Bayesian Hierarchical Modeling for Problems in Computational Biology

by

Hyung Won Choi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2009

Doctoral Committee:

Assistant Professor Zhaohui Qin, Co-Chair
Associate Professor Debashis Ghosh, Penn State University, Co-Chair
Assistant Professor Alexey I. Nesvizhskii
Assistant Professor Sinae Kim

ACKNOWLEDGEMENTS

I would like to thank the academic advisors Drs. Debashis Ghosh and Zhaohui Qin for their invaluable guidance for the completion of this thesis. I am also grateful to Dr. Sinae Kim for her helpful comments on the thesis. Special thanks are due to Dr. Alexey Nesvizhskii who has shown me the merit of rigourous reasoning and the importance of solid scientific principles in an elegant manner.

The work presented in this thesis would not have been possible if it were not for Dr. Jun Li, who introduced me into contemporary molecular biology and has taught me through the discipline since I worked in his lab at Stanford University. Drs. Johan Lim and Kyu Hahn have always supported me through critical moments during my graduate studies since we met in Palo Alto.

Mike Tyers lab in Samuel Lunenfeld Research Institute generously allowed me to use their data in the development of protein-protein interactome analysis. Timothy Green at the Gayle Morris Sweetland Writing Center deserves a mention for his fabulous job in proofreading this thesis.

This thesis is dedicated to my inspiring parents, lovely wife Yoon Kyung and her family, and younger sister Hyung Mee. The recent company of Arthur-Lee, my son, has been the greatest joy and achievement during the graduate program. It is only through them that I realize the values of my professional career.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

1.1 Statistical Challenges in Integrative Analysis

This thesis addresses the application of Bayesian hierarchical models to integrative data analysis in computational biology, where consistent biological signals need to be identified from multiple subjects and distinguished from experimental noise featuring multiple levels of molecular biology. Differentiating biological signals from experimental noise is challenging in the presence of heterogeneity across subjects and varying experimental protocols. The data analysis methods developed here illustrate that hierarchical models with Bayesian inference are useful in the presence of limited sample size and mounting computational complexity commonly observed in the integrative analysis of large-scale experimental datasets.

Every chapter in the thesis addresses distinct challenges unique to the analysis of high-throughput genomic and proteomic experiments. These experiments allow us to monitor thousands of molecules simultaneously, and thus facilitate the study of structural and functional characterization of the cell. With the spread of these technologies applicable in a common lab setting, an increasing number of large-scale datasets have been generated from genomic assays [1, 2], mass spectrometry-based shotgun proteomics experiments [3, 4, 5, 6], and a number of other "omics"

experiments [7, 8] that fortify a systems view of the biological investigation.

While the new technologies may offer unprecedented opportunities, analyzing numerical data and interpreting the results in an appropriate biological context is difficult because data are observed without knowledge of the relationship between individual genes and proteins. It should be remembered that measurement of abundance in individual molecules reflects not only their own activities, but also the influence of other related molecules.

In practice, many data analysis methods were developed under the assumption that individual molecules are statistically independent and the multivariate models were not explored explicitly. However, the functional mechanism in a given biological problem can be properly described only when individual pieces of information specific to genes and proteins are assembled into a reasonable context, and it is therefore desirable to develop inferential methods that reflect the connection between molecules.

This task is extraordinarily challenging due to the limitation of existing experimental protocols and perhaps the absence of appropriate computational tools. Previous studies addressing the multivariate modeling have noted this difficulty as well, including co-expression networks [9], regulatory modules [10, 11, 12, 13], or protein-protein interactions [14, 15, 16, 17, 18]. It is in this context that hierarchical models with Bayesian inference appear to be an attractive alternative [19].

1.2 Bayesian Hierarchical Models in Computational Biology

Hierarchical modeling is a useful approach for constructing statistical models of high complexity [20]. From a theoretical perspective, the classical use of exchangeability [21, 22] and partial exchangeability [23] along with de Finetti's theorem [24] and its generalization by Hewitt and Savage [25] altogether provide a solid theoret-

ical basis for hierarchical modeling. From a practical standpoint, the availability of the standard posterior sampling-based inference via Markov chain Monte Carlo (MCMC) has boosted the popularity of hierarchical models in a variety of settings such as spatial statistics [26] and longitudinal analysis [27].

For the problems in computational biology, these theoretical assumptions imply that genes and proteins are exchangeable, i.e. the joint distribution of their measurements is invariant under random ordering of genes and proteins. Such an assumption that genes are interchangeable can be regarded arguably strong with respect to biological interpretation. Despite this problem, hierarchical models can still be a good inferential framework for addressing the statistical challenges mentioned above. Borrowing statistical strength across all the molecules profiled in a dataset is a particularly important feature of Bayesian hierarchical models.

Most high-throughput experiments are rarely replicated in sample sizes sufficient to infer individual parameters. Given d dimensional data $Y_j = \{y_{ij}\}_{i=1}^{d}$ for $j = 1, \ldots, n$ and parameters $\Theta = \{\theta_i\}_{i=1}^{d}$, a general form of a hierarchical model is

$$y_{ij} \sim \pi(\cdot|\theta_i) \tag{1.1}$$

$$\theta_i \sim \pi(\cdot|\gamma) \quad i = 1, \ldots, d, \quad j = 1, \ldots, n \tag{1.2}$$

In this setting, posterior inference for the parameter of a molecule i is based on the marginal predictive posterior distribution

$$\pi(\theta_i|Y) \propto \int \int \prod_{k=1}^{d} \left\{ \left(\prod_{j=1}^{n} \pi(y_{kj}|\theta_k) \right) \pi(\theta_k|\gamma) \right\} \pi(\gamma) d\Theta_{-i} d\gamma \tag{1.3}$$

$$\propto \int \left(\prod_{j=1}^{n} \pi(y_{ij}|\theta_i) \right) \pi(\theta_i|\gamma) g(Y_{-i}, \gamma) \pi(\gamma) d\gamma \tag{1.4}$$

where Y_{-i} and Θ_{-i} are the entire data and parameter set excluding the i-th compo-

nent and

$$g(Y_{-i},\gamma)=\int\prod_{k\neq i}\left\{\left(\prod_{j=1}^{n}\pi(y_{kj}|\theta_k)\right)\pi(\theta_k|\gamma)\right\}d\Theta_{-i} \tag{1.5}$$

From the expression in (1.4), it must be noted that integrating the integrand including $g(Y_{-i},\gamma)$ with respect to γ is the key characteristic of hierarchical Bayes where information across different molecules are pooled together. It follows that the statistical inference on each parameter θ_i will be notably more robust than marginal inference based on $\pi(\theta_i|y_{i1},\ldots,y_{in})$ with small n, as long as exchangeability assumption holds between θ_i and θ_j for $i\neq j$.

In practice, this modeling framework has been used in a broad spectrum of high-throughput genomic data analysis, including the methods for differential expression analysis [28, 29, 30, 31, 32, 33], gene expression clustering [34, 35, 36, 37], multiple testing correction [38, 39, 40], and Bayesian network analysis [41, 42].

1.3 Relevant Statistical Topics

Each chapter in this thesis utilizes hierarchical models where structure is imposed on the prior distribution (with the exception of Chapter 3) and heavily relies on parameter estimation algorithms via posterior sampling-based MCMC. The development of MCMC constitutes a large majority of the literature of Bayesian computation [43, 44, 45, 46, 47, 48, 49, 50], and interested readers are referred to [51, 52, 53] for further comprehensive review.

In addition to the reliance on MCMC, this thesis makes use of a few existing classes of models. First, mixture modeling forms the basis for the methods developed in Chapters 2,3, and 5. Mixture model-based posterior probabilities give a natural classification criterion. Bayesian mixture models make use of latent variables, often associated with physical or biological interpretation, and these have been widely used

for distinguishing differential expression from experimental noise in gene expression data analysis [29, 32, 54]. In addition, Chapter 2 features a copy number segmentation algorithm, which is a change-point problem requiring a dimension-switching posterior sampling algorithm called reversible jump MCMC. Copy number data typically shows a considerable degree of local correlation and traditional Bayesian solutions to change-point problems are well suited to this data [52, 55, 56]. Lastly, Chapter 4 employs model selection method using Bayes factors, the computation of which has occupied a long range of statistical literature [57, 58, 59, 60].

1.4 Outline

This thesis consists of four chapters, each addressing independent statistical problems in different molecular biology investigations.

Chapter 2 describes a model-based method for analyzing joint profiles of DNA copy number and mRNA transcript expression datasets from array comparative genomic hybridization (aCGH) and cDNA/oligonucleotide microarrays. In order to address the local correlation in the copy number data, a Bayesian change point estimation procedure is combined with sampling methods for a two-stage mixture model.

Chapter 3 presents a hierarchical hidden Markov model as a tool that can effectively summarize multiple parallel sequential data. Particularly, the model is used to construct a genome-wide map of transcription factor binding sites using chromatin immunoprecipitation data from multiple experimental platforms. It is shown that the hierarchical Bayesian inference for multiple HMMs gives improved performance in terms of classification accuracy compared to competing methods.

Chapters 4 and 5 discuss novel statistical methods for mass spectrometry-based

quantitative proteomics data analysis. A model-based method called *QSpec* is developed to detect differentially expressed proteins using label-free absolute quantification while addressing the limited sample size issue in the typical experimental data with hierarchical Bayes. Moreover, a novel method called *Significance Analysis of Interactome*, or SAInt, is proposed as a model-based filter for protein-protein interactions observed in large-scale affinity purification - mass spectrometry (AP-MS) experiments.

CHAPTER II

A Double-Layered Mixture Model for the Joint Analysis of DNA Copy Number and mRNA Expression Data

2.1 DNA Copy Number and Gene Expression in Cancer Genomics

Gene expression has been considered as a proxy for investigating the source of phenotypic variation in human populations [61]. Lately, genomic alterations [62] including copy number variants (CNV), inversions, and tandem repeats, have been found to be associated with mRNA expression phenotype in a number of studies [61, 63], although questions remain as to the consistency and the strength of the association in comparison to other small genetic variants such as single nucleotide polymorphism (SNP) and large chromosomal events such as aneuploidy, rearrangements, and fragile sites [64].

Most of the large-scale surveys of the paired profiles of DNA copy number and mRNA expression phenotype have been conducted in healthy individuals from well-studied populations, e.g. HapMap samples [63]. Nonetheless, it is in cancer genomics that the correlation between the two datasets has been more extensively investigated using array-based comparative genomic hybridization, or array CGH [65] and mRNA expression microarrays. For example, Pollack et al [66] was one of the earliest to investigate direct association between the two data in breast cancer tissues and cell lines. Hyman et al [67] also showed nearly half the amplification events in breast

cancer cell lines associated with elevated mRNA expression and found that similar results could be replicated in tumor tissue samples. The association between the two data has been reported in other types of cancer as well, such as the lung cancer data in Tonon et al [68]. A large number of genome-wide surveys like these have established that some degree of association is present between the two data, and also that the variability in mRNA expression was not entirely accounted for by the copy number changes alone.

The increasing motivation for joint analysis in cancer genomic datasets results from work that has characterized copy number changes as a hallmark of cancer. A great deal of effort has been devoted to development of statistical methods for determining breakpoints between genomic segments of homogenous copy numbers. Popular algorithms include circular binary segmentation [69], hidden Markov Models [70, 71, 72, 73], hierarchical clustering-based algorithms [74], information criteria-based change point model application [75], and a mixture model-based dynamic programming algorithm [76]. Comparison of some of these algorithms has been provided in recent reviews [77, 78, 79].

In previous examples of joint analysis, the segmentation algorithms were valuable tools for detecting segmental amplification and deletion sites because one can look for copy number-associated expression changes within the segments of aberrant copy number level without the need to interrogate the entire genome once these segments are selected beforehand. Tonon et al [68] and Kim et al [80], for example, performed linear correlation analysis and differential expression hypothesis testing after applying segmentation algorithms to identify focal regions. However, lacking a model-based connection between the decisions made in each dataset, their gene selection methods are inevitably subject to multiple sources of error and it is dif-

ficult to accurately estimate overall error rates due to the stepwise application of hard thresholds (e.g. copy number ratio greater than 0.2, gene expression t-statistic greater than 3).

To date, few systematic approaches are available for the joint analysis of copy number and gene expression. Lipson et al [81] developed a regional analysis called genomic continuous submatrix, or GCSM, using an extended definition of standard correlation measures to account for the local similarity in correlation patterns between copy number and mRNA expression. A recent paper by van Wieringen and van de Viel [82] proposes a non-parametric hypothesis testing framework for finding changes in mRNA expression distribution conditional on the probability of calls for gain and loss in copy number data (NPtest hereafter). Although their formulation specifically targets relevant sources of variation, the model is still conditional on the fact that copy number status is known *a priori*, and therefore fails to completely account for the uncertainty involved in making copy number calls and subsequent regional analysis.

This problem will be addressed with an approach called *double-layered mixture model* (DLMM), which probabilistically scores the association between the paired copy number and gene expression data. Copy number data are considered as a series of random variables whose mean parameters form a stochastic process along the genome with a finite number of jumps and follow a mixture distribution representing differential copy number status. DLMM not only makes probabilistic calls for amplification and deletion events, but also searches for the best arrangement of copy number boundary points using an advanced sampling algorithm. Meanwhile, mRNA expression levels are considered as observations from mixture models in individual genes. In this process, elevated or reduced expression measurements are

scored only if they are accompanied by corresponding copy number changes. DLMM simultaneously computes the (marginal) probability of copy number changes and the joint probability that the copy number change is coherently matched by gene expression change. The joint modeling feature of DLMM removes the burden of stepwise analysis typically used in the literature.

The organization of this chapter is as follows. In Section 2.2, the statistical model DLMM is described in three parts: model for copy number data, model for mRNA expression data, and scoring algorithm. Section 2.3 explains the posterior distributions for model parameters and Markov chain Monte Carlo sampling scheme for inference. Section 2.4 presetns the application of the proposed methodology to a breast cancer dataset by Pollack et al [66] with comparison to two competing methods by Lipson et al [81] and van Wieringen and van de Viel [82]. Section 2.5 concludes the chapter.

2.2 Statistical Model

DLMM is composed of two main parts, one for copy number and the other for mRNA expression data respectively. For the clarity of presentation, a grahical representation is provided to show the conditional independence structure among the model parameters in Figure 2.1.

2.2.1 Model for Copy Number Data

Suppose that the copy number data $X = \{x_{gs}\}$ is observed for genes $g = 1, \ldots, G$ in samples $s = 1, \ldots, N$. For the sake of simplicity, tumor-only analysis is discussed throughout this work, but the methodology can easily be extended to two group comparison such as tumor versus normal tissue comparison. Let N denote the number of tumor specimens.

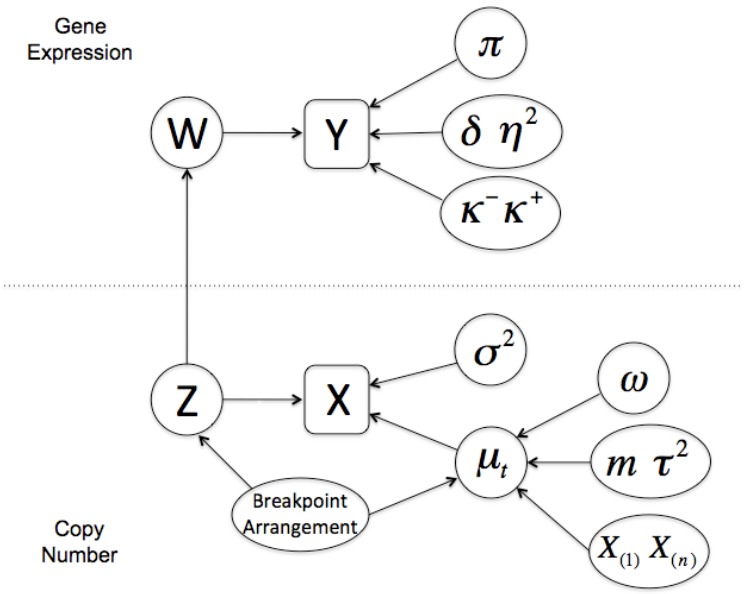


Figure 2.1: Conditional independence graph of DLMM. X and Y denote the observed copy number and expression respectively. Z and W are the calls of significant copy number and gene expression. $X_{(1)}$ and $X_{(n)}$ denote the minimum and the maximum copy number in each sample respectively. Note that the mixture models in the copy number data are sample-specific, while those in the gene expression data are gene-specific. Given these parameters, the two datasets are independent.

The copy number data, e.g. log-scaled CGH ratio, in sample s is modeled as a series of random observations from Gaussian distribution with mean parameters forming a stochastic process along the chromosome, represented in piecewise constant functions. Each chromosome of sample s is divided into T_s segments, with T_s itself being a random variable from a Poisson distribution with mean λ_s. The parameter λ_s is assumed to follow Gamma distribution $\mathcal{G}(k_1, k_2)$. The Poisson-Gamma mixture leads to negative binomial prior for λ_s, which can be considered as a flexible prior partially accounting over-dispersion. In this setting, there are $(T_s - 1)$ boundary points between adjacent segments and two fixed points on the start and the end positions of the chromosome, to give $(T_s + 1)$ in total. These *breakpoints* are denoted by $(p_{s0}, p_{s1}, \ldots, p_{sT_s})$ so that their locations vary by sample. The segment $\mathcal{S}_t$ flanked by $(p_{s(t-1)}, p_{st})$ contains at least one gene, and the copy number data in sample s in

a given segment is a series of independent observations from a Gaussian distribution with mean μ_{ts} and variance σ_s^2.

To be specific, the model for copy number data can be written as follows. For each sample s with segment configuration $\{\mathcal{S}_t\}_{t=1}^{T_s}$,

$$\begin{aligned}
x_{gs} &\sim \mathcal{N}(\mu_{t(g),s}, \sigma_s^2), \quad g = 1, \ldots, G \\
\mu_{t(g)s} &\sim \omega_s \mathcal{U}(X_{s_{(1)}}, X_{s_{(G)}}) + (1 - \omega_s)\mathcal{N}(m_s, \tau_s^2), \quad t = 1, \ldots, T_s \\
T_s &\sim \mathcal{P}(\lambda_s) \\
\lambda_s &\sim \mathcal{G}(k_1, k_2) \\
(p_{s0}, p_{s1}, \ldots, p_{sT_s}) &\stackrel{d}{\equiv} (U_{(1)}, U_{(2)}, \ldots, U_{(T_s)})
\end{aligned}$$

where $t(g)$ indexes the segment containing gene g, and $U_{(i)}$ denote i-th order statistic of $(T_s + 1)$ Uniform random variables on $(0, L)$. The mean process has a Uniform-Gaussian mixture prior distribution, and $(X_{s(1)}, X_{s(G)})$ are the minimum and the maximum copy number intensity ratios in sample s respectively. In the mixture distribution of μ_{ts}, latent variables are introduced for the sampling procedure. Define the latent variables Z_{ts} as follows:

$$\begin{aligned}
Z_{ts} = 1 \ \text{ if } \ \mu_{ts} &\sim \mathcal{U}(X_{s_{(1)}}, X_{s_{(G)}}) \\
Z_{ts} = 0 \ \text{ if } \ \mu_{ts} &\sim \mathcal{N}(m_s, \tau_s^2)
\end{aligned}$$

The genome-wide mean copy number m_s is assumed to follow $\mathcal{N}(\nu, \zeta^2)$ prior distribution. The variance components in the likelihood and the prior are assumed to have standard inverse Gamma distributions $\sigma_s^2 \sim \mathcal{IG}(b_1, b_2)$ and $\tau_s^2 \sim \mathcal{IG}(a_1, a_2)$, and the mixing proportion parameter ω_s has Uniform prior distribution $\mathcal{U}(0, 1)$.

2.2.2 Model for Gene Expression Data

Suppose that mRNA expression is measured for some of the G genes, which is denoted by $Y = \{y_{gs}\}$ with parallel indexing of gene IDs. For example, if gene g has both the copy number and the expression data, $\{(x_{t(g)s}, y_{gs})\}_{s=1}^{N}$ denotes the pair across the N samples. To keep the notations tractable, it will be assumed that every gene in the data has both copy number and mRNA expression measurements, i.e. $t(g) = g$ for all g. Extending to the case where the copy number data has higher density of coverage than the expression data is trivial, and so is incorporting multiple chromosomes. $\{y_{gs}\}_{s=1}^{N}$ is modeled as observations from Uniform-Gaussian mixture distribution, where the uniform component corresponds to the distribution of aberrant copy number-associated expression, and the Gaussian component corresponds to either gene expression not associated with aberrant copy number levels or gene expression in samples with normal copy number levels. If the data contains non-tumor samples, all measurements from those samples will belong to the Gaussian component, guiding the estimation of the mixture in a semi-supervised way. In sum, the mixture formulation attempts to quantify the enrichment of copy number-associated expression levels in the tail of the entire expression distribution in each gene.

More specifically, a hierarchical Uniform-Gaussian mixture model is fitted to the mRNA expression data:

$$y_{gs} \sim \pi_g \mathcal{U}(l_g - \kappa_g^-, u_g + \kappa_g^+) + (1 - \pi_g)\mathcal{N}\left(\delta_g, \eta_g^2\right), \qquad s = 1, \ldots, N$$

The difference is that this time mixture models are fitted in individual genes (unlike in individual samples for copy number data). This particular model specification has been previously used in cancer classification work by Parmigiani et al [54]. A set of

latent variables W_{gs} are defined as follows:

$$W_{gs} = 1 \quad \text{if} \quad y_{gs} \sim \mathcal{U}(l_g - \kappa_g^-, u_g + \kappa_g^+)$$
$$W_{gs} = 0 \quad \text{if} \quad y_{gs} \sim \mathcal{N}\left(\delta_g, \eta_g^2\right)$$

respectively, where (l_g, u_g) denote the minimum and the maximum expression values of gene g across the samples, and (κ_g^-, κ_g^+) are the extended tail parameters for the Uniform component for under and over expression of gene g. Priors for the Gaussian component representing normal expression levels are given as $\delta_g \sim \mathcal{N}(\theta, \psi^2)$ and $\eta_g^2 \sim \mathcal{IG}(d_1, d_2)$ Those for the Uniform component are the following: $\kappa_g^+ \sim \mathcal{E}(\rho_+)$, $\kappa_g^- \sim \mathcal{E}(\rho_-)$, and $\pi_g \sim \mathcal{U}(0, 1)$.

