0.002	0.15	-
13	FRASOR ER UP	2.64E-157	2.12E-01	0	253.83	29	741.37	741.34	0.03	6	0	0.23	+
14	CIS XPC UP	1.12E-157	9.01E-02	0	93.43	131	740.51	723.93	16.58	144	0.315	0.81	+
15	LEI MYB REGULATED GENES	2.57E-158	2.06E-02	0	19.86	302	739.04	738.96	0.08	438	0.512	0.67	-
16	RUTELLA HEMATOGFSNDCS DIFF	2.01E-158	1.62E-02	0	15.51	505	738.79	738.78	0.01	192	0.119	0.34	-
17	UVB NHEK3 C7	2.57E-159	2.06E-03	0	1.95	50	736.74	736.67	0.07	46	0.012	0.21	-
18	VANTVEER BREAST OUTCOME GOOD VS POOR UP	1.92E-159	1.54E-03	0	1.46	20	736.44	736.37	0.07	3	0	0.05	+
19	GREENBAUM E2A UP	1.14E-159	9.19E-04	0	0.87	25	735.93	735.84	0.09	97	0.089	0.25	-
20	ZHAN MM CD138 PR VS REST	1.07E-159	8.59E-04	0	0.81	26	735.86	735.79	0.07	31	0.014	0.22	-
21	BRCA PROGNOSIS POS	7.32E-160	5.88E-04	0	0.56	26	735.48	735.45	0.03	2	0	0.64	+
22	MIDDLEAGE DN	5.25E-160	4.22E-04	0	0.4	13	735.15	735.13	0.02	50	0.019	0.2	-
23	IRITANI ADPROX LYMPH	4.78E-160	3.84E-04	0	0.36	121	735.05	724.14	10.92	132	0.025	0.29	-
24	LEE TCELLS3 UP	1.96E-160	1.58E-04	0	0.15	63	734.16	733.98	0.18	15	0.014	0.27	-
25	KLEIN PEL DN	1.68E-160	1.35E-04	0	0.13	57	734.01	733.6	0.41	43	0.03	0.23	-

7.2.3 Comparison with GSEA

This data set is also analyzed with GSEA. Totally, 29 gene sets are significantly enriched at nominal p-value less than 0.01, including all the interesting ones described above except the undifferentiated cancer signature and the Myb pathway signature. Table 7.1 exibits the NES rank, nominal p-value, FDR q-value and correlation status with ER reported by GSEA for each of the top 25 PROPA gene sets. In general, those gene sets clearly relevant to the breast tumor ER status are correctly identified by both PROPA and GSEA. Some of the other top PROPA gene sets are also identified by GSEA with relatively low false discovery rate (FDR q-value < 0.25). In conclusion, the pathway annotation results provided by PROPA and GSEA are generally consistent, showing similar performance of these two methods for the analysis of this data set in terms of detecting relevant pathways.

The gene sets, which are found in the top list given by PROPA, but not by GSEA due to the insignificance according to nominal p-value, FDR q-value and NES rank, tend to be composed of both positively and negatively associated genes. For example, A_{14}, A_{15} and A_{16} are non-significant according to GSEA. The association probabilities of genes in A_{15} has been shown in the histogram in Figure 7.3(g), and the histograms of A_{14} and A_{16} are in Figure 7.4(a) and (b).

Although they contain many genes highly associated with ER status, these gene sets are considered not enriched by GSEA; This is because GSEA performs one-way tests. A gene set appears to be significant only when a majority of genes are correlated with the phenotype in the same way. Therefore, GSEA only works properly on gene sets curated as upregulation and downregulation sets exclusively. PROPA does not currently use the sign information on the association between each gene and the phenotype. Instead, it leaves this information to the posterior investigation of the top gene sets. This is why PROPA detects the gene set representing Myb pathway,

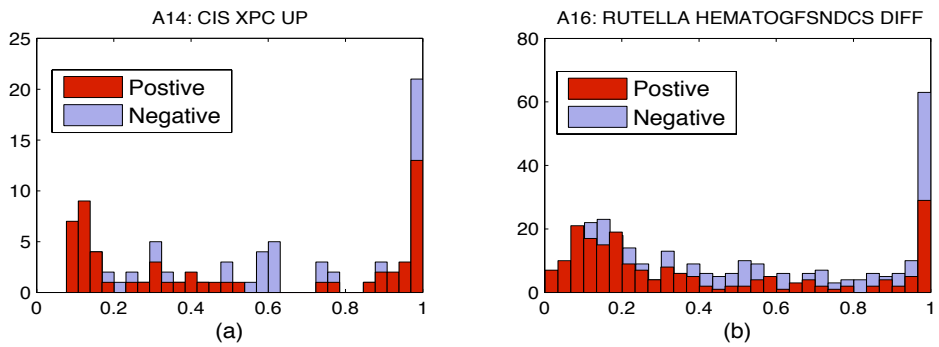


Figure 7.4: Association probability histograms of two gene sets highly ranked by PROPA, but not by GSEA. These gene sets contain both positive and negative genes in terms of correlation with ER status. GSEA cannot identify such gene sets, because it ranks gene sets by one-way NES.

which is critically biologically related to breast tumor ER regulation. In such cases, those gene sets exclusively containing upregulated or downregulated genes in certain contexts (for example, A_{14}) may not be of biological interest in terms of the pathway name they represent.

The ability of PROPA to identify gene sets comprised of both upregulated and downregulated genes offers advantage when gene sets are complicated and the expression regulation direction information is not available.

7.2.4 Influence of Hyper-parameters and Data

PROPA hyper-parameters r_A and r_B represent the prior means of the true positive rate and false negative rate of gene pathway membership specified by the reference gene set A. The specification of r_A and r_B depends on *á priori* knowledge on how precisely a gene set represents the theoretical pathway in the context being studied. These two parameters are involved in the evaluation of marginal likelihoods for the pathways in comparison. In a Bayesian analysis, ideally, the priors should have the flexibility to incorporate existing knowledge or justified beliefs. Meanwhile, moderate change of their values should not have dominant effects on the model inference. The

values of r_A and r_B that have been used in the above analysis are 0.7 and 0.005. To demonstrate the influence of these two hyper-parameters, I replace their values with 0.9 and 0.01 and observe the change of the pathway annotation results.

Association probabilities, the data in PROPA, come from the gene expression analysis using BFRM (refer to chapter 2 and 3). The sparsity modeling that aims to minimize the false positive rate in biomarker identification pushes the association probabilities toward 0 or 1. This forms the basis of the beta mixture modeling of the association probability for each gene in PROPA. Without *á priori* knowledge of pathway activities under the phenotype, PROPA ranks the pathways according to the data marginal likelihood given the allocation of genes specified by the reference gene set. Here the robustness of PROPA is tested on the simulated data generated by manipulating the true association probabilities in the ER example. The simulated association probability of gene g is generated by $\pi'_g = F^{-1}(R(\pi_g)/(p+1))$, where p is the number of genes, R is the ranking function, and F is a polynomial cumulative density function, explicitly, $F(x) = \frac{1}{6}x^3 - \frac{1}{2}x^2 + \frac{3}{2}x$. In this manipulation, the association ranks of all the genes are unchanged, while the distribution of association probability is much more diffused. The histogram of $\{\pi'_g\}_{g\in\mathcal{G}}$ is shown in Figure 7.5. Then the pathways related to ER status are analyzed with PROPA based on these simulated association probabilities.

I focus on the top 30 pathways in each analysis. The level of log marginal likelihood generally is not changed when hyper-parameters change, but drops largely in the analysis of the simulated data, meaning that the manipulated data poorly fit the PROPA model in general. In spite of the levels of log marginal likelihoods, these two analyses show considerable consensus with the original one in terms of the identified top pathway gene sets. In the top 30 lists, three analyses have fifteen gene sets in common, including nine ER related significant gene sets that have been discussed

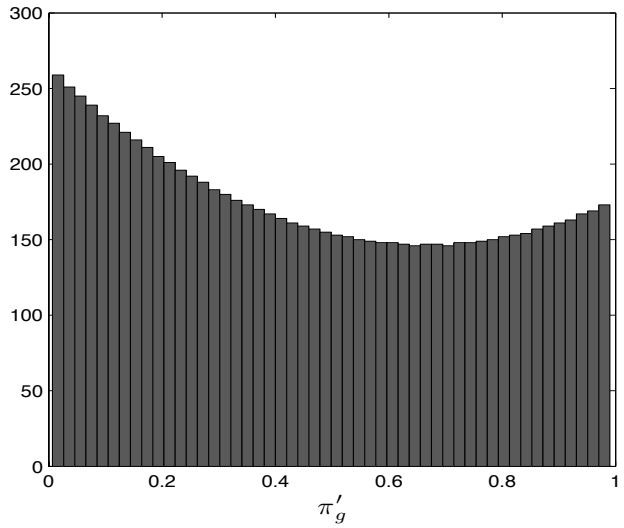


Figure 7.5: Histogram of the simulated association probabilities in the ER example.

in Section 7.2.2 and shown in Figure 7.3. The ER negative and positive signatures are still the most significant ones in all analyses. This consensus of results indicates the robustness of PROPA to the choice of hyper-parameters and data distribution. Particularly, the analysis of the simulated data implies that the gene-phenotype *association ranks* have a driving effect on the annotation results.

Further observation on the sizes of the top gene sets may help us in understanding the roles of the hyper-parameters and data in the analyses. Figure 7.6 summarizes the sizes of the top 30 gene sets in each of the three analyses. As illustrated, in this example, increasing r_A and r_B results in more findings of smaller pathway gene sets, while diffused data tend to give higher probabilities to large pathway gene sets. In the original data distribution in Figure 7.1, one can see that a relatively large number of genes are involved in the breast tumor ER phenotype. Increasing the value of r_A imposes tighter constrains on a gene set in terms of its accuracy as a representative of the corresponding pathway. Meanwhile, a larger value of r_B allows for adding more genes in the superficial pathway signatures in the currently exam-

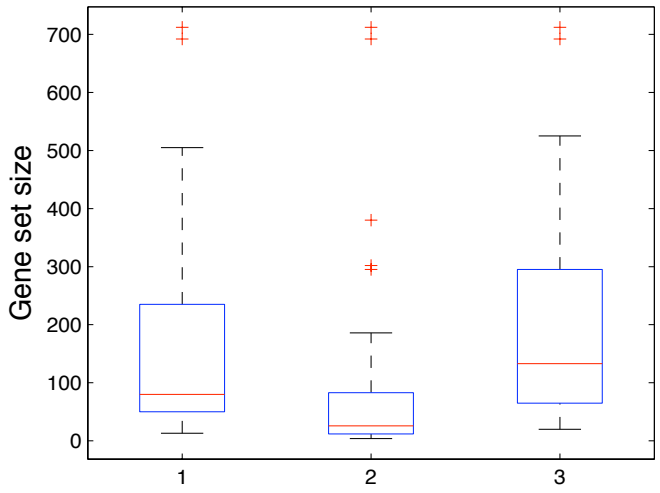


Figure 7.6: Box plots of the sizes of top 30 ER related pathway gene sets identified by PROPA in three analyses. 1 is the analysis of the original data with $r_A = 0.7$ and $r_B = 0.005$; 2 is the analysis of the original data with $r_A = 0.9$ and $r_B = 0.01$; 3 is the analysis of the simulated data with $r_A = 0.9$ and $r_B = 0.005$.

ined biological context, hence relaxing the size constraint on the gene set from data likelihood. Therefore, more small gene sets that contain large portions of highly ER-associated genes are identified under such a setting of hyper-parameters. The effect of data diffusion is rather obvious. Diffusing the the distribution of gene association probabilities makes the differences between genes ambiguous. Large gene sets tend to benefit from this ambiguity, because the flatness of the beta distribution of association probability for pathway member genes allows a larger number of genes in a gene set to have low association probabilities. Hence, while the gene ranks have a fundamental effect on pathway ranking, the distinct beta mixture distribution of data is essentially important for PROPA to remove the gene set size effect and ensure its sensitivity.

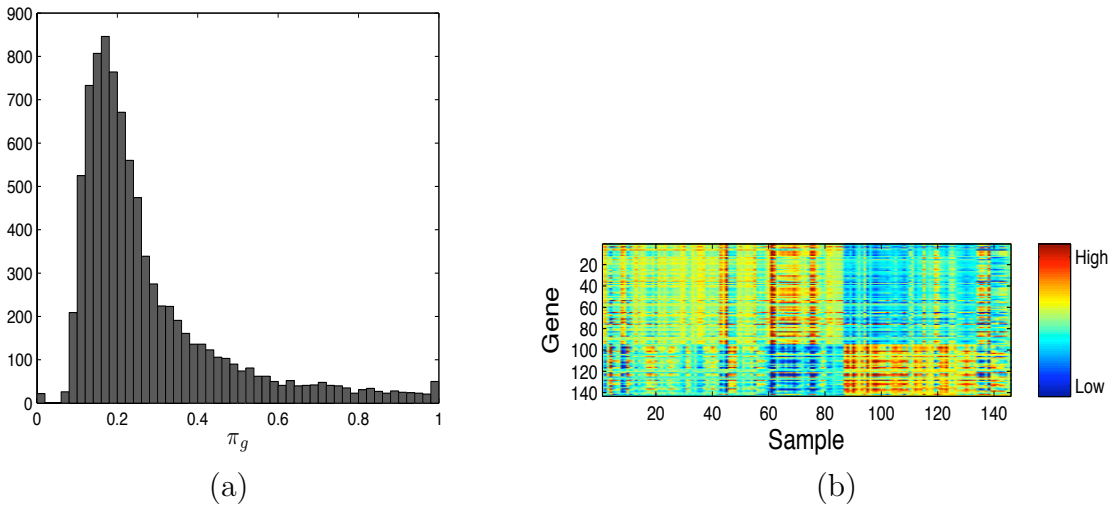
7.3 Breast Tumor ErbB2 Pathway Annotation

ErbB2, also called HER2 or Neu, is in the same transmembrane receptor family as epidermal growth factor receptor (EGFR or HER1), ErbB3 and ErbB4. Upon epidermal growth factor (EGF) binding, these proteins can form homo- or hetero-dimers, which recruit signaling molecules and activate specific cell growth signaling pathways. The structure of the ErbB2 protein allows it to interact with the other ErbB family members, especially with ErbB3, to form hetero-dimers in the absence of ligand binding. Such hetero-dimers formed with ErbB2 appear to be remarkably efficient in signal transduction. The downstream signaling pathways lead to cell proliferation, differentiation, survival and migration. There is also evidence of ErbB2 translocation to nucleus regulating the expression of certain pro-oncogenic genes. Additionally, ErbB2 also plays a role in controlling translation of specific proteins. About 20-25% of breast cancers have over-expression of ErbB2. This over-expression is primarily due to gene amplification, which seems to be the major cause of ErbB2 pathway deregulation in breast cancers (Ménard *et al.* 2003; Badache and Gonçalves 2006). ErbB2 has been identified as the therapeutic target of ErbB2-positive breast cancer and clinically treated with Herceptin. Here I use PROPA to detect the pathways associated with the breast tumor ErbB2 status, then focus on demonstrating gene pathway membership inference.

7.3.1 Data and Gene Sets

Pathways associated with ErbB2 status are analyzed by using the 146 primary breast tumor samples from Carvalho *et al.* (2007) (a subset of the original data set with sample ErbB2 status available). Among these tumors, 86 are ErbB2-postive, i.e., have ErbB2 over-expression, while the other 60 are ErbB2-negative. The genome-wide gene expression profiles are obtained through microarray experiments on Affymetrix

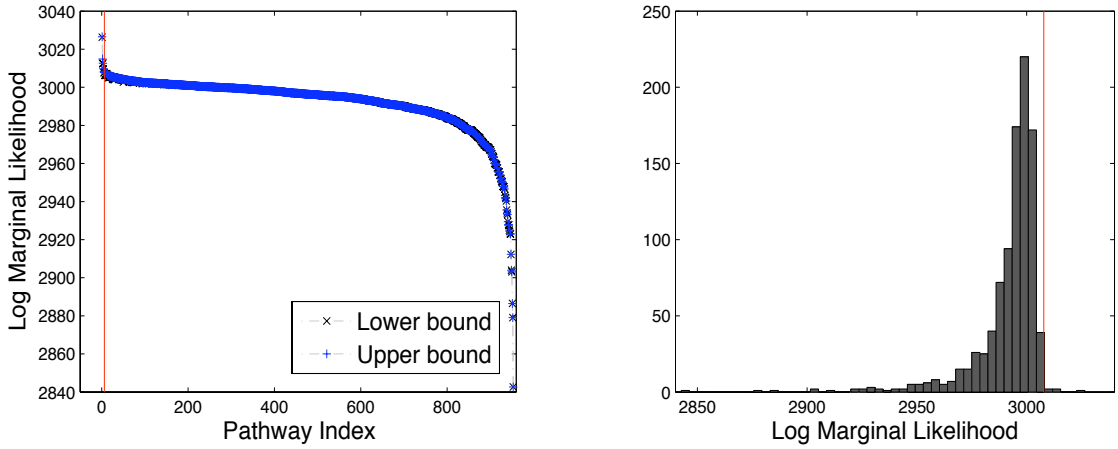
Human genome U95Av2 chips. The probability and status (positive or negative) of association between each gene expression and tumor ErbB2 status are obtained from a regression analysis with BFRM. Figure 7.7(a) is the histogram of the association probabilities of the 8,764 unique genes. The gene signature of the hypothetical pathway under study is a relatively small set of genes whose expression profiles are shown in Figure 7.7(b).



(a) (b)

Figure 7.7: Association probabilities and expression profiles of the genes correlated with ErbB2 status of breast tumors. (a) Histogram of association probabilities π_g; (b) Expression profiles of 143 genes whose association probabilities are greater than 0.9.

The 956 human pathway gene sets drawn from the MSigDB do not include pathway signatures explicitly linked to breast tumor ErbB2 status. To validate the effectiveness of PROPA, I curate two gene sets from literatures representing ErbB2 associated pathway signatures in breast cancer. The first gene set, called "molecular portrait" of ErbB2-positive breast tumors for convenience, consists of several genes that are mainly located at the chromosome 17 and have been identified as a cluster corresponding to ErbB2 over-expression through gene clustering analyses of breast tumor microarray data sets (Perou *et al.* 2000; Sørlie *et al.* 2001). The second gene



(a) (b)

Figure 7.8: Log marginal likelihoods for 958 pathways in breast tumor ErbB2 status pathway annotation. (a) Pathway log marginal likelihood upper bounds (blue +) and lower bounds (black ×); pathways are sorted in a decreasing order of log marginal likelihoods; the pathways on the left side of the red line are the 6 pathways with the largest marginal likelihoods. (b) Histogram of the pathway log marginal likelihood upper bounds; the bars on the right side of the red line correspond to the top 6 pathways.

set is curated from the ErbB2 gene expression signature defined by Bertucci *et al.* (2004). It includes the genes that differentially expressed in tumors and cell lines with vs. without over-expression of ErbB2 protein. Notably, among the 24 genes, three are negatively correlated with ErbB2 status. Although many positive correlated genes in the signature gene set locate in chromosome 17, only two of them (ErbB2 and GRB7) overlap with those in the molecular portrait gene set.

7.3.2 Pathway Annotation

Figure 7.8(a) shows the optimal upper bound and corresponding lower bound of log marginal likelihood for each pathway. The pathways are arranged in a decreasing order of log marginal likelihood upper bounds. Figure 7.8(b) is the histogram of the upper bounds. The first five or six pathways may be of particular interest,

because the log marginal likelihoods drop slowly after them. As mentioned above, the regression analysis indicates that only a very small set of genes appear to be strongly associated with this ErbB2 status in this breast tumor samples, suggesting either the pathway activities related to ErbB2 over-expression are weakly reflected on the transcriptional level, or the heterogeneity of tumors within the two categories defined by ErbB2 status makes the related changes of gene expression too subtle to be detected. The distribution of the evidence (or posterior probabilities) for the pathways under test, as shown in Figure 7.8, corroborates with this observation on the overall gene-phenotype association, and implies that only a few pathways are prominently associated with the breast tumor ErbB2 status.