In the scoring method explained above, note that $Z_{t(g),s} = 0$ implies $W_{gs} = 0$, meaning that the definition of over or under expression is relative to the expression distribution in samples with no aberrant copy numbers. Thus even if a gene is highly expressed in many samples, this gene will not be considered as over-expressed so long as this is not related to the concordant amplification. In terms of model parameters, this implies that δ_g can be far from zero, requiring appropriate elicitation of prior. This definition of W therefore highlights the gene and the sample with expression changes specifically associated with copy number changes, (partially) removing the influence of other potential confounders that may drive both measures in the same direction.

2.2.3 Probabilistic Scoring and Criterion-based Gene Selection

DLMM reports two sets of probability scores: (1) copy number calls summarizing Z, and (2) probability score of copy number-associated expression changes. For the latter, the probability that y_{gs} is associated with the Uniform component, i.e. $P(W_{gs} = 1)$ for all g and s such that $Z_{t(g),s} = 1$ is calculated. This enables us to eval-

uate the joint probability $P_{gs} = P(W_{gs} = 1)$. Since the event $W_{gs} = 1$ may represent either over or under expression, it is also checked whether the direction of changes in both types of data is coherent, i.e. copy number gain should be associated with over-expression, and copy number loss should be associated with under-expression. This is equivalent to calculating an over-expression score P^u_{gs} and an under-expression score P^d_{gs}, where

$$P^u_{gs} = P(W_{gs} = 1, \mu_{t(g),s} > m_s, y_{gs} > \delta_g)$$

$$P^d_{gs} = P(W_{gs} = 1, \mu_{t(g),s} < m_s, y_{gs} < \delta_g)$$

where the event $Z_{t(g),s} = 1$ is omitted in each expression because it is a necessary condition for $W_{gs} = 1$. One can summarize the two-dimensional score into a signed score $P_{gs} = P^u_{gs} - P^d_{gs}$ or follow the two scores separately. This calculation results in signed probability for a gene in a specific sample, and positive and negative scores of large magnitude indicate the strength of evidence for copy number-associated expression change for the given gene in the sample.

Since this probability score is the joint probability of aberrant copy number and gene expression, this number can range from a very small number to a value close to 1, depending on multiple factors such as the sample size, the prevalence of copy number changes, and the separation of copy number-associated expression from expression distribution in samples with normal copy numbers. Thus it is important to establish an objective criterion to select genes based on the estimated model parameters and the model fit. The L-measure introduced by Ibrahim et al [29] seems well-suited for the purpose. Computation of the L-measure at probability threshold p^* is achieved by taking average of the following quantity over the posterior samples used for the

inference:

$$L(p^*) = \sum_{g=1}^{G}\sum_{s=1}^{N}\left[U_{gs}(p^*) + N_{gs}(p^*) + D_{gs}(p^*)\right]$$

where

$$\begin{aligned}
U_{gs}(p^*) &= 1\{P_{gs} > p^*\}\left[\frac{1}{12}(u_g + \kappa_g^+ - (l_g - \kappa_g^-))^2 + \nu\left(\frac{u_g + \kappa_g^+ + \delta_g}{2} - y_{gs}\right)^2\right] \\
N_{gs}(p^*) &= 1\{-p^* \le P_{gs} \le p^*\}\left[\eta_g^2 + \nu(y_{gs} - \delta_g)^2\right] \\
D_{gs}(p^*) &= 1\{P_{gs} < -p^*\}\left[\frac{1}{12}(u_g + \kappa_g^+ - (l_g - \kappa_g^-))^2 + \nu\left(y_{gs} - \frac{l_g - \kappa_g^+ + \delta_g}{2}\right)^2\right]
\end{aligned}$$

The weighting constant ν of the squared bias with respect to the predictive mean relative to the predictive variance was set at 0.5, following the theoretical justification of Ibrahim et al [83]. Genes are selected using the threshold yielding minimal L-measure.

2.3 Inference

Statistical inference for the model parameters is performed through sampling from the posterior distributions and summarizing the distribution from the output. Due to the segmentation feature in the copy number data, a part of the posterior sampling involves transdimensional moves guided by reversible jump MCMC. Samples are drawn from the appropriate posterior distributions in the following order: [Copy Number Parameters] → [Copy Number Segment Arrangement] → [Gene Expression Parameters].

2.3.1 Gibbs Update for Copy Number Parameters

Given a fixed segmentation arrangement, the segmental mean μ_{ts} is drawn from

$$\mu_{ts}|\cdot \propto L(X_s|\mu_{ts}) \cdot \left(\omega_s \mathcal{U}(X_{s(1)}, X_{s(G)}) + (1-\omega_s)\mathcal{N}(m_s, \tau_s^2)\right)$$

by Metropolis-Hastings sampling, where $L(X|\mu_{ts}) = \prod_{g:t(g)=t} \mathcal{N}(X_i; \mu_{ts}, \sigma_s^2)$. Next, the latent variables Z_{ts} are drawn by sampling from Bernoulli random variable with the success probability

$$\frac{\omega_s \mathcal{U}(X_{s(1)}, X_{s(G)})}{\omega_s \mathcal{U}(X_{s(1)}, X_{s(G)}) + (1-\omega_s)\mathcal{N}(m_s, \tau_s^2)}$$

for every segment. The rest of the parameter updates are done through Gibbs sampling based on closed form distributions. The variance component for segmental means has the following distribution

$$\tau_s^2|\cdot \sim \mathcal{IG}\left(a_1 + \sum_{t=1}^{T_s}(1-Z_{ts}), a_2 + \sum_{t=1}^{T_s}(1-Z_{ts})(\mu_{ts}-m_s)^2\right)$$

The variance of observation is drawn from

$$\sigma_s^2|\cdot \sim \mathcal{IG}\left(b_1 + G/2, b_2 + \sum_{g=1}^{G}(x_{gs}-\mu_{t(g)s})^2/2\right)$$

Finally, the mixing proportion parameter is drawn from

$$\omega_s|\cdot \propto \prod_{t=1}^{T_s}\left\{\omega_s \mathcal{U}(\mu_{ts}; X_{s(1)}, X_{s(G)}) + (1-\omega_s)\mathcal{N}(\mu_{ts}; m_s, \tau_s^2)\right\}.$$

2.3.2 Gibbs Update for Expression Parameters

Mean and variance of the Gaussian component are updated from

$$\delta_g|\cdot \sim \mathcal{N}\left(\frac{\frac{\sum_{s=1}^{N}(1-W_{gs})Y_{gs}}{\eta_g^2} + \frac{\theta}{\psi^2}}{\frac{\sum_{s=1}^{S}(1-W_{gs})}{\eta_g^2} + \frac{1}{\psi^2}}, \frac{1}{\frac{\sum_{s=1}^{N}(1-W_{gs})}{\eta_g^2} + \frac{1}{\psi^2}}\right)$$

$$\eta_g^2|\cdot \sim \mathcal{IG}\left(d_1 + \sum_{s=1}^{N}(1-W_{gs})/2, d_2 + \frac{1}{2}\sum_{s=1}^{N}(1-W_{gs})(Y_{gs}-\delta_g)^2/2\right)$$

The extended tail parameters in the Uniform component is updated as follows:

$$\kappa_g^+|\cdot \propto \prod_{s=1}^{N_t}\left(\frac{1}{(u_g+\kappa_g^+)-(l_g-\kappa_g^-)}\right)^{W_{gs}} \rho^+ e^{-\rho^+\kappa_g^+}$$

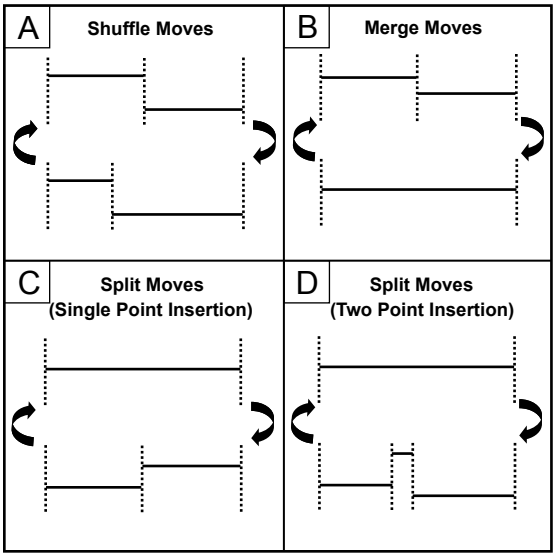


Figure 2.2: Copy number boundary updates in reversible jump MCMC. Four types of moves are suggested. Types B,C,D involve trans-dimensional moves.

$$\kappa_g^-|\cdot \propto \prod_{s=1}^{N_t} \left(\frac{1}{(u_g + \kappa_g^+) - (l_g - \kappa_g^-)} \right)^{W_{gs}} \rho^- e^{-\rho^- \kappa_g^-}$$

and the mixing proportion parameter is drawn from

$$\pi_g|\cdot \quad \propto \quad \prod_{s=1}^{N_t} \left\{ \pi_g \mathcal{U}(Y_{gs}; l_g - \kappa_g^-, u_g + \kappa_g^+) + (1 - \pi_g) \mathcal{N}\left(Y_{gs}; \delta_g, \eta_g^2\right) \right\}$$

Finally, the latent variable (for differential expression calls) are updated from Bernoulli distribution with a success probability

$$\frac{\frac{\pi_g}{(u_g+\kappa_g^+)-(l_g-\kappa_g^-)}}{\frac{\pi_g}{(u_g+\kappa_g^+)-(l_g-\kappa_g^-)} + \frac{1-\pi_g}{(\eta_g^2)^{1/2}} \exp\left(-\frac{(Y_{gs}-\delta_g)^2}{2\eta_g^2}\right)}$$

2.3.3 Breakpoint Arrangement Update by Reversible Jump MCMC

More challenging part of the sampling steps is altering the segment arrangement in the copy number data because this involves taking transdimensional moves.

Four types of arrangement changes have been used : (A) shuffling of existing breakpoints, (B) merging of two adjacent segments, (C) spliting of an existing segment by single point insertion, and (D) spliting of an existing segment by two point

insertion. These moves will be attempted at randomly chosen locations with corresponding probability of $(0.1, 0.4, 0.1, 0.4)$. The choice of these probabilities was made in a way that will give more chances for transdimensional moves, increasing the acceptance rates in the sampling. In our test runs, it was found that the two point insertion move is able to capture short length segments spanning five or less genes better than the single point insertion. This move resembles the operation of circular binary segmentation algorithm [69] where an arc is chosen from a circular band, i.e. chromosome with both ends tied to one another, for testing of differential copy number changes.

Shuffle Move

One can move an existing boundary point left or right, altering membership of the genes on the border line into either side of the two adjacent segments. This move will solely change the likelihood without changing the dimension of the parameter space, since it retains the same number of breakpoints. See Figure 2.2A. For this update, an existing boundary point is randomly selected, and a new location is proposed by randomly shifting the current location. The acceptance criteria is simply the likelihood ratio of the two adjacent segments (Metropolis-Hastings).

Split and Merge Moves

The more challenging updates are adding and removing boundary points. These moves are called split and merge moves (Figure2.2B-Figure2.2D). Since merge moves work exactly the opposite way split moves operate, only the split moves will be elaborated. There are two types of split moves, one in which a single boundary point is added inside a randomly chosen segment, and another in which two points are added so that resulting range flanked by the two new points form a new segment,

giving three daughter segments for the chosen segment. Single point insertions will add one additional mean parameter and one additional breakpoint, increasing the model parameter dimension by two, while two point insertions will add twice as many parameters, adding the dimension by four.

Split Move by Single Point Insertation

The single point insertion is discussed first. Updates are attempted at randomly chosen locations within each easmple. A new point p^* is poposed so that $p_{s(t-1)} < p^* < p_{st}$ for some $t \in \{1, 2, \ldots, T\}$. This additional point divides an existing segment with mean copy number μ_{ts} into two distinct daughter segment means, requiring the specification of two new mean copy number $\mu_{t_1 s}$ and $\mu_{t_2 s}$ in place of μ_{ts}. As there is an increment of dimension by one parameter in each sample, the two new mean values are proposed so as to satisfy

$$\begin{aligned} \mu_{t_2 s} - \mu_{t_1 s} &= \xi \\ (p^* - p_{s(t-1)})\mu_{t_1 s} + (p_{st} - p^*)\mu_{t_2 s} &= (p_{st} - p_{s(t-1)})\mu_{ts} \end{aligned}$$

where ξ is a random number generated from a Gaussian proposal $\mathcal{N}(0, k\sigma_s^2)$ for dimension matching purposes, where the constant k is selected in a way that will retain a minimal rate of acceptance. This update complies with the detailed balance condition of the reversible jump MCMC [52]. This proposal is equivalent to specifying the mean values for the two daughter segments:

$$\begin{aligned} \mu_{t_1 s} &= \mu_{ts} - \frac{p_{st} - p^*}{p_{st} - p_{s(t-1)}}\xi \\ \mu_{t_2 s} &= \mu_{ts} + \frac{p^* - p_{s(t-1)}}{p_{st} - p_{s(t-1)}}\xi \end{aligned}$$

This inverse relationship is used for the opposite move for merging. Notice that this transdimensional move has a unit Jacobian since the transformation $(\mu_{ts}, \xi) \mapsto$

$(\mu_{t_1 s}, \mu_{t_2 s})$ is orthonormal. Then the Metropolis-Hastings ratio for the acceptance of the new proposal becomes

$$\min\left\{(\mathrm{LR})\frac{\mathcal{P}(T_s+1;\lambda_s)}{\mathcal{P}(T_s;\lambda_s)}\frac{d_{T_s+1}(p_{T_s+1}-p_{s0})}{b_{T_s}(T_s+1)}\frac{f(\mu_{t_1 s})f(\mu_{t_2 s})}{f(\mu_{ts})},1\right\}$$

where LR denotes likelihood ratio and $f(\cdot)$ refers to the Uniform-Gaussian prior distribution for the segmental means.

Split Move by Double Point Insertion

The second type of split move proceeds by randomly selecting a segment and proposes two middle points p_1^* and p_2^* in a way that every one of the three resulting segments (p_{st}, p_1^*), (p_1^*, p_2^*), and $(p_2^*, p_{s(t+1)})$ contains at least one probe. This split move creates three segments, hence a single mean parameter needs to be divided into three daughter means, namely $\mu_{t_1 s}$, $\mu_{t_2 s}$, and $\mu_{t_3 s}$, such that

$$\begin{aligned}\mu_{t_1 s} &= \mu_{ts}+\xi_1\\ \mu_{t_2 s} &= \mu_{ts}+\xi_2\\ \mu_{t_3 s} &= \mu_{ts}+\xi_3\end{aligned}$$

subject to

$$\frac{p_1^*-p_{st}}{p_{s(t+1)}-p_{st}}\xi_1+\frac{p_2^*-p_1^*}{p_{s(t+1)}-p_{st}}\xi_2+\frac{p_{s(t+1)}-p_2^*}{p_{s(t+1)}-p_{st}}\xi_3=0$$

As in the previous case, this relationship can be inversely translated into

$$\mu_{t_3 s}=\mu_{ts}-\frac{p_1^*-p_{st}}{p_{t+1}^2-p_2^*}\xi_1-\frac{p_2^*-p_1^*}{p_{t+1}^2-p_2^*}\xi_2$$

Unlike in the single point insertion case, this parametrization comes with a non-unit Jacobian $(p_{s(t+1)}-p_{st})/(p_{s(t+1)}-p_2^*)$. With proposal of (ξ_1,ξ_2) from Gaussian kernel, the Metropolis-Hastings ratio for the acceptance of new proposal becomes

$$\min\left\{(\mathrm{LR})\frac{\mathcal{P}(T_s+2;\lambda_s)}{\mathcal{P}(T_s;\lambda_s)}\frac{(T_s+2)T_s L_{s0}}{2L^2}\frac{d_{T_s+2}(p_{s(T_s+1)}-p_{s0})}{b_{T_s}}\frac{f(\mu_{t_1 s})f(\mu_{t_2 s})f(\mu_{t_3 s})}{f(\mu_{ts})}|J|,1\right\}$$

where $f(\cdot)$ again refers to the Uniform-Gaussian prior distribution for the segmental means, and (L_{s0}, L) are the lengths of the chosen segment in sample s and the whole chromosome respectively, and $|J|$ is the Jacobian.

2.4 Breast Cancer cDNA Microarray Data

The proposed method was applied to the breast cancer data in Pollack et al [66]. A set of 5581 genes meeting 30% missing value criteria in both copy number and gene expression data were selected. This is slightly more stringent filtering compared to the procedure in van Wieringen and van de Viel [82]. Median centering was applied to both copy number and gene expression data. Standard deviation of each sample was adjusted to the median standard deviation across the 37 tumor samples in the gene expression data (sd=0.27). Using Pollack data has several advantages. First, the data was generated using the same cDNA microarray platform of ∼8,000 clones (300-500 bp long on average) for both copy number and gene expression profiles. Second, the clones in this array platform represent Unigene clusters and their homologue EST sequences, with average inter-clone distance of 0.5 million bp, providing genome-wide coverage at a modest resolution. More interestingly, nearly half the known oncogenes reported in cancer gene census of Futreal et al [84] are included in this set, and thus one can assess the impact of copy number on expression changes in the context of cancer studies. Third, previously proposed methodologies have been tested on this dataset [81, 82], and therefore it serves as a good benchmark dataset to compare the performance.

2.4.1 Prior Elicitation and Convergence of MCMC

Noninformative priors were used wherever possible. To be precise, priors for the variance parameters in the copy number data were set at $b_1 = b_2 = a_1 =$

$a_2 = 0.01$. Prior parameters for the genome-wide mean copy number parameter m_s for all samples were set at $\nu = 0$ and $\zeta = 1$. This prior can be considered as noninformative considering the fact that all copy number profiles have been equally scaled with standard deviation 0.27. Priors in the expression data were set at $\theta = 0$ and $\psi^2 = 100$ for the mean, $d_1 = d_2 = 0.01$ for the variance, $\rho+$ and $\rho-$ were set equal to the standard deviation of each gene for the tail of the Uniform component. All priors can be considered noninformative since the variability of prior has been set wider than the estimates from the raw data.

A very important prior distribution is regarding the mean number of copy number segments λ_s. When noninformative prior was given, i.e. $k_1 = 0.01$ and $k_2 = 0.01$, segmentation results varied widely across the samples, and the need for elaborate prior elicitation was noted. A relatively large value was preferred for this value in Pollack data because each clone in this cDNA microarray represents large chromosomal segments (resolution of $\sim$500K bp) and thus the variation of each clone may represent a segmental change. For this reason, $k_1 = G/100$ and $k_2 = 0.1$ were specified, where G is the number of genes on a chromosome. However, it is noted that (k_1, k_2) must be adjusted in individual datasets. In a high-resolution dataset such as high-throughput SNP array data, one can adjust the prior moderately high for λ_s, which also affects the computation time because the number of parameter updates is proportional to the number of segments.

For inference, samples were drawn from the posterior distribution using Markov chain Monte Carlo. 10,000 iterations were run with 1,000 initial period of burn-in. For Pollack data with 5581 genes, the entire algorithm takes around 30 minutes. One can reduce the computation time even further if some of the nuisance parameters are integrated out or plugged in with embedded MLE estimates using the EM algorithm,

e.g. variance parameters whose posterior distribution has a closed form solution, but this was not further pursued it in this work. Visual convergence diagnostic was conducted to monitor the convergence for randomly selected 50 copy number and gene expression parameters, namely $(\mu_{ts}, \sigma_s^2, m_s, \tau_s^2)$ and (δ_g, η_g^2). In five repeated runs, all selected parameters showed quick convergence to reasonable range of values within 200 initial burn-in period (not shown).

2.4.2 Regions with Aberrant Copy Number

The probabilistic copy number calls of DLMM were validated by benchmarking its cross-sample average copy number probabilities against the average copy number profile of 1136 breast cancer cases stored in Progenetix CGH database [85]. The latter can be regarded as a well-established copy number profile of breast cancer cases since the data consists of 40 independent studies of varying sample sizes (from 2 to 80). Figure 2.3 shows the graphical comparison between the two profiles. DLMM copy number profile shown in the first panel has at most 937 clones with aberrant copy number probabilities (0.2 in absolute value) concentrated in cytobands 1p32-p34, 1q, 8p21 (deleted), 8q21-24 (amplified), 16p11-12, 17q11 and 17q21-25, and 20q11-13. Copy number aberration in these regions have been reported in more than 20% of the samples across studies in Progenetix.

The copy number probabilities of DLMM have also been compared to those computed by CGHcall [86], which is one of the most recent methods that calculate similar posterior probability of amplification or deletion events using raw measurements and pre-existing segmentation calls. The bottom panel of Figure 2.3 shows the average copy number probabilities computed from CGHcall. Although the average profiles in DLMM and CGHcall overlapped in most chromosomes, some of the calls for deletion events in CGHcalls were not found in DLMM, e.g. chromosomes 4, 5, 10, 15, and 18.

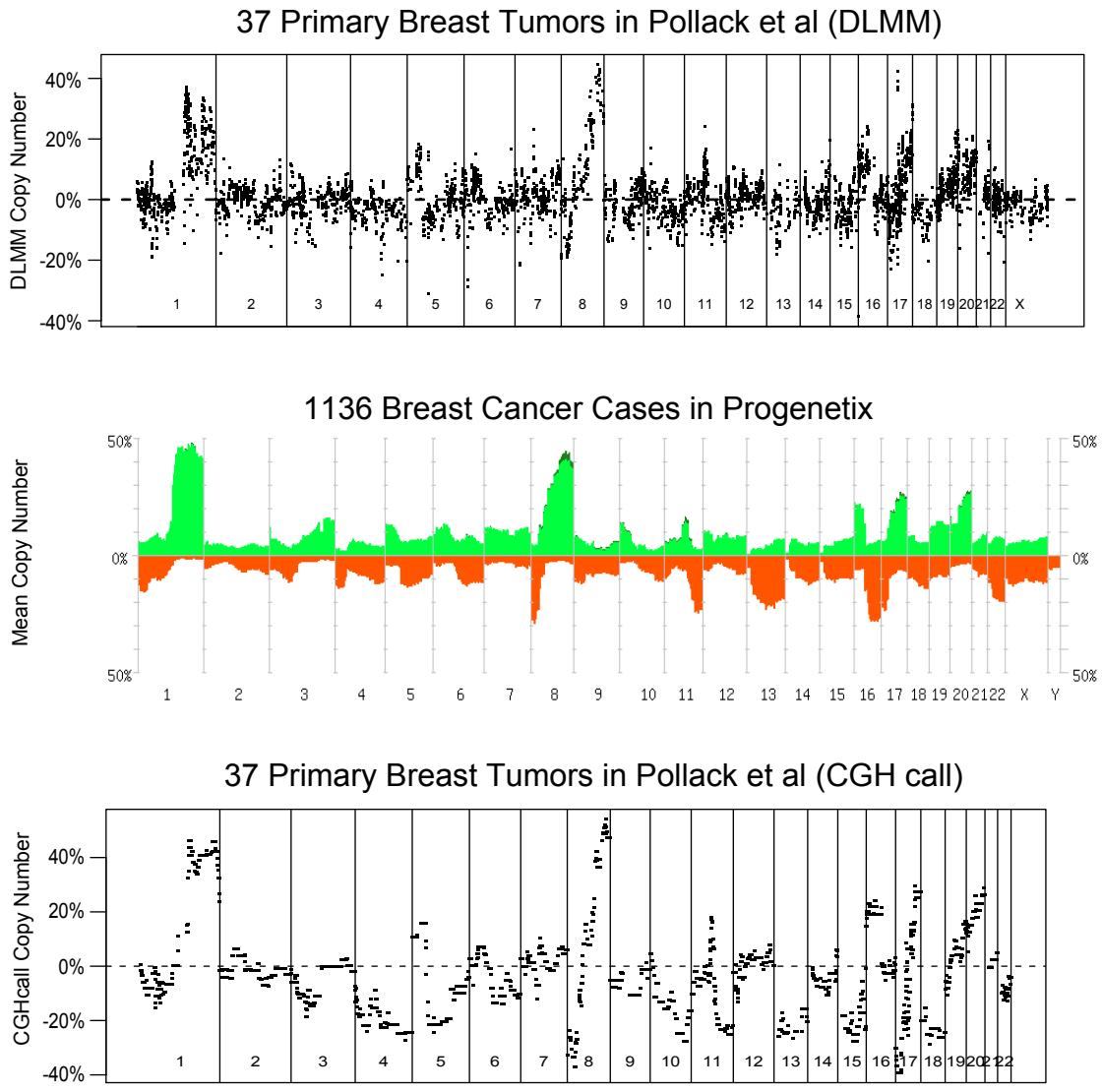


Figure 2.3: Copy number probability calls against breast cancer cases in Progentix. Average copy number probabilities of DLMM in Pollack et al [66], Progenetix data of 1136 breast cancer cases, and mean copy number calls by CGHcall method of van de Viel et al [86].

Table 2.1: Gene selection criterion of DLMM using L-measure. L-measure values were calculated with $\nu = 0.5$ for selecting copy number-associated gene expression changes. Decimal points were rounded off.

Threshold	0.01	0.02	0.03	0.04	0.05	0.10	0.20	0.50
L-measure	121587	119772	119384	119350	119385	119496	119686	119857

However, no pronounced deletion events were found in all five chromosomes in the Progenetix data. By contrast, a Progenetix record of 20% deletion event in chromsome 13 was recovered more clearly by CGHcall than DLMM. Unless the benchmark set fails to represent the general breast cancer population, the overall comparison shows that DLMM and CGHcall make similar copy number calls with a caveat that the latter method can be more prone to false positive calls.

2.4.3 Copy Number-Associated Gene Expression Changes

Using the probabilistic scoring and the criterion-based gene selection, genes were selected if the score was 0.04 and above in absolute value in each sample separately. The threshold score 0.04 was chosen based on the minimal L-measure across multiple candidate cutoff points shown in Table 2.1. Following this step, 203 genes with copy number-associated over or under expression in near 10% frequency (3 out of 37 samples) were selected. The set of selected genes will be called DLMM signature from here on.

Congruent with the results reported in Lipson et al [81] and van Wieringen and van de Viel [82], a large proportion of the selected genes were found on the amplified regions on chromosomes 1, 8, and 17. Eight genes from the cancer gene census [84] were included in the list: APC, FGFR1, EXT1, MYC, FANCA, MLLT6, ERBB2, and CLTC. As a clear demonstration of how the probabilistic scoring works, Figure 2.4 shows the case of ERBB2 on the cytoband 17q11, where 8 (22% of 37) samples shows clear amplification events. All these samples were assigned the joint probability

score 0.4 and above as shown in the top right panel of the figure. The other seven genes all show similar patterns (data not shown). Even though the proportion of the actual oncogenes included in the DLMM signature is low, it is interesting to observe that 152 genes (75%) are located within 500K bp distance from at least one oncogene, indicating a high degree of proximity of DLMM signature to the oncogenes. The observation that the expression of oncogenes themselves is not largely influenced by the copy number changes should not be surprising since the oncogenes are targets of more direct regulation controlled by other oncogenes and tumor suppressors than cytogenetic events.

In order to strengthen the biological interpretation of the DLMM signature, DAVID [87] was used to examine the enrichment of Gene Ontology (GO) biological processes in the DLMM signatures. Table 2.2 lists the GO terms with the highest statistical significance. As expected, the genes related to the regulatory activities regarding cell death and cell cycle are deemed to be target of copy number-driven expression changes. Despite its small number of hits, It is interesting to observe the term 'positive regulation of epithelial cell proliferation,' as primary breast epithelial cells are the major target for carcinogenesis. It is noted that this interpretation is quite different from the analysis of van Wieringen and van de Viel [82], where a significantly greater number of genes (1225) were selected based on their hypothesis testing framework, or NPtest as is called in short in this work.

In addition to the GO term analysis, the DLMM signature seems to be highly correlated with the clinical indicators of breast cancer provided in Pollack et al [66]. To see this, the frequency of having a score above the threshold was calculated in each sample, i.e. $\sum_{g=1}^{G} 1\{|P_{gs}| \geq 0.04\}$, resulting in an enrichment index of copy number-associated gene expression changes. This index was compared with the tumor grade,

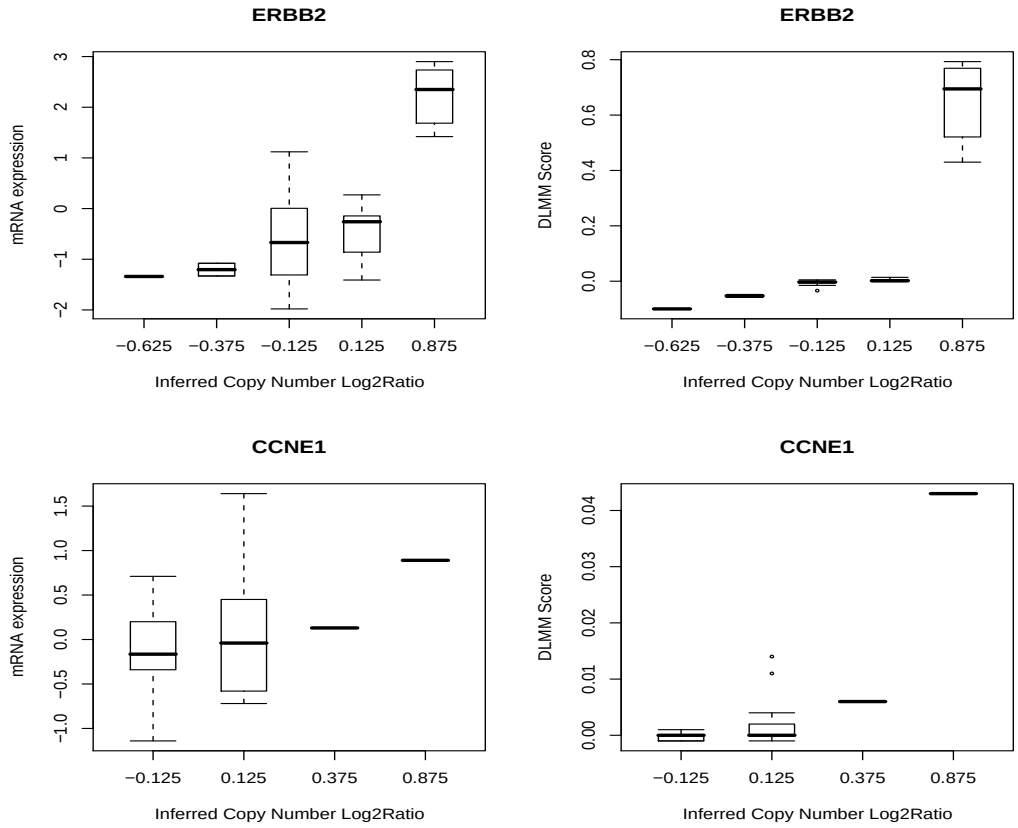


Figure 2.4: Oncogenes ERBB2 and CCNE1. Observed gene expression and DLMM score against copy number probabilities in the two genes. ERBB2 was selected by DLMM, GCMS, and nonparametric tests as top candidate, while CCNE1 was selected only by the nonparametric tests. Other rank-based tests did not pick up CCNE1 either. However, copy number probabilities in CCNE1 is significantly high in only one out of 37 tumor samples.

Table 2.2: Gene Ontology (Biological Process) terms enriched in the DLMM signature

Function	Counts	p-value	FDR
apoptosis	22	8.8e-05	0.2%
cell death	22	2.1e-04	0.4%
regulation of apoptosis	16	6.4e-04	1.1%
regulation of progression through cell cycle	14	4.2e-03	7.2%
negative regulation of progression through cell cycle	8	7.2e-03	12.1%
integrin-mediated signaling pathway	5	7.4e-03	12.3%
positive regulation of epithelial cell proliferation	3	2.1e-02	31.3%
negative regulation of apoptosis	7	3.7e-02	49.0%
cell morphogenesis	11	4.2e-02	53.1%

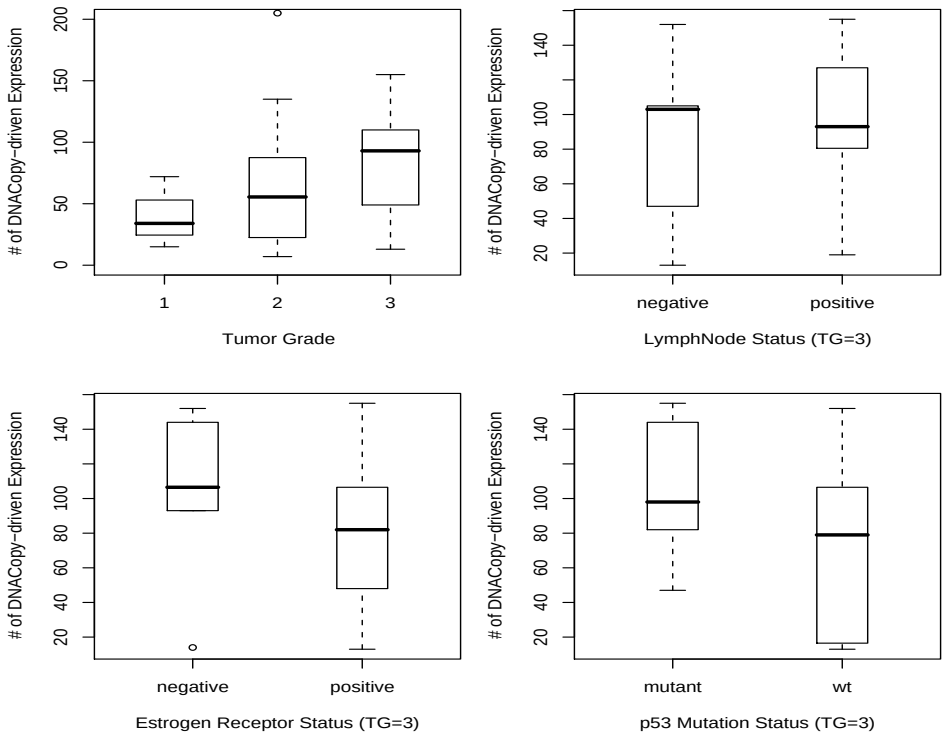


Figure 2.5: DLMM score and clinico-pathological information. Sample-specific enrichment index of copy number associated gene expression is correlated with tumor grade and other clinico-pathological information related to breast cancer. The index was compared against lymph node status, estrogen receptor status, and p53 mutation information for the tumors in grade 3 only due to biased sampling of low grade tumors with respect to the distribution of lymph node and estrogen receptor status.

as well as lymph node status, estrogen receptor (ER) status, and p53 gene mutation status. The latter three comparison was conducted only for tumor grade 3 samples because almost all lymph node negative (and thus ER negative) samples were at lower tumor grade. Figure 2.5 shows the result. The top left panel illustrates that the samples in higher tumor grade has increased enrichment of the target genes. The top right panel shows that lymph node (metastasis) positive samples tend to have more copy number-associated gene expression changes, while the bottom left panel shows the similar trend for ER negative samples relative to ER positive ones. Also, the bottom right panel indicates that the somatic mutation in p53 gene is also positively correlated with the number of copy number-associated expression changes.

2.4.4 Comparison

In order to comparatively assess the performance of DLMM and highlight its distinct features, the DLMM signature was compared to the genes selected by using the GCSM method of [81] and the NPtest method proposed in [82].

Comparison with GCSM

GCSM searches for genes whose expression levels are linearly correlated with raw copy number levels in the local neighborhood of the gene itself. [81] reported 174 genes with the GCSM score above 40, and this list includes five oncogenes reported in [84] (PRCC, SET, MLLT6, ERBB2, MYH9). Comparing the signatures, 53 genes (26% of DLMM) overlaps with DLMM signature, implying that there is a significant discrepancy between the two gene selection criteria. This is expected since the analysis in DLMM is one-to-one correspondence between the two data without regional analysis.

It was found that many genes unique to the GCSM signature is from the regions where probabilistic copy number profiles show little aberrant behavior, which means that high linear correlations can still be observed in regions with few significantly aberrant copy number changes. Figure 2.6 clearly shows this result. The top left panel shows the average (signed) copy number probabilities in all 5581 genes, and the top right and the bottom left panels show those in DLMM and GCSM signatures respectively. These figures illustrate that the DLMM signature is enriched in regions with higher prevalence of significant copy number changes than GCSM, enhancing its specificity of copy number-associated expression changes in the former. However, it is also important to note that the regional analysis feature of GCSM can still be valuable since it has recovered three new oncogenes (PRCC, SET, MYH9) that were

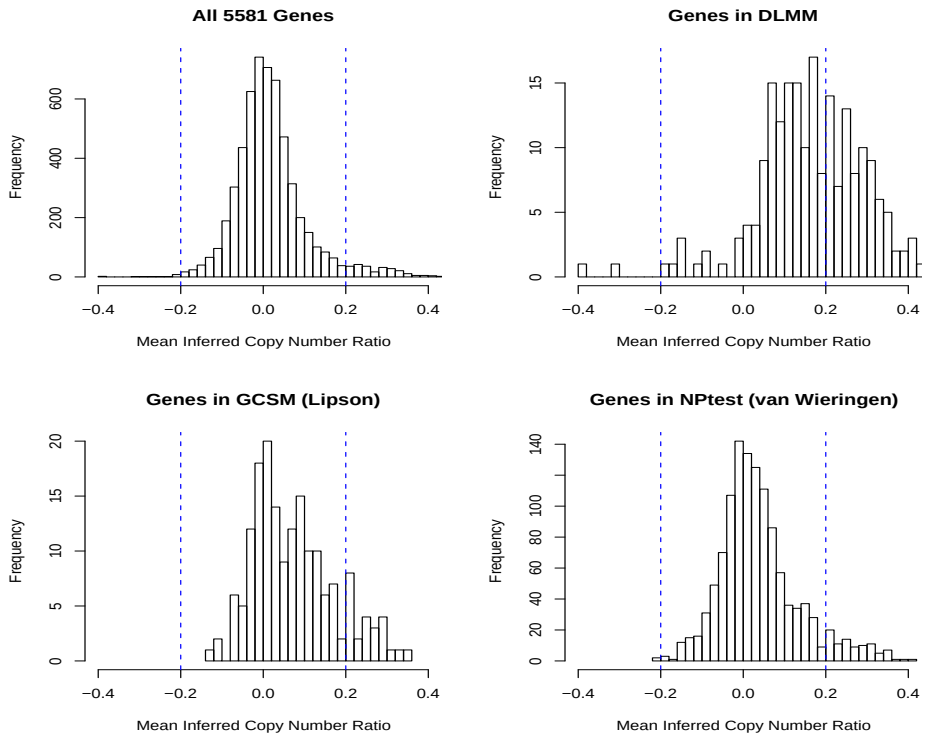


Figure 2.6: Comparison of DLMM against GCSM and NPtest. DLMM signature was compared with GCSM signature of [81] and NPtest signature of [82] in terms of average copy number profiles. Many selected genes in the latter two sets are not enriched in regions with aberrant copy numbers.

not recovered by DLMM.

Comparison with NPtest

van Wieringen and van de Viel [82] propose a modified Cramér-Von Mises test and another test based on weighted Mann-Whitney statistic to test the equality of gene expression distribution between samples with and without copy number gains or losses. Significance of test statistics are computed based on permutations and probability weights computed by CGHcall are used as weighting factors in this process. NPtest method has reported a total of 1225 genes (22% of 5581 genes) meeting FDR 10% criterion. These include 37 genes from cancer gene census, which accounts for 3% of the total. DLMM signature shares 125 (61% of DLMM) genes with this

set, which is more than double the overlap with the GCSM signature. All eight oncogenes in the DLMM signature are in the common signature as well.

Despite the close overlap, the two gene signatures are vastly different in terms of size. Note that the size of the NPtest signature (1225) seems to be surprisingly large to represent the group of genes with copy number-associated expression because the number of clones with average copy number probabilities (0.2 or above in absolute value) was no more than 973 in DLMM analysis. In order to see this, histogram of the mean copy numbers of all 1225 genes were drawn in the bottom right panel of Figure 2.6. The plot shows that the distribution of mean copy number is almost identical to the entire set of 5581 genes, without significant enrichment in aberrant copy number levels (e.g. 0.2 and above in absolute value).

To investigate this observation more closely, the relationship between estimated copy number calls and raw gene expression was examined, and that between estimated copy number calls and DLMM scores was also compared for the ten genes used for the power study in van Wieringen and van de Viel [82]. These genes were included in the NPtest signature and they were selected for the power study because these genes are candidates known to be associated with the development of breast cancer in the literature. Thus the assumption made in their work is that these genes serve as the gold standard where copy number associated expression changes are supposed to be observed. Surprisingly, DLMM selected ERBB2 gene only (a few genes were filtered out in the missing data filter). However, when the copy number profiles were revisited for the remaining nine genes inferred from both CGHcall method and DLMM, it was observed that either the proportion of samples with high copy number probability calls was low, or the expression distribution was not clearly separable between samples with and without aberrant copy number changes in probability.

This also corroborates with the previous observation that the majority of oncogenes reported in cancer gene census [84] were not directly associated with copy number-driven expression changes. See the example of CYCLINE gene (CCNE1) shown in the bottom panels of Figure 2.4. Although the pattern exhibits positive correlation between the two data, only two samples have probability calls above 0.2 (out of 37 samples). It is obvious that such a small proportion cannot robustly represent the group of copy number aberrant samples, and therefore the test may result in a false positive call.

2.5 Discussion

In this work, a model-based method DLMM has been proposed for identifying coherent signals in the paired profiles of copy number and mRNA expression. DLMM consists of probabilistic copy number calling method and mixture model-based differential expression analysis that incorporates the copy number calls in it. The method achieves the goal by computing the joint probability of aberrant copy number and concordant differential gene expression between samples with and without copy number changes, and thus accounts for uncertainty in both data simultaneously. The analysis of breast cancer data from Pollack et al [66] has shown that the copy number probability profile estimated by DLMM is largely congruent with a large-scale repository of breast cancer cases, and the selected signature of genes showing evidence of copy number associated expression are located in the vicinity of known oncogenes while many oncogenes themselves were not directly under the influence. The sample index constructed from the selected genes was also positively correlated with the clinicopathological information, highlighting the potential of this gene signature as a diagnostic or prognostic measure in cancer.

Joint inference of these two datasets is challenging particularly because copy number data should be analyzed within each sample while gene expression data analysis is a comparison across samples. The reason the copy number data analysis is specific to individual samples is that, unlike properly normalized gene expression data, copy number ratios of a gene are not directly comparable across samples for two main reasons. First, every tumor specimen is a mixture of tumor and normal cells and the ratio of this mixture varies by sample. Thus with a common reference sample used in competitive hybridization, the copy number level in each sample is affected by the proportion of tumor cells in the specimen, especially for genes in aberrant copy number. Hence approaches that take the raw copy number data as measurements comparable across the samples, e.g. Lipson et al [81], may be subject to unexpected errors and this was indirectly shown in the analysis of Pollack data. Secondly, relative copy number levels can be inferred more accurately if one considers the segmental patterns present in the copy number data, particularly because the signal-to-noise ratio is not often very high and the level of noise can be learned well from the genome-wide profile of each sample. For example, the median sample standard deviation in local windows of 100 clones was around 0.12, while the copy number ratio in the regions most frequently reported as amplified (1q, 8q) was as small as 0.35 in relevant samples. Thus those methods using sample-specific copy number probability calls such as DLMM and NPtest seem to be more relevant than linear correlation analysis.

Despite the differences of inferential techniques in DLMM and NPtest, the two methods share a common principle of distinguishing gene expression distributions between samples with and without aberrant copy number levels. NPtest adopts a non-parametric hypothesis testing framework for the hypothesis that the expression

distributions are equal in the two groups of samples by incorporating uncertainty of copy number calls in each sample with a tuning algorithm for unbalanced grouping. However, it was shown that, through the examples of the oncogenes in the Pollack data, the method still selects genes whose copy number calls are high in few samples only even after applying the tuning algorithm proposed in their work. DLMM takes a different approach, where the scores of copy number associated expression levels is computed for individual gene in each sample and the frequency that the score is above a chosen threshold across the samples is used for final gene selection. This approach seems more relevant than both NPtest and GCSM in the joint analysis because copy number associated gene expression changes is a relatively rare event compared to our direct gene regulation activities.