Table 7.2 contains the summary of the top six pathways in terms of association with ErbB2 status of the 146 breast tumor samples. PROPA identifies the two ErbB2 signatures as the top one and four in the whole list of 958 pathway signatures. In Figure 7.9(a) and (c), the association probabilities of genes in the two sets are categorized by the sign of the correlation. Concordantly, all the genes except one in the ErbB2 portrait gene set are positively correlated with ErbB2 status, while the three negative signature genes in the ErbB2 signature gene set fall right in the negative correlation category.

Additionally, almost all the genes in the third pathway gene set have positive correlation with the ErbB2 status. This gene set contains genes upregulated in multiple myeloma cells treated with pro-proliferative cytokine IL-6 (Croonquist *et al.* 2003). Most of these IL-6 upregulated genes are involved in cell cycle progression; ErbB2 has been known for triggering cell G1/S transition by activating Ras/Erk signaling cascade through Shc/Grb-2 recruitment, and consequently increasing cell proliferation (Badache and Gonçalves 2006; Ricci *et al.* 1995). This effect between IL-6 treatment on myeloma cells and ErbB2 over-expression on breast tumor cell may provide the

rationale for this pathway phenotype association.

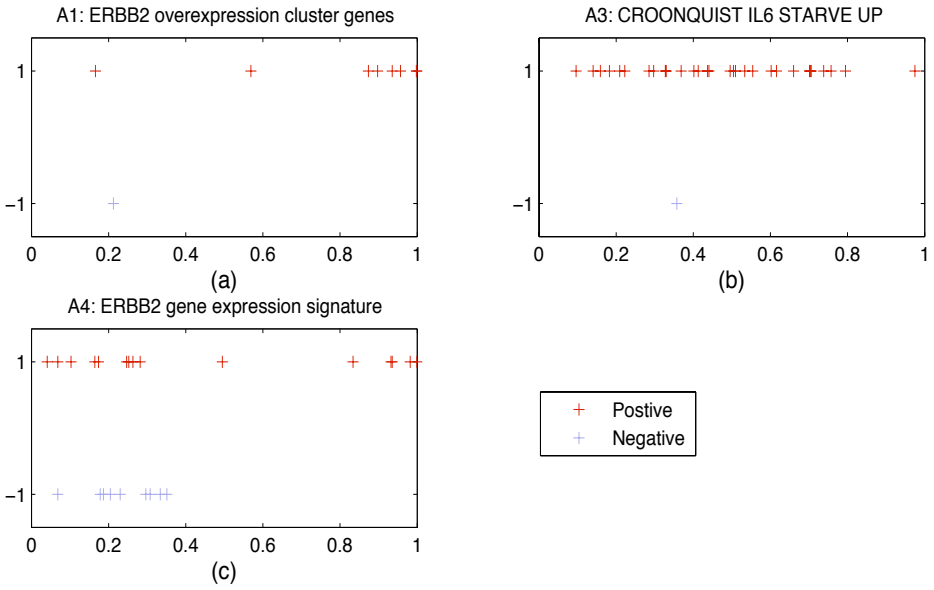


Figure 7.9: Association probability plots of the ErbB2-related pathway gene sets identified by PROPA. In each plot, x-axis is probability, and y-axis has two states, -1 and 1, representing negative and positive correlation with ErbB2, respectively. For each gene set, the association probabilities of the genes positively correlated with tumor ErbB2 status are in red, while those of the genes negatively correlated with tumor ErbB2 status is in blue.

The same data set and pathway genes sets are analyzed by GSEA. According to GSEA report, *none* of the 958 gene sets are enriched in the ErbB2 cancer phenotype with FDR q-value less than 25%. If thresholded with p-value 0.01, four gene sets are significantly downregulated in ErbB2-postive tumors, but no gene sets are significantly upregulated. Table 7.3 gives the GSEA summary of the gene sets with highest NES. The two gene sets that have been specifically curated and added to the database are identified by GSEA in top upregulated gene set list (top four and six). However, neither of them can be considered with strong significance. In this example, compared with GSEA, PROPA presents better sensitivity and specificity when transcriptional evidence of phenotype-pathway association are relatively weak.

Table 7.2: Summary of the top six pathways identified by PROPA as being related to breast tumor ErbB2 status

Rank	Pathway	Post. Pr	Post.Pr (sub)	BF	BF (sub)	Size	LogML (UB)	LogML (LB)	UB-LB
1	ERBB2 overexpression cluster genes	1.00E+00		98428243.91		9	3026.53	3026.29	0.24
2	HUMAN TISSUE KIDNEY	8.61E-06	8.90E-01	0.0082	7684.2458	11	3014.87	3012.41	2.46
3	CROONQUIST IL6 STARVE UP	9.28E-07	9.59E-02	0.0009	100.8729	31	3012.64	3012.59	0.05
4	ERBB2 gene expression signature	5.25E-08	5.43E-03	0.0001	5.1949	24	3009.77	3009.73	0.04
5	HDACI COLON CUR16HRS DN	1.37E-08	1.41E-03	0	1.344	8	3008.42	3007.66	0.76
6	MMS HUMAN LYMPH LOW 4HRS DN	7.63E-09	7.89E-04	0	0.7512	16	3007.84	3007.81	0.03

Table 7.3: Top pathways (ranked by NES) identified by GSEA that are related to ErbB2 status in Miller breast tumor data set

ErbB2 + vs. -	NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
Up-regulated	FATTY ACID SYNTHESIS	13	-0.60603946	-1.5583651	0.030927835	1	0.957
	HUMAN TISSUE KIDNEY	11	-0.75427413	-1.5481495	0.016877636	1	0.966
	CMV HCMV TIMECOURSE 4HRS DN	34	-0.43504864	-1.5229394	0.028225806	1	0.976
	ERBB2 gene expression signature	23	-0.64703345	-1.5219238	0.026871402	1	0.978
	CROONQUIST IL6 RAS DN	24	-0.702116	-1.5067064	0.09727626	1	0.983
	ERBB2 overexpression cluster genes	9	-0.902758	-1.4867045	0.011673152	1	0.992
Down-regulated	H2O2 CSBRESCUED C2 UP	10	0.69896984	1.669812	0.004273505	1	0.728
	ZHAN MM CD138 LB VS REST	19	0.661346	1.6618054	0.006036217	1	0.752
	ROSS MLL FUSION	61	0.5561933	1.6345505	0.025157232	1	0.828
	BRG1 SW13 UP	34	0.5690319	1.5945125	0.006369427	1	0.893
	ZHAN MMPC LATEVS	43	0.6071136	1.569556	0.008492569	1	0.931

7.3.3 Pathway Membership Inference

Here I demonstrate gene pathway membership inference in PROPA focused on refining the ErbB2 molecular portrait gene set. The pathway represented by this gene set has strongest association with breast tumor ErbB2 status among the 958 pathways from the database. According to the description of the context, this pathway signature should only include the genes located in the chromosomal region where ErbB2 gene is, and amplified together with ErbB2 on DNA on the transcription level. I illustrate gene pathway membership inference with a simulation followed by discussion of refining this gene set in its real biological context.

Simulation Study

The ErbB2 molecular portrait gene set, as described in the previous section, has nine genes including ErbB2, the key signature gene whose amplification is the substantial factor activating the cancer signaling cascades in ErbB2-positive breast tumors. PROPA has identified this gene set from over 900 gene sets in the database in pathway annotation analysis of breast tumor ErbB2 status. I choose this gene set as an example, slightly manipulate the members of the set and observe how PROPA infers gene pathway membership.

Denote the ErbB2 portrait gene set by A_1. Generate a new gene set A_2 by excluding the ErbB2 gene from A_1 and adding another gene g_r, which is randomly chosen from the full gene list $\mathcal{G}$ and has small association probability. In the data set, ErbB2 and g_r have association probability 0.94 and 0.11, respectively. Run PROPA on these two gene sets. This predefined membership swap between ErbB2 and g_r results in a four to five points drop of the log marginal likelihood, which does not change the rank of this pathway signature in the whole list. For each gene $g \in \mathcal{G}$, the Bayesian estimate of its pathway membership probability $\hat{\beta}_g^*$ is obtained

from the posterior mean of ρ_g in (4.6), and the pathway membership Bayes factor $BF_{g\in\mathcal{A}}$ is indicated in (4.7) and (4.8). The pathway membership evidence in nats, defined as $\log(BF_{g\in\mathcal{A}})$, is plotted in Figure 7.10(a) and (b) for gene sets A_1 and A_2, respectively. Similarly, the pathway membership evidence in decibans (dBs), defined as $10\log_{10}(BF_{g\in\mathcal{A}})$, is plotted in Figure 7.11. The genes in the gene set are represented by stars, while the genes not in the gene set are dots. Genes are sorted with respect to their association probabilities. The yellow curve is generated through polynomial curve fitting to the dots and somehow shows the expected values of pathway membership evidence given π_g for $g \notin A_x$ ($x = 1$ or 2).

In each scatter plot, the genes located above the horizontal zero line have evidence greater than 0, suggesting these genes might be the members of the true pathway; vice versa for the genes below the line. Within a category (in A_x or not in A_x), stronger belief goes to those genes far away from the zero line. Such a scatter plot gives us a summary of gene pathway membership inference results. The blue dots with large positive membership evidence values correspond to the genes that are likely to be true pathway members but are missed by the pathway gene set defining process (false negatives); the red stars with large negative membership evidence values represent the genes that may have been incorrectly defined as pathway members (false positives).

The ErbB2 gene, which is a member of gene set A_1 but not a member of A_2, is shown as a green star in Figure 7.10(a) and Figure 7.11(a), and as a green dot in Figure 7.10(b) and Figure 7.11(b); inversely, gene g_r is shown as a green dot in Figure 7.10(a) and Figure 7.11(a), and a green star in Figure 7.10(b) and Figure 7.11(b). Under both models (corresponding to A_1 and A_2), ErbB2 has pathway membership evidence near 20 dBs, very strongly suggesting that ErbB2 is a true member of the ErbB2 portrait signature; on the contrary, g_r has substantially large negative evidence

values in both cases confirming that it is not a member of this pathway signature.

Refining the ErbB2 "Portrait"

As discussed in Section 4.2, the pathway membership of a gene is largely driven by its association probability. Nevertheless, this relation is pathway specific and inferred through the posterior simulation. The monotonically increasing yellow curve in Figure 7.10(a) or 7.11(a) displays the positive association between posterior pathway membership probability and association probability. In real applications of PROPA, one is interested in a pathway with relatively high posterior probability of association with the investigated phenotype. Pathway membership inference provides the basis for decision-making on whether certain genes potentially are the members of this pathway. Given the biological context in which the pathway is defined, pathway membership inference can help to refine the pathway gene sets by highlighting those false positive and false negative genes.

Table 7.4 gives the information as well as the pathway membership inference summary on the genes in the ErbB2 molecular portrait gene set A_1. Six genes located in the chromosomal regions 17q11-q12 and 17q21 have relatively high probabilities of positive association with breast tumor ErbB2 status. The posterior pathway membership probabilities of these genes are also high enough to confirm their membership. The other three genes with relatively low association probabilities are inferred by PROPA as false positive genes. Their posterior pathway membership probabilities are zero. Notably, gene MMP15 is located at 16q13-q21. It was included in the ErbB2 portrait gene set by a gene clustering analysis based on microarray data. I conclude that MMP15 should not be a true member of this ErbB2 pathway.

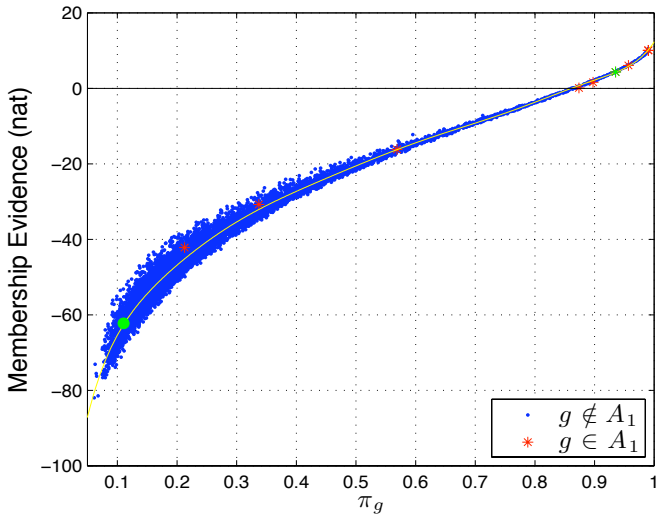
Table 7.5 lists the genes not in the ErbB2 portrait gene set that have posterior pathway membership probabilities greater than 0.5. As can be seen, these genes

have association probabilities higher than 0.97. Several genes (G6PC, ERAL1, OMG, RPL19, CRKRS) are located in the regions 17q11-q12 and 17q21, and have positive correlation with ErbB2 status. Their pathway membership evidence are greater than 8 nats (greater than 34 dBs), decisive evidence for these genes being false negatives, i.e. members of the true ErbB2 portrait signature.

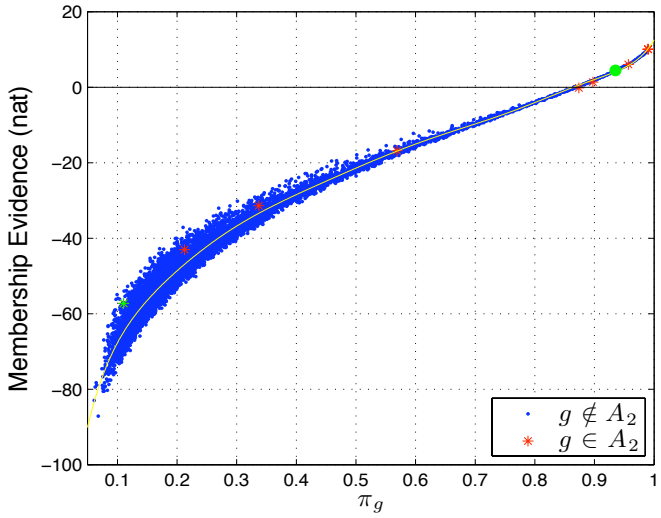
This example has presented how PROPA identifies those false positive and false negative genes in terms of pathway membership. Naturally, the interpretation based on the knowledge of individual genes and the understanding of the pathway is essentially important in this process. If the pathway signature is defined in a biological context different from the one being studied, which is usually the case in reality, such inference may also convey the difference of biological processes in the two contexts. Those genes in Table 7.5 other than those identified as false negatives can be interpreted as being involved in the biological processes beyond the ErbB2 chromosomal region amplification, for example, the downstream signaling pathway members. These genes may also be the members of other significant pathways. For example, RPL19 is in the ErbB2 gene expression signature, and OIP5 is in the IL-6 up-regulated gene signature.

From another point of view, this gene pathway membership inference is potentially useful for identifying potential cancer biomarkers. In most cases, a biological phenotype under investigation involves complex biological processes; this is reflected in the pathway annotation as several distinct biological pathways may appear to be of interest. It is not surprising that some genes with high probability would be suggested as members of many of these pathways with or without considering the precise biological contexts. Without interaction information among signature genes, such results in a larger sense do not conflict with the understanding of biological pathways, the artificial dissections of the underlying dynamic gene regulatory network. These path-

ways are modules in the whole network and linked through gene regulation activities that have or have not been discovered or indicated in the database. In such cases, a key function of pathway membership inference is to identify potential biomarkers of the biological phenotype based on the prior knowledge on pathway gene sets and beliefs in the pathway activities under the phenotype. Hence, those genes inferred as false negatives are of particular interest.

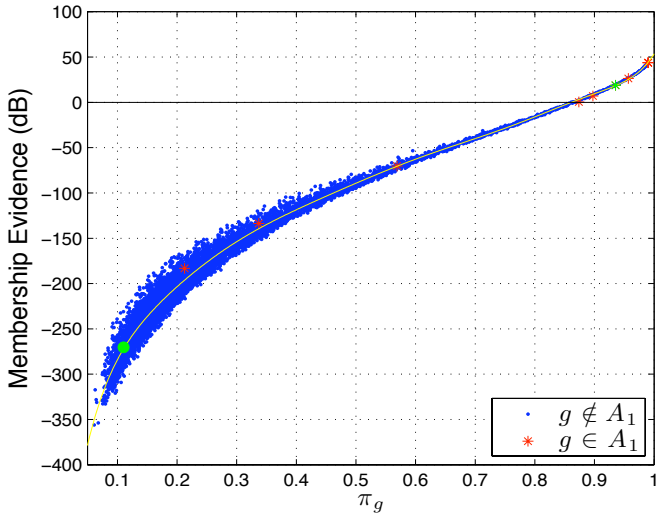


(a)

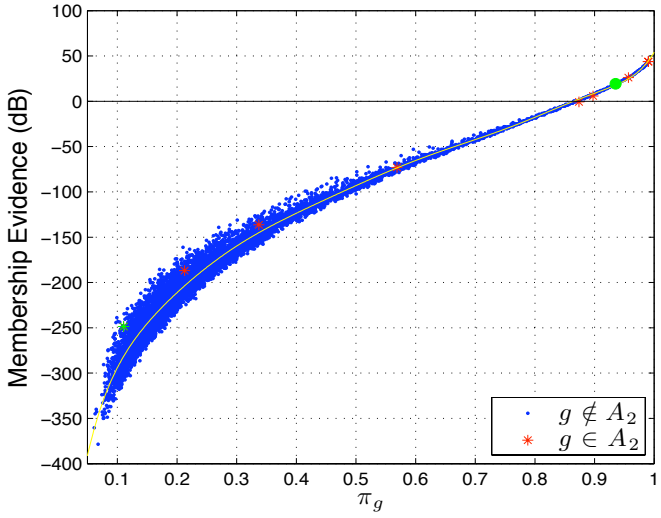


(b)

Figure 7.10: Gene pathway membership probability inference for ErbB2 molecular portrait: scatter plots of membership evidence in nats ($\log(BF_{\beta_g^*})$) vs. association probability π_g for each gene $g \in \mathcal{G}$. Dots and stars correspond to genes in and not in pathway signature gene set A_j ($j = 1$ or 2), respectively. The genes not in A_j and with large positive evidence values, corresponding to the blue dots in the upper right corner, potentially are true members of the theoretical pathway $\mathcal{A}$, i.e. false negatives; the genes in A_j (red stars) with large but negative evidence potentially are not true members of $\mathcal{A}$, i.e. false positives. (a) A_1 is the original ErbB2 expression signature gene set; the green star represents ErbB2 gene, while the gene dot represents the randomly picked gene g_r. (b) A_2 shares the same set of genes with A_1 except that the gene set-membership of ErbB2 and gene g_r are exchanged; the green dot represents ErbB2 gene, while the gene star is g_r.



(a)



(b)

Figure 7.11: Gene pathway membership probability inference for ErbB2 molecular portrait: scatter plots of membership evidence in decibans ($10\log_{10}(BF_{\beta_g^*})$) vs. association probabitlity π_g for each gene $g \in \mathcal{G}$.