DLMM can easily be extended to tumor-normal comparisons or comparisons between different types of tumors by changing the way the final joint probability is calculated from the latent variables Z and W. In tumor-normal case, one can perform a semi-parametric estimation by making the normal samples contribute to the estimation of parameters in the mixture component for samples without aberrant copy number levels, i.e. (δ_g, η_g^2) with 100% chance by fixing $W = 0$ since normal cells are supposed to have little copy number aberration. In tumor-tumor comparisons, one should keep track of copy number changes in the two groups separately, and select genes whose copy number-driven expression changes are unique in either group. DLMM can also be used for the data where more than a single copy number probe or clone can be mapped to a gene in the expression data. The segmentation applies to high-resolution arrays exactly the same way the Pollack data was analyzed in this work, and one can still score the coherent signal in the two data by defining multiple W variables for each pair of copy number probe and gene expression probe.

CHAPTER III

Hierarchical Hidden Markov Model with Application to Joint Analysis of ChIP-seq and ChIP-chip data

3.1 Study of Transcriptional Regulation by ChIP-experiments

Chromatin immunoprecipitation (ChIP) is a powerful method for isolating a particular protein bound to DNA sequences *in vivo* [88, 89]. In ChIP experiments, cells are first treated with reagents such as formaldehyde inducing protein-DNA crosslinks, and DNA is isolated and fragmented afterwards. An antibody specific to a target protein is added to precipitate the interacting pairs, and their crosslinks are reversed. The resulting DNA fragments are direct evidence for physical interactions between the target protein and the genes. These DNA segments can be simultaneously mapped to the genome with array-based hybridization, which is known as ChIP-chip [90, 91]. This technology has been widely used for the identification of transcription factor binding sites (TFBS). Recently, ChIP experiment coupled with massively parallel sequencing [92], or ChIP-seq, has been proposed as an alternative to ChIP-chip [93, 94, 95, 96, 97, 98, 99]. ChIP-seq offers genome-wide coverage in a single basepair resolution at low cost [100].

Although a number of previous studies have demonstrated the power of ChIP-seq, it has also been shown that different mapping strategies may identify mutually exclusive peak regions as candidate binding sites. For instance, Robertson et al [94]

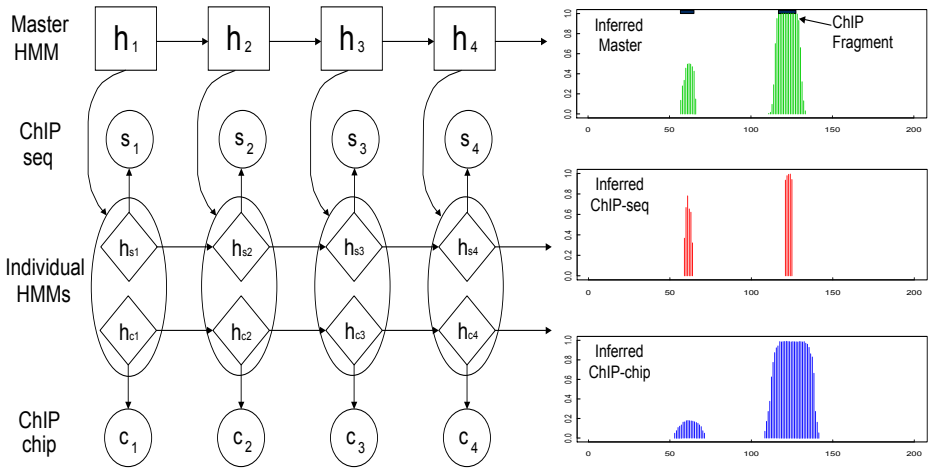


Figure 3.1: Hierarchical hidden Markov model framework. The master process in the top layer summarizes multiple individual processes in the bottom layer. The hidden states in ChIP-seq and ChIP-chip data are considered as emission from the master process.

reported that the overlap between STAT1 ChIP-enriched regions identified by ChIP-seq and ChIP-chip is around 60%. Euskirchen et al [101] found that ChIP-chip and ChIP-PET [102, 103], a sequencing-based method, are frequently complementary to each other in identifying validated targets when the signal is not sufficiently strong. The evidence by Robertson et al [94] suggests that massively parallel sequencing may not work well for all DNA fragments uniformly. For example, the sequencing can be biased toward certain parts of the genome due to the complex chromatin structure of DNA molecules in their native form. Also, sequence reads may also have a reduced sensitivity in the genomic regions where repeat sequences exist. For those DNA fragments, other mapping methods not reliant on direct sequencing, e.g. ChIP-chip, can be a valuable source to complement the weakness of the sequencing technology.

For many of the existing ChIP-seq data, ChIP-chip experiments have also been conducted and the data are publicly available. It is desirable to take advantage of existing ChIP-chip data sets to assist in TFBS identification using ChIP-seq. While such a joint analysis has promise, it is a challenging statistical task to account for

the heterogeneity of data from the ChIP-chip and ChIP-seq platforms. This is because the two technologies show vastly different behavior in terms of sensitivity and specificity. Specifically, the peaks identified by ChIP-seq are expected to form regions that are much sharper than those in ChIP-chip due to its superior resolution, whereas ChIP-chip tends to report broader regions of moderate significance including potential false positives. Hence the signals from the two channels have to be appropriately weighted in order to keep the overall false positive rates low in the joint analysis.

To this end, a hierarchical hidden Markov model (HHMM), a collection of multiple individual-level hidden Markov models (HMMs) governed by a population or master-level HMM, is proposed. The structure of the model is illustrated in Figure 3.1. In this method, individual-level HMMs function as a de-noising filter that converts the raw data into inferred binary hidden states representing ChIP-enrichment and background noise, and the master-level HMM uses individual-level hidden states as a basis to infer the underlying *true* states. In this process, individual-level HMMs serve as a buffer to reduce the heterogeneity present in raw ChIP-chip and ChIP-seq data, and the master-level HMM summarizes their ChIP-enrichment status to produce the final probability score.

Development of HHMMs has been proposed previously in the literature [104, 105]. Recently, Shah et al [106] used this class of models for accurately detecting boundary points of copy number changes across multiple samples in genome-wide array-comparative genomic hybridization (aCGH) data. In their model, hidden states in the individual samples exchange mutual feedback with the hidden state in the master level. By contrast, for our problem, data from each channel is represented as an individual HMM, whose inferred hidden states are then modeled as the bivariate

emission probabilities of the master-level HMM. Thus the inference on the master level is based on the status of ChIP-enrichment in the two channels.

3.2 ChIP-chip and ChIP-seq Data

Data generated from ChIP-chip and ChIP-seq experiments are different. ChIP-chip data are intensity levels from probes on tiling arrays which are assumed to be proportional to the quantity of DNA fragments matching the probe. Probes on tiling arrays are usually 36-50 basepair long. Elevated intensity levels from multiple probes indicate ChIP-enrichment. By contrast, raw ChIP-seq data are sequencing reads with alignment position on the genome, where each sequence read is assumed to be copied from a DNA fragment in the sample. High frequency of sequence reads indicates ChIP-enrichment.

Because a HMM framework was adopted, the data is first summarized into fragment counts in units of windows of fixed size (25 nucleotides in this study and adjustable) along the genome. Dissecting chromosomes into windows of equal length has been used previously in the ChIP-seq literature [96]. Since the start and end positions of ChIP-chip probes do not match to these windows precisely, ChIP-chip probe boundaries were redefined so as to match the ChIP-seq windows (later in the master-level HMM). With typical probes having length greater than 25 nucleotides, one ChIP-chip probe can be mapped to multiple windows.

3.3 HHMM Model

Basic notations are introduced first. Let the ChIP-seq count data and the ChIP-chip intensity data denoted by $S = (s_1, s_2, \ldots, s_T)$ and $C = (c_1, c_2, \ldots, c_T)$ respectively, for a chromosome that has been divided into T windows, assuming the number of windows is identical in the two data. It is assumed that each channel

follows its own independent HMM. Their respective hidden states are denoted by $h_s = (h_{s1}, h_{s2}, \ldots, h_{sT})$ and $h_c = (h_{c1}, h_{c2}, \ldots, h_{cT})$. As shown in Figure 3.1, these hidden states are modeled as bivariate random variables in the emission of master HMM, whose hidden states are denoted by $h = (h_1, h_2, \ldots, h_T)$. Hidden states in both the individual level (h_s, h_c) and the master level h consist of either ChIP enriched (denoted 1) or background (denoted 0) states. Note that the ultimate goal of HHMM is to infer the master-level hidden states h.

The model parameters are now specified in the individual level first. The three main components of HMM – the initial probabilities, transition probabilities and emission probabilities [107] – are defined. The initial state distribution $\pi(h_{s1})$ and $\pi(h_{c1})$ and the transition probabilities $A^s_{x,y} = P(h_{st} = x, h_{s(t+1)} = y)$ and $A^c_{x,y} = P(h_{ct} = x, h_{c(t+1)} = y)$ for $x, y = 0, 1$ and $t = 1, \ldots, T$ can either be fixed or estimated from the data. In the latter case, one can assume each row of A^s and A^c follows multinomial distribution and estimate the probabilities from the frequency of relevant moves in the inference of h_s and h_c respectively. Parametric models are used to describe emission probabilities in different states.

3.3.1 HMM in ChIP-chip

In ChIP-chip data, Gaussian and Uniform distributions are used to model the observed hybridization intensities from the array probes. To be precise, $C_t|h_{ct} = 1 \sim \mathcal{U}_{\theta_{c1}}(\cdot)$ and $C_t|h_{ct} = 0 \sim \mathcal{N}_{\theta_{c0}}(\cdot)$, where $\mathcal{U}$ and $\mathcal{N}$ denote uniform and normal distributions in the ChIP-enriched and the background states respectively. Gaussian-Uniform mixture model has been previously used to detect differentially expressed genes in microarray data analysis [54]. The uniform distribution parameters θ_{c1} are fixed as the minumum and maximum of intensities $\{C_t\}_{t=1}^T$, and the mean and variance parameters of the normal distribution $\theta_{c0} = (\mu_c, \sigma_c^2)$ will be estimated.

The likelihood for the ChIP-chip HMM model can be written as

$$\pi(h_{c1}) \prod_{t=2}^{T} \pi(c_t|h_{ct}, \theta_{h_{ct}}) \pi(h_{ct}|h_{c(t-1)}, A_c)$$

A Bayesian inference was implemented for the HMM [108], which iteratively repeats imputation and posterior sampling steps: (1) Imputation: in this step, the path of hidden states is drawn from $h_c^{(i+1)} \sim \pi(h_c|C, \theta_{c0}, \theta_{c1}, A_c)$ using the forward-backward algorithm [107, 108], and (2) Posterior sampling: in this step, the parameters for the emission probabilities are drawn from $\theta_{cj}^{(i+1)} \sim \pi(\theta_{cj}|C, h_c^{(i+1)}, A_c)$ for $j = 0, 1$. A conjugate prior is assumed for the emission in the background state, i.e. $\mu_c \sim \mathcal{N}(\nu_c, \tau_c^2)$ and $\sigma_c^{-2} \sim \mathcal{G}(\alpha_c, \beta_c)$. All hyperpriors were set to be non-informative prior. Recall that the distribution for ChIP-enriched class has been specified as uniform distribution with fixed minimum and maximum parameters, so it remains to update parameters for the background state only. To be precise, the mean and variance parameters are drawn from

$$\begin{aligned} \mu_c &\sim \mathcal{N}\left(\frac{\sigma_c^{-2}\sum_{\mathcal{T}_0} c_t + \tau_c^{-2}\nu_c}{\sigma_c^{-2}|\mathcal{T}_0| + \tau_c^{-2}}, \quad \frac{1}{\sigma_c^{-2}|\mathcal{T}_0| + \tau_c^{-2}}\right) \\ \sigma_c^{-2} &\sim \mathcal{G}\left(\alpha_c + |\mathcal{T}_0|/2, \quad \beta_c + \sum_{\mathcal{T}_0}(c_t - \mu_c)^2/2\right) \end{aligned}$$

where the set $\mathcal{T}_0$ denotes $\{t = 1, \ldots, T : h_{ct} = 0\}$ and $|\mathcal{T}_0|$ is the cardinality of the set. Given all emission parameters, $\{h_{ct}\}_{t=1}^T$ were sampled using the algorithm in Scott [108]. The posterior probabilities for $\{q_{ct}\}_{t=1}^T$, where $q_{ct} = P(h_{ct} = 1|C)$, were calculated by averaging posterior samples of h_{ct} across all t.

3.3.2 HMM in ChIP-seq

In ChIP-seq data, in order to account for over-dispersion and higher proportion of zero counts, generalized Poisson (GP) and zero-inflated Poisson (ZIP) are used to model read counts in ChIP-enriched and background states respectively [109, 110].

There are two parameters (λ_1, λ_2) in the GP distribution, with the probability mass function of the observed count s_t at location t

$$P(s_t; \lambda_1, \lambda_2) = \frac{\lambda_1(\lambda_1 + \lambda_2 s_t)^{s_t - 1}}{s_t!} e^{-\lambda_1 - \lambda_2 s_t}$$

On the other hand, the ZIP distribution is a mixture distribution of a point mass at zero and a Poisson distribution:

$$P(s_t; \alpha_0, \rho) = \alpha_0 \delta_0(s_t) + (1 - \alpha_0) e^{-\rho} \frac{\rho^{s_t}}{s_t!}$$

where $\delta_0(\cdot)$ is a point mass at zero. For inference purposes, a latent variable Z_t is defined for sequence count s_t at location t such that $Z_t = 0$ if $s_t = 0$ and s_t is generated from the point mass at zero, and $Z_t = 1$ otherwise (thus it is always the case $Z_t = 1$ if $s_t > 0$).

The likelihood for the HMM model can be written as

$$\pi(h_{s1}) \prod_{t=2}^{T} \pi(s_t | h_{st}, \theta_{h_{st}}) \pi(h_{st} | h_{s(t-1)}, A_s)$$

where θ_{s1} and θ_{s0} are parameters of GP and ZIP distributions respectively. A Bayesian inference was implemented with the following posterior sampling distributions in the emission. For the emission in the background state, non-informative Uniform prior is assumed for α_0 for the mixing proportion parameter. Then the posterior distribution for α_0 is proportional to the likelihood, i.e.

$$\alpha_0 \propto \prod_{t \in \mathcal{T}_0} P(s_t; \alpha_0, \rho)$$

where the set $\mathcal{T}_0$ denotes $\{t = 1, \ldots, T : h_{st} = 0\}$ as in ChIP-chip. Now a non-informative gamma prior $\mathcal{G}(0.01, 0.01)$ is assumed for the Poisson mean ρ. Then the posterior distribution for ρ is a gamma distribution

$$\rho \sim \mathcal{G}(0.01 + \sum_{t \in \mathcal{T}} Z_t s_t, 0.01 + \sum_{t \in \mathcal{T}} Z_t).$$

After updating these parameters, Z_t is sampled for those windows with zero count ($s_t = 0$) from Bernoulli distribution with probability $\frac{e^{-\rho}}{1+e^{-\rho}}$ and set $Z_t = 1$ for all other windows with positive counts ($s_t > 0$).

For the emission in ChIP-enriched states, the likelihood is re-written with respect to $\lambda' = \frac{\lambda_1}{1-\lambda_2}$ and $\theta' = \frac{1}{(1-\lambda_2)^2}$ after change of variables. With non-informative priors $P(\lambda') \propto 1$ for $\lambda' > 0$ and $P(\theta') \propto 1$ for $\theta' < 1$, the posterior distribution for these two parameters is proportional to the Generalized Poisson likelihood of all observations with $h_{st} = 1$, i.e.

$$P(\lambda', \theta'|s_t) \propto \prod_{t \in \mathcal{T}_1} P(s_t; \lambda', \theta')$$

where the set $\mathcal{T}_1$ denotes $\{t = 1, \ldots, T; h_{st} = 1\}$. With this likelihood, it is easy to proceed to construct Metropolis-Hastings update. For this part, only the data points in the ChIP-enriched states ($h_{st} = 1$) are used. Both parameters are updated by rejection sampling with symmetric Gaussian proposals. In both the ChIP-enriched and the background states, parameter updates were subject to the following identifiability constraint: $E(s_t|h_{st} = 1) \geq E(s_t|h_{st} = 0)$. This constraint generally avoids the label switching problem in two component mixture models such as two-state HMM [111].

Given all emission parameters, $\{h_{st}\}_{t=1}^T$ were sampled using the algorithm in Scott [108]. The posterior probabilities for $\{q_{st}\}_{t=1}^T$, where $q_{st} = P(h_{st} = 1|S)$, were calculated by averaging posterior samples of h_{st} across all t.

3.3.3 Master HMM

In the master level, the initial state distribution $\pi(h_1)$ and transition probabilities $A_{x,y} = P(h_t = x, h_{t+1} = y)$ for $x, y = 0, 1$, $t = 1, \ldots, T$ are defined the same way as in the individual level. For the emission, the data (h_s, h_c) are modeled with two multi-

nomial distributions, i.e. $(h_{st}, h_{ct})|h_t = 1 \sim \mathcal{M}_{\theta_1}(\cdot)$ and $(h_{st}, h_{ct})|h_t = 0 \sim \mathcal{M}_{\theta_0}(\cdot)$, where the distribution for the enriched state $\mathcal{M}$ denotes multinomial distribution, and $\theta_1 = (p^1_{00}, p^1_{01}, p^1_{10}, p^1_{11})$ and $\theta_0 = (p^0_{00}, p^0_{01}, p^0_{10}, p^0_{11})$ are their parameters for ChIP-enriched and background states respectively. These parameters are given a conjugate Dirichlet prior with parameters $(\gamma^1_{00}, \gamma^1_{01}, \gamma^1_{10}, \gamma^1_{11})$ and $(\gamma^0_{00}, \gamma^0_{01}, \gamma^0_{10}, \gamma^0_{11})$ respectively.

Given the posterior probability pairs (q_{st}, q_{ct}) at all positions $t = 1, \ldots, T$ estimated in the individual-level HMMs, hidden states in the master level are inferred as follows. Had (h_{st}, h_{ct}) been observed directly, the likelihood for the master level HMM would be

$$\pi(h_1) \prod_{t=2}^{T} \pi(h_{st}, h_{ct}|h_t, \theta_{h_t}) \pi(h_t, h_{t-1}, A)$$

From the inference of individual HMM, $\{(q_{st}, q_{ct})\}_{t=1}^T$ are computed, but the actual hidden states $\{(h_{st}, h_{ct})\}_{t=1}^T$ remain unknown still. Treating this as a missing data problem, the likelihood is integrated over all four possibilities of (h_{st}, h_{ct}) based on the marginal weights (q_{st}, q_{ct}), i.e.

$$\pi(h_1) \prod_{t=2}^{T} \left[\sum_{(h_{st}, h_{ct})} g_t \cdot \pi(h_{st}, h_{ct}|h_t, \theta_{h_t}) \right] \pi(h_t, h_{t-1}, A)$$

where $g_t = (q_{st})^{h_{st}}(1 - q_{st})^{(1-h_{st})}(q_{ct})^{h_{ct}}(1 - q_{ct})^{(1-h_{ct})}$. This multiplicative factor g_t weights the four possible cases of (h_{st}, h_{ct}) based on the product of their corresponding marginal posterior probabilities in ChIP-seq and ChIP-chip at position t, as an approximate solution to the missing data problem.

With this likelihood, imputation and posterior sampling steps are iterated as in the ChIP-chip case: (1) Imputation: draw $h^{(i+1)} \sim \pi(h|h_s, h_c, \theta_0, \theta_1, A)$ using the forward-backward algorithm, and (2) Posterior sampling: draw $\theta_j^{(i+1)} \sim \pi(\theta_j|h_s, h_c, h^{(i+1)}, A)$ for $j = 0, 1$. With the multinomial likelihood and the Dirichlet prior, the posterior is again Dirichlet distribution, thus $\theta_j = (p^j_{00}, p^j_{01}, p^j_{10}, p^j_{11})$ are drawn from

$\mathcal{D}(\gamma^j_{00}+H^j_{00}, \gamma^j_{01}+H^j_{01}, \gamma^j_{10}+H^j_{10}, \gamma^j_{11}+H^j_{11})$ where $H^j_{kl} = \sum_t 1\{h_{st} = k, h_{ct} = l, h_t = j\}$ for $k, l = 0, 1$ and $j = 0, 1$.

Prior was elicited to reflect the known technological difference between ChIP-seq versus ChIP-chip in terms of precision and sensitivity. The ideal posterior distribution is one that scores peaks from ChIP-seq higher than those from ChIP-chip on average, but a non-informative prior is not able to achieve such weighting because signals are more abundant in the latter. This reinforces the need to elicit an informative prior for the master-level HMM. In fact, there are ways to conjecture the optimal posterior distribution in real data. For example, if one is aware of the false positive rates in ChIP-seq and ChIP-chip, then the posterior can be set so that the ratio p^1_{10}/p^1_{01} is inversely proportional to the ratio of false positives. One can also learn this knowledge from preliminary motif search in TFBS identified in ChIP-seq and ChIP-chip and reflect the sensitivity ratio in p^1_{10}/p^1_{01} and p^1_{11}/p^1_{10}. Through multiple simulations and real data analysis, it was found the following prior works well: $\gamma^1_{11} = M/2$, $\gamma^1_{10} = M/5$, and $\gamma^1_{01} = M/10$ in the ChIP-enriched windows, and $\gamma^0_{kl} = 1$ in the background windows. This specification leads to $(p^0_{00}, p^0_{01}, p^0_{10}, p^0_{11}) = (0.48, 0.36, 0.15, 0.01)$ and $(p^1_{00}, p^1_{01}, p^1_{10}, p^1_{11}) = (0.001, 0.001, 0.01, 0.988)$ in our simulation study, and the same prior is used in the real data analysis presented later as well.

The elicited prior results in the posterior probability ratios $1 < r_{01} < r_{10} < r_{11}$ where $r_{kl} = p^1_{kl}/p^0_{kl}$. This requirement is important since the noise in ChIP-chip will substantially increase counts H^1_{01} and as a result the posterior probability for regions identified by ChIP-chip only will be higher than the regions identified by ChIP-chip only if a noninformative prior were used. Admission of ChIP-chip unique signals with higher frequency than ChIP-seq unique signals is likely to result in elevated false positive rates. It should be noted that, however, when the data

supports such relationship without strong prior elicitation, this setting is unnecessary and a noninformative prior should be used instead.