Table 7.4: Genes in the ErbB2 molecular portrait gene set

Index	Symbol	Description	Gene ID	Chr. Loc.	Memb. Pr	Memb. log(BF)	Asso. Pr	Corr.
1	STARD3	START domain containing 3	10948	17q11-q12	1	10.0675	0.99	+
2	GRB7	growth factor receptor-bound protein 7	2886	17q12	1	10.0647	0.99	+
3	THRAP4	thyroid hormone receptor associated protein 4	9862	17q21.1	0.999	6.0859	0.9569	+
4	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	2064	17q11.2-q12| 17q21.1	0.9944	4.3388	0.9352	+
5	TRAF4	TNF receptor-associated factor 4	9618	17q11-q12	0.9207	1.6042	0.8975	+
6	FLOT2	flotillin 2	2319	17q11-q12	0.7218	0.1062	0.8737	+
7	PCGF2	polycomb group ring finger 2	7703	17q12	0	-16.1889	0.5688	+
8	MMP15	matrix metallopeptidase 15 (membrane-inserted)	4324	16q13-q21	0	-30.7798	0.3371	+
9	SMARCE1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1	6605	17q21.2	0	-42.1872	0.2126	-

Table 7.5: Genes not in the ErbB2 molecular portrait gene set that have posterior pathway membership probabilities greater than 0.5

Index	Symbol	Description	Gene ID	Chr. Loc.	Memb. Pr	Memb. log(BF)	Asso. Pr	Corr.
1	SLC2A2	solute carrier family 2 (facilitated glucose transporter), member 2	6514	3q26.1-q26.2	0.9475	10.494	0.99	+
2	G6PC	glucose-6-phosphatase, catalytic subunit	2538	17q21	0.9438	10.4218	0.99	+
3	DNAJC6	DnaJ (Hsp40) homolog, subfamily C, member 6	9829	1pter-q31.3	0.9372	10.3026	0.99	+
4	EZH1	enhancer of zeste homolog 1 (Drosophila)	2145	17q21.1-q21.3	0.9333	10.2384	0.99	-
5	CNKSR2	connector enhancer of kinase suppressor of Ras 2	22866	Xp22.12	0.9269	10.1407	0.99	+
6	PPP1R10	protein phosphatase 1, regulatory (inhibitor) subunit 10	5514	6p21.3	0.9266	10.1361	0.99	+
7	ERAL1	Era G-protein-like 1 (E. coli)	26284	17q11.2	0.9263	10.1314	0.99	+
8	TOX	thymocyte selection-associated high mobility group box	9760	8q12.1	0.9262	10.1299	0.99	+
9	SLC34A1	solute carrier family 34 (sodium phosphate), member 1	6569	5q35	0.9257	10.1228	0.99	+
10	EFEMP1	EGF-containing fibulin-like extracellular matrix protein 1	2202	2p16	0.9238	10.0957	0.99	-
11	SCARB1	scavenger receptor class B, member 1	949	12q24.31	0.9225	10.077	0.99	-
12	CRTAP	cartilage associated protein	10491	3p22.3	0.9221	10.0714	0.99	-
13	SOX11	SRY (sex determining region Y)-box 11	6664	2p25	0.9214	10.0615	0.99	+
14	CPN2	carboxypeptidase N, polypeptide 2	1370	3q29	0.9201	10.0442	0.99	+
15	MAPK8IP3	mitogen-activated protein kinase 8 interacting protein 3	23162	16p13.3	0.9199	10.0418	0.99	-
16	OMG	oligodendrocyte myelin glycoprotein	4974	17q11.2	0.9192	10.0319	0.99	+
17	MLC1	megalencephalic leukoencephalopathy with subcortical cysts 1	23209	22q13.33	0.9185	10.0221	0.99	+
18	EHHADH	enoyl-Coenzyme A, hydratase/3-hydroxyacyl Coenzyme A dehydrogenase	1962	3q26.3-q28	0.9177	10.0123	0.9899	+

(Continued on next page)

Continuing Table 7.5 [1]

Index	Symbol	Description	Gene ID	Chr. Loc.	Memb. Pr	Memb. log(BF)	Asso. Pr	Corr.
19	TTR	transthyretin (prealbumin, amyloidosis type I)	7276	18q12.1	0.9175	10.0088	0.99	+
20	LOC54103	hypothetical protein LOC54103	54103	7q11.23	0.9167	9.9984	0.99	-
21	FLJ13236	hypothetical protein FLJ13236	79962	12q13.12	0.9151	9.9775	0.99	+
22	PDE10A	phosphodiesterase 10A	10846	6q26	0.9116	9.9342	0.99	+
23	TMEM16C	transmembrane protein 16C	63982	11p14.2	0.9064	9.871	0.9879	+
24	NCAPH	non-SMC condensin I complex, subunit H	23397	2q11.2	0.8952	9.7456	0.9878	+
25	CNTN5	contactin 5	53942	11q21-q22.2	0.8942	9.7344	0.9884	+
26	POU4F2	POU domain, class 4, transcription factor 2	5458	4q31.2	0.8933	9.7256	0.989	+
27	MYST3	MYST histone acetyltransferase (monocytic leukemia) 3	7994	8p11	0.8927	9.7187	0.9881	-
28	COL4A4	collagen, type IV, alpha 4	1286	2q35-q37	0.8884	9.6746	0.9887	+
29	STK17B	serine/threonine kinase 17b	9262	2q32.3	0.8816	9.6077	0.9888	+
30	NAB1	NGFI-A binding protein 1 (EGR1 binding protein 1)	4664	2q32.3-q33	0.8748	9.5445	0.9873	-
31	PTENP1	phosphatase and tensin homolog (mutated in multiple advanced cancers 1), pseudogene 1	11191	9p21	0.87	9.5013	0.9878	+
32	EPB49	erythrocyte membrane protein band 4.9 (dematin)	2039	8p21.1	0.869	9.4922	0.9864	-
33	CXCL6	chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2)	6372	4q21	0.8685	9.4882	0.9863	+
34	PF4V1	platelet factor 4 variant 1	5197	4q12-q21	0.8438	9.2875	0.9856	+
35	USP34	ubiquitin specific peptidase 34	9736	2p15	0.8411	9.2665	0.9851	-
36	PCDH17	protocadherin 17	27253	13q21.1	0.8348	9.2205	0.985	+
37	P18SRP	P18SRP protein	285672	5q12.3	0.833	9.2072	0.9857	+
38	MYF5	myogenic factor 5	4617	12q21	0.8265	9.1615	0.9836	+
39	CD302	CD302 molecule	9936	2q24.2	0.8223	9.1323	0.9852	-

(Continued on next page)

Continuing Table 7.5 [2]

Index	Symbol	Description	Gene ID	Chr. Loc.	Memb. Pr	Memb. log(BF)	Asso. Pr	Corr.
40	SLC39A14	solute carrier family 39 (zinc transporter), member 14	23516	8p21.3	0.7925	8.9404	0.9845	-
41	FNTA	farnesyltransferase, CAAX box, alpha	2339	8p11	0.779	8.8605	0.9831	-
42	HLXB9	homeobox HB9	3110	7q36	0.766	8.7865	0.9827	+
43	APOBEC1	apolipoprotein B mRNA editing enzyme, catalytic polypeptide 1	339	12p13.1	0.7578	8.741	0.983	+
44	CTSF	cathepsin F	8722	11q13	0.7493	8.6954	0.9813	-
45	RPL19	ribosomal protein L19	6143	17q11.2-q12	0.7455	8.6753	0.982	+
46	IFNA16	interferon, alpha 16	3449	9p22	0.7359	8.6254	0.9814	+
47	GPATCH8	G patch domain containing 8	23131	17q21.31	0.7296	8.5932	0.9813	-
48	MBL2	mannose-binding lectin (protein C) 2, soluble (opsonic defect)	4153	10q11.2-q21	0.7261	8.5752	0.981	+
49	C11orf58	chromosome 11 open reading frame 58	10944	11p15.1	0.6962	8.4299	0.9795	-
50	CRKRS	Cdc2-related kinase, arginine/serine-rich	51755	17q12	0.6127	8.0591	0.9776	+
51	VPS11	vacuolar protein sorting 11 homolog (S. cerevisiae)	55823	11q23	0.6006	8.0082	0.9756	-
52	SLC23A2	solute carrier family 23 (nucleobase transporters), member 2	9962	20p13	0.5973	7.9945	0.9756	-
53	TRIAD3	TRIAD3 protein	54476	7p22.1	0.5746	7.9012	0.9753	+
54	OIP5	Opa interacting protein 5	11339	15q15.1	0.5544	7.8191	0.9748	+
55	OR7E24	olfactory receptor, family 7, subfamily E, member 24	26648	19p13.2	0.5476	7.7915	0.9737	+
56	OR2F1	olfactory receptor, family 2, subfamily F, member 1	26211	7q35	0.5046	7.6189	0.9729	+

Chapter 8

Analysis of Lactic Acidosis Response in Breast Tumors

In this chapter, the application of PROPA in decomposing the complexity of the mechanisms of cancer development is demonstrated with a study of cellular response to lactic acidosis in breast cancers.

One characteristic of solid tumor micro-environment is the high production of lactate and extracellular acidity, called lactic acidosis. This physiological change with the other common traits of malignant extracellular environment – oxygen depletion (hypoxia) and nutrient deprivation – is initially caused by insufficient and inappropriate vascular supply, and poor tissue perfusion in solid tumor micromilieu due to unregulated proliferation and rapid growth of cells (Vaupel *et al.* 1989; Vaupel 2004). When tissue oxygenation is inadequate, cells obtain energy through anaerobic glucose metabolism (glycolysis) and produce lactic acid. Therefore, hypoxia is commonly thought to be the primary cause of lactic acidosis. It has been demonstrated that hypoxia indeed promotes tumor cell glycolysis by upregulating genes encoding glucose transporters and glycolytic enzymes (Ebert *et al.* 1996; Elson *et al.* 2000). However, even in the presence of oxygen, tumor cells still rely on glycolysis for energy production, a phenomenon known as aerobic glycolysis or the "Warburg effect" (Warburg 1956).

Many studies have shown that tumor glycolytic phenotype may be achieved through oncogenic activation or stabilization of transcription factors such as HIF-1 through mechanisms other than hypoxia (Kim and Dang 2006). Several oncogenes and tumor suppressors involved in cell respiration circuits may directly contribute to

the acquisition of a tumor cell glycolytic phenotype. The AKT oncogene is able to enhance glucose uptake over cell membrane and trigger intracellular glucose trapping and phosphorylation leading to enhanced glycolytic flux. The MYC transcription factor, which is widely activated in human cancers, can activate glycolytic enzyme genes and cause mtDNA mutations that in turn cause enhanced glycolysis and aerobic respiration dysfunction of mitochondria. Expression loss of the p53 tumor suppressor may inactivate the aerobic respiratory chain and cause the switch of cellular respiration to glycolysis. Besides the direct oncogenic activation, HIF-1 stabilization is a factor significantly contributing to aerobic glycolysis. The transcription factor HIF-1 can activate glycolytic enzyme genes as well as PDK1, which in turn inactivates the mitochondrial pyruvate dehydrogenase (PDH) and inhibits mitochondrial function and aerobic respiration. HIF-1, a heterodimer of HIF-1α and HIF-1β, cannot be formed in normal cells under nonhypoxic conditions because of the sensitivity of HIF-1α to oxygen. However, in some human tumor cells HIF-1 can be stabilized in the presence of adequate oxygen when certain oncogenic events occur, such as the activation of RAS and SRC, and the repression of VHL, SDH and FH.

Gatenby and Gillies (2004) proposed that the tumor glycolytic phenotype is a result of active selection processes and must confer a significant growth advantage necessary for evolution of invasive human cancers. Through somatic evolution, tumor cell populations manage to survive in lactic acidosis environment and become resistant to acid-induced toxicity. Furthermore, through upregulation of glycolysis, tumor cells worsen the extracellular environment by increasing the acidity to a toxic level for normal populations; the consequent destruction of adjacent normal issues and degradation of the extracellular matrix facilitate tumor cell invasion and angiogenesis. This suggests increasing glycolysis and resistance to extracellular lactic acidosis of tumor cells is a malignant phenotype associated with cancer aggressiveness and

a driving force for the evolution to invasiveness of primary cancers as well, which may further be crucial in cancer metastasis. Studies of lactic acidosis through *in vivo* measurement of acidity and lactate in a variety of human cancers have provided evidence for such association (Walenta *et al.* 2000; Brizel *et al.* 2001). Hence, understanding the fundamental cellular response to lactic acidosis in tumors may facilitate cancer risk prediction and lead to novel cancer therapies to improve patients outcomes. However, compared with that on glycolytic phenotype induction, the existing knowledge on tumor lactic acidosis response is rather limited.

The primary goal of the study of Chen *et al.* (2007) is to analyze the cellular response to lactic acidosis in human breast cancers using genome-wide gene expression data. This study has two steps. The first is to observe gene expression variation in normal mammary cells induced by *in vitro* exposure in a condition of lactic acidosis. This gene expression variation characterizes mammary cell transcriptional response to lactic acidosis in general. This may give us opportunities to examine the existence of association between lactic acidosis and breast cancer phenotypes. This approach is based on the premise that normal and tumor cells from same type of tissue may have common traits of behavior on certain levels. Similar approaches have previously been used to study the association of cellular wound healing and hypoxia response with tumor progression (Chang *et al.* 2004; Chi *et al.* 2006; Bild *et al.* 2006). The second step is to evaluate the molecular mechanisms and prognostic roles of cellular response to lactic acidosis in breast cancers *in vivo*. To determine the lactic acidosis response phenotype of breast cancers, a general molecular signature of this phenotype is defined from cultured cell *in vitro* study on normal mammary cells and used to impute the response level of a heterogeneous set of breast tumor samples (Lucas *et al.* 2007; Chen *et al.* 2007). Using evolutionary factor analysis and probabilistic pathway annotation, this aim to deconvolute molecular mechanisms of breast tumor lactic

acidosis response and detect possible connectivities with other clinical phenotypes of breast cancers. Figure 8.1 shows a overall workflow diagram of this analysis. The detail is described in the following sections.

8.1 HMEC Lactic Acidosis Response Annotation

To characterize the transcriptional response of mammary cells to lactic acidosis, Chen *et al.* (2007) cultured human mammary epithelial cells (HMECs) in lactic acidosis environment (25mM lactic acid, pH = 6.7), then measured global gene expression of each of the treated and control samples with Affymetrix GeneChip U133 plus 2.0 array, which contains more than 54,000 probe sets corresponding to about 20,000 unique genes. The microarray data set attained from this experiment includes 12 HMEC samples – six controls and six lactic acidosis samples. A one-way ANOVA with BFRM generates the posterior probability of association with lactic acidosis status π_g^* as well as the posterior mean of loading β_g for each gene g (refer to equation (2.4)). The posterior association probabilities of the whole gene list appear in the histogram in Figure 8.2(a). A gene expression signature of HMEC lactic acidosis response can be defined as a gene set comprising about 200 genes with association probabilities near 1 and highest absolute loadings on lactic acidosis covariate (Lucas *et al.* 2007). A preliminary functional annotation of these signature genes using the Gene Oncology tool GATHER (Chang and Nevins 2006) show that lactic acidosis induces genes involved in G-protein coupled receptor signaling, antigen processing and presentation and cellular catabolism, and represses genes involved in cell cycle, RNA metabolism and RNA processing.

Pathway annotation is then applied to further reveal the biological processes underlying mammary cell lactic acidosis response. I use PROPA to compare the phenotype association of 965 pathway gene sets that include 956 human pathway gene

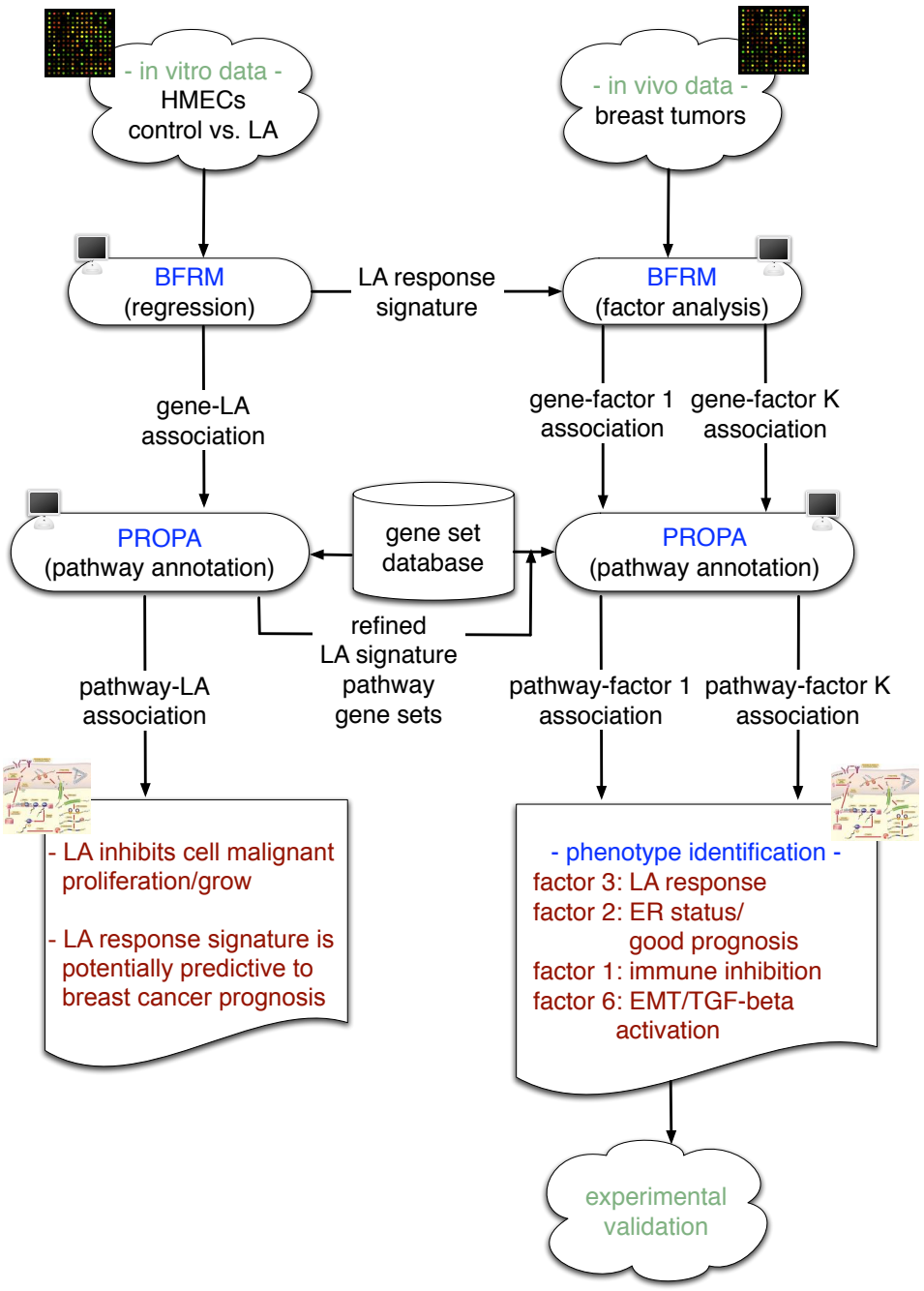
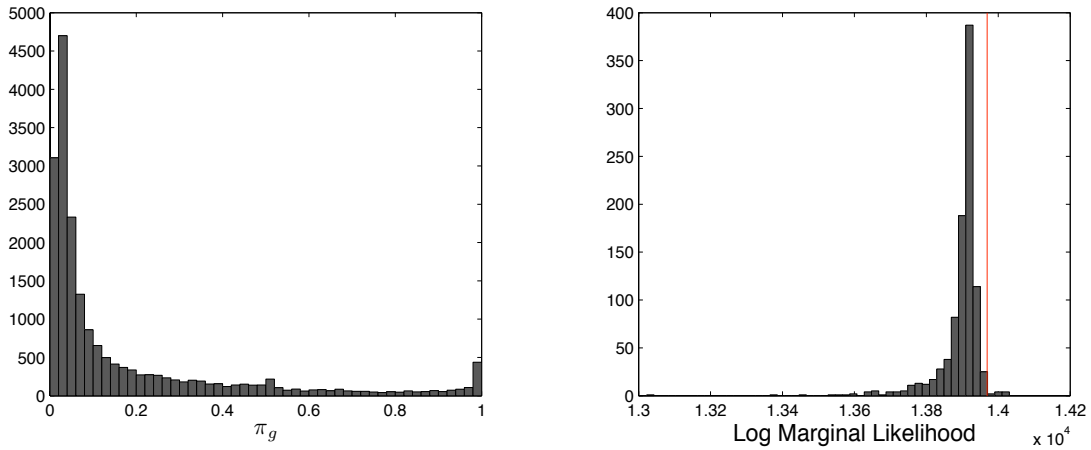


Figure 8.1: Workflow diagram of the analysis of cellular response to lactic acidosis in breast tumors

sets from the MSigDB database and nine oncogenic pathway signature gene sets. Five of these oncogenic pathway signatures, corresponding to the activation of RAS, MYC, E2F3, SRC and β-catenin, are curated from the paper by Bild *et al.* (2006). The other four, corresponding to the activation of AKT, P110, E2F1 and P63, are curated from unpublished data sets. The log marginal likelihoods are distributed as Figure 8.2(b). I only focus on the top 12 pathways summarized in Table 8.1. These top pathways link to cell proliferation and cancer phenotypes, giving abundant evidence to the nature of lactic acidosis as a potential signal and predictor of cancer development, and are reviewed as follows:



(a) (b)

Figure 8.2: Distributions of association probabilities and pathway log marginal likelihoods in HMEC transcriptional response analysis. (a) Histogram of association probabilities. (b) Histogram of pathway log marginal likelihood upper bounds; the bars on the right side of the red line correspond to the top 12 pathways.

- Pathways linked to other traits of cancer micro-environment

  Gene set A_1 includes the human genes downregulated in response to glutamine deprivation. Glutamine is the principal energy, carbon, and nitrogen source for mammalian cells. Peng *et al.* (2002) used murine CTLL-2 T lymphocytes to generate transcription profiles in response to glutamine deprivation, and the hu-

man homologs of the perturbed genes were collected in a gene set to impute the human cell transcriptional response to the same intervention. The histogram of association probabilities in Figure 8.3(a) shows that most of the genes downregulated by glutamine deprivation are also downregulated in HMECs exposed in lactic acidosis environment. Some genes (corresponding to the red bars) appear to be upregulated by lactic acidosis. However, a majority of them have large uncertainty in their association with lactic acidosis. The downregulation gene set in response to leucine deprivation (A_7), which was generated in a similar experiment by Peng *et al.* (2002), and largely overlaps with glutamine deprivation downregulated gene set, is also in the top pathway list. This reveals that the cells perceive lactic acidosis as an energy deficient status. Gene set A_4 was generated by Manalo *et al.* (2005) from the examination of gene profile variation in human pulmonary endothelial cells exposed in hypoxia or with hypoxia-inducible factor 1 (HIF-1α) activated. It contains the downregulated genes under either conditions. The activation of the genes involved in nutrient deprivation and hypoxia signaling pathways under the condition of lactic acidosis is consistent with the fact that nutrient starvation, hypoxia and lactic acidosis are commonly coexisting/interacting conditions in tumor micro-environment.

- Pathways linked to cancer progression

  PROPA identifies that lactic acidosis can shut down transcription of the genes whose activation are associated with cancer progression. Several gene signatures directly characterize wound healing, or are linked to certain biological processes during wound healing. Wound healing is a program initiated by extravasated blood coagulation and involving a complex pathway activities to restore tissue integrity, including immune cell recruitment, fibroblast and epithelial cell proliferation, cell migration and angiogenesis induction. The presence of deregulated

wound healing resembles cancer invasion and metastasis, and is predictive of breast, lung and gastric cancer metastasis (Chang *et al.* 2004).

A_2 and A_8 contain genes periodically expressed (cell-cycle dependent) and consistently upregulated (not cell-cycle dependent), respectively, following exposure to serum in human fibroblasts from a variety of anatomic sites; A_3 contains the genes commonly upregulated in these fibroblasts under serum stimulation. A fibroblast is a type of cell that synthesizes and maintains the extracellular matrix of many animal tissues and plays a critical role in wound healing. These gene sets characterizes the underlying transcriptional activities of wound healing. Genes in A_5 were found upregulated in myeloma cells when supplemented with IL-6. Gene set A_{12} has a relatively large overlap with A_5 (Croonquist *et al.* 2003). IL-6 is one of the key cytokines increasing endothelial cell proliferation and strongly associated with tumor angiogenesis (Kishimoto 1989), a hallmark of cancers that correlates with the risk of metastasis, recurrence and progression in a variety of cancers such as lung, breast, espophagus and prostate cancers. A_{10} contains genes with higher expression in less mature T cells than in more mature T cells. The biological functions of these genes are mainly linked to cell cycle regulation, cell cycle progression, mitosis, DNA replication, recombination or repair (Lee *et al.* 2004). The association between this gene set and cell lactic acidosis response implicates the impact of lactic acidosis on immune cells, which regulate angiogenesis, tissue homeostasis and wound healing. A majority of the genes in these sets are downregulated in HMECs as response to lactic acidosis (as shown in Figure 8.1(b), (c), (e), (h), (j) and (l)), indicatiing that lactic acidosis may have a general inhibiting effect on the wound healing program in cells.

- Pathways linked to cancer prognosis

Three cancer prognosis signatures are found linked to lactic acidosis response in HMECs. A_6 includes genes identified by Rhodes *et al.* (2004) that are commonly upregulated in undifferentiated cancer (correlating with poor prognosis) relative to well-differentiated cancer (correlating with good prognosis). A_{11} includes genes expressed higher in breast cancers with poor outcomes than in those with good outcomes (van't Veer *et al.* 2002). A_9 includes genes upregulated in gastric cancer cell lines resistant to doxorubicin, a chemotherapeutic agents, compared to chemosensitive cell lines (Kang *et al.* 2004). Upregulation of these sets of genes is related to aggressiveness in a variety of cancers. Here, these genes appear to be generally downregulated in HMECs in presence of lactic acidosis (Figure 8.3 (f), (i) and (k)).

In conclusion, from the transcriptional response to lactic acidosis in HMECs, the pathway annotation by PROPA has identified pathways relevant to the common biological processes and traits of cancer development, such as energy metabolism modulation, progression through collaborative interactions between diverse cell types, and clinical outcomes. The association between these pathways and cell lactic acidosis response strongly suggests the molecular mechanism by which lactic acidosis, as an important feature of tumor micro-enviroment, modulates cellular behaviors and links with cancer phenotypes. The annotation results also show that the genes in these pathways, which have been found induced during the biological processes involved in cancer development, are mostly downregulated in HMECs under the condition of lactic acidosis. This implies the function of lactic acidosis as a direct environmental factor that suppresses cell malignant proliferation and growth, thereby potentially leads to favorable clinical outcomes of cancers. Surprisingly, among the top significant pathway gene sets identified in this analysis, no biologically interpretable pathways contain large number of genes upregulated by lactic acidosis in this experiment. In

other words, the genes with high probabilities of being positively correlated with lactic acidosis status do not appear in any top pathway gene sets that can lead to reasonable biological interpretation. The most plausible reason is that the pathway gene sets included in the database do not yet adequately represent the range of biological processes involved in the cellular response to lactic acidosis.

8.1.1 Pathway Signature of HMEC Lactic Acidosis Response

Twelve gene sets were generated based on the 12 top pathways identified in this annotation analysis of HMEC response to lactic acidosis. Each of these gene sets contains the true positive genes in the original pathway gene set inferred by PROPA during the pathway annotation, and represents the intersection between the annotating pathway signature and HMEC lactic acidosis response signature. For example, gene set A_1 in the HMEC data pathway annotation result, as shown in Table 8.1, is named "PENG GLUTAMINE DN" and represents the genes downregulated in the cells under glutamine deprivation. I refine this gene set by selecting the genes with pathway membership evidence greater than 20dB, i.e. the refined gene set

$$A_{HMEC,1} = \{g : g \in A_1, 10 \log_{10} BF_{g \in \mathcal{A}_1} > 20\},$$

where $BF_{g \in \mathcal{A}_1}$ is the Bayes factor of pathway membership computed as (4.7). I name the pathway represented by this refined gene set $A_{HMEC,1}$ as "HMEC LA DN: PENG GLUTAMINE DN". This represents a certain sub-module of transcriptional activity in cell glutamine deprivation response that is linked to HMEC lactic acidosis response. In the same way, the pathways are defined on the other corresponding refined gene sets. These new pathway gene sets, listed in Table 8.2, compose a "pathway signature" for the lactic acidosis response of HMECs. The member genes of each pathway in this signature, as have been shown in the last section, are generally

downregulated by lactic acidosis.

8.2 Lactic Acidosis Response Analysis of In Vivo Gene Expression Programs of Breast Tumors

8.2.1 Signature Dissection and Enhancement

The breast tumor data set from Miller *et al.* (2005) are used to dissect the *in vivo* lactic acidosis response. This data set includes 251 primary invasive breast tumor samples. The expression profiles of more than 44,000 probe sets, corresponding to ∼18,000 unique genes, have been measured with the Affymetrix GeneChip U133 set arrays.

Lucas *et al.* (2007) used the 200 signature genes of HMEC lactic acidosis response as seeds for BFRM in a setting of evolutionary factor analysis over the Miller expression data set. This analysis decomposes the lactic acidosis response signature into a number of factors according to the latent expression patterns of the signature genes, and recruits new genes from the full gene list to enhance the expression signal in each factor and bring in new factors. Through such iteration, ten latent factors linked to lactic acidosis response of the breast tumor samples were extracted from gene expression profiles. These represent ten factor-phenotypes for PROPA analysis. Meanwhile, the posterior probability of association between each gene and each factor was estimated. Based on the regression of gene signature predefined in the HMEC lactic acidosis data analysis, each tumor sample achieves a score of lactic acidosis response. Among the ten factors derived from the lactic acidosis signature, some are highly related to the lactic acidosis response variation across the tumor samples, while the others are more likely to be subtler phenotypes linked to lactic acidosis response identified through the evolutionary factor analysis.

Table 8.1: Summary of the top 12 pathways identified by PROPA as being related to HMEC lactic acidosis status

Rank	Pathway	Post. Pr	Post.Pr (sub)	BF	BF (sub)	Size	LogML (UB)	LogML (LB)	UB-LB
1	PENG GLUTAMINE DN	9.87E-01		74900.0828		250	14031.01	14030.31	0.7
2	SERUM FIBROBLAST CELLCYCLE	9.37E-03		9.0337		134	14026.46	14025.87	0.59
3	CHANG SERUM RESPONSE UP	3.22E-03		3.0842		145	14025.38	14024.98	0.4
4	MANALO HYPOXIA DN	9.68E-09	8.62E-01	0	5966.3374	77	14012.37	14011.97	0.4
5	CROONQUIST IL6 STARVE UP	1.55E-09	1.38E-01	0	152.8603	32	14010.64	14010.24	0.4
6	CANCER UNDIFFERENTIATED META UP	1.12E-14	9.99E-07	0	0.001	66	13998.43	13998.4	0.03
7	PENG LEUCINE DN	1.02E-15	9.13E-08	0	0.0001	141	13996.2	13995.88	0.31
8	SERUM FIBROBLAST CORE UP	5.38E-16	4.79E-08	0	0	199	13995.72	13995.5	0.21
9	DOX RESIST GASTRIC UP	7.27E-21	6.47E-13	0	0	44	13984.33	13984.22	0.11
10	LEE TCELLS3 UP	1.18E-24	1.05E-16	0	0	100	13976.11	13975.43	0.68
11	VANTVEER BREAST OUTCOME GOOD VS POOR DN	2.68E-27	2.39E-19	0	0	65	13969.68	13969.14	0.54
12	CROONQUIST IL6 RAS DN	1.93E-27	1.72E-19	0	0	23	13969.55	13968.99	0.57

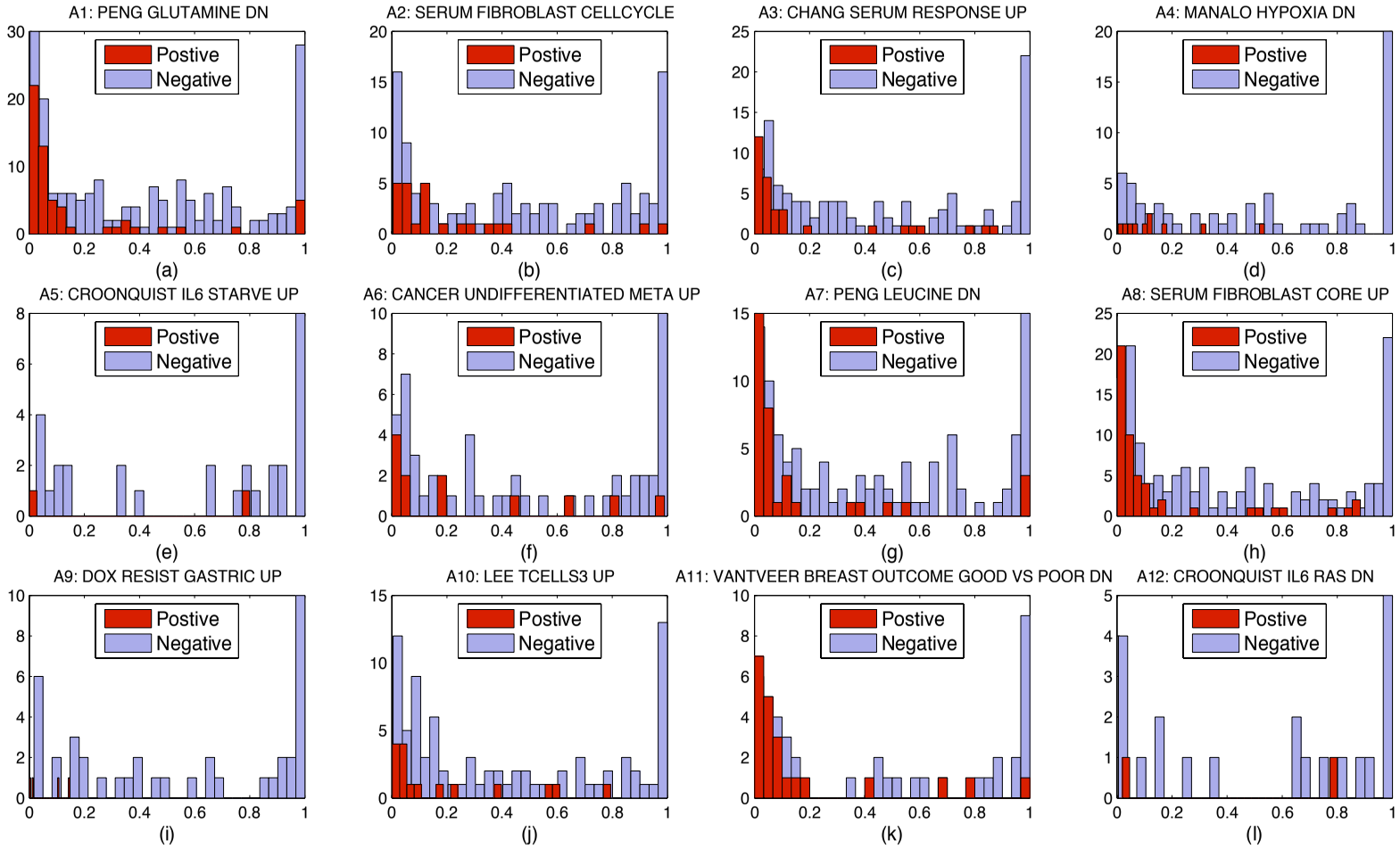


Figure 8.3: Association probability histograms of the lactic acidosis status related pathway gene sets identified by PROPA. For each gene set, the association probability histogram of the genes higher expressed in lactic acidosis-positive HMECs than in lactic acidosis-negative HMECs (i.e.positively correlated with lactic acidosis status) is in red, while that of the genes higher expressed in lactic acidosis-negative HMECs than in lactic acidosis-positive HMECs (i.e. negatively correlated with lactic acidosis status) is in blue.

Table 8.2: Signature pathways of lactic acidosis response in HMECs

Index	Pathway	Size	Genes
1	HMEC LA DN: PENG GLUTAMINE DN	52	ACAT2 AK2 C1QBP CCNB1 CDC25A CLIC1 CTPS DDX1 DHX9 DHCR7 DHCR24 DKC1 SLC29A1 ERF GCDH HRAS DNAJA1 HSPA8 LDHA LDHB LMNB1 NCL ODC1 PFN1 PHB PRODH PRPS1 PSMB6 PSMD1 SFRS1 SFRS2 SRM SURF5 TAP1 TCP1 TUBG1 UBE2N IFRD2 PIK3R3 SYNGR2 TRIP13 PTDSS1 SF3B4 NOL5A AHSA1 TXNIP RUVBL2 EBNA1BP2 ADRM1 ALG8 RRP9 TUBA4A
2	HMEC LA DN: SERUM FIBROBLAST CELLCYCLE	37	CDC6 CDC25A CENPA CKS2 FEN1 FOXM1 KIFC1 LMNB1 MCM4 MCM5 MCM6 PCNA PLK1 PRIM1 MAPK13 RFC2 RFC4 RRM2 EXO1 TRIP13 ABCC5 TUBB2C NCAPD3 NCAPH UBE2T UHRF1 DONSON TIPIN CDCA8 DKFZP762E1312 ASF1B CENPM FAM83D CDCA5 TUBB C13ORF3 TUBA4A
3	HMEC LA DN: CHANG SERUM RESPONSE UP	37	SLC25A5 CHEK1 EIF4EBP1 MCM3 MCM7 MT1F NUDT1 MYBL2 PFN1 PLAUR PLOD2 POLE2 PSMC3 PSMD2 SFRS2 SFRS10 SNRPB SNRPC SNRPD1 SRM TPI1 TUBG1 IFRD2 JTV1 RUVBL1 GGH PCSK7 RNASEH2A EBNA1BP2 TNFRSF12A CDCA4 NOLA2 DCLRE1B IPO4 LMNB2 MRTO4 TUBA4A
4	HMEC LA DN: MANALO HYPOXIA DN	26	ATIC CDC6 CTPS MCM2 MCM3 MCM4 MCM5 MCM6 POLE2 PRPS1 RFC2 RFC5 RRM2 SFRS2 SNRPD1 CDC7 PRC1 CCNE2 TRIP13 PSME3 NOL5A ZWINT ORC6L MCM10 MRPL40 RRP9
5	HMEC LA DN: CROONQUIST IL6 STARVE UP	19	BUB1B CCNB1 CDC6 CENPA CENPE CHEK1 FOXM1 KIF11 KIFC1 LIG1 MYBL2 RRM2 TK1 WEE1 CDC45L GINS1 ZWINT CSNK1G1 SPC25
6	HMEC LA DN: CANCER UNDIFFERENTIATED META UP	22	BIRC5 CCNB1 CDC6 CENPA CKS2 FOXM1 GARS MCM2 MCM3 MCM6 NUDT1 MYBL2 PCNA PSMD2 RFC4 TAP1 SLC7A5 GGH TRIP13 ADRM1 NCAPH TUBB
7	HMEC LA DN: PENG LEUCINE DN	31	ACAT2 CCNB1 DDX1 DHX9 DHCR24 SLC29A1 GCDH DNAJA1 HSPA8 LDHA LDHB MCM4 NCL PSMD1 PSMD2 SFRS2 SRM TCP1 TK1 UBE2N PIK3R3 TRIP13 EI24 SF3B4 NOL5A AHSA1 SPINT2 ADRM1 ALG8 DHRS1 TUBA4A
8	HMEC LA DN: SERUM FIBROBLAST CORE UP	44	SLC25A5 CHEK1 EIF4EBP1 MCM3 MCM7 MT1F NUDT1 MYBL2 PFN1 PLAUR PLOD2 POLE2 PSMC3 PSMD2 SFRS2 SFRS10 SNRPB SNRPC SNRPD1 IFRD2 JTV1 RUVBL1 GGH PCSK7 RNASEH2A RPP40 RUVBL2 EBNA1BP2 DNAJC9 C12ORF24 TNFRSF12A CDCA4 NOLA2 DCLRE1B WDR77 IPO4 C19ORF48 LMNB2 NUP35 PTPLB HYLS1 C18ORF24 MRTO4 TUBA4A
9	HMEC LA DN: DOX RESIST GASTRIC UP	16	CDC6 CENPA FOXM1 LMNB1 MCM4 PLK1 RFC4 RRM2 PRC1 GINS1 KIF4A NCAPG2 CDCA8 DKFZP762E1312 ASF1B CENPM
10	HMEC LA DN: LEE TCELLS3 UP	23	BIRC5 BUB1B CCNB1 CDC25A CENPA CKS2 KIF11 MCM2 MCM4 RRM2 GGH PRC1 GINS1 KIF20A TUBB2C UBE2T UHRF1 DTL FAIM NEIL3 MCM10 CDCA3 TUBB
11	HMEC LA DN: VANTVEER BREAST OUTCOME GOOD VS POOR DN	18	BIRC5 CENPA CKS2 CTPS MCM6 PSMD2 RFC4 RRM2 TK1 PIR GGH PRC1 CCNE2 TRIP13 PTDSS1 ORC6L DKFZP762E1312 VEGFA
12	HMEC LA DN: CROONQUIST IL6 RAS DN	9	BUB1B CCNB1 CDC6 CENPA FOXM1 MCM3 MYBL2 WEE1 SPC25

To demonstrate the application of PROPA in such type of analysis, I focus on five factors and use PROPA to identify the pathway activities under each of them. To be consistent with the description given by Lucas *et al.* (2007), I maintain their original indices: factor 3, factor 2, factor 7, factor 9 and factor 6. In addition to the 956 gene sets from MSigDB and the nine curated oncogenic pathway signature gene sets, the 12 signature pathways of HMEC lactic acidosis response, as listed in Table 8.2, are added in this annotation analysis.