Given the parameter updates by sampling, the forward-backward equations are computed in order to sample the hidden states $h = \{h_t\}_{t=1}^{T}$ sequentially. The posterior probability is then estimated by the average of the posterior output. At the last step, contiguous blocks of windows above a fixed threshold probability are identified as ChIP-enriched regions, where in this study the minimum length of a ChIP-enriched region was set at 100bp.

3.3.4 Regions with Missing Data

Because of technology limitations and repetitive regions, neither ChIP-seq nor ChIP-chip is able to survey all bases of the human genome. Regions that are inaccessible from both are marked and skipped. There are also regions on the genome that is accessible by ChIP-seq only or ChIP-chip only. When data from one source is missing, the inference of the hidden states at the upper level in HHMM will rely on the other data source alone. That is, using the marginal distribution (Bernoulli) of the joint distribution to model the observed (non-missing) data.

3.4 Simulation Study

A simulation study was conducted in order to assess the performance of HHMM. The posterior probabilities were generated instead of the raw signals, as the focus of this simulation study is the assessment of master level HMM, where the information from both data sources are combined.

First, the master-level hidden states h in a chromosome containing a hundred thousand probes (25M bp chromosome) were simulated from a stationary Markov

chain with a transition probability matrix

$$A = \begin{pmatrix} 0.99 & 0.01 \\ 0.15 & 0.85 \end{pmatrix}$$

Hidden state 1 denotes ChIP-enrichment. ChIP-enriched states have been accepted only when the probes formed a contiguous block, i.e. all 'singletons' of the enriched state have been converted to the background state. This generates the baseline 'truth' where the true ChIP enriched sites are 150bp long on average (range from 75bp to 1375bp and IQR of 100bp $\sim$ 250bp).

Given a value of hidden state $h_t = 1$ at each locus t, posterior probabilities $P(h_{st} = 1|S)$ and $P(h_{ct} = 1|C)$ have been generated from Beta distributions with mean 0.9 and 0.8 respectively. In order to reflect higher resolution in ChIP-seq over ChIP-chip, data was generated so that each true ChIP-enriched region is almost exactly covered by a ChIP-seq peak region with ChIP-chip signals surrounding it. Negative signals ($h_t = 0$) have been placed as follows. Reflecting the actual false positive rates of less than 5% in ChIP-seq and 25% in ChIP-chip previously reported in analyses of real data sets [94], these false positive signals were planted in blocks of 3-8 windows with probability 0.05 and 0.25 in the two data sets respectively.

Datasets with four possible sampling behaviors (p_s, p_c) have been simulated. Sampling behavior here refers to the sensitivity of each data source producing signal within the true ChIP-enriched regions. Case I ($p_s = 0.75, p_c = 0.9$) and Case II ($p_s = 0.6, p_c = 0.8$) represent scenarios where ChIP-chip signals appear with a greater frequency (with a greater error rate) than ChIP-seq, which may represent the cases where the sequencing depth is low and hence a number of real ChIP enriched regions are missed by ChIP-seq. Case III ($p_s = 0.75, p_c = 0.75$) and Case IV ($p_s = 0.9, p_c = 0.9$) represent scenarios where both data sources cover real binding

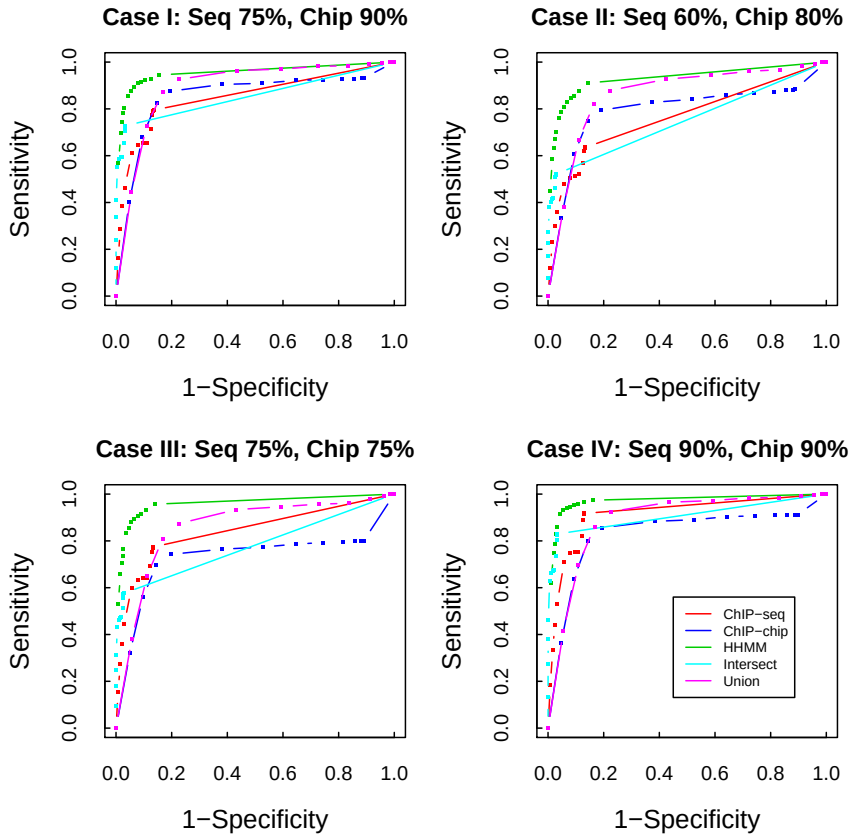


Figure 3.2: Receiver operating characteristic in the simulation study

site motif regions with a good sensitivity and some of the platform specific regions host a good number of real motifs. Other scenarios of a varying range of combinations with the fixed p_s/p_c ratio have also been simulated, and the results were consistent.

HHMM was compared with four other ways to identify ChIP-enriched regions with high probability. (1) ChIP-seq only: peak regions from ChIP-seq HMM; (2) ChIP-chip only: peak regions from ChIP-chip HMM; (3) Intersection: common peak regions in both sources; and (4) Union: peak regions from either source. Figure 3.2 shows the receiver operating characteristic (ROC). In all examples, HHMM is the best performing method in terms of sensitivity followed by Union, outperforming both single-source analyses. More importantly, HHMM keeps the specificity higher

than Union for nearly all decision points (square dots). The fact that the ROC curve bent to the right significantly for high specificity decision points in the Union indicates that blind picking of all signals would result in high false positive rates at a fixed specificity, mostly due to ChIP-chip data. HHMM removes most of the low-key negative signals, which can be seen in the upper left corner of the ROC curves. Despite the high specificity, Intersection is the least competitive with respect to the ROC in the simulated data.

In sum, the results in Cases I through IV indicate that the area under the curve of the ROC is the highest in HHMM followed by Union and the number of positive calls is almost always highest in HHMM at fixed specificity. In all scenarios where either the more advanced mapping platform misses some of the true signals or both platforms complement the identification for each other, HHMM has the potential to collect the highest number of binding sites and, at the same time, keep the false discovery rates lower than blind picking of all signals.

Meanwhile, another dataset was simulated with $(p_s, p_c) = (0.8, 0.6)$, where the better performing platform ChIP-seq covers most of the signals picked up by ChIP-chip. Examination of ROC curve shows that HHMM, ChIP-seq only, and Union methods perform equivalent to one another, indicating that there is no additional benefit earned by HHMM as expected. Also, this is consistent with the fact that the ROC improved the least in Case IV out of the four scenarios, where the number of overlapping signals in ChIP-seq and ChIP-chip is the largest among all. This is an intuitive result since the more advanced platform captures more of real signals in the data, and thus repetitive identification in another platform with reduced resolution does not improve the results.

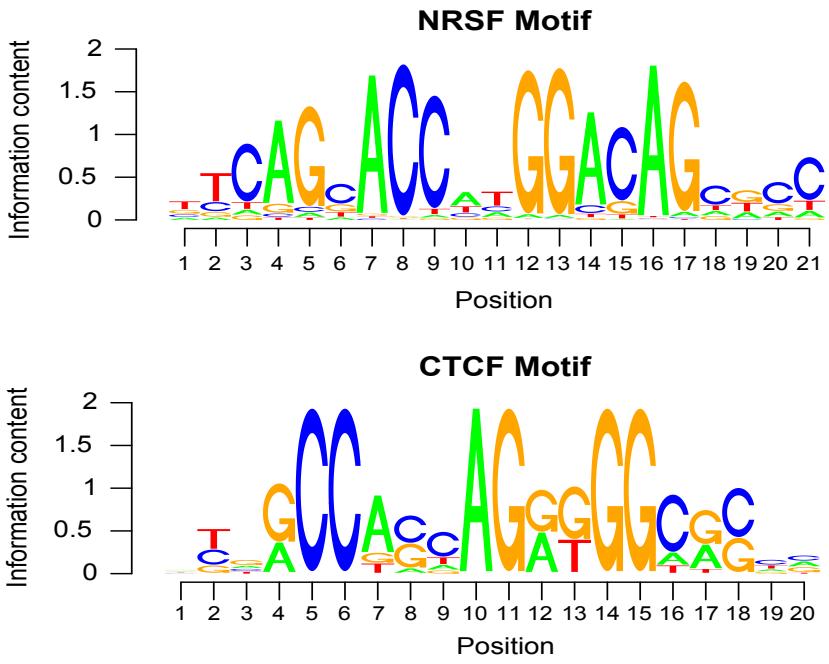


Figure 3.3: Binding site motifs of NRSF and CTCF. Sequence logos are shown for both transcription factors

3.5 Application to NRSF and CTCF Data

HHMM was applied to real datasets to evaluate the performance in terms of the statistical significance of motif enrichment in two well-studied transcription factors. For regions identified by each method, MatInspector [112, 113] in Genomatix (http://www.genomatix.de) was used to find TF binding sites. In alignment scoring, the position weight matrix (PWM) stored in MatBase of Genomatix for the NRSF data, or a custom built PWM reported in Kim et al [114] for the CTCF data were used. See Figure 3.3. All default parameter settings were used throughout. For comparison of different peak selection methods, the enrichment of TFBS motifs was tested by chi-squared test in a contingency table setting. The rows of the 2×2 table indicate whether the motif search was done in the original sequence or the permuted sequence, and the columns indicate whether the sequences contain motifs or not. Match rate was also computed, which is defined by the difference in the number of

motifs found in the original sequence and the permuted sequence per 1kb, which can be used a measure of average resolution of TFBS identification in different methods.

3.5.1 NRSF Data

In a recent study, Johnson et al [93] used the ChIP-seq technique to study genome-wide mapping of binding sites of NRSF, a neuron-restrictive silencer factor NRSF, known for its negative regulation of many neuronal genes in non-neuronal cells [115], were mapped to approximately 2,000 locations in the human genome using ChIP-seq. ChIP-seq data was available from reference [93] and an unpublished ChIP-chip data for NRSF in the Nimblegen ENCODE array platform was also available in Gene Expression Omnibus (GSE7372). Since the Nimblegen ENCODE tiling arrays do not cover the whole genome, this section focuses on the 10 ENCODE regions each spanning 5 million bps, i.e. approximately 1% of the human genome.

The regions identified in ChIP-seq and ChIP-chip data covered roughly 28.8K basepairs and 4.7M basepairs respectively. Among these, 422 windows were overlapping, which represents 37% of ChIP-seq. The posterior probabilities were then combined into a parallel matrix as mentioned previously, and master-level HMM was fit. Peak regions were selected with five different probability thresholds (0.2,0.4,0.6,0.8,0.9) using all the methods compared in the simulation study.

NRSF binding site motifs were searched within the selected peak regions. In the search, PWM of the 21bp-long motif reported in Schoenherr et al [116] was used in Genomatix (See the top panel of Figure 3.3). The sequences have also been randomly permuted and the motif search was reiterated, providing a reference to assess the significance of the hits. It is important to realize that the peak regions selected with a certain fixed threshold is not necessarily comparable from one method to another. Rather, one should examine the overall performance across all five thresholds when

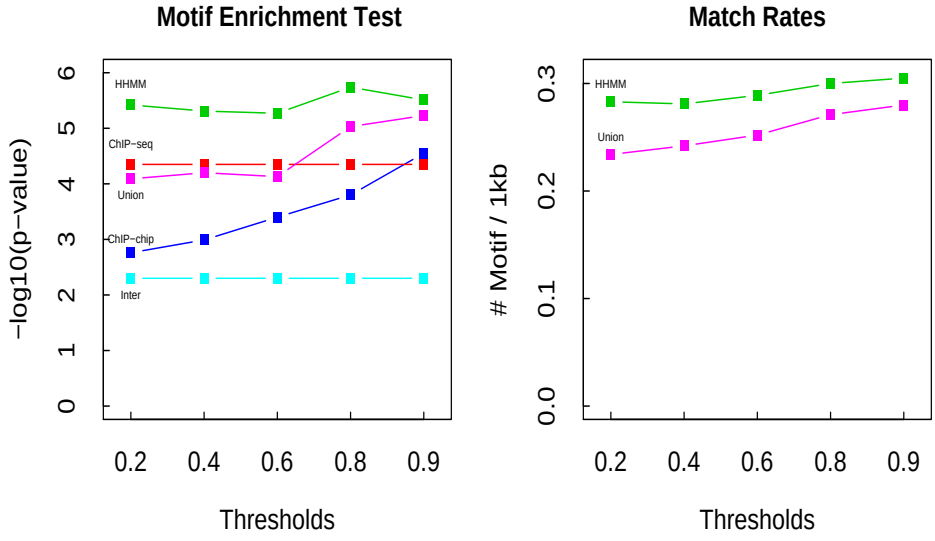


Figure 3.4: Motif enrichment across multiple probability thresholds. Significance analysis of motif enrichment by the five methods at multiple probability thresholds (0.2,0.4,0.6,0.8, 0.9). The left panel shows the negative log10 transformed p-value in chi-squared tests. The right panel shows the match rates, the number of motifs per 1 kilobases, in HHMM and Union.

comparing different selection methods.

Figure 3.4 shows the result of the analysis. The left panel shows that the chi-squared test of motif enrichment gives the highest statistical significance in the selected regions reported from HHMM among all methods, followed by Union or ChIP-seq depending on the choice of the thresholds. The right panel compares the match rates, i.e. the number of binding site motifs per 1 kb, in the two best competing methods Union and HHMM. It is observed that HHMM shows a higher match rate across all thresholds, indicating higher resolution of identified TFBS's. For a more detailed illustration, see the case of a high probability threshold 0.9. Table 3.1 shows that the total number of motifs is the highest as 67 in the Union method, but the HHMM picks up 46 motifs while keeping the false positives less than half of the Union, indicating improved control of false positive rates, at the expense of a fewer low ranking signals.

Method	#Match	#Permute	#Peaks	Coverage(bp)	OR	-$\log_{10}(p)$	Match Rate
HHMM	46	11	416	177.0K	4.58	5.51	0.20
Union	67	24	817	274.5K	2.95	5.23	0.16
ChIP-seq	25	4	53	25.9K	10.94	4.35	0.81
ChIP-chip	52	17	788	254.9K	3.20	4.55	0.14
Intersect	10	1	24	6.2K	16.43	2.30	1.45

Table 3.1: Binding site identification results in NRSF data. Regions containing at least one peak with probability 0.9 and above were selected and motifs with position weight matrix corresponding to NRSF motif were searched. Match rate is defined as (#Match - #Permute) / 1kb.

Method	#Match	#Permute	#Peaks	Coverage(bp)	OR	χ^2	Match Rate
HHMM	23,772	4,815	65,808	30.31M	7.16	16,057.36	0.63
Union	26,788	6,200	83,325	40.08M	5.89	16,018.71	0.51
ChIP-seq	16,771	1,836	25,372	9.33M	25.00	18,926.85	1.59
ChIP-chip	16,599	5,134	69,246	33.83M	3.94	7,172.77	0.34
Intersect	6,310	719	9,576	3.06M	23.80	7,023.18	1.83

Table 3.2: Binding site identification results in CTCF data. Peak regions that contains a signal with probability 0.9 and above were selected and motifs with position weight matrix corresponding to CTCF motif were searched. Match rate is defined as (#Match - #Permute) / 1kb.

3.5.2 CTCF Data

As a second example, the binding sites of CTCF were mapped on a genome-wide basis using the ChIP-chip data from Kim et al [114] and the ChIP-seq data from Barski et al [95]. CTCF is a zinc finger protein that has a multivalent character as a transcription factor [117, 118] capable of participating in both repression and activation due to the combinatorial use of its 11 zinc fingers. CTCF zinc fingers can be selectively utilized based on the different needs of target genes, and thus the binding sites are likely to be more variable than other transcription factors. For instance, Kim et al [114] has reported 62 genes for which multiple CTCF binding sites were identified. See the bottom panel of Figure 3.3 for the binding site motif reported in the study.

Individual HMM fits in this data showed that 419,457 windows in ChIP-seq and

around 3.4 million probes (7.1 million windows worth) in ChIP-chip had positive posterior probabilities, where 152,025 windows overlapped with each other (37% of ChIP-seq). Among these, around 1.5 million windows had posterior probabilities 0.9 in at least one channel in the individual level, and around 1.2 million of these had 0.9 and above probability in the master level.

Table 3.2 presents the motif enrichment test results based on the analysis with the probability threshold 0.9. It is easy to see that HHMM and Union are the two methods that collects the highest number of TFBS motifs, but the number of random hits in the permuted sequences show almost a 3 to 4 ratio, indicating that the relative significance of motif search results is improved in HHMM.

Since the number of hits in a genome-wide data is extremely large, all chi-squared tests gave p-value of flat zero, and the statistics themselves are of a similar magnitude as most methods demonstrates a certain degree of enrichment. However, the odds ratio of observing motifs in the selected regions was higher in HHMM (7.16) than in Union (5.89), and the match rate was also higher in HHMM (0.98/1,000 basepairs) than in Union (0.84/1,000 basepairs). This improvement is an obvious consequence of the fact that the regions picked by HHMM (30M basepairs) is far narrower than those picked by Union (40M basepairs) on average.

On the other hand, ChIP-seq data from Barski et al [95] seems to demonstrate the ultra-performance of ChIP-seq, where 62% of the motifs found in Union were identified, but the search regions are so specific that the number of random hits is low (16,711/26,788) and therefore the odds ratio and the match rates are high. But it is the goal of HHMM to find a compromise between Union and ChIP-seq only analysis, in which an extra 7,000 motifs were saved by allowing some of the most significance ChIP-chip specific regions at the expense of a reduced overall statistical

significance of motif enrichment.

3.6 Discussion

The availability of multiple experimental datasets profiling the activity of a specific transcription factor is an important asset for delineating regulatory mechanisms. The proposed HHMM method not only identifies more binding sites with increased specificity, but also serves as an assessment of agreement and discrepancy between both technologies. It is noted that the proposed methodology may not be optimal when the best performing experimental platform (ChIP-seq in this case) reports most of the data from the other platforms, since additional information with a decreased precision will do nothing but dilute the signal with little contribution to finding extra binding site motifs. Nevertheless, it is difficult to expect that the new sequencing technology will always be able to provide perfect coverage of the genome in practice, and thus the previously deposited ChIP-chip data sets may be of significant value in improving TFBS identification in most cases.

Although our method is designed for combining ChIP-chip and ChIP-seq data, the HHMM framework is rather general and can be applied to other scenarios where information collected from multiple sources may be integrated. The opportunities for this type of joint analyses frequently arise in biomedical research. With the rapid development of new technologies, there are often multiple assays co-existing, measuring the same or closely related quantities of interest. As an example, both microarray and SAGE can be used to measure gene expression levels. Each method has its own advantages and disadvantages. Even for microarray, there are multiple platforms available such as cDNA spotted microarray, Affymetrix GeneChip and Agilent long oligo array. Also for measuring protein-DNA binding, a series of assays

have been developed, e.g. ChIP-PCR, ChIP-chip, ChIP-PET and ChIP-seq. Since these assays often have different sensitivity and specificity, straightforward combinations such as union and intersection do not work well. HHMM, on the other hand, is built under a coherent probability framework that is able to handle heterogeneity in sensitivity and specificity from the individual channels, and therefore allows for easy incorporation of data from multiple experimental platforms. With the technologies constantly changing, the advantage of this joint modeling approach is expected to have far-reaching implications.

CHAPTER IV

Significance Analysis of Quantitative Proteomic Data using Spectral Counts

4.1 Label-Free Quantitative Proteomics by Spectral Counting

The following two chapters discuss hierarchical modeling in the context of quantitative proteomic data analysis. Although there is a significant difference between genomics and proteomics datasets, properly processed datasets of either experimental platform can be modeled using Bayesian hierarchical models since they naturally address the small sample size, which is the common problem.

Mass spectrometry (MS) - based large-scale shotgun proteomics is currently the most commonly used approach for the identification and quantification of proteins in large-scale studies [119, 120, 121]. A variety of mass spectrometry-driven protein quantification methods have been proposed including conventional 2D-gel electrophoresis followed by the MS-based identification, as well as the methods involving stable isotope labeling of proteins or peptides coupled with tandem mass spectrometry (MS/MS) sequencing, e.g. Isotope-Coded Affinity Tags (ICAT) and multiplexed quantification using isobaric tagging reagents (iTRAQ) [122, 123, 124, 125, 126].

In the recent years, the so-called label-free methods have received increasing attention as a promising solution that automatically waives some of the disadvantages of the stable isotope labeling methods. Developments in that area have focused on

the analysis of two-dimensional images (spectrogram) of ion intensities across the span of retention time and mass-to-charge (m/z) ratio from a liquid chromatography (LC) - MS or LC-MS/MS run, where peak intensities are used as the abundance measure [127, 128, 129]. Despite the rich physico-chemical information contained in the spectrogram, the computational effort required for processing the data, including background filtering, peak detection and alignment, is often daunting.

Another viable yet much simpler quantification strategy is spectral counting, where the number of spectra matched to peptides of a protein is used as a surrogate measure of protein abundance. A number of recent studies have demonstrated that spectral counting can be as comprehensive as ion peak intensities in terms of detection range, while retaining linearity [130]. A number of groups have proposed a variety of normalized scores based on transformed spectral counts, including methods that explore weighted scoring by peptide match score [131], normalized by the number of potential peptide matches [132] or peptide sequence length and overall experiment-wide abundance [133], or models that incorporate the probability of identification into counting [134]. There are also other works that not only attempt to refine the abundance index, but also propose to adopt standard statistical tests on the raw/transformed counts to analyze the protein expression data. See Fu et al and Zhang et al [135, 136].