8.2.2 Factor Pathway Annotation

Here I exhibit some interesting moleculer phenotypes of breast cancers identified by PROPA that are potentially related to lactic acidosis cellular response. To show the identification of each factor, I present the top 30 pathways ranked by PROPA without using any cut-offs. In contrast to the earlier analyses, the goal of this annotation analysis for each specific latent factor is to inversely identify a tumor factor phenotype rather than explore the pathway activity under a known phenotype.

Sometimes, the biological themes represented by a factor are not easily recognizable due to the limitation of sample size. Each tumor sample is a combination of numerous entangled molecular phenotypes. Although dissecting this complexity is exactly the purpose of the analysis, the tumor sample size is almost always limited. One or two hundred sample is not large when dealing with the complexity of tumor progression. Some strong/well distinguishable phenotypes, such as ER status in breast cancers, are easier to detect, but some others are vague. One needs to examine the top pathway gene sets, and see whether they are enriched for a certain biological theme. In this sense, the phenotype identification relies on the evaluation of pathway sets enrichment.

Factor 3: Inverse lactic acidosis response factor

Table 8.3 lists the top 30 pathway gene sets associated with factor 3. As can be seen, among the top pathways, HMEC lactic acidosis response signature pathways are highly enriched. Totally eight of the 12 show up in the top 30 pathway list, while six of them are in top eight. Figure 8.4 presents the association probability histograms of each top pathway gene sets. The member genes of each signature pathways, as shown in the previous section, are downregulated in HMECs in presence of lactic acidosis. Here these genes concordantly have positive correlation with factor 3. Hence, factor 3 is believed to inversely represent the lactic acidosis response phenotype of the tumor samples.

A set of gene sets in this top list – A_{15}, A_{18}, A_{19}, A_{26} and A_{29} – are related to the activation of the p21 pathway, either independent or dependent on the p53 tumor supressor gene (Wu *et al.* 2002). p21 has been reported as a cell cycle regulator that can mediate the p53-dependent cell cycle G1 and G2/M phase arrest and apoptosis of tumor cells in response to stress stimuli. The expression of p21 gene is tightly controlled by protein p53. The genes included in these gene sets are downregulated following transduction of p21 in ovarian cancer cells, but in general are positively correlated with factor 3. It may implies a positive correlation between tumor lactic acidosis response and cell apoptosis.

Factor 2: ER/good prognosis factor

Figure 8.5 shows the association probability distributions of the pathway gene sets that are most associated with factor 2 (Table 8.4) and can be biologically interpreted. Several pathway gene sets predefined by van't Veer *et al.* (2002) are identified. These gene sets have been mentioned in Section 7.2 in pathway annotation for breast tumor ER phenotype, the status of which is highly associated with breast cancer outcomes.

The top two gene sets A_1 and A_2 contain genes lower-expressed and higher-expressed, respectively, in ER positive tumors vs. ER negative tumors. A_{11} and A_{18} contain genes lower-expressed and higher-expressed, respectively, in breast tumors with vs. without BRCA1 mutants, which are highly associated with ER phenotype. The expression of genes in A_{14} and A_{27} are negatively and positively correlated with breast cancer good prognosis, respectively. Another prognosis-related gene set A_{17} defined by Rhodes *et al.* (2004) contains genes higher-expressed in undifferentiated cancers vs. well-differentiated cancers. The significant association with these gene sets gives substantial evidence that factor 2 is positively linked to breast tumor ER phenotype and good prognosis.

Some oncogenic pathway gene sets are identified being associated with factor 2, including Ras, p110-α, E2F3, Myc and β-catenin pathways. Although Ras mutations are infrequent in breast cancers (less than 5%), considerable evidence suggests that Ras pathways are deregulated in breast cancer cells (Clark and Der 1995). It has been reported that gene Ha-ras, the normal prototype of Ras, is over-expressed in 50% breast cancers, contributing to tumorigenesis. Pethe and Shekhar (1999) have shown evidence of the existence of estrogen-mediated Ha-ras upregulation in breast tumor cells. p110-α (or PI3K) activity has been reported to be associated with the activation of ER pathway in breast cancers (Fry 2001; Baldi *et al.* 1986; Simoncini *et al.* 2000; Ahmad *et al.* 1999; Razandi M 2000; Campbell *et al.* 2001). Myc pathway controls cell proliferation and cell fate decisions. Sears *et al.* (2000) have demonstrated that Ras can enhance the accumulation of Myc activity by stabilizing a normally short-lived Myc protein. Leone *et al.* (2001) have shown that the Myc transcription factor induces transcription of the E2F3 gene, a member of E2F transcription factor family, whose activities are an essential component of the Myc pathway. The concurrence of these oncogenic pathways in the ER factor may be due

to such interactions. Further examination of the correlation (positive or negative) between these oncogenic pathway genes and the factor will give the confirmation of whether these pathways are activated or inactivated in the factor.

The identification of this factor links tumor lactic acidosis response to ER status, cancer prognosis as well as oncogenic pathways. It has been mentioned that oncogenic pathways may play a role in tumor glycolytic phenotype formation, and it is also possible that tumor cells response to lactic acidosis through evolution including oncogenic mutation. These oncogenic mutation then affect the ER pathway. Of course, it is also possible that lactic acidosis directly perturbs the ER pathway, which can be a testable biological hypothesis.

Factor 7 and factor 9: Immune function factors

Factor 7 and factor 9 are enriched for immune function pathway gene sets (Table 8.5 & 8.6, Figure 8.6 & 8.7). For example, the top gene set in both factors contains immune function genes relevant to cancer development (Brentani *et al.* 2003); A_2 in factor 7 is a collection of genes whose upregulation are associated with transplanted kidney rejection (Flechner *et al.* 2004); A_6 in factor 7 or A_4 in factor 9 contains cytolytic effector genes induced during antigen activation of CD8+ T cells (Goldrath *et al.* 2004). A majority of pathway gene sets in these two factors are related to immune activities in many different contexts. Generally, the over-expression of the genes in those immune activation pathways are negatively correlated with factor 7 and factor 9, suggesting that these two factors reflect inhibitory immune function status of the tumor samples.

The identification of the factors related to immune function inhibition implies the connection between tumor lactic acidosis response and immune activity perturbation, a key process in cancer development. A recent study by Fischer *et al.* (2007) has

provided evidence for the inhibitory effect of lactic acid on tumor-reactive T cell proliferation and cytokine production. The cytotoxtic/immune activity of T cells rely on glycolysis and efficient secretion of lactic acid. The increase of extracellular lactic acid concentration diminishes the pH gradient between cytoplasmic and extracellular environment and blocks lactic export in T cells, thereby results in impaired immune function.

Factor 6: TGF-β induced EMT factor

Factor 6 is enriched for pathways related to epithelial-to-mesenchymal transitions (EMTs), especially, transforming growth factor-β (TGF-β) pathways (Table 8.7, Figure 8.8). EMT is a process involved in wound healing whereby fully differentiated epithelial cells undergo transition to a mesenchymal phenotype giving rise to fibroblasts and myofibroblasts (Vincent-Salomon and Thiery 2003). TGF-β is a family of multifunctional cytokines that plays an important role in the regulation of epithelial cell growth, differentiation and apoptosis. Evidence show that TGF-β inhibits epithelial cell growth in early stage of breast tumorigenesis, and induces EMT in a later stage of carcinogenesis (Knabbe *et al.* 1987; Colletta *et al.* 1991; Dannecker *et al.* 1996; Rodriguez *et al.* 2002; Zavadil *et al.* 2004). A_{19} is a upregulated signature gene set of EMT in tumor regression (Jechlinger *et al.* 2003). A_1, A_4, A_{10}, A_{11}, A_{18} and A_{30} together include genes induced by TGF-β in skin fibroblasts (Verrecchia *et al.* 2001). These gene signatures in general are positively correlated with factor 6, suggesting that factor 6 may correspond to TGF-β induced EMT phenotype in tumors. The factor analysis thus links together the tumor lactic acidosis response and TGF-β/EMT phenotype, a connection which has not been made before. It provides a foundation for a novel and testable hypothesis that lactic acidosis has an effect on breast tumor EMT through changing the expression of TGF-β. Moreover, it is important to note that hypoxia has been shown to induce TGF-β (Falanga *et al.*

2002), trigger the EMT (Manotham *et al.* 2004), and link to poor clinical outcomes (Chi *et al.* 2006). Hence, the regulation effect of lactic acidosis on TGF-β and EMT, if can be validated, may consequently provides an explanation for its link with cancer prognosis.

8.2.3 Summary

This tumor analysis takes advantage of the heterogeneity in a large set of tumor samples to dissect the pathway activities involved in cancerous cellular response to a biological intervention. Each factor derived from this analysis represents a molecular phenotype that may be relevant to the intervention. PROPA plays a central role in identifying these molecular phenotypes. In this lactic acidosis response dissection example, it is found that lactic acidosis may be a signal to breast cancer ER, immune, and TGF-β induced EMT pathways. The ER pathway is a prominent phenotype of breast cancers that involves complex pathway activities. This analysis connects the ER pathway activation with some oncogenic pathway through the lactic acidosis response signature. The detection of immune function pathway provides evidence for the role of lactic acidosis in immune cell function perturbation during cancer development. A more intriguing finding of the TGF-β induced EMT pathway forms a more concrete hypothesis that can be tested through experimental investigation.

This study demonstrates how to define a "pathway signature" for a phenotype through gene set refinement, and how to use this pathway signature to identify the molecular phenotype represented by a derived factor. A pathway signature is a set of pathway gene sets. If one believes each biological phenotype can be represented by a specific set of pathways, then a unknown molecular phenotype can be identified through the evaluation of pathway set enrichment. Inspired by pathway annotation, I call this approach *phenotype annotation*. PROPA ranks all the pathways in the

database according to their posterior probabilities of association with a factor. By using a non-parametric method similar to those for gene set-based pathway annotation, one can compute the enrichment of the pathway set representing a known phenotype, thereby identifying the unknown phenotype and its association with the known phenotypes. Although not quantitively, this approach has been used to identify the latent phenotypes in the tumor samples.

This application also demonstrates how PROPA works with BFRM to dissect the complexity of tumor phenotype in tumorigenesis studies. PROPA is based on likelihood evaluation, and can directly compare the phenotype association of all pathways based on gene-phenotype association. In contrast, GSEA is limited in such applications – it depends on sample permutation to generate null distribution and so loses power when sample size is small. This limitation is shown in the pathway annotation for the HMEC lactic acidosis response. The sample size is 12, hence, the annotation output is generally informative. Moreover, GSEA is not applicable in the multiple regression setting in the tumor lactic acidosis response decomposition by BFRM.

Table 8.3: Top 30 pathways associated with factor 3 in Miller breast tumors lactic acidosis response analysis

Index	Pathway	Description	Size
1	HMEC LA DN: SERUM FIBROBLAST CELL-CYCLE	HMEC lactic acidosis signature pathway	35
2	DOX RESIST GASTRIC UP	Upregulated in gastric cancer cell lines reistant to doxorubicin, compared to parent chemosensitive lines	43
3	HMEC LA DN: LEE TCELLS3 UP	HMEC lactic acidosis signature pathway	22
4	CROONQUIST IL6 STARVE UP	Genes upregulated in multiple myeloma cells exposed to the pro-proliferative cytokine IL-6 versus those that were IL-6-starved.	32
5	HMEC LA DN: CANCER UNDIFFERENTIATED META UP	HMEC lactic acidosis signature pathway	20
6	HMEC LA DN: DOX RESIST GASTRIC UP	HMEC lactic acidosis signature pathway	16
7	HMEC LA DN: CROONQUIST IL6 STARVE UP	HMEC lactic acidosis signature pathway	18
8	HMEC LA DN: VANTVEER BREAST OUTCOME GOOD VS POOR DN	HMEC lactic acidosis signature pathway	16
9	DAC FIBRO DN	Downregulated by DAC treatment in LD419 fibroblast cells	11
10	CROONQUIST IL6 RAS DN	Genes dowmregulated in multiple myeloma cells exposed to the pro-proliferative cytokine IL-6 versus those with N-ras-activating mutations.	22
11	HMEC LA DN: CROONQUIST IL6 RAS DN	HMEC lactic acidosis signature pathway	8
12	LAMB CYCLIN D3 GLOCUS	E2F target genes highly correlated with cyclin D3 expression (p = 0.002)	14
13	MIDDLEAGE DN	Downregulated in fibroblasts from middle-age individuals, compared to young	13
14	GOLDRATH CELLCYCLE	Cell cycle genes induced during antigen activation of CD8+ T cells	28
15	P21 P53 MIDDLE DN	Down-regulated at intermediate timepoints (12-16 hrs) following ectopic expression of p21 (CDKN1A) in OvCa cells, p53-dependent	25
16	ZHAN MM CD138 PR VS REST	50 top ranked SAM-defined over-expressed genes in each subgroup PR	43
17	BREAST DUCTAL CARCINOMA GENES	Genes upregulated in breast tumors that are identified as high tumor grade and that are progressing from preinvasive ductal carcinoma in situ (DCIS) to invasive ductal carcinoma (IDC)	18
18	P21 P53 EARLY DN	Down-regulated at early timepoints (4-8 hrs) following ectopic expression of p21 (CDKN1A) in OvCa cells, p53-dependent	12
19	P21 MIDDLE DN	Down-regulated at intermediate timepoints (12-16 hrs) follwing ectopic expression of p21 (CDKN1A) in OvCa cells	15
20	CANCER UNDIFFERENTIATED META UP	Sixty-nine genes commonly upregulated in undifferentiated cancer relative to well-differentiated cancer, from a meta-analysis of the OncoMine gene expression database	66
21	KAMMINGA EZH2 TARGETS	Putative targets or partners of Ezh2 in HSCs	33
22	HMEC LA DN: MANALO HYPOXIA DN	HMEC lactic acidosis signature pathway	26
23	OXSTRESS BREASTCA DN	Downregulated by H2O2, Menadione and t-BH in breast cancer cells	10
24	MRNA BINDING ACTIVITY	Genes with mRNA binding activity	9
25	CHEOK MP UP	Genes upregulated by mercaptopurine (MP) treatment	5
26	P21 P53 ANY DN	Down-regulated at any timepoint (4-24 hrs) following ectopic expression of p21 (CDKN1A) in OvCa cells, p53-dependent	48
27	XU ATRA PLUSNSC DN	Examples of genes down-regulated by ATRA with further enhancement of effect due to addition of NSC	14
28	CMV IE86 UP	Upregulated by expression of cytomegalovirus IE86 protein in primary human fibroblasts	50
29	P21 EARLY DN	Down-regulated at early timepoints (4-8 hrs) follwing ectopic expression of p21 (CDKN1A) in OvCa cells	13
30	CHEOK MP DN	Genes downregulated by mercaptopurine (MP) treatment	4

Table 8.4: Top 30 pathways associated with factor 2 in Miller breast tumors lactic acidosis response analysis

Index	Pathway	Description	Size
1	BRCA ER NEG	Genes whose expression is consistently negatively correlated with estrogen receptor status in breast cancer - higher expression is associated with ER-negative tumors	906
2	BRCA ER POS	Genes whose expression is consistently positively correlated with estrogen receptor status in breast cancer - higher expression is associated with ER-positive tumors	519
3	LEE TCELLS2 UP	Transcripts enriched in more mature cells (SP4, CB4, and AB4) more than 3-fold, with average signal value differences of at least 100 between less mature (ITTP, DP) and more mature (SP4, CB4, and AB4) cells	1088
4	ALZHEIMERS DISEASE UP	Upregulated in correlation with overt Alzheimer's Disease, in the CA1 region of the hippocampus	1445
5	ALZHEIMERS DISEASE DN	Downregulated in correlation with overt Alzheimer's Disease, in the CA1 region of the hippocampus	1111
6	RAS1 pathway signature		308
7	P110 pathway signature		222
8	E2F3 pathway signature		150
9	FLECHNER KIDNEY TRANSPLANT REJECTION DN	Genes downreglated in acute rejection transplanted kidney biopsies relative to well functioning transplanted kidney biopsies from stable, immunosuppressed recipients.	527
10	MYC1 pathway signature		149
11	BRCA BRCA1 POS	Genes whose expression is consistently positively correlated with brca1 germline status in breast cancer - higher expression is associated with sporadic tumors	102
12	HINATA NFKB UP	Genes upregulated by NF-kappa B	107
13	SERUM FIBROBLAST CELLCYCLE	Cell-cycle dependent genes regulated following exposure to serum in a variety of human fibroblast cell lines	135
14	BRCA PROGNOSIS NEG	Genes whose expression is consistently negatively correlated with breast cancer outcomes - higher expression is associated with metastasis and poor prognosis	95
15	LINDSTEDT DEND UP	Genes up-regulated in DC stimulated for 8 and 48 h	50
16	LINDSTEDT DEND 8H VS 48H UP	Genes up-regulated in DC stimulated for 8 h as compared to DC stimulated for 48 h	64
17	CANCER UNDIFFERENTIATED META UP	Sixty-nine genes commonly upregulated in undifferentiated cancer relative to well-differentiated cancer, from a meta-analysis of the OncoMine gene expression database	66
18	BRCA BRCA1 NEG	Genes whose expression is consistently negatively correlated with brca1 germline status in breast cancer - higher expression is associated with BRCA1 tumors	154
19	FLECHNER KIDNEY TRANSPLANT REJECTION UP	Genes upreglated in acute rejection transplanted kidney biopsies relative to well functioning transplanted kidney biopsies from stable, immunosuppressed, recipients	81
20	IRITANI ADPROX VASC	BLOOD VASCULAR EC	148
21	KLEIN PEL DN	Genes downregulated in AIDS-related primary effusion lymphoma (PEL) cells compared to normal B cells and other tumor subtypes.	57
22	BREASTCA TWO CLASSES	Gene set that can be used to differentiate BRCA1-linked and BRCA2-linked breast cancers	136
23	UVB NHEK3 C7	Regulated by UV-B light in normal human epidermal keratinocytes, cluster 7	51
24	BCAT pathway signature		38
25	CARIES PULP HIGH UP	Highly up-regulated (¿4-fold) in pulpal tissue from extracted carious teeth (cavities), compared to tissue from extracted healthy teeth	91
26	ZHAN MM CD1 VS CD2 DN	158 genes commonly dysregulated in CD-1 and CD-2 groups downregulated	49
27	BRCA PROGNOSIS POS	Genes whose expression is consistently positively correlated with breast cancer outcomes - higher expression is associated with good prognosis	40
28	ROSS MLL FUSION	Genes that distinguish pediatric acute myeloid leukemia (AML) subtypes with MLL chimeric fusion genes.	74
29	CARIES PULP UP	Up-regulated in pulpal tissue from extracted carious teeth (cavities), compared to tissue from extracted healthy teeth	202
30	VERHAAK AML NPM1 MUT VS WT DN	Description Genes that are downregulated in AML NPM1 mutant versus AML NPM1 wild type	240

Table 8.5: Top 30 pathways associated with factor 7 in Miller breast tumors lactic acidosis response analysis

Index	Pathway	Description	Size
1	BRENTANI IMMUNE FUNCTION	Cancer related genes involved in immune function	50
2	FLECHNER KIDNEY TRANSPLANT REJECTION UP	Genes upreglated in acute rejection transplanted kidney biopsies relative to well functioning transplanted kidney biopsies from stable, immunosuppressed, recipients	81
3	ROTH HTERT UP	Genes significantly up-regulated in cells that overexpress hTERT	14
4	YANG OSTECLASTS SIG	Relative gene expression for osteoclast-associated genes, chemokines, and chemokine receptors	37
5	BENNETT SLE UP	Genes Significantly up-regulated in SLE patient Blood Mononuclear Cells	27
6	GOLDRATH CYTOLYTIC	Genes associated with cytolytic effector function induced during antigen activation of CD8+ T cells.	16
7	DAC IFN BLADDER UP	Interferon-regulated genes upregulated by DAC treatment in T24 bladder carcinoma cells	16
8	GRANDVAUX IFN NOT IRF3 UP	Genes up-regulated by interferon-alpha,beta but not by IRF3 in Jurkat (T cell)	14
9	ZHAN MMPC PC	Microarray-derived expression levels of genes differentially expressed during PC development	21
10	CMV HCMV TIMECOURSE 12HRS UP	Up-regulated in fibroblasts following infection with human cytomegalovirus	26
11	INFLAMMATORY RESPONSE PATHWAY	Inflammatory response pathway	29
12	ROTH HTERT DN	Genes significantly down-regulated in cells that overexpress hTERT	4
13	LINDSTEDT DEND UP	Genes up-regulated in DC stimulated for 8 and 48 h	50
14	DAC BLADDER UP	Upregulated by DAC treatment in T24 bladder carcinoma cells	28
15	CMV 8HRS UP	Upregulated at 8hrs following infection of primary human foreskin fibroblasts with CMV	31
16	SARCOMAS HISTIOCYTOMA UP	Top 20 positive significant genes associated with malignant fibrous histiocytomas, versus other soft-tissue tumors	13
17	TNFA NFKB DEP UP	Up-regulated at any timepoint following TNFa treatment, only with functional NFkB	18
18	HUMAN TISSUE SALIVARY	Genes expressed specifically in human salivary gland tissue	12
19	DER IFNA UP	Genes up-regulated by interferon-alpha in HT1080 (fibrosarcoma)	65
20	IFNALPHA NL HCC UP	Upregulated by interferon alpha treatment in both normal primary hepatocytes and Hep3B hepatocellular carcinoma cells	18
21	IFNA UV-CMV COMMON HCMV 6HRS UP	Up-regulated in fibroblasts at 6 hours following either infection with UV-inactivated CMV or interferon-alpha	28
22	NF90 UP	Upregulated by ectopic expression of NF90 in GHOST(3)CXCR4 cells	24
23	GRANDVAUX IRF3 UP	Genes up-regulated by IRF3 in Jurkat (T-cell)	13
24	GREENBAUM E2A DN	Transcripts down-regulated 3-fold or greater in the E2A-deficient cell lines	17
25	IFNA HCMV 6HRS UP	Up-regulated in fibroblasts at 6 hours following treatment with interferon-alpha	52
26	HUMAN TISSUE KIDNEY	Genes expressed specifically in human kidney tissue	11
27	KUROKAWA 5FU IFN SENSITIVE VS RESISTANT UP	Genes highly expressed in hepatocellular carcinoma sensitive to 5-Fluorouracil + interferon	5
28	IFNALPHA NL UP	Upregulated by interferon alpha treatment in normal primary hepatocytes	27
29	MOREAUX TACI BAFF UP	Genes increased by BAFF/APRIL deprivation in primary myeloma cells common with the TACI-related gene signature	6
30	HEDVAT ELF DN	MEF Regulates IL-8 Expression, genes down-regulated by more than 2 fold	3

Table 8.6: Top 30 pathways associated with factor 9 in Miller breast tumors lactic acidosis response analysis

Index	Pathway	Description	Size
1	BRENTANI IMMUNE FUNCTION	Cancer related genes involved in immune function	50
2	MYC1 pathway signature		149
3	TSA CD4 DN	Down-regulated in mouse CD4+ T-cells following 4 hour treatment with 100 nM trichostatin A	18
4	GOLDRATH CYTOLYTIC	Genes associated with cytolytic effector function induced during antigen activation of CD8+ T cells	16
5	ZHAN MMPC PC	Microarray-derived expression levels of genes differentially expressed during PC development	21
6	LEE TCELLS4 UP	Transcripts enriched in more mature cells (SP4, CB4, and AB4) more than 3-fold, with average signal value differences of at least 100 between less mature (ITTP, DP) and more mature (SP4, CB4, and AB4) cells	55
7	GOLUB ALL VS AML UP	Genes highly correlated with acute lymphoblastic leukemia	19
8	FSH HUMAN GRANULOSA UP	Up-regulated by FSH in primary human granulosa cells	14
9	LIZUKA L1 SM G1	Genes highly expressed in well differentiated hepatocellular carcinoma vs. non-tumor liver with hepatits	10
10	FERRANDO T CELL DIFFERENTIATION PATHWAY	Genes involved with T cell differentiation which are upregulated in samples positive for HOX11, TAL1, or LYL1 by RT-PCR	15
11	HUMAN TISSUE THYMUS	Genes expressed specifically in human thymus tissue	16
12	SRC1 pathway signature		53
13	FERRANDO TAL1 NEIGHBORS	Genes associated with the expression of the T-cell oncogene TAL1+	12
14	HUMAN TISSUE SALIVARY	Genes expressed specifically in human salivary gland tissue	12
15	PITUITARY FETAL DN	Down-regulated in human fetal pituitary tissue, compared to adult pituitary tissue	12
16	ROTH HTERT DN	Genes significantly down-regulated in cells that overexpress hTERT	4
17	MOREAUX TACI XG 13 UP	Genes increased by BAFF/APRIL in XG-13 HMCL common with the TACI-related gene signature	7
18	FERRANDO MLL T ALL UP	Top 100 nearest neighbor genes positively associated with MLL T-ALL cases	83
19	LINDSTEDT DEND 8H VS 48H UP	Genes up-regulated in DC stimulated for 8 h as compared to DC stimulated for 48 h	64
20	INFLAMMATORY RESPONSE PATHWAY	Inflammatory response pathway	29
21	TSADAC HYPOMETH HYPERAC OVCA UP	Genes with basally hypomethylated promoters upregulated by the combination of TSA and DAC in ovarian carcinoma (CP70) cells, with hyperacetylated promoters upon activation	14
22	ZHAN MMPC LATEVS	Late differentiation genes top 50 differentially expressed genes in comparison of CD138-enriched tonsil PCs and CD138-enriched bone marrow PCs	45
23	HASLINGER B CLL MUTATED	Top 25 Genes Differentially Expressed in the Respective B-Cell Lymphocytic Leukemia Subgroups VH-mutated-unmutated	14
24	MENSE HYPOXIA DN	List of Hypoxia-suppressed genes found in both Astrocytes and HeLa Cells	4
25	TCA	Tricarboxylic acid related genes	15
26	BCNU GLIOMA NOMGMT 48HRS UP	Up-regulated in an MGMT-deficient glioma cell line (A172) at 48 hours following treatment with BCNU	18
27	GPCRS CLASS A RHODOPSIN LIKE 2	GPCR class A rhodopsin genes	10
28	UV ESR WS UNREG	Genes involved in the environmental stress response that were not regulated following treatment of Werner syndrome fibroblasts with UV light	10
29	KUROKAWA 5FU IFN SENSITIVE VS RESISTANT UP	Genes highly expressed in hepatocellular carcinoma sensitive to 5-Fluorouracil + interferon.	5
30	HEATSHOCK OLD UP	Upregulated after heat shock in lymphocytes from old individuals, compared to young	13

Table 8.7: Top 30 pathways associated with factor 6 in Miller breast tumors lactic acidosis response analysis

Index	Pathway	Description	Size
1	TGFBETA ALL UP	Upregulated by TGF-beta treatment of skin fibroblasts, at any timepoint	79
2	CROONQUIST IL6 STROMA UP	Genes upregulated in multiple myeloma cells exposed to the pro-proliferative cytokine IL-6 versus those co-cultured with bone marrow stromal cells	37
3	CROONQUIST RAS STROMA DN	Genes downregulated in multiple myeloma cells with N-ras-activating mutations versus those co-cultured with bone marrow stromal cells	20
4	TGFBETA EARLY UP	Upregulated by TGF-beta treatment of skin fibroblasts at 30 min	46
5	AGEING KIDNEY SPECIFIC UP	Up-regulation is associated with increasing age in normal human kidney tissue from 74 patients, and expression is higher in kidney than in whole blood	181
6	ADIP HUMAN DN	Down-regulated in primary human adipocytes, versus preadipocytes	27
7	CORDERO KRAS KD VS CONTROL UP	Genes upregulated in kras knockdown vs control in a human cell line	75
8	VEGF MMMEC 6HRS UP	Up-regulated at 6hrs follwing VEGF treatment of human myometrial microvascular endothelial cells	50
9	TSA HEPATOMA CANCER UP	Cancer-related genes up-regulated in any of four human hepatoma cell lines following 24-hour treatment with 200ng/mL of trichostatin A	39
10	TGFBETA C1 UP	Upregulated by TGF-beta treatment of skin fibroblasts	17
11	TGFBETA C2 UP	Upregulated by TGF-beta treatment of skin fibroblasts	17
12	GILDEA BLADDER UP	Top 30 genes differentially expressed in metastatic (T24T) and nonmetastatic (T24) human bladder cancer cell lines	29
13	ZELLER MYC DN	Genes down-regulated by MYC in > 3 papers	7
14	DORSEY DOXYCYCLINE UP	Genes upregulated by doxycycline in K562-derived Gab2WT-6 and Gab2WT-7 cells	29
15	CMV 24HRS DN	Downregulated at 24hrs following infection of primary human foreskin fibroblasts with CMV	71
16	IL6 SCAR FIBRO DN	Downregulated following IL-6 treatment in hypertrophic scar fibroblasts	10
17	O6BG RESIST MEDULLOBLASTOMA UP	Up-regulated in a medulloblastoma cell line resistant to both BCNU and O6-BG, compared to sensitive parent line	23
18	TGFBETA C4 UP	Upregulated by TGF-beta treatment of skin fibroblasts	11
19	JECHLINGER EMT UP	Genes upregulated for epithelial plasticity in tumor progression	54
20	BAF57 BT549 UP	Up-regulated following stable re-expression of BAF57 in Bt549 breast cancer cells that lack functional BAF57	236
21	SARCOMAS HISTIOCYTOMA UP	Top 20 positive significant genes associated with malignant fibrous histiocytomas, versus other soft-tissue tumors	13
22	MKK6EE UP	Upregulated by expression of constitutively active MKK6	10
23	KANNAN P53 DN	Target genes down regulated by p53	16
24	SCHRAETS MLL UP	Expression profile of Mll wild-type cells Top 40 list includes Genbank accession codes (GAC#), gene names (target gene), and log2 factors of differential gene expression	33
25	INFLAMMATORY RESPONSE PATHWAY	Inflammatory response pathway	29
26	CMV HCMV TIMECOURSE 10HRS UP	Up-regulated in fibroblasts following infection with human cytomegalovirus (at least 3-fold, with Affymetrix change call, in at least two consectutive timepoints), with maximum change at 10 hours	11
27	IRITANI ADPROX DN	BEC-specific suppressed by AdProx-1	59
28	AS3 FIBRO DN	Downregulated by sodium arsenite in fibroblasts	31
29	RORIE ES PNET UP	The 30 genes showing the greatest increase in expression in NBa Ews/Fli-1 infectants	26
30	TGFBETA LATE UP	Upregulated by TGF-beta treatment of skin fibroblasts only at 1-4 hrs	33

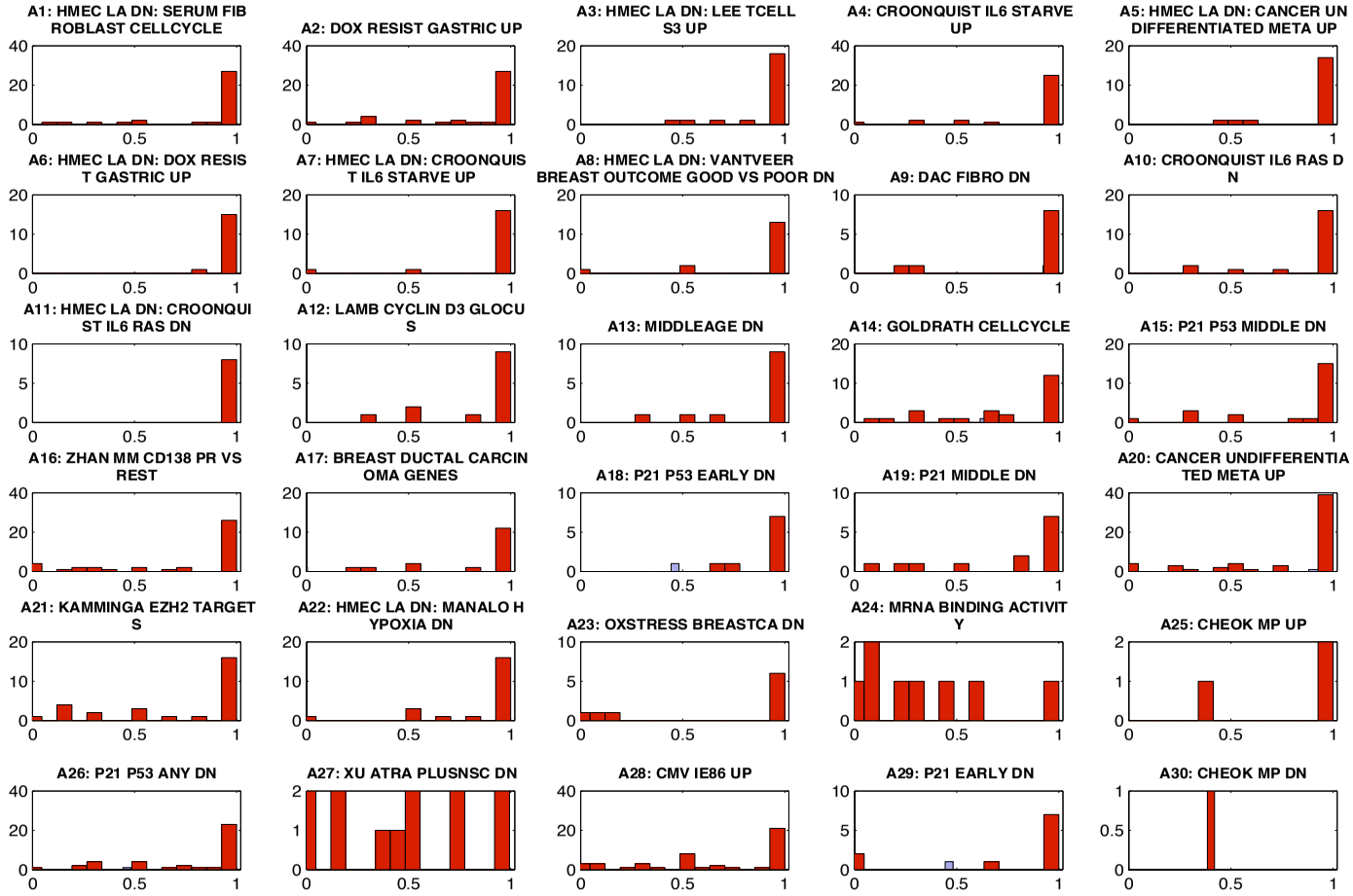


Figure 8.4: Association probability histograms of top 30 pathway gene sets associated with factor 3 (positive correlation in red, negative correlation in blue).

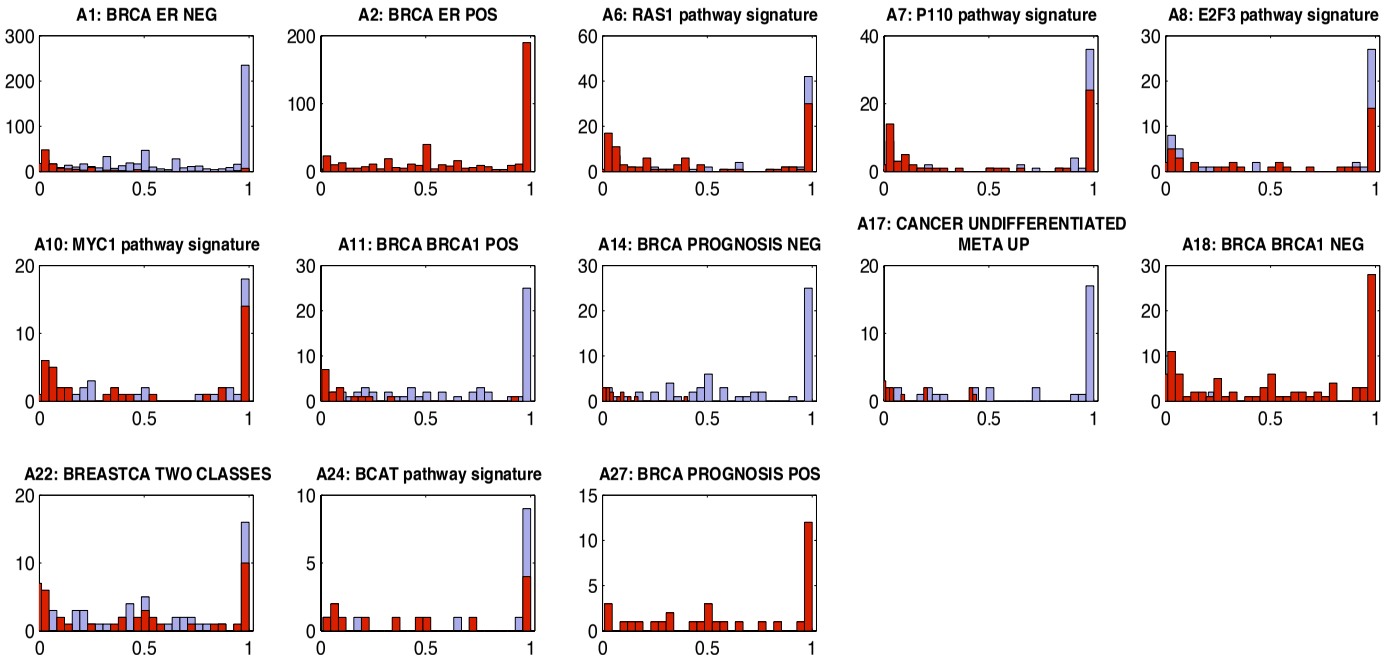


Figure 8.5: Association probability histograms of top pathway gene sets associated with factor 2 (positive correlation in red, negative correlation in blue).

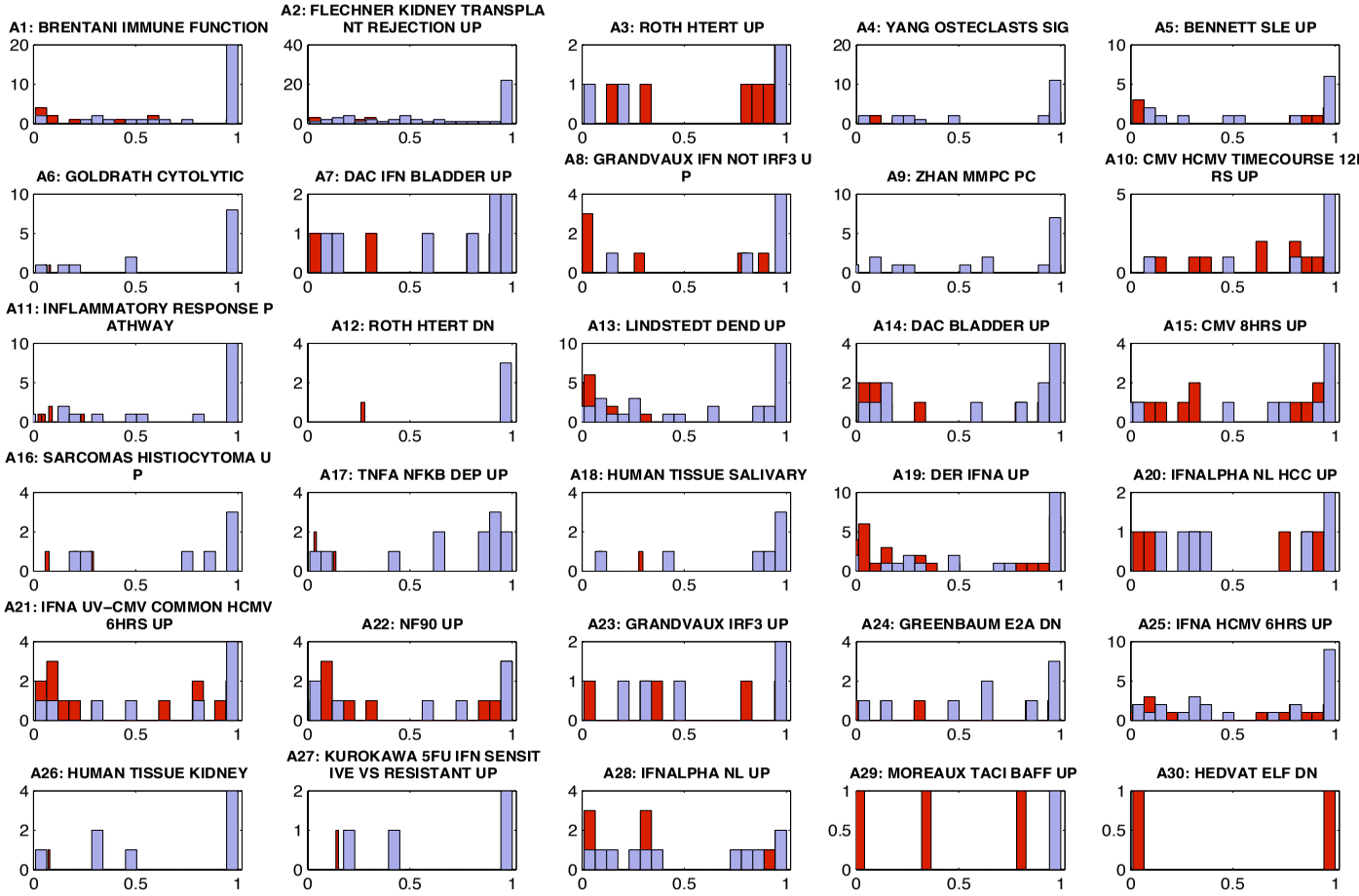


Figure 8.6: Association probability histograms of top 30 pathway gene sets associated with factor 7 (positive correlation in red, negative correlation in blue).