Despite many published successful examples, there is a lack of standard computational and statistical methods for analyzing this type of data as well established as in the gene expression data, including differential expression analysis such as significance analysis of microarray data (SAM) [137], clustering and classification, and network analysis [138, 139, 140]. In fact, most studies demonstrating the use of spectral counts have resorted to data-driven corrections of conventional signal-to-noise

ratio statistics such as mean-variance model adjustment [141] and detection rate adjustment [135]. These adjustments are primarily targeted to correct the bias in the statistic that favors large differences in highly abundant proteins. The technical challenges for modeling quantitative proteomics data are distinct in their own right and can be described in two ways. First, neither ion peak intensities nor spectral counts can be easily modeled with standard distributions as in the gene expression data. This increases the burden of finding the appropriate statistical model and estimation methods. Second, due to the cost and effort considerations, profiling two or more replicates is rarely done in the comparison of distinct biological conditions. This makes it difficult to perform robust estimation and inference on the model parameters since it is not feasible to observe evidences consistently over many samples in homogeneous biological condition. Even the permutation-based method for generating reference distributions will not work well unless more than 4 or 5 replicates are generated for each condition.

In this chapter, a general statistical framework for analyzing spectral count data is described. The method can not only address the issue of the appropriate probability distribution for count data, but also tackle the lack of statistical information due to the absence of replicate samples. In this work, hierarchical Bayes estimation is implemented assuming a generalized linear mixed effects model (GLMM) [142] for the entire matrix data, where the spectral counts of a protein are considered to be random numbers from a large population of proteins and hence the model parameters are directly shared within replicates and across proteins. This comprehensive modeling strategy is more powerful than calculating the signal-to-noise ratio type of differential expression test statistics per protein basis and referencing them to an approximate null distribution, especially when the number of replicates is limited.

This section is organized as follows. First, overall modeling framework is described and its applicability to a wide variety of experimental designs is discussed. The performance of the model-based protein selection is evaluated using synthetic datasets with comparison to the methods using signal-to-noise ratio statistics, particularly in terms of the power to detect differentially expressed proteins at fixed error rates and the property of the detected proteins such as abundance. A experimental dataset taken from Pavelka et al [141] is re-analyzed for the comparison of proteomic profiles of a Yeast strain at two different phases in cell growth. Annotation enrichment analysis is performed to compare the biological functions highlighted by the protein signature detected by the proposed method and the conventional signal-to-noise method, and related computational and statistical issues are discussed.

4.2 Statistical Model

4.2.1 QSpec: Hierarchical Bayesian Model

For a dataset with n samples and p proteins, a model-based method is proposed to select proteins whose absolute abundance changes by a statistically significant amount in different biological conditions. Spectral counts of a protein are modeled as observations from the Poisson distribution and the expected counts as a linear function of normalizing factors, treatment or disease status, and other experimental information. Unlike in gene expression datasets, typical proteomics dataset have data over a few replicates or samples only, and as a result, fitting Poisson regression model for individual proteins separately is nearly infeasible as a consequence.

A seemingly reasonable approach might be to use fold change ratios as indicators of relative protein changes, and follow up with a proper global error rate correction. However, this approach selects proteins based solely on the effect size without incorporating the variability, and therefore it may introduce a number of false positive

calls in low abundance proteins where a small difference may result in a large fold change ratio. To face the challenges, a statistical methodology called hierarchical Bayes, which pools the statistical information on the regression models across proteins, can be used. Considering each protein as a member of the population of all identified proteins, the regression parameters for each protein is modeled as random effects.

More specifically, suppose that a spectral count data matrix $X = [X_{ij}]$ is given. Assuming that X_{ij} are observations from a Poisson distribution with expected count μ_{ij}, for $i = 1, 2, , p$, consider the model for the expected count matrix as a generalized linear mixed model (GLMM)

$$\log(\mu_{ij}) = \log(L_i \cdot N_j) + a_0 + b_{0i} + b_{1i}T_j \tag{4.1}$$

where μ_{ij} is the expected count for protein i in replicate j, L_i is the sequence length of protein i, N_j is the normalizing constant of replicate j, a_0 is the baseline abundance, and b_{0i} and b_{1i} are the protein specific abundance and differential expression parameters for protein i. The first term in the right hand side of the Equation (4.1) is a fixed normalizing term, often referred to as offset in regression analysis. The protein sequence length L_i adjusts for the bias in the count for longer proteins, and the normalizing constant N_j of replicates adjust for the overall abundance of each replicate or sample. For the latter constant, the aggregate count across all proteins in each sample is used to reflect the total abundance of all proteins identified in each MS/MS experiment. Most importantly, the treatment effect is defined as follows: $T_j = 1$ if replicate j is in treatment, and $T_j = 0$ otherwise. If the treatment effect b_{1i} were a redundant parameter, then the model in Equation 4.1 reduces to

$$\log(\mu_{ij}) = \log(L_i \cdot N_j) + a_0 + b_{0i} \tag{4.2}$$

Given the model setup, the probability distribution is specified for the model parameters as follows. Since M_R is a nested model of M_F, it suffices to write the model specification for M_F. Even though the expected spectral counts are expressed in the form of a GLMM, the connection across the model parameters in different proteins has yet to be established. To this end, assume the likelihood with Poisson distribution $X_{ij} \sim \text{Poisson}(\mu_{ij})$ where μ_{ij} is a linear function of a_0, b_{0i}, and b_{1i}. It is assumed the conventional Gaussian prior $a_0 \sim N(0, \sigma_a^2)$ and $(b_{0i}, b_{1i}) \sim \mathcal{N}(\mathbf{0}, \text{diag}(\sigma_0^2, \sigma_1^2))$ for all $i = 1, \ldots, P$, and Inverse Gamma prior for the variance component $\sigma_j^2 \sim \text{IG}(\alpha, \beta)$ for $j = 0, 1$.

4.2.2 Tests for Differential Expression

The full model in Equation (4.1) and the reduced model in Equation (4.2) are named as M_F and M_R respectively. If the evidence from the spectral count data supports M_F over M_R, the protein will be considered as differentially expressed. Comparing the goodness of fit by M_F and M_R leads to the selection of differentially expressed proteins because the model with differential expression parameter fits the data better than the model without it if the protein is indeed differentially expressed. The exact protein selection method will be described in the next section more precisely.

The strategy for determining whether each protein is differentially expressed between the two conditions is straightforward. For each protein, Bayes factor [143] is calculated

$$B_i = \frac{p(X_i|M_F)}{p(X_i|M_R)} = \frac{\int p(\Theta_F|M_F)p(X|\Theta_F, M_F)d\Theta_F}{\int p(\Theta_R|M_R)p(X|\Theta_R, M_R)d\Theta_R} \tag{4.3}$$

In Equation 4.3, the numerator and the denominator are essentially the likelihoods of observing the counts under M_F and M_R respectively. Thus if this ratio is large,

the data supports the model with the differential expression parameter over the model without, providing a probabilistic evidence that the protein is differentially expressed.

Conventionally, Bayes Factor greater than 10 suggests a strong evidence for the model in the numerator, and Bayes Factor greater than 30 suggests a very strong evidence for the same model according to Jeffreys [143]. However, these conventional cutoffs do not work efficiently in the high-throughput datasets due to the need to apply the multiple testing correction. Applying a sole Bayes factor threshold, however, may have its own minor drawback when there are low quality replicates. Empirically it was found that Bayes factor can be over-estimated due to the heterogeneity in counts that cannot be explained by a single expected count rather than the real differential expression, especially in extremely high abundance proteins. In this case, the averaged likelihood in the model without the differential expression parameter tends to be penalized more than the model with the parameter. In order to address this issue, it was enforced that the selected proteins have a fold change by no less than 50%. In the subsequent data analysis, it was also found that almost all proteins filtered by this step are in the high abundance range, and the number of these proteins is quite small.

4.3 Inference

Model parameters were estimated by taking average of posterior samples generated from the random walk Metropolis-Hastings algorithm with Gaussian kernel. That is, for each parameter θ, a random sample θ' was drawn from Gaussian distribution with mean equal to the current value θ^c, and the proposal was accepted with

probability

$$\min\left\{1, \frac{\text{Likelihood}(\theta'|\text{data})\text{Prior}(\theta')}{\text{Likelihood}(\theta^c|\text{data})\text{Prior}(\theta^c)}\right\}$$

. The order of parameter update was as follows:

- Update $a_0 \propto \mathcal{N}(0, \tau^2) \cdot P\left(\{X\}_{ij}; \mu(a_0)\right)$.
- For every i, update $(b_{0i}, b_{1i}) \propto \mathcal{N}\left(\mathbf{0}, \text{diag}(\sigma_0^2, \sigma_1^2)\right) \cdot P\left(\{X\}_{i\cdot}; \mu(b_{0i}, b_{1i})\right)$.
- For $j = 0, 1$, update $\sigma_j^{-2} \sim \mathcal{G}\left(\alpha + P/2, \beta + \sum_i b_{ji}^2/2\right)$

4.4 Simulation Study

The first dataset can be considered as a control dataset of 8 replicates of the BY4741 strain in yeast, four cultures grown in ^{14}N and the other four in ^{15}N media. The MudPIT analysis was performed on the four pools of labeled proteins mixed at 1:1 ratio in the experiment, and the spectral count matrix was generated for ^{14}N and ^{15}N labeled proteins in the four mixtures, giving 8 samples total. This control data was used for generating simulated datasets by inserting fold changes to a pre-selected set of the proteins.

Using the control dataset, two groups of synthetic datasets have been generated from the control dataset. Total of 1307 proteins have been identified at least once in one of the eight samples. Since the cultures were grown in ^{14}N and ^{15}N media and then mixed into four pools at 1:1 ratio before the mass analysis, in effect this data has no real signals between the differentially labeled samples in all proteins. In order to create synthetic datasets with non-trivial differential expression, the rows of the data were shuffled to ensure that the distribution of high and low abundance proteins is uniform across the rows. Then the first 200 proteins in the matrix were selected, and 2 fold changes were inserted to the selected proteins, generating the

first synthetic dataset (FC2). The second synthetic dataset (FC4) were generated by inserting 4 fold changes to the selected proteins. By construction, the signals present in the first 200 proteins are stronger in FC4 than FC2.

One interesting issue ahead is to determine how the number of replicates affects the power in detecting differentially expressed proteins. Thus further variants of the two datasets (FC2, FC4) were derived by varying the size of replicates as follows. The first column was taken from each group of four identically labeled replicates and the two columns were saved as the dataset with no replicate (FC2-1SPL, FC4-1SPL). Likewise, The first two and three columns from each group of four replicates were saved into separate files as the datasets with 2 replicates (FC2-2SPL, FC4-2SPL) and 3 replicates (FC2-3SPL, FC4-3SPL) respectively. The original data with all four replicates of each group was considered as the dataset with 4 replicates, named (FC2-4SPL, FC4-4SPL) accordingly.

To assess the performance of the proposed method from a comparative viewpoint, the proposed method and the conventional signal-to-noise ratio statistics coupled with false discovery rate (FDR) control were used. Particularly, the variance adjustment of t-statistics by the power law global error model (PLGEM) was reported to have improved the detection of interesting proteins in [141], hence their method was used in place of the conventional t-statistic. Raw spectral count matrix was converted into NSAF values [133], and ran the PLGEM model to calculate moderated t-statistics, and obtained permutation-based p-values with the FDR control. Then proteins were selected using various cutoffs in order to examine the power over a wide range of FDRs. Using the outputs from both methods, the comparisons were made based on the power of detection at a fixed error rate in both methods. The number of proteins called as differentially expressed by the two methods were compared

across the span of FDR. It is important to note, however, that the signal-to-noise ratio statistics require the calculation of variance, thus the methods like PLGEM StN cannot be applied to datasets that have less than 3 replicates such as FC2-1SPL/2SPL/SUM and FC4-1SPL/2SPL/SUM datasets. Therefore the comparisons below are shown for FC2-3SPL/4SPL and FC4-3SPL/4SPL datasets only, although QSpec was applied to all datasets.

Figures 4.1 A and B illustrate the comparison. The two figures correspond to the synthetic datasets FC2 and FC4 respectively. In both figures, it is easily seen that the two methods pick up more proteins with more replicates at a fixed FDR point. Also, the performance improves as one moves from FC2 to FC4 for every curve included in the figures. Comparing the two methods for the data with the same number of replicates and the fold change, QSpecs model based protein selection clearly outperforms that of the signal-to-noise ratio statistic with PLGEM variance adjustment (PLGEM StN hearafter) across the board. For example, in the FC2-4SPL dataset, QSpec selects 50 proteins (25%) at FDR 10% while PLGEM StN selects 24 proteins (12.5%). In the four replicate FC4-4SPL dataset, QSpec collects 193 proteins (96.5%) at the same FDR level, while the other method selects 167 (83.5%). Furthermore, it is worth noting that QSpecs protein selection from the single replicate FC2-1SPL and FC4-1SPL datasets performs no worse than PLGEM StNs selection from the three replicate FC2-3SPL dataset, and QSpec in the two replicate FC4-2SPL data is equivalent to PLGEM in the three replicate FC4-3SPL data.

Meanwhile, it was found that, in the aggregate sum FC2-SUM and FC4-SUM datasets, QSpecs model performed equally well with the same model applied to the four replicate FC2-4SPL and FC4-4SPL datasets. As discussed earlier, this can be

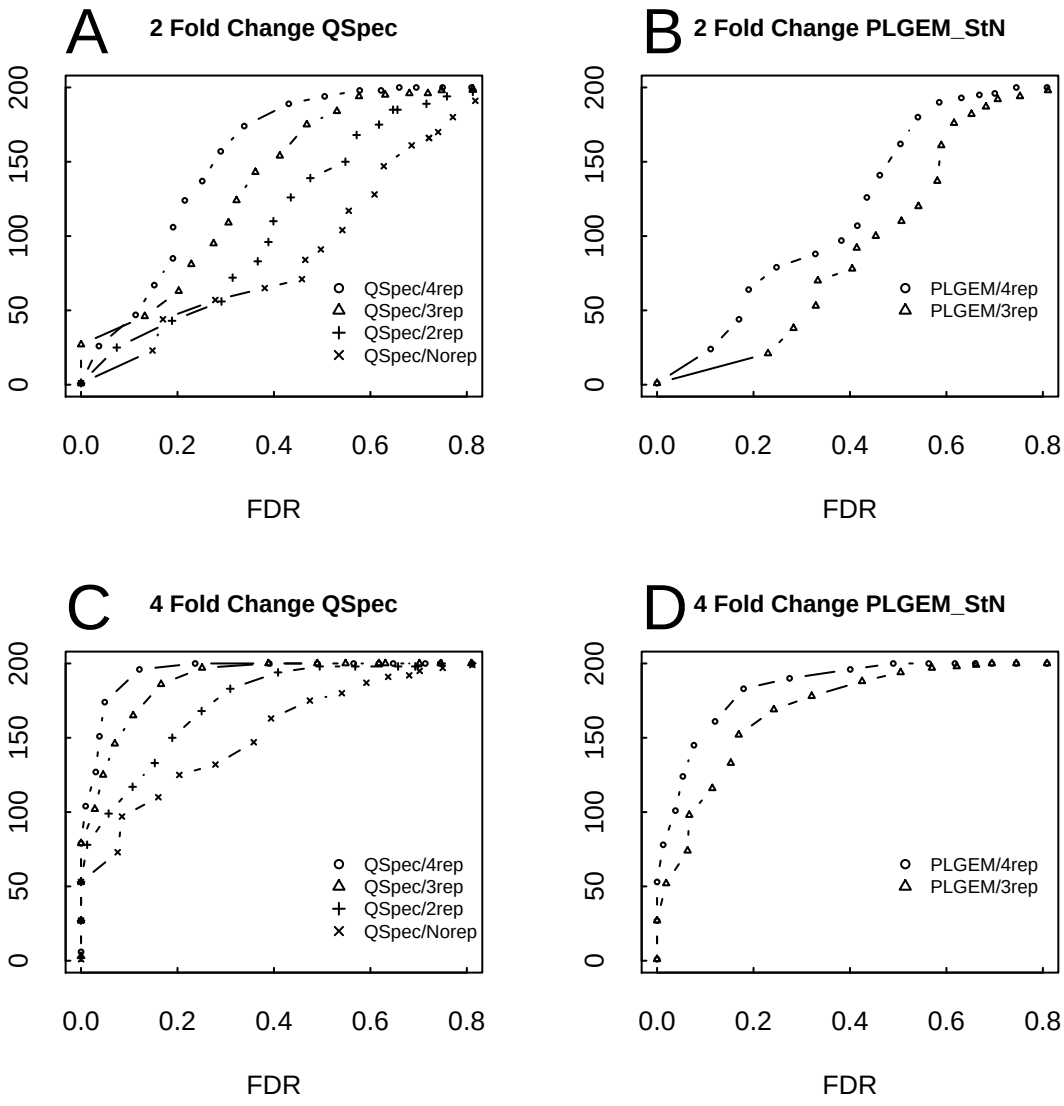


Figure 4.1: QSpec and PLGEM StN methods on simulation datasets. The number of true positive proteins identified by QSpec and PLGEM StN at fixed FDRs in synthetic datasts with known fold changes (two fold data in 2A and four fold data in 2B). The datasets with the same number of replicates, but analyzed by different methods, were marked with the same point marks. The lines from QSpec were colored in black, those from PLGEM StN were colored in red. The line from a QSpec run on the count sum dataset was colored in blue.

explained by the fact that all the information used to fit the Poisson model was summarized in the count sum (sufficient statistic). More precisely, this is so because the Poisson model assumes that the expected count is equal to the variability of the counts (variance) due to its parametrization, so the model does not have a separate variance parameter. This feature of Poisson models becomes problematic when there exists a considerable degree of heterogeneity within replicates of the same condition that cannot be corrected by simple normalization procedures, and the remedy to this problem will be discussed later.

4.5 Comparative Growth Analysis

A dataset generated from the mass analysis of four replicates of the same strain grown up to the logarithmic and stationary phases in ^{14}N medium was re-analyzed for a comparative growth phase analysis. Using this dataset, it will be shown that the protein selection by the propose method may lead to more relevant biological interpretation of the data than the conventional data analysis methods.

In [141], the authors applied PLGEM StN to the selected subset of 511 proteins that were consistently identified across most replicates grown in the logarithmic (LP) and stationary (SP) phases, and annotated the selected hundred proteins with the highest signal-to-noise ratio using Gene Ontology in order to interpret the protein signature in the context of the cool-down of biosynthetic processes and translation activities as the cell growth moves into the stationary phase.

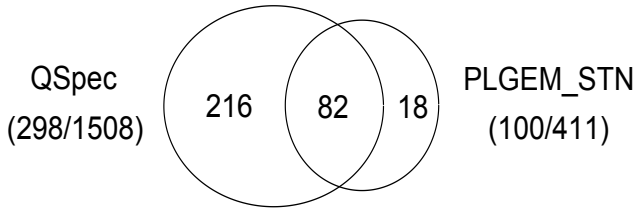
QSpec selected 298 proteins with Bayes Factor above 9.8, and considering all proteins satisfying this criterion as significantly differentially expressed would introduce on average 5% FDR or less according to the mixture model-based error estimation. Out of 298 proteins, 121 were over-expressed in the stationary phase and 177 were

over-expressed in the log phase, whereas 34 of the 100 proteins from PLGEM were over-expressed in the stationary phase and the remaining 66 were over-expressed in the log phase. The breakdown of expressed proteins in the two phases shows that a larger number of differentially expressed proteins are expressed in the log phase, but does not necessarily indicate that there are more proteins active in the log phase. Note that 82 of the top 100 proteins from PLGEM were also in the list of 298 proteins from QSpec, which implies that the top list from PLGEM was almost completely recovered by QSpec. One should be alerted caution, however, in comparing the sensitivity of the two methods based on this result since the top 100 proteins from PLGEM was not a selection based on the control of FDR to the same degree as in the selection of 298 proteins from QSpec. The top lists between QSpec and PLGEM were overlapped using the entire 1508 proteins.

The GO annotations and their significance measures were given by FATIGO+, and the most significant terms located in a reasonably high hierarchy of the GO are shown in Figures 4.2. Table A (QSpec) and B (PLGEM) in Figure 4.2 lists the biological functions (FDR corrected p-value less than 0.05) enriched in the two lists above. It was found that the list obtained from QSpec, reported in Table A in Figure 4.2, highlights almost all these functions reported in Table B with enhanced statistical significance measures. Biological processes such as translation and cellular biosynthetic process are the common top significant terms in the both QSpec and PLGEM lists of proteins overexpressed in the log phase, with the multiple testing corrected p-values much lower in QSpec annotation table (higher significance), giving a high confidence explanation for the slow-down of biosynthesis machinery in the stationary phase of cell growth. Meanwhile, a large number of terms selected only in QSpec annotation were found to be enriched in the list of proteins overexpressed

A **Functional Annotation for Proteins Reported To Be Up&Down Regulated in Log Phase by QSpec**

(Over)ExpressedIn	Category	GO Term	FDR-adjusted p-value
LogPhase	Biological Process	translation	5.55E-15
		macromolecule biosynthetic process	8.72E-12
		cellular biosynthetic process	4.36E-08
		biosynthetic process	3.06E-07
	Cellular Component	ribosome	6.75E-15
		ribonucleoprotein complex	5.02E-13
		intracellular non-membrane-bound organelle	6.70E-07
		small ribosomal subunit	3.84E-06
	Molecular Function	structural constituent of ribosome	3.96E-13
StationaryPhase	Biological Process	amino acid and derivative metabolic process	2.00E-10
		organic acid metabolic process	2.00E-10
		acetyl-CoA catabolic process	6.41E-07
		aerobic respiration	7.03E-07
		cofactor catabolic process	1.27E-06
		tricarboxylic acid cycle	1.27E-06
		glutamine family amino acid metabolic process	1.92E-06
		cellular respiration	1.60E-05
		energy derivation by oxidation of organic compounds	6.98E-05
		generation of precursor metabolites and energy	7.39E-05
	Cellular Component	mitochondrial part	1.43E-09
		cytoplasm	6.18E-07
	Molecular Function	oxidoreductase activity	1.43E-09



B **Functional Annotation for Proteins Reported To Be Up&Down Regulated in Log Phase by PLGEM**

(Over)ExpressedIn	Category	GO Term	FDR-adjusted p-value
LogPhase	Biological Process	macromolecule biosynthetic process	3.74E-09
		translation	3.74E-09
		cellular biosynthetic process	6.72E-07
		biosynthetic process	1.98E-06
	Cellular Component	cytosolic part	1.04E-08
		ribonucleoprotein complex	1.23E-08
		ribosome	5.80E-06
		small ribosomal subunit	2.38E-04
		cytosol	5.14E-04
	Molecular Function	structural constituent of ribosome	3.74E-09

Figure 4.2: Protein signature comparison between QSpec and PLGEM StN. Venn diagram is shown for the selected proteins from QSpec with all 1508 proteins and PLGEM StN with the subset of 511 proteins. Tables A and B correspond to the significantly enriched Gene Ontology terms in the protein list identified by QSpec and PLGEM StN respectively.

in the stationary phase, and they were mostly child terms of (especially glycolysisrelated) catabolism and cellular respiration, and oxidoreductase activity. This finding extends the biological interpretation beyond what was given in the above: as cell growth process cools down in the stationary phase, the focus of molecular activities shifts to breaking down large molecules into smaller units and releasing energy, potentially creating energy required for chemical reactions in anabolism, or more generally the maintenance of the cell.