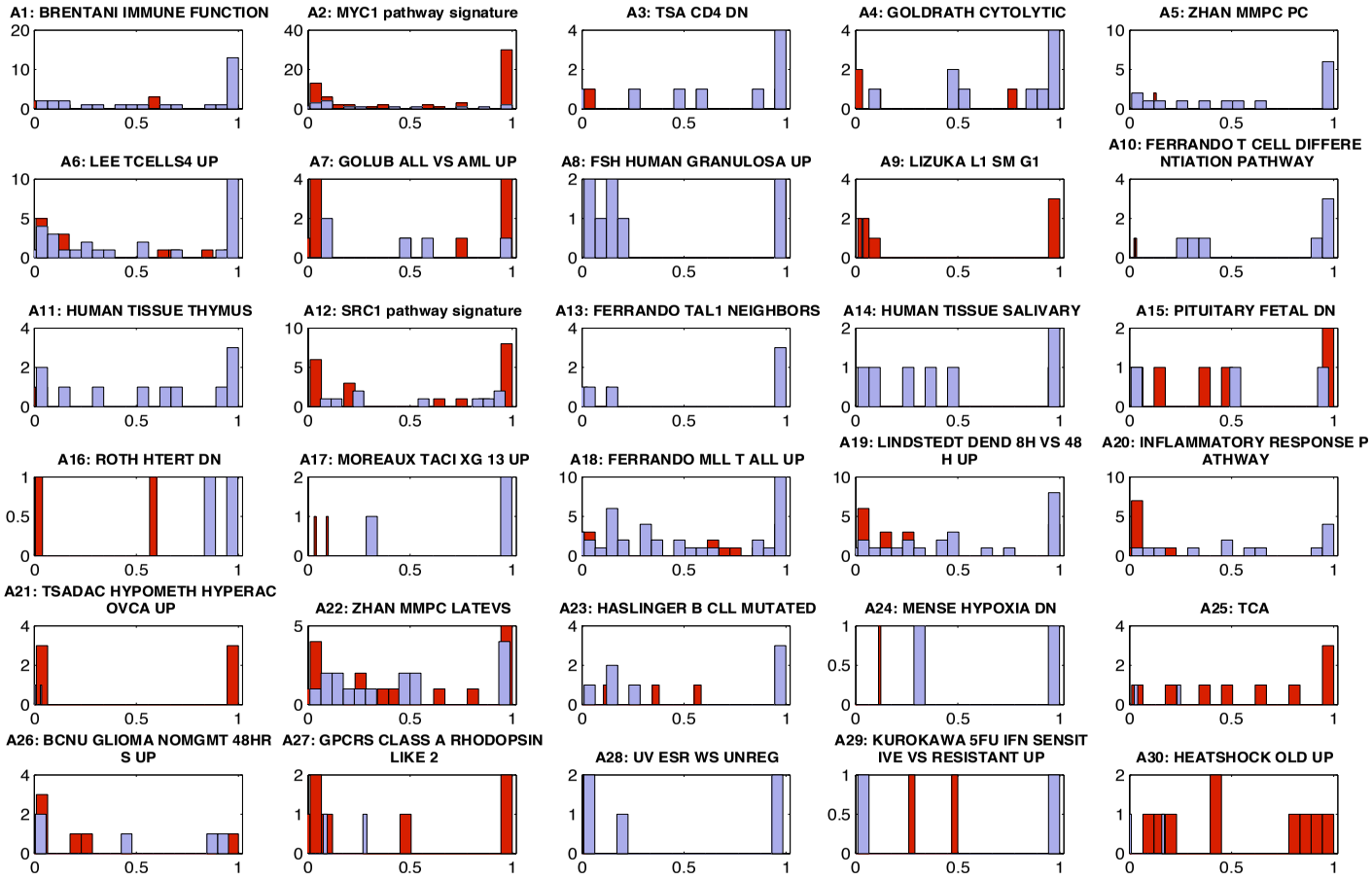


Figure 8.7: Association probability histograms of top 30 pathway gene sets associated with factor 9 (positive correlation in red, negative correlation in blue).

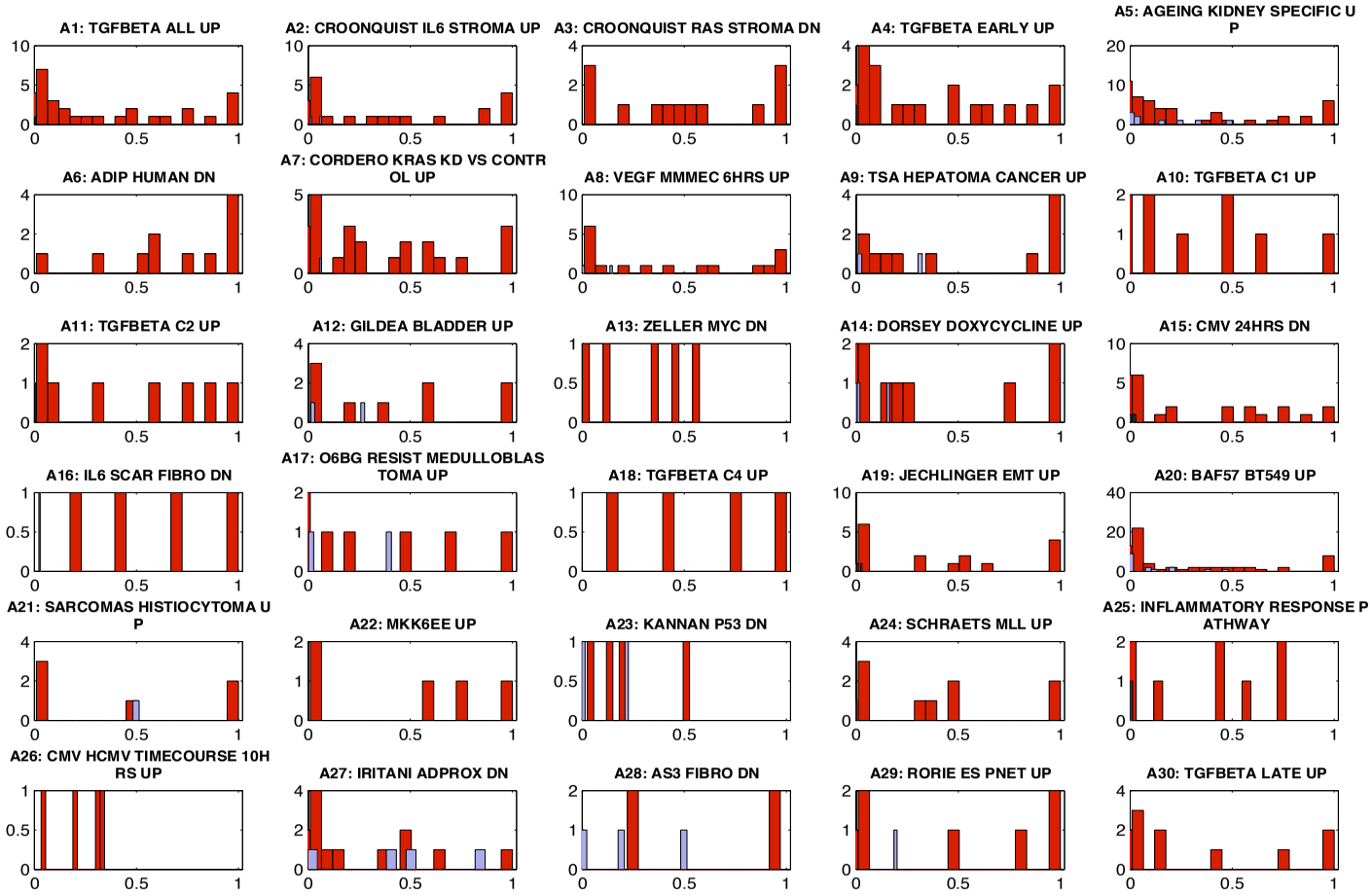


Figure 8.8: Association probability histograms of top 30 pathway gene sets associated with factor 6 (positive correlation in red, negative correlation in blue).

Chapter 9

Concluding Remarks and Future Directions

This dissertation presents a methodology for studying cancer signaling pathways using genome-wide gene expression profiles. The major contribution is innovative statistical modeling and computational methodology for the pathway annotation problem from a Bayesian perspective. This is the first time that understanding of data and knowledge from the biological perspective has been incorporated into formal statistical modeling of the uncertainty in pathway analysis. The development of this methodology involves advanced high-dimensional computation for biological modeling. Analysis of the models uses MCMC methods and novel variational methods for statistical computation. My work generates innovation in these areas of statistical methodology as well as in the cancer genomics applications. The application of probabilistic pathway annotation is demonstrated in a series of examples and studies, including a study of the cellular response to a cancerous micro-environment. The work represents a successful combination of advanced statistical modeling and computation with modern cancer biology research.

The method is currently used in many other ongoing cancer genomics research projects, one of which focuses on oncogenic signature dissection in different type of tumors and cancer cell lines. Although it is developed and discussed in the context of cancer research, this method can be applied to any other biological studies based on genome-wide gene expression profiling data.

Technically, the methodology developed in this dissertation can be extended in terms of statistical modeling, biological applications as well as statistical computa-

tion.

Model extension

One of the key idea in PROPA is the modeling of knowledge uncertainty. In the current model, all the gene sets presumably have the same accuracy (false positive and false negative rate) as the representation of corresponding pathways. In fact, the existing pathway gene set databases contain heterogeneous noise, either due to the experimental and statistical defining procedures or the database curation processes. This generally reduces the reliability of analysis with any pathway annotation methods. Apparently, improving database quality will solve this problem. An alternative solution may be derived from my method through the modeling of the uncertainty related to the pathway gene set databases. The corresponding model refinement involves the modeling of the prior gene pathway membership probabilities β_g (refer to Section 3.3).

In the current model, the β_g *á priori* depend on the hyper-parameter r_g, which have two states, r_A and r_B, determined by the relationship between g and A. The most straightforward extension is to allow more states of r_g to incorporate additional information on the relationship between genes and true pathway $\mathcal{A}$. As a simple and practical example, two gene sets A_1 and A_2 in a database – either defined via different experiments or generated by different statistical procedures – represent the same pathway $\mathcal{A}$. This uncertainty in knowledge can simply be modeled by introducing an extra states of r_g. Explicitly,

$$(\beta_g | g \in A_1, g \in A_2, \mathcal{F} = \mathcal{A}) \sim \mathrm{Be}(\beta_g; r_A),$$

$$(\beta_g | g \notin A_1, g \notin A_2, \mathcal{F} = \mathcal{A}) \sim \mathrm{Be}(\beta_g; r_B),$$

$$(\beta_g | g \in A_1, g \notin A_2, \mathcal{F} = \mathcal{A}) \sim \mathrm{Be}(\beta_g; r_C),$$

$$(\beta_g | g \notin A_1, g \in A_2, \mathcal{F} = \mathcal{A}) \sim \mathrm{Be}(\beta_g; r_C),$$

where r_C is the prior mean of β_g for g in one of A_1 and A_2, and $r_A > r_C > r_B$, representing deterministic *á priori* belief in gene pathway membership according to the voting by these two reference gene sets. Such an extension does not increase the complexity of the PROPA model. The computational methods for model inference and comparison remain the same.

More sophisticated extensions would involve specifying priors of r_g that incorporate general uncertainty and the factors affecting the thinking about the accuracy of reference gene sets. A direct extension is to give r_g the priors in some general forms, such as beta distributions, that allow r_g to vary in specific ranges. Furthermore, if the influencing factors can be identified, the r_g can be modeled with priors parameterized with these factors. The size of a reference gene set, which can be associated with the biological characteristics of the true pathway as well as the defining process, may be a potential factor to consider. Importantly, the posterior estimates of r_g in such extended models can be instrumental in refining the understanding of the database and facilitate further analyses through a Bayesian procedure. However, such model refinement will involve more complicated and intensive computation that requires further study and methodology development.

There are also some methodological issues related to the specification of model priors across the pathway spaces that have only been partially studied in my work. So far I have been treating all the pathways represented by corresponding reference gene sets as *á priori* equally possible to be associated with a phenotype. How to link this prior specification to biological thinking about pathways and databases is worth further research efforts and will shed light on the statistical modeling for other similar type of knowledge-based analysis approaches and application contexts.

Gene set refinement

One of the major components of my pathway annotation methodology is the inference on gene pathway membership. In Chapter 2, I have discussed the significance of this membership inference for the studies of cancer signaling pathways, further demonstrated with the breast tumor ErbB2 pathway analysis in Chapter 7 and lactic acidosis response pathway dissection analysis in Chapter 8. This inference generates a formal statistical refining process for pathway gene sets that aims to increase pathway specificity, identify new biomarkers or correct defining errors in gene sets. With such a refining process, pathway annotation can be applied reciprocally as an *in silico* approach to defining new context-specific pathway gene sets. Such gene sets, combining the original and current biological contexts, are more informative, and can be added into pathway databases to facilitate future studies.

Phenotype annotation

An important application area of pathway annotation is to identify the biological themes related to substructures of an experimental phenotype. This application is essential to the analyses of the gene expression data sets containing heterogeneous cancer phenotypes. Such data sets are enriched in information on cancer developmental mechanisms and usually depend on unsupervised learning approaches to decompose the phenotype complexity. A major component of this type of study is the identification of the molecular phenotypes emerging from the decomposition. Chapter 8 has demonstrated this phenotype identification analysis using the pathway annotation approach combined with BFRM. In this example, I identified some biological processes or cancer phenotypes in the latent factors of gene expression profiles by observing the annotation pathways. As has been mentioned, it is potentially valuable to formalize this phenotype annotation with a quantitative "pathway enrichment

analysis". My pathway annotation method has already formed a foundation for this analysis by ranking the pathways from the database according to their uncertainty of association with the unidentified molecular phenotype. This ranked list of pathways can easily be used to test the enrichment of pathway sets that represent certain biological themes. The current difficulty lies in the lack of such pathway sets databases. The preliminary study in Chapter 8 presented the possibility of this approach. Intensive research is needed to address the pathway sets definition problem to bring this analysis approach into reality.

Application of Monte Carlo variational method

The Monte Carlo variational method developed in this dissertation successfully solves the computational problem in probabilistic pathway annotation. The innovative double-sided bounding technique improves the approximation of marginal likelihoods in Bayesian model comparisons. Intriguingly, this new method is generally applicable in models analyzed using MCMC. Particularly, the framework of this approach demonstrated in PROPA can be directly generalized for estimation of marginal likelihoods in the finite mixture models. This type of statistical models are frequently employed in analysis of many types of biological data, such as gene expression, flow cytometry data, mass spectrum and biological images, showing considerable promise and utility of this new variational method in computational biology.

Software

PROPA is implemented in MATLAB. The major functional modules include pathway gene set database curation, pre-processing of gene-phenotype association probabilities, pathway annotation, gene set refinement and result presentation. Future work will involve the development of a software package in C++, integrating these

functional modules, and providing a higher-speed, user-friendly, and web-available analysis platform for use of the method by the research community.

Bibliography

Ahmad, S., Singh, N., and Glazer, R. I. (1999). Inhibition of mitogen-activated protein kinase and phosphatidylinositol 3-kinase activity in MCF-7 cells prevents estrogen-induced mitogenesis. *Biochem Pharmacol* **58**, 425–430.

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* **7**, 55–65.

Badache, A. and Gonçalves, A. (2006). The ErbB2 signaling network as a target for breast cancer therapy. *Journal of Mammary Gland Biology and Neoplasia* **11**, 13–25.

Baldi, A., Boyle, D. M., and Wittliff, J. L. (1986). Estrogen receptor is associated with protein and phospholipid kinase activities. *Biochem Biophys Res Commun* **135**, 597–606.

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London.

Beal, M. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics*, vol. 7. Oxford University Press.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Berriz, G. F., King, O. D., Bryant, B., Sander, C., and Roth, F. P. (2003). Characterizing gene sets with FuncAssociate. *Bioinformatics* **19**, 18, 2502–2504.

Bertucci, F., Borie, N., Ginestier, C., Groulet, A., Charafe-Jauffret, E., Adélaïde, J., Geneix, J., Bachelart, L., Finetti, P., Koki, A., Hermitte, F., Hassoun, J., Debono, S., Viens, P., Fert, V., Jacquemier, J., and Birnbaum, D. (2004). Identification and validation of an ERBB2 gene expression signature in breast cancers. *Oncogene* **23**, 14, 2564–2575.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Chasse, A. P. D., Joshi, M., Harpole, D., Lancaster, J. M., Berchuck, A., Olson Jr., J. A., Marks, J. R., Dressman, H. K., West, M., and Nevins, J. R. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357.

Brentani, H., Caballero, O. L., and et al, A. A. C. (2003). The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc Natl Acad Sci U S A* **100**, 23, 13418–23.

Brizel, D. M., Schroeder, T., Scher, R. L., Walenta, S., Clough, R. W., Dewhirst, M. W., and Mueller-Klieser, W. (2001). Elevated tumor lactate concentrations predict for an increased risk of metastases in head-and-neck cancer. *Int J Radiat Oncol Biol Phys.* **51**, 2, 349–353.

Campbell, R. A., Bhat-Nakshatri, P., Patel, N. M., Constantinidou, D., Ali, S., and Nakshatri, H. (2001). PI3 kinase/AKT-mediated activation of estrogen receptor alpha: a new model for anti-estrogen resistance. *J Biol Chem* **276**, 9817–9824.

Carvalho, C., Lucas, J., Wang, Q., Chang, J., Nevins, J., and West, M. (2007). High-dimensional sparse factor modelling – Applications in gene expression genomics. *Department of Statistical Science, Duke University, Discussion Paper 05-15.* **(Submitted for publication)**.

Cavenee, W. K., Dryja, T. P., Phillips, R. A., Benedict, W. F., Godbout, R., Gallie, B. L., Murphree, A. L., Strong, L. C., and White, R. L. (1983). Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature* **305**, 27, 779–784.

Celeux, G. and Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports* **41**, 127–146.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving observations. *J. Amer. Statist. Assoc.* **90**, 242–252.

Chang, H. Y., Sneddon, J. B., Alizadeh, A. A., Sood, R., West, R. B., Montgomery, K., Chi, J., van de Rijn, M., Botstein, D., and Brown, P. O. (2004). Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. *PLoS Biol.* **2**, 2, e7.

Chang, J. T. and Nevins, J. R. (2006). GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics* **22**, 23, 2926–2933.

Cheadle, C., Becker, K. G., Cho-Chung, Y. S., Nesterova, M., Wood III, T. W. W., Prabhu, V., and Barnes, K. C. (2007). A rapid method for microarray cross platform comparisons using gene expression signatures. *Molecular and Cellular Probes* **21**, 35–46.

Cheeseman, P. and Stutz, J. (1996). Bayesian classification (Autoclass): Theory and results. In U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, eds., *Advances in Knowledge Discovery and Data Mining*, 153–180. AAAI Press/MIT Press., Menlo Park, CA.

Chen, J. L., Lucas, J. E., Schroeder, T., Mori, S., Nevins, J., Dewhirst, M., West, M., and Chi, J. A. (2007). Genomic analysis of response to lactic acidosis in human cancers. *Department of Statistical Science, Duke Unversity, Discussion Paper* **(Submitted for publication)**.

Chi, J. T., Wang, Z., Nuyten, D. S., Rodriguez, E. H., Schaner, M. E., Salim, A., Wang, Y., Kristensen, G. B., Helland, A., Borresen-Dale, A. L., Giaccia, A., Longaker, M. T., Hastie, T., Yang, G. P., van de Vijver, M. J., and Brown, P. O. (2006). Gene expression programs in response to hypoxia: Cell type specificity and prognostic significance in human cancers. *PLoS Med.* **3**, 3, e47.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90**, 773–795.

Clark, G. J. and Der, C. J. (1995). Aberrant function of the ras signal transduction pathway in human breast cancer. *Breast Cancer Res Treat.* **35**, 1, 133–144.

Colletta, A. A., Wakefield, L. M., Howell, F. V., Baum, D. D. M., and Sporn, M. B. (1991). The growth inhibition of human breast cancer cells by a novel synthetic progestin involves the induction of transforming growth factor beta. *J Clin Invest.* **87**, 1, 277–283.

Corduneanu, A. and Bishop, C. M. (2001). variational Bayesian model selection for mixture distributions. In T. Jaakkola and T. Richardson, eds., *Artificial Intelligence and Statistics*, 27–34. Morgan Kaufmann.

Croonquist, P. A., Linden, M. A., Zhao, F., and Van Ness, B. G. (2003). Gene profiling of a myeloma cell line reveals similarities and unique signatures among IL-6 response, n-ras-activating mutations, and coculture with bone marrow stromal cells. *Blood* **102**, 7, 2581–2592.

Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2002). GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics* **31**, 1, 19–20.

Dannecker, C., Possinger, K., and Classen, S. (1996). Induction of TGF-beta by an antiprogestin in the human breast cancer cell line. *Annals of Oncology* **7**, 391–395.

Davis, P. J. and Rabinowitz, P. (1984). *Methods of Numerical Integration*. Academic Press, New York, 2nd edn.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics* **27**, 1, 94–128.

Deroo, B. J. and Korach, K. S. (2006). Estrogen receptors and human disease. *J Clin Invest.* **116**, 3, 561–570.

Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S. A., and Tainsky, M. A. (2003). Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.* **31**, 13, 3775–3781.

Ebert, B. L., Gleadle, J. M., O'Rourke, J. F., Bartlett, S. M., Poulton, J., and Ratcliffe, P. J. (1996). Isoenzyme-specific regulation of genes involved in energy metabolism by hypoxia: similarities with the regulation of erythropoietin. *Biochem J.* **313**, Pt 3, 809–814.

Efron, B. and Tibshirani, R. (2006). On testing the significance of sets of genes. *Annals of Applied Statistics* **1**, 1, 107–129.

Elson, D. A., Ryan, H. E., Snow, J. W., R, R. J., and Arbeit, J. M. (2000). Coordinate up-regulation of hypoxia inducible factor (HIF)-1α and HIF-1 target genes during multi-stage epidermal carcinogenesis and wound healing. *Cancer Research* **60**, 6189–6195.

Esteller, M. and Herman, J. G. (2002). Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *Journal of Pathology* **196**, 1, 1–7.

Evan, G. I. and Vousden, K. H. (2001). Proliferation, cell cycle and apoptosis in cancer. *Nature* **411**, 342–348.

Falanga, V., Zhou, L., and Yufit, T. (2002). Low oxygen tension stimulates collagen synthesis and COL1A1 transcription through the action of TGF-β_1. *J Cell Physiol* **191**, 42–50.

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S. A., Nobel, A. B., van't Veer, L. J., and Perou, C. M. (2006). Concordance among gene-expression-based predictors for breast cancer. *New England Journal of Medicine* **355**, 6, 560–569.

Fischer, K., Hoffmann, P., Voelkl, S., Meidenbauer, N., Ammer, J., Edinger, M., Gottfried, E., Schwarz, S., Rothe, G., Hoves, S., Renner, K., Timischl, B., Mackensen, A., Kunz-Schughart, L., Andreesen, R., Krause, S. W., and Kreutz, M. (2007). Inhibitory effect of tumor cell–derived lactic acid on human T cells. *Blood* **109**, 9, 3812–3819.

Flechner, S. M., Kurian, S. M., Head, S. R., Sharp, S. M., Whisenant, T. C., Zhang, J., Chismar, J. D., Horvath, S., Mondala, T., Gilmartin, T., Cook, D. J., Kay, S. A., Walker, J. R., Rhodes, D. R., and Salomon, D. R. (2004). Kidney transplant rejection and tissue injury by gene profiling of biopsies and peripheral blood lymphocytes. *Am J Transplant* **4**, 9, 1475–1489.

Friend, S. H., Bernards, R., Rogelj, S., Weinberg, R. A., Rapaport, J. M., Albert, D. M., and Dryja, R. P. (1986). A human DNA segement with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 16, 643–646.

Fry, M. J. (2001). Phosphoinositide 3-kinase signalling in breast cancer: How big a role might it play? *Breast Cancer Res* **3**, 304–312.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo – stochastic simulation for Bayesian inference*. Texts in Statistical Science. Chapman & Hall/CRC, 2nd edn.

Gatenby, R. A. and Gillies, R. J. (2004). Why do cancers have high aerobic glycolysis? *Nat Rev Cancer.* **4**, 11, 891–899.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

Gerstner, T. and Griebel, M. (1998). Numerical integration using sparse grids. *Numer. Algorithms* **18**, 209–232.

Goldrath, A. W., Luckey, C. J., Park, R., Benoist, C., and Mathis, D. (2004). The molecular program induced in T cells undergoing homeostatic proliferation. *Proc Natl Acad Sci U S A* **101**, 48, 16885–16890.

Hahn, W. C., Counter, C. M., Lundberg, A. S., Beigersbergen, R. L., Brooks, M. W., and Weinberg, R. A. (1999). Creation of human tumour cells with defined genetic elements. *Nature* **400**, 29, 464–468.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* **100**, 7, 57–70.

Hodges, L. C., Cook, J. D., Lobenhofer, E. K., Li, L., Bennett, L., Bushel, P. R., Aldaz, C. M., Afshari, C. A., and Walker, C. L. (2003). Tamoxifen functions as a molecular agonist inducing cell cycle-associated genes in breast cancer cells. *Mol. Cancer Res.* **1**, 300–311.

Humphreys, K. and Titterington, D. (2000). Approximate Bayesian inference for simple mixtures. In *Proceedings in Computational Statistics*. COMPSTAT'2000, Springer-Verlag.

Jaakkola, T. and Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**, 25–37.

Jaakkola, T. S. and Jordan, M. I. (1997). Bayesian logistic regression: A variational approach. In *In Proceedings on the 1997 Conference on Artificial Intelligence and Statistics*, 283–294, Fort Lauderdale, FL.

Jechlinger, M., Grunert, S., Tamir, I. H., Janda, E., Lüdemann, S., Waerner, T., Seither, P., Weith, A., Beug, H., and Kraut, N. (2003). Expression profiling of epithelial plasticity in tumor progression. *Oncogene* **22**, 46, 7155–7169.

Johnson, V. E. (1992). A technique for estimating marginal posterior densities in hierarchical models using mixtures of conditional densities. *J. Amer. Statist. Assoc.* **87**, 419, 852–860.

Jones, P. A. and Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics* **3**, 415–428.

Jordan, M. (2004). Graphical models. *Statist. Science* **19**, 140–15.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, K. (1999). An introduction to variational methods for graphical models. *Machine Learning* **37**, 2, 183–233.

Kang, H. C., Kim, I. J., Park, J. H., Shin, Y., Ku, J. L., Jung, M. S., Yoo, B. C., Kim, H. K., and Park, J. G. (2004). Identification of genes with differential expression in acquired drug-resistant gastric cancer cells using high-density oligonucleotide microarrays. *Clinical Cancer Research* **10**, 1 Pt 1, 272–284.

Kim, J. and Dang, C. V. (2006). Cancer's molecular sweet tooth and the warburg effect. *Cancer Research* **66**, 18, 8927–8930.

Kim, S. and Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* **6**.

Kishimoto, T. (1989). The biology of interleukin-6. *Blood* **74**, 1, 1–10.

Knabbe, C., Lippman, M. E., Wakefield, L. M., Flanders, K. C., Kasid, A., Derynck, R., and Dickson, R. B. (1987). Evidence that transforming growth factor-beta is a hormonally regulated negative growth factor in human breast cancer cells. *Cell* **48**, 3, 417–428.

Knudson, A. G. (1971). Mutation and cancer: Statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 4, 820–823.

Kufe, D. W., Bast, R. C., Hait, W. N., Hong, W. K., Pollock, R. E., Weichselbaum, R. R., Holland, J. F., and Frei III, E., eds. (2006). *Cancer Medicine*. American Association for Cancer Research, 7th edn.

Kushner, H. J. and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edn.

Lee, M. S., Hanspers, K., Barker, C. S., Korn, A. P., and McCune, J. M. (2004). Gene expression profiles during human CD4+ T cell differentiation. *International Immunology* **16**, 8, 1109–1124.

Lei, W., Rushton, J. J., Davis, L. M., Liu, F., and Ness, S. A. (2004). Positive and negative determinants of target gene specificity in Myb transcription factors. *J. Biol. Chem.* **279**, 28, 29519 – 29527.

Leone, G., Sears, R., Huang, E., Rempel, R., Nuckolls, F., Fielda, S. J., Thompson, M. A., Yang, H., Fujiwara, Y., Greenberg, M. E., Orkin, S., DeGregoria, J., Smith, C., and Nevins, J. R. (2001). Myc requires distinct E2F proteins to induce S phase and apoptosis. *Mol Cell Biol.* **8**, 105–114.

Lewis, S. M. and Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *J. Amer. Statist. Assoc.* **92**, 648–655.

Lim, K. and Counter, C. M. (2005). Reduction in the requirement of oncogenic ras signaling to activation of pi3k/akt pathway during tumor maintenance. *Cancer Cell* **8**, 381–392.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., and West, M. (2006). Sparse statistical modelling in gene expression genomics. In P. M. K.A. Do and M. Vannucci, eds., *Bayesian Inference for Gene Expression and Proteomics*, 155–176. Cambridge University Press.

Lucas, J. E., Carvalho, C. M., Chen, L., Chi, J., and West, M. (2007). Bench-to-bedside and cross-study projections of genomic biomarkers: An evaluation in breast cancer genomics. *Department of Statistical Science, Duke University, Discussion Paper* **(Submitted for publication)**.

MacKay, D. (1995). Developments in probabilistic modelling with neural networks ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks*, Nijmegen, Netherlands.

Manalo, D. J., Rowan, A., Lavoie, T., Natarajan, L., Kelly, B. D., Ye, S. Q., Garcia, J. G., and Semenza, G. L. (2005). Transcriptional regulation of vascular endothelial cell responses to hypoxia by HIF-1. *Blood* **105**, 2, 659–669.

Manoli, T., Gretz, N., Gröne, H., Kenzelmann, M., Eils, R., and Brors, B. (2006). Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics* **22**, 20, 2500–2506.

Manotham, K., Tanaka, T., Matsumoto, M., Ohse, T., Inagi, T., Miyata, T., Kurokawa, K., Fujita, T., Ingelfinger, J. R., and Nangaku, M. (2004). Transdifferentiation of culture tubular cells induced by hypoxia. *Kidney Int.* **65**, 871–880.

Maynard, P., Davies, C. J., Blamey, R., Elston, C. W., Johnson, J., and Griffiths, K. (1978). Relationship between oestrogen-receptor content and histological grade in human primary breast tumours. *Br J Cancer* **38**, 6, 745–748.

McGrory, C. A. and Titterington, D. M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis* **51**, 11, 5352–5367.

Ménard, S., Pupa, S. M., Campiglio, M., and Tagliabue, E. (2003). Biologic and therapeutic role of her2 in cancer. *Oncogene* **22**, 6570–6578.

Meng, X. L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6**, 831–860.

Miller, D. L., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., and Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A* **102**, 38, 13550–13555.

Minka, T. (2001). *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis, MIT.

Moggs, J. G. and Orphanieds, G. (2001). Estrogen receptors: Orchestrators of pleiotropic cellular responses. *EMBO Rep.* **2**, 9, 775–781.

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-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**, 3, 267–273.

Newton, M. A., Quintana, F. A., den Boon, J. A., Sengupta, S., and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Annals of Applied Statistics* **1**, 1, 85–106.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B* **56**, 3–48.

Peng, T., Golub, T. R., and Sabatini, D. M. (2002). The immunosuppressant rapamycin mimics a starvation-like signal distinct from amino acid and glucose deprivation. *Mol Cell Biol.* **22**, 15, 5575–5584.

Perou, C. M., Sorlie, T., Eisen, M. B., van deRijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature* **406**, 6797, 747–752.

Pethe, V. and Shekhar, P. V. M. (1999). Estrogen inducibility of c-Ha-ras transcription in breast cancer cells. *J Biol Chem, Vol.* **274**, 43, 30969–30978.

Pichon, M. F., Broet, P., Magdelenat, H., Delarue, J. C., Spyratos, F., Basuyau, J. P., Saez, S., Rallet, A., Courriere, P., Millon, R., and Asselain, B. (1996). Prognostic value of steroid receptors after long-term follow-up of 2257 operable breast cancers. *Br J Cancer* **75**, 12, 1545–1551.

Razandi M, Pedram A, L. E. (2000). Plasma membrane estrogen receptors signal to antiapoptosis in breast cancer. *Mol Endocrinol* **14**, 1434–1447.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A* **101**, 25, 9309–14.

Ricci, A., Lanfrancone, L., Chiari, R., Belardo, G., Pertica, C., Natali, P. G., Pelicci, P., and Segatto, O. (1995). Analysis of protein-protein interactions involved in the activation of the Shc/Grb-2 pathway by the ErbB-2 kinase. *Oncogene* **11**, 8, 1519–1529.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics* **22**, 400–407.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, 2nd edn.

Rodriguez, G. C., Nagarsheth, N. P., Lee, K. L., Bentley, R. C., Walmer, D. K., Cline, M., Whitaker, R. S., Isner, P., Berchuck, A., Dodge, R. K., and Hughes, C. L. (2002). Progestin-induced apoptosis in the macaque ovarian epithelium: Differential regulation of transforming growth factor-beta. *J Natl Cancer Inst* **94**, 1, 50–60.

Schena, M., Shalon, D., Davis, R. W., and Brown Jr., P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 5235, 467–470.

Sears, R., Nuckolls, F., Haura, E., Taya, Y., Tamaia, K., and Nevins, J. R. (2000). Multiple Ras-dependent phosphorylation pathways regulate Myc protein stability. *Genes and Development* **14**, 2501–2514.

Seo, D. M., Goldschmidt-Clermont, P. J., and West, M. (2007). Of mice and men: Sparse statistical modelling in cardiovascular genomics. *Annals of Applied Statistics* **1**, 1, 152–178.

Shen, H., Ji, C., and West, M. (2007). Monte carlo variational approximation of marginal likelihoods for finite mixture models. Tech. rep., Department of Statistical Science, Duke University.

Simoncini, T., Hafezi-Moghadam, A., Brazil, D. P., Ley, K., Chin, W. W., and Liao, J. K. (2000). Interaction of oestrogen receptor with the regulatory subunit of phosphatidylinositol-3-OH kinase. *Nature* **407**, 538–541.

Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K. V., Gazdar, A. F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P., and Børresen-Dale, A. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 19, 10869–10874.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lauder, 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 U S A* **102**, 43, 15545–15550.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., and Park, I. S. K. P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A* **102**, 38, 13544–13549.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**, 82–86.

Ueberhuber, C. W. (1997). *Numerical Computation 2: Methods, Software, and Analysis*. Springer-Verlag, Berlin.

Ueda, N. and Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks* **15**, 1223–1241.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

Vaupel, P. (2004). Tumor microenvironmental physiology and its implications for radiation oncology. *Semin Radiat Oncol.* **14**, 3, 198–206.

Vaupel, P., Kallinowski, F., and Okunieff, P. (1989). Blood flow, oxygen and nutrient supply, and metabolic microenvironment of human tumors: A review. *Cancer Research* **49**, 6449–6465.

Verrecchia, F., Chu, M., and Mauviel, A. (2001). Identification of novel TGF-beta /Smad gene targets in dermal fibroblasts using a combined cDNA microarray/promoter transactivation approach. *J Biol Chem* **276**, 20, 17058–17062.

Vincent-Salomon, A. and Thiery, J. P. (2003). Host microenvironment in breast cancer development: Epithelial–mesenchymal transition in breast cancer development. *Breast Cancer Res* **5**, 2, 101–106.

Virtaneva, K., Wright, F. A., Tanner, S. M., Yuan, B., Lemon, W. J., Caligiuri, M. A., Bloomfield, C. D., de la Chapelle, A., and Krahe, R. (2001). Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc Natl Acad Sci U S A* **98**, 3, 1124–1129.

Vogelstein, B. and Kinzler, K. W. (1993). The multistep nature of cancer. *Trends Genetics* **9**, 138–141.

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and thepathways they control. *Nature Medicine* **10**, 8, 789–799.

Walenta, S., Wetterling, M., Lehrke, M., Schwickert, G., Sundfør, K., Rofstad, E. K., and Mueller-Klieser, W. (2000). High lactate levels predict likelihood of metastases, tumor recurrence, and restricted patient survival in human cervical cancers. *Cancer Research* **60**, 4, 916–921.

Wang, B. and Titterington, D. (2004). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In M. Chickering and J. Halpern, eds., *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*.

Warburg, O. (1956). On the origin of cancer cells. *Science* **123**, 3191, 309–314.

West, M. (2003). Bayesian factor regression models in the"large p, small n" paradigm. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, eds., *Bayesian Statistics 7*. Oxford University Press.

West, M., Blanchette, C., H., Dressman, Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson Jr., J. A., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* **98**, 20, 11462–11467.

Westfall, P. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley-Interscience.

Wu, Q., Kirschmeier, P., Hockenberry, T., Yang, T. Y., Brassard, D. L., Wang, L., McClanahan, T., Black, S., Rizzi, G., Musco, M. L., Mirza, A., and Liu, S. (2002). Transcriptional regulation during p21WAF1/CIP1-induced apoptosis in human ovarian cancer cells. *J. Biol. Chem.* **277**, 39, 36329–36337.

Zahn, J. M., Sonu, R., Vogel, H., Crane, E., Mazan-Mamczarz, K., Rabkin, R., Davis, R. W., Becker, K. G., Owen, A. B., and Kim, S. K. (2006). Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS Gen.* **2**, 7, e115.

Zavadil, J., Cermak, L., Soto-Nieves, N., and Böttinger, E. P. (2004). Integration of TGF-β/Smad and Jagged1/Notch signalling in epithelial-to-mesenchymal transition. *The EMBO Journal* **23**, 1155–1165.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, 4, R28.

Zhong, S., Li, C., and Wong, W. H. (2003). ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis. *Nucleic Acids Res.* **31**, 13, 3483–3486.

Biography

Haige Shen was born on December 2, 1973 in Guizhou, China. She received a B.S. degree in Electronics Engineering in May of 1995 and a Ph.D. degree in Electronics Engineering in July of 2000, both from Beijing Institute of Technology in China. During the following one and a half years, she worked as a research fellow with the Bioinformatics Center at National University of Singapore, and started her exploration in bioinformatics. In August 2003, she decided to pursue her interest in computational biology and statistics and enter the Ph.D. program at Duke University.