4.6 Discussion

At present, many studies that utilize spectral counting for relative quantification still rely on simple data analysis methods such as filtering based on fold change ratios. Such an approach selects proteins based solely on the effect size without incorporating the variability, and therefore it may introduce a number of false positive calls in low abundance proteins where a small difference may results in artificially large fold change ratios. Moreover, the limited number, or total absence of replicates makes it difficult to find a robust method to assign significance to these statistics and reasonably control global false discovery rates. For example, in the popular method of referencing observed statistics to the permutation distribution, the number of possible permutations is 70 at most when there are 4 replicates in each comparison group, which gives a low-resolution permutation distribution, vulnerable to outlying observations.

The method presented in this work has several advantages. It can be applied to a variety of situations including the comparative experiments that feature either a small number of replicates or none at all within each biological condition. In contrast to other methods, the Poisson model of QSpec faces no issues with the absence

of replicates. Since the parameters are modeled across all proteins in the dataset as random numbers generated from the same population distribution, it effectively pools statistical information needed for robust estimation and provides a simple way to filter proteins based on Bayes Factors.

In any case, hierarchical Bayes estimation will effectively pool the statistical information across the proteins from different fractions for more robust parameter estimation and attempt to overcome the paucity of information due to the small sample size. Another advantage of the method is the flexibility for possible extensions to more complicated data structures. This class of GLMMs with hierarchical Bayes estimation can be applied to even more general data analysis scenarios, such as a longitudinal profiling study without the comparative design (no differential expression), a replicate analysis where the reproducibility of quantification is studied by comparing the within and between replicate variability, and a protein-protein interaction study with a large number of pull-down experiments where the strength of interaction between pairs of proteins is validated based on the number of spectra corresponding to the interaction partners.

Yet there remain a number of areas for improvements in this modeling strategy. One well-known problem with Poisson models is the potential violation of the assumption of the equal mean-variance relationship, the so-called over-dispersion problem. In datasets with many replicates, for instance, the observed data can include very heterogeneous counts across replicates even within the same biological condition. In that case, the Poisson model with conventional assumptions may not work as efficiently. Furthermore, aggregating counts over replicates in a dataset, when analyzed using the presented model, will produce largely identical results, as in the case of applying it to the same dataset but with replicates represented in it

as separate experiments. In effect, this observation shows the drawback of the plain Poisson model from a different angle in that the model does not make full use of the variability observed in the data efficiently. The remedy to this would be to extend the model specification. Another possibility is to use alternative distributions such as negative binomial models replacing the Poisson model. The latter model has a natural connection to Bayesian modeling through mixture model specification.

Another area of potential improvement is the inclusion of known protein properties in the model in addition to sequence length normalization. By specifying more flexible distributions, such as mixture distribution in place of single Gaussian distribution in the hyper-prior reflecting the baseline abundance of proteins, more adaptive models, which efficiently account for the differences in spectral count patterns attributable to this abundance property, can be fitted. This characterization, in turn, may provide more concrete description of the heterogeneity in spectral count distribution in the population of all proteins, which may itself be an interesting summary.

Finally, the discussion in this work was limited to spectral counts defined as the number of MS/MS spectra identified for each protein. However, related metrics such as the number of unique peptides or the percentage of protein sequence covered by the identified peptides are likely to contain additional useful information. Future work should involve detailed analysis of these different protein abundance parameters and their relative performance in different applications. To this end, the future efforts should focus on designing multivariate statistical approaches that can effectively combine different abundance metrics leading to improved statistical ability to detect differential proteins.

The statistical methodology presented in this work is a proteome-wide model-

based assessment of differential expression using GLMM, equipped with a hierarchical Bayes estimation procedure that borrows statistical strengths across all proteins. Unlike the conventional methods using ad-hoc data transform, signal-to-noise ratio, and post-hoc data-driven adjustments, the proposed method is more powerful in finding differentially expressed proteins, and robust to the variation due to the limited number of biological replicates at the individual protein level. The model showed superior performance in terms of sensitivity of detection over existing alternatives. The real data analysis examples have also illustrated the important advantages of handling the challenges due to the limited number of replicates, and of providing flexibility of extension of the same model to more complicated study designs. It is expected that the computational framework presented in this work will be useful in a wide range of applications in shotgun proteomics.

CHAPTER V

Significance Analysis of Protein-Protein Interaction

5.1 Quantitative Proteomics for Protein-Protein Interactions

Another important area of proteomics research is the investigation of protein-protein interactions (PPI). With two-hybrid system/tandem affinity purification and mass spectrometry (AP/MS)-based protein identification, physical interactions can be profiled in a near proteome coverage. For example, recent works based on high-throughput AP/MS experiments have generated more than ten thousand interactions between proteins in budding yeast [14, 15, 16]. These data have been accumulated in large-scale functional databases including BIND [144], DIP [145], MIPS [146], and bioGRID [147]. The availability of these datasets has enabled biologists to delineate signaling pathways in crucial cellular processes [148, 149, 150, 151, 152], understand the topology sustained by physiologically important hub proteins and their evolutionary characteristic [153], and identify biological features such as scaffolds, cellular localization, expression, and substrate specificity [154, 155].

Constructing PPI network (interactome) from multiple purification experiments is a challenging task for several reasons. First, non-specific bindings frequently occur in purification experiments, including spurious co-purification of background contaminants. Second, affinity purification has limited coverage of detection depend-

ing on the tags, and thus some interactions may be less reproducible than others. Third, proteins in the low abundance range are usually lost in the mass spectrometry, thus transient interactions may be detected with reduced sensitivity. Last, post-translational modifications may also influence the identification of interactions and the structural composition of the interactome. In consequence, PPIs reported in one research lab are not necessarily reproduced elsewhere, which is well illustrated in the poor overlap between studies catalogued in the PPI databases.

In the literature, few studies have paid attention to validating individual PPIs using a computational method. Because PPI data are a collection of binary calls with no repeated observations, computational approaches usually learn the confidence of interactions from the relative position and the role of the proteins in the topology of network, to produce what can be collectively named as *affinity scores*. Socio-affinity index (SAI) by Gavin et al [15], for example, is the odds that proteins identify each other or co-purify when another protein is expressed. Purification Enrichment (PE) score by Collins et al [17] is a refinement of SAI, which accounts for the case of repeat purifications and the reciprocity of acquisition in the likelihood calculation. The graph-theory and likelihood-based approach by Scholtens and Gentleman [156, 157] is another example that exploits the topology information statistically. It is important to note, however, that the common weakness in these methods is the dependence on the topology without a direct reference to the strength of physical interaction.

A recent surge of quantitative proteomics is gradually changing the landscape. Quantitative MS analysis has been shown to be powerful in capturing real interactions using either stable isotope labeling or label-free quantification approaches. Quantitative measurements provide a direct access to the strength of interactions in a given experiment, and thus filtering interactions is now expected to show an

increased specificity. It is worth noting that few computational methods have successfully capitalized on this benefit. Currently, Sardiu et al [158] is the only published work that has proposed a naïve Bayes posterior probability based on the normalized spectral abundance index, where probabilities are proportional to the spectral count in individual baits. However, it is possible to make use of the quantitative information more efficiently in assessing confidence, since the strength of one interaction is related to that of other interactions involving the same bait or the same prey proteins. A statistical model with a sparse set of model parameters can be devised to share this information between relevant interactions. In this context, it is the goal of this chapter to propose a model-based approach for the Significance Analysis of INTeractome, or SAINT. SAINT performs a frequency-based filtering of background contaminants and translates the spectral count data into probability scores for individual interactions.

The rest of the section is organized as follows. Section 5.2 introduces the two-layered Poisson mixture model of SAINT and Section 5.3 elaborates on the estimation steps based on Markov chain Monte Carlo. Section 5.4 illustrates the methodology in a dataset generated for the complete set of 131 yeast kinases, proteins involved in the phosphorylation of other proteins as a part of cell signaling. Section 5.5 provides the summary and the discussion for further methodological development.

5.2 Statistical Model

5.2.1 Significance Analysis of Interactome

Suppose that large-scale AP/MS experiments were conducted, and spectral count profiles from multiple bait purifications were generated. A proteome-wide statistical model is now proposed to differentiate real interactors (called preys) from background contaminants. The underlying principle is that, if preys are identified with

a sufficiently large spectral counts in purification of a bait, then the interaction between the bait and the prey is likely to be a real interaction. This goal can be achieved in a probabilistic manner using the Poisson mixture model. Here the Poisson distribution was chosen for the convenience of modeling the count data, and it is not a requirement. It is assumed that the proteome-wide count distribution can be described as a mixture of three Poisson distributions. First, a prey can be a background contaminant that was co-purified not as a result of an interaction with the bait, but a non-specific binding to the affinity tag or other experimental noise. Second, a prey can also be a non-contaminant, but the prey does not interact with the bait at all or the strength of interaction between the two proteins is weak. Lastly, a non-contaminant prey could have a real interaction with the bait, which should be indicated by a high spectral count relative to others.

To describe the model, a few basic notations will be helpful. For the data, the numbers of baits and preys are denoted by n_b and n_p respectively. Also let $N = \{N_{ij}\}_{i=1,j=1}^{i=n_p,j=n_b}$ denote the spectral count matrix, where N_{ij} is the spectral count of prey i identified in the purification experiment for bait j. It is helpful to define latent variables for background contaminants and real interactions at this point. Let $\{Y_i\}_{i=1}^{n_p}$ be the indicator for background contaminants, which is defined as follows: $Y_i = 0$ if prey i is a background contaminant, or $Y_i = 1$ otherwise. For preys with $Y = 1$, consider a sub-indicator for real interactions $\{Z_{ij}\}$ for $i = 1, \ldots, n_p$ and $j = 1, \ldots, n_b$. This variable is define as: $Z_{ij} = 1$ if the interacting pair (i, j) is a real interaction, and $Z_{ij} = 0$ otherwise.

5.2.2 Poisson Mixture Model

For the sake of convenience, suppose now that the contaminants were already identified and that the main task is to differentiate real interactions from the experi-

mental noise. For a non-contaminant prey i, it is assumed that N_{ij}, the spectral count of the interaction between prey i and bait j N_{ij}, follows the Poisson distribution with mean λ_{ij}^s:

$$N_{ij} \mid \lambda_{ij}^s, Z_{ij} = 1, Y_i = 1 \quad \sim \quad \mathcal{P}(\lambda_{ij}^s) \tag{5.1}$$

if i and j are partners of a real interaction, where the mean count can be written as

$$\begin{aligned} \log \lambda_{ij}^s &= \log l_i + \log a_i + \log c_j + \beta_0 + \alpha_{ij} \\ &= \log l_i + \log a_i + \log c_j + \beta_0 + \underbrace{\alpha_j^b + \alpha_i^p}_{\text{multiplicative model}} . \end{aligned}$$

In equation (5.1), the coefficient β_0 denotes the baseline abundance of prey proteins interacting with baits, l_i is the sequence length of prey i, a_i is the PeptideAtlas counts [159] as a surrogate baseline abundance prior to bait enrichment, and c_j is the bait coverage measured by the spectral count of bait itself respectively. More importantly, the parameter α_{ij} indicates the strength of interaction between prey i and bait j. In order to effectively pool the information across preys and baits and to prevent over-parametrization, an multiplicative model $\exp(\alpha_{ij}) = \exp(\alpha_j^b + \alpha_i^p)$ is assumed, where the strength of interaction is a sum of parameters of the bait(α_j^b) and the prey(α_i^p). Through this parametrization, the parameters are standardized and shared across preys and baits, giving a proteome-wide model that borrows statistical strength between relevant interactions. For the identifiability of the model, it was assumed that the interaction potential parameters add up to zero.

As a counterpart to the mixture component representing real interactions, a mixture component for the experimental noise (non-real interactions) is set to be a zero-inflated Poisson distribution with a small mean count $\lambda_j^{s_0}$. In notation,

$$N_{ij} \mid Z_{ij} = 0, Y_i = 1 \quad \sim \quad r_{j0}\delta_0(\cdot) + (1 - r_{j0})\mathcal{P}(\lambda_j^{s_0}) \tag{5.2}$$

The proportion of r_{j0} differs by experiment and bait, depending on the purification quality and the abundance of the protein itself.

Meanwhile, for the case of background contaminants, the prey-specific affinity was assumed to remain constant across all baits, i.e. $\alpha_{ij} = \mu_i$ for a given purification j. This results in a mixture component for the cases with $Y_i = 0$,

$$N_{ij} \mid \lambda_{ij}^c, Y_i = 0 \quad \sim \quad \mathcal{P}(\lambda_{ij}^c) \tag{5.3}$$

where the mean count can be written as

$$\log \lambda_{ij}^c \quad = \quad \log l_i + \log a_i + \log c_j + \gamma_0 + \mu_i.$$

with γ_0 being the baseline abundance of all contaminant preys. The interaction parameter μ_i is also subject to the identifiability constraint of zero sum.

Finally, the proportion of data explained by each mixture component is defined as

$$\begin{aligned} p_s^* = (1-p_c)p_s \quad &\equiv \quad E\left\{1(Z_{ij}=1, Y_i=1)\right\} \\ p_{s_0}^* = (1-p_c)p_{s_0} = (1-p_c)(1-p_s) \quad &\equiv \quad E\left\{1(Z_{ij}=0, Y_i=1)\right\} \\ p_c \quad &\equiv \quad E\left\{1(Y_i=0)\right\} \end{aligned}$$

In sum, the mixture components (5.1,5.2) and (5.3) are two competing models for every prey i in the data, and the first two components are again two competing models for every interaction N_{ij} involving the prey i. Now the likelihood of this model can be written as

$$\pi(N) = \prod_{i=1}^{i=n_p} \left\{ p_c \prod_{j=1}^{j=n_b} \underbrace{\mathcal{P}(N_{ij}|\lambda_{ij}^c)}_{\text{contaminants}} + (1-p_c) \prod_{j=1}^{j=n_b} \underbrace{\left(p_s \mathcal{P}(N_{ij}|\lambda_{ij}^s) + (1-p_s)\mathcal{ZP}(N_{ij}|r_{j0}, \lambda_{ij}^{s_0})\right)}_{\text{non-contaminants}} \right\} \tag{5.4}$$

where $\mathcal{ZP}$ denotes the zero-inflated Poisson distribution. For the convenience of the reader, the notations introduced above are summarized in Table 5.1.

N_{ij}	Spectrum count of prey i for bait j
$\alpha_{ij} = \alpha_i^p + \alpha_j^b$	Post-enrichment interaction effect of real interactors
α_i^p	Interaction potential of prey i
α_j^b	Interaction potential of bait j
μ_i	Post-enrichment interaction effect of contaminant prey i
a_i	Pre-enrichment baseline abundance of prey i before enrichment
l_i	Sequence length of prey i
c_j	Bait coverage for bait j
Y_i	$= 1$ if prey i is non-contaminant
	$= 0$ otherwise
Z_{ij}	$= 1$ if interaction is present given $Y_i = 1$
	$= 0$ if interaction is not present given $Y_i = 1$
p_s	Proportion of Specific Interactions with Non-contaminant Preys
p_c	Proportion of Common Contaminants among Preys

Table 5.1: Key parameters in the mixture model of SAINT.

5.2.3 Background Contaminants and Real Interactions

The operation of SAINT is explained in further details here. First, the probability of a prey is a background contaminant is

$$\mathbb{P}(Y_i = 0) = \frac{p_c \prod_{j=1}^{n_b} \left(\mathcal{P}(N_{ij}|\lambda_{ij}^c)\right)}{(1-p_c)\prod_{j=1}^{n_b}\left(p_s\mathcal{P}(N_{ij}|\lambda_{ij}^s) + (1-p_s)\mathcal{Z}\mathcal{P}(N_{ij}|r_{j0}, \lambda_j^{s_0})\right) + p_c\prod_{j=1}^{n_b}\left(\mathcal{P}(N_{ij}|\lambda_{ij}^c)\right)}.$$

This probability is estimated by taking the average of samples of the latent variable Y_i drawn from the posterior distribution

$$\widehat{\mathbb{P}}(Y_i = 0) \approx T^{-1}\sum_{t=1}^{T}\left(1 - Y_i^{(t)}\right)$$

where T is the number of iteration of the sampler and the index t runs through the iterations. In this implementation, preys with $\mathbb{P}(Y_i = 0) \geq 0.1$ were flagged as background contaminants.

It is noted that such a filter can be applied either in an entirely model-based way, or can be modified as a hybrid process that combines model output and data-dependent threshold. Sometimes this model-based filtering criterion is not able to distinguish every contaminant case perfectly, since the distinction of contaminant and non-contaminant could be ambiguous in the count distribution of some preys

in the observed data. This usually happens when the spectral count distribution follows neither a mixture model with completely distinguishable components (non-contaminant) nor a unimodal Poisson mixture model (contaminant). The most realistic strategy here is to calculate the frequency of real interactions for a given prey, i.e. $\mathbb{P}(Z_{ij} = 1)$ for $j = 1, 2, \ldots, n_b$ (explained below), and then filter the prey as a contaminant if and only if

$$n_b^{-1} \sum_{j=1}^{n_b} \mathbb{P}(Z_{ij} = 1) \geq p^*$$

where $p^* = 0.1$ was set. This threshold is data-dependent and will therefore have to be adjusted elsewhere. In addition to the model-based filter, this extra filtering removes the ambiguous cases where on average more than 10% of the observed interactions are thought to be real. An empirical threshold filter of this kind has been previously used [16], and effectively removes contaminant preys not captured by the model-based filter.

For non-contaminant preys, the probability of real interactions with non-contaminant preys can be computed as

$$\mathbb{P}(Z_{ij} = 1) = \frac{p_s \mathcal{P}(N_{ij}|\lambda_{ij}^s)}{p_s \mathcal{P}(N_{ij}|\lambda_{ij}^s) + (1 - p_s)\mathcal{Z}\mathcal{P}(N_{ij}|r_{j0}, \lambda_j^{s_0})}$$

If the purification was repeated, for a total of R repeat purifications $\{j_1, j_2, \ldots, j_R\}$ of a bait protein, a single probability is calculated per bait

$$\mathbb{P}(Z_{ij} = 1) = \left(\frac{p_s \prod_{r=1}^{R} \mathcal{P}(N_{ij_r}|\lambda_{ij_r}^s)}{p_s \prod_{r=1}^{R} \mathcal{P}(N_{ij_r}|\lambda_{ij_r}^s) + (1 - p_s) \prod_{r=1}^{R} \mathcal{Z}\mathcal{P}(N_{ij_r}|r_{j0}, \lambda_j^{s_0})} \right)$$

The estimation is straightforward as in the comtaminant probability, using the posterior sample average

$$\widehat{\mathbb{P}}(Z_{ij} = 1) \approx T^{-1} \sum_{t=1}^{T} Z_{ij}^{(t)}$$

Then all interactions with the probability above 0.9 were selected.

5.3 Inference

In this section, the appropriate posterior distributions for model parameters are derived and sampling steps of Markov chain Monte Carlo are suggested.

5.3.1 Prior Distributions

Recall the likelihood (5.4). If the indicators Z and Y were known *a priori*, the complete likelihood would have been

$$\begin{aligned}\pi(N|\cdot, Z, Y) \quad \propto \quad & \prod_{i=1,j=1}^{i=n_p,j=n_b} \left(\mathcal{P}(N_{ij}|\lambda_{ij}^{s})\right)^{1\{Z_{ij}=1,Y_i=1\}} \\ & \times \left(\mathcal{Z}\mathcal{P}(N_{ij}|r_{j0}, \lambda_j^{s_0})\right)^{1\{Z_{ij}=0,Y_i=1\}} \\ & \times \left(\mathcal{P}(N_{ij}|\lambda_{ij}^{c})\right)^{1\{Y_i=0\}}\end{aligned}$$

With the complication of repeat purifications and a large number of baits and preys, a direct maximization of the likelihood with respect to the large number of parameters becomes intractable, which motivates the use of Markov chain Monte Carlo. The following prior distributions are assumed:

$$\begin{aligned}\beta_0 \sim \mathcal{N}(0, \sigma_\beta^2), \quad \alpha_i^p \sim \mathcal{N}(0, \sigma_\alpha^2), \quad \alpha_j^b \sim \mathcal{N}(0, \sigma_\alpha^2). \\ \gamma_0 \sim \mathcal{N}(0, \sigma_\gamma^2), \quad \mu_i \sim \mathcal{N}(0, \sigma_\mu^2) \\ r_{j0} \sim \mathcal{U}(0, 1), \quad \lambda_j^{s_0} \sim \mathcal{G}(\epsilon, \kappa) \\ (p_s^*, p_{s_0}^*, p_c) \sim \mathcal{D}(a_s, a_{s_0}, a_c)\end{aligned}$$

where the prior distribution is specified as non-informative as possible. Based on the likelihood and the prior, the posterior distribution can be obtained up to a scaling factor, which naturally leads to the construction of Metropolis-Hasting sampler.

5.3.2 Gibbs Sampling with Embedded Metropolis-Hastings

Parameters were iteratively drawn in the order of: $[\beta,\alpha^p,\alpha^b] \rightarrow [\gamma,\mu^p] \rightarrow [\lambda_j^{s_0}] \rightarrow [Y,Z] \rightarrow [p_s^*,\ p_c]$. For the parameters whose posterior distribution is not easy to sample from, a standard Metropolis-Hastings algorithm is used. The posterior sample are thinned out by taking every 20th sample in the chain, which is expected to reduce the auto-correlation in the sequentially drawn samples. The entire chain was run for 50,000 iterations in total, with additional burn-in steps prior to the main iterations.

Beginning with the parameter updates in the mixture component for the background contaminants,

$$\begin{aligned} p(\gamma_0|\cdot) &\propto \prod_{\{(i,j):Y_i=0,Z_{ij}=1\}} \left[e^{-\lambda_{ij}^c}(\lambda_{ij}^c)^{N_{ij}}\right] \times (\sigma_\gamma)^{-1/2} e^{-\frac{\gamma_0^2}{2\sigma_\gamma^2}} \\ p(\mu_i|\cdot) &\propto \prod_{\{j:Z_{ij}=1\}} \left[e^{-\lambda_{ij}^c}(\lambda_{ij}^c)^{N_{ij}}\right] \times (\sigma_\mu^2)^{-1/2} e^{-\frac{(\mu_i)^2}{2\sigma_\mu^2}} 1\{Y_i=0\} \end{aligned}$$

The changes in (γ,μ) alter λ_{ij}^c, which is important to note in calculating Metropolis-Hastings ratio.

For the parameter updates in the mixture component for the real interactions,

$$\begin{aligned} p(\beta_0|\cdot) &\propto \prod_{\{(i,j):Y_i=1,Z_{ij}=1\}} \left[e^{-\lambda_{ij}^s}(\lambda_{ij}^s)^{N_{ij}}\right] \times (\sigma_\beta^2)^{-1/2} e^{-\frac{\beta_0^2}{2\sigma_\beta^2}} \\ p(\alpha_i^p|\cdot) &\propto \prod_{\{j:Z_{ij}=1\}} \left[e^{-\lambda_{ij}^s}(\lambda_{ij}^s)^{N_{ij}}\right] \times (\sigma_\alpha^2)^{-1/2} e^{-\frac{(\alpha_i^p)^2}{2\sigma_\alpha^2}} 1\{Y_i=1\} \\ p(\alpha_j^b|\cdot) &\propto \prod_{\{i:Y_i=1,Z_{ij}=1\}} \left[e^{-\lambda_{ij}^s}(\lambda_{ij}^s)^{N_{ij}}\right] \times (\sigma_\alpha^2)^{-1/2} e^{-\frac{(\alpha_j^b)^2}{2\sigma_\alpha^2}}. \end{aligned}$$

Notice that the changes in $(\beta,\alpha^p,\alpha_b)$ alter λ_{ij}^s.

Lastly, for the parameter updates in the mixture component for the rest of the

interaction pairs, $\lambda_j^{s_0}$ are drawn from Gamma distribution

$$p(\lambda_j^{s_0}|\cdot) \propto \mathcal{G}\left(\epsilon + \sum_{Y_i=1,Z_{ij}=0}^{\text{row } i} N_{ij}O_{ij} \quad \kappa + \sum_{Y_i=1,Z_{ij}=0}^{\text{row } i} O_{ij}\right).$$

where $O_{ij} = 1$ if N_{ij} is an observation from non-zero Poisson distribution and 0 if it is from the point mass at zero. O_{ij} is also sampled from Bernoulli distribution with probability $e^{-\lambda_j^{s_0}}/(1+e^{-\lambda_j^{s_0}})$ if $N_{ij} = 0$, or set to 1 if $N_{ij} > 0$. These updates complete the sampling step for all mixture components except for the latent variables.

Given these updates, the membership of the preys and the bait-prey pairs to the three categories of interactions are now updated. First, the latent variable Y_i, indicator of prey i being a background contaminant, is updated using the Bayes rule

$$Y_i|\cdot \quad \sim \quad \mathcal{B}\left\{\frac{(1-p_c)\prod_{j=1}^{n_b}\left(p_s\mathcal{P}(N_{ij}|\lambda_{ij}^s) + (1-p_s)\mathcal{Z}\mathcal{P}(N_{ij}|r_{j0},\lambda_j^{s_0})\right)}{(1-p_c)\prod_{j=1}^{n_b}\left(p_s\mathcal{P}(N_{ij}|\lambda_{ij}^s) + (1-p_s)\mathcal{Z}\mathcal{P}(N_{ij}|r_{j0},\lambda_j^{s_0})\right) + p_c\prod_{j=1}^{n_b}\left(\mathcal{P}(N_{ij}|\lambda_{ij}^c)\right)}\right\}.$$

where the letter $\mathcal{B}$ stands for Bernoulli distribution. If $Y_i = 0$, there is no reason to draw Z_{ij} since a contaminant cannot be involved in real interactions. If $Y_i = 1$, then Z_{ij} is drawn from

$$Z_{ij}|Y_i = 1, \cdot \sim \mathcal{B}\left\{\frac{p_s\mathcal{P}(N_{ij}|\lambda_{ij}^s)}{p_s\mathcal{P}(N_{ij}|\lambda_{ij}^s) + (1-p_s)\mathcal{Z}\mathcal{P}(N_{ij}|r_{j0},\lambda_j^{s_0})}\right\}$$

Finally, the mixture proportions are sampled from the Dirichlet distribution using Gibbs sampling

$$(p_s^*, p_{s_0}^*, p_c) \quad \sim \quad \mathcal{D}(\delta_s, \delta_{s_0}, \delta_c)$$

where $\delta_s = a_s + \sum_{i,j} 1(Z_{ij} = 1, Y_i = 1)$, $\delta_{s_0} = a_{s_0} + \sum_{i,j} 1(Z_{ij} = 0, Y_i = 1)$, and $\delta_c = a_c + n_b \sum_i 1(Y_i = 0)$. This completes a full iteration of parameter updates in the MCMC.

5.4 Analysis of Kinase Network Data

Epitope tagged kinase alleles (HA, FLAG) were transiently expressed from the GAL1 promoter in small-scale cultures. Kinase complexes were recovered from cell extracts on pre-coupled Protein A magnetic beads and rapidly washed before on-bead trypsin digestion. Samples were analyzed by nano-scale liquid chromatography on a C18 gradient column coupled to an LTQ mass spectrometer. Proteins were identified using the Mascot search engine, and spectral counting was performed all identified proteins.

5.4.1 Selected Interactions

Using two different affinity tags (HA and FLAG), 131 known yeast kinases were expressed and purified by affinity chromatography, and all of their interaction partners were identified by mass spectrometry analysis. More than 26,500 interactions have been reported from 267 purifications using the two tags (113 HA, 154 FLAG). Among these, 89 baits were expressed with both tags, while about 39 baits were expressed more than once using the same tag. SAINT handles repeated measurements by calculating a probability for multiple purification of each bait. The model was applied to the HA and FLAG data separately in order to account for the differences in the tags, and the union of real interactions from both datasets was considered as the final set of interactions.

As mentioned before, a prey was considered to be a contaminant if the probability of being contaminant was greater than or equal to 0.1. SAINT has identified 238 and 210 contaminants from HA and FLAG tags respectively. After the contaminant filter, SAINT has selected 1509 unique high confidence interactions (probability 0.9 and above) out of more than 6,800 total interactions (654/2855 HA, 987/4002

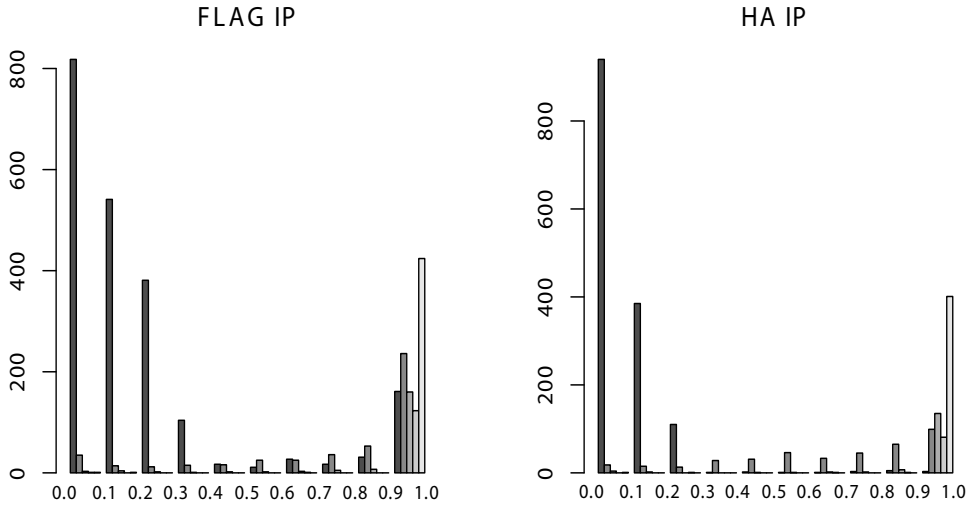


Figure 5.1: Distribution of SAINT probabilities versus observed spectral counts. Distribution of the estimated probabilities is shown against actual spectral counts in three tag data. In each bin, five bars were drawn, in the order of 1, 2 to 3, 4 to 5, 6 to 10, and 10 counts and above. Typically, four or more counts lead to high probability, i.e. above 0.9, while two or three counts may result in a variety of significance decisions. However, large counts such as 5 or more does not always give significance calls, since more convincing counts may be necessary for highly abundant proteins.

FLAG). Figure 5.1 shows the relationship between the actual spectral counts and the estimated probability. It is easy to see that having a single spectral count does not give a high confidence on the interaction in most cases, although probabilities over 0.9 were occasionally assigned to such interactions. Having two or three counts leads to most of the borderline decisions between significant and non-real interaction, and for every interaction pair in this case, the call will be affected by the count distribution of all other interactions involving the same bait or the same prey. For instance, if two or three count was the maximum count in repeat purifications of a given bait, then the interaction is likely to score high probability. On the other hand, if there was another prey with a large spectral count, then the prey with small count is likely to be a noise, therefore assigned a low probability. These individual cases illustrate the subtlety and the circumstantial nature of these borderline decisions across all

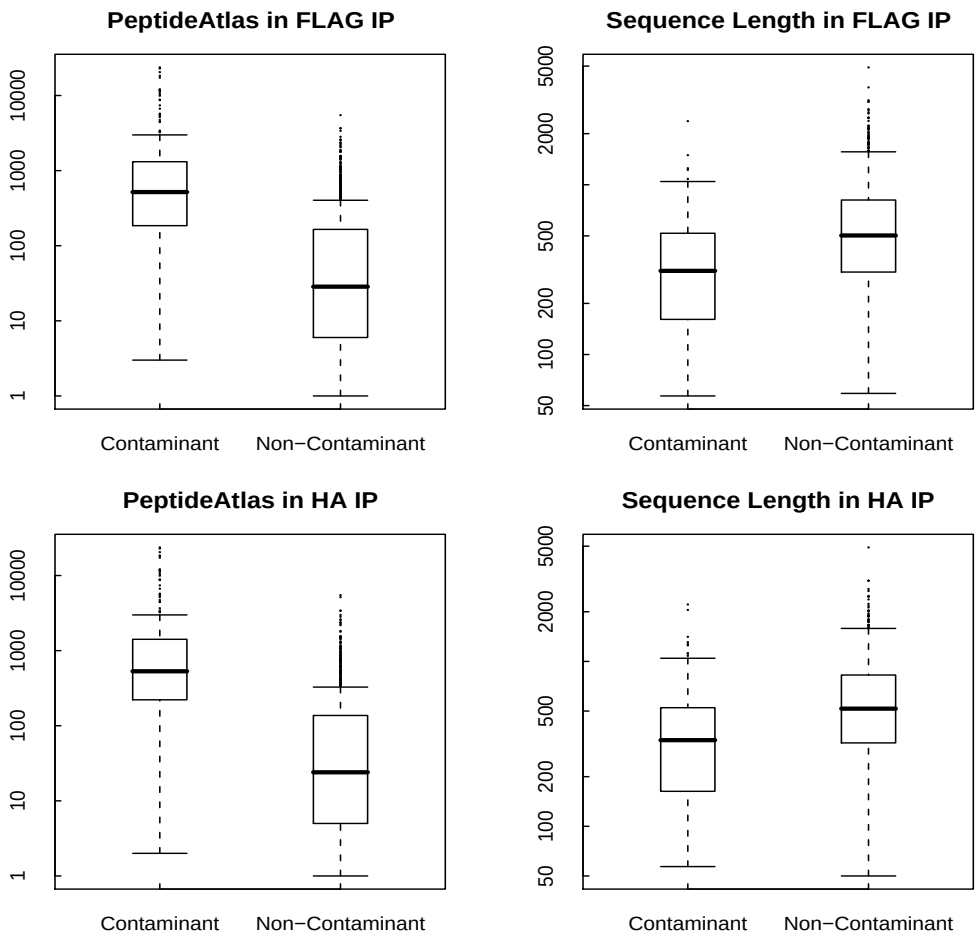


Figure 5.2: Characterization of background contaminants. Distribution of PeptideAtlas counts and sequence length are shown for contaminant and non-contaminant preys. Across the three tags, contaminant proteins had higher PeptideAtlas counts and their sequence length was shorter than non-contaminants.

three data and these decisions are well handled by SAINT, whereas a simple count threshold will not be able to distinguish them efficiently.

5.4.2 Effect of Normalization in SAINT

In the previous section, it was pointed out that SAINT normalizes raw spectral counts by multiple factors. Figure 5.2 illustrates the difference between contaminants and non-contaminants in terms of those factors. First, contaminant proteins identified by SAINT tend to have large PeptideAtlas counts. This is consistent with the common notion that the background contaminants are usually most abundant

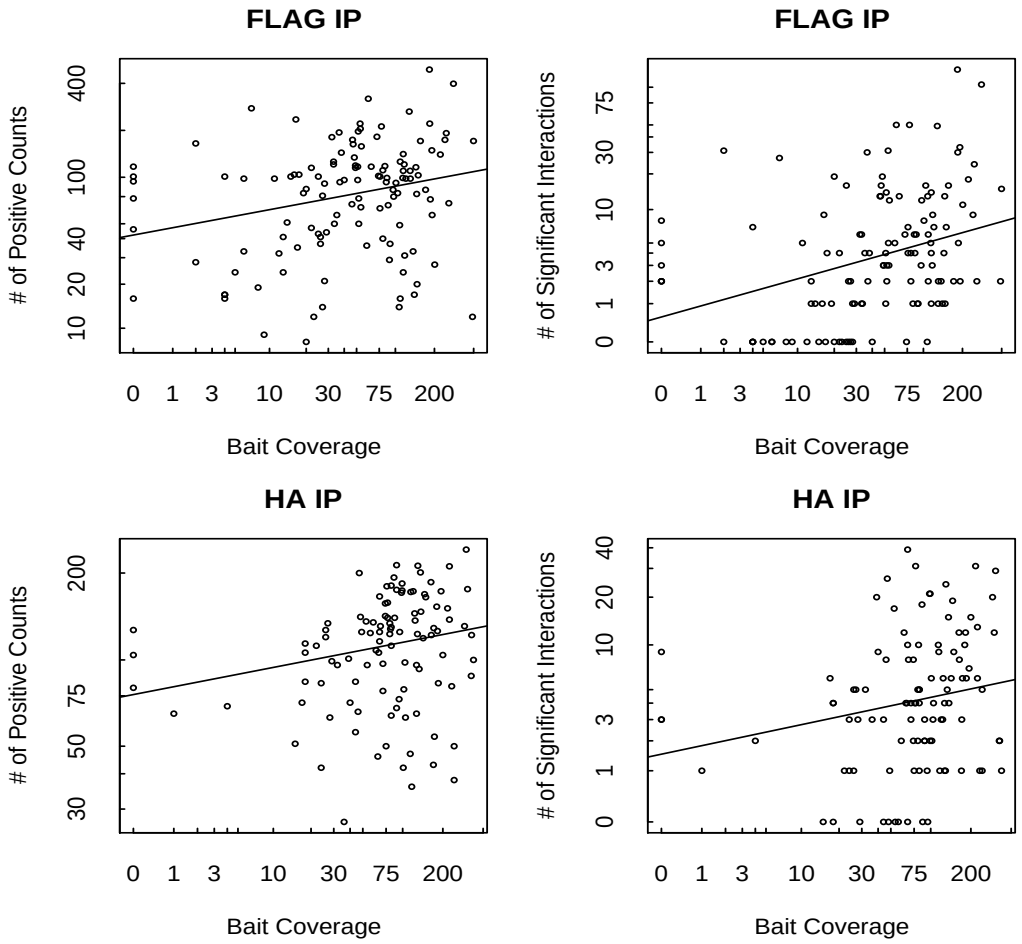


Figure 5.3: Correlation between purification quality and the number of real interactions. Shown is the bait coverage versus the number of identified interactions before filtering and the number of real interactions after filtering. A mild degree of correlation exists between both quantities and the bait coverage.

proteins in the cell. In the model, dividing spectral counts by larger PeptideAtlas counts lowers the abundance and thus helps the model consider proteins with low normalized abundance in a large number of baits as prototypes of contaminants. The figure also shows that contaminants tend to have short sequence length, and the manner in which this is reflected in the model is about the same as in the case of PeptideAtlas counts.

Meanwhile, SAINT also divides spectral counts by the bait coverage, i.e. the count of bait itself. Since bait coverage is a bait-wise definition, the effect of this normal-

ization was examined across baits. However, this normalization does not seriously influence the sensitivity of detection. Figure 5.3 shows that bait coverage has a mild correlation with the number of real interactions, indicating that high quality purifications generate more real interactions than low quality ones do. The figure shows the relationship between bait coverage and the number of all identified interactions (before probability filtering) in the left panels, and the relationship between the bait coverage and the number of real interactions (after probability filtering) in the right panels. Although the magnitude of correlation with the bait coverage improves, the variability accounted for by the bait coverage is ignorable ($R^2 \leq 0.05$ in all cases).

5.4.3 Network Construction

By combining the list of real interactions acquired from the two tags, a network was composed and visualized in Cytoscape [160]. Figure 5.4 shows the entire network involving 130 kinase baits. It may be misleading to conclude that the two tags generated a non-overlapping set of interactions from this figure, since not all baits were purified using both tags. However, when the set of baits profiled using both tags were examined, at most 20% of the real interactions were found in both tags, indicating that these two tags are complementary to one another in terms of the coverage of the network. Meanwhile, it is also worth noting that nearly half of these kinases are linked with one another either directly or indirectly, indicating that there exists a systematic network of signaling activities via phosphorylation, the biology of which is beyond the scope of this chapter. See Figure 5.5 for the visualization of the inter-linked property of kinases.

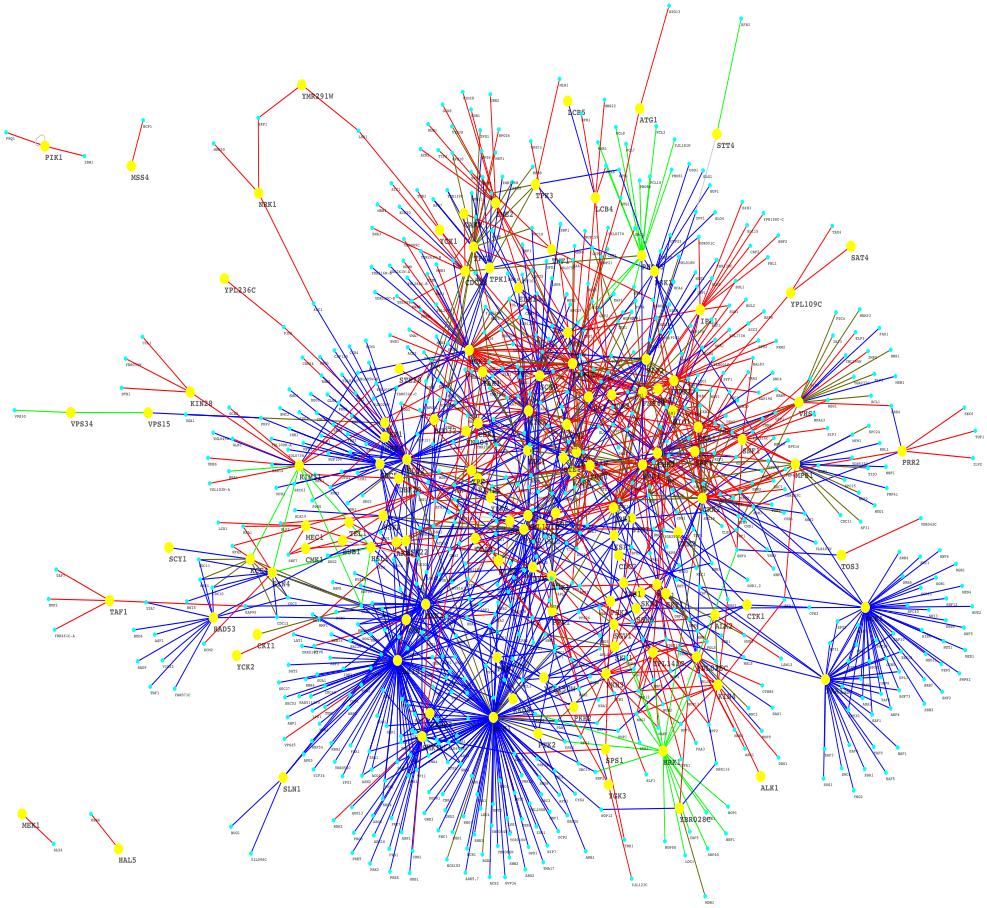


Figure 5.4: Network view of the SAINT-filtered yeast kinome. Entire yeast kinase network constructed by SAINT was visualized in Cytoscape. Large yellow nodes are 130 baits and small cyan nodes are 700 preys. Edges are color-coded in order to differentiate the contribution of each tag to the construction of the entire network (red - HA, blue - FLAG, Dard Brown-HA and FLAG). It must be reminded that the seemingly low overlap between the two tags is the result of the bait selection in each tag that led many baits to be profiled using one tag only, not necessarily a reflection of tag real interactions.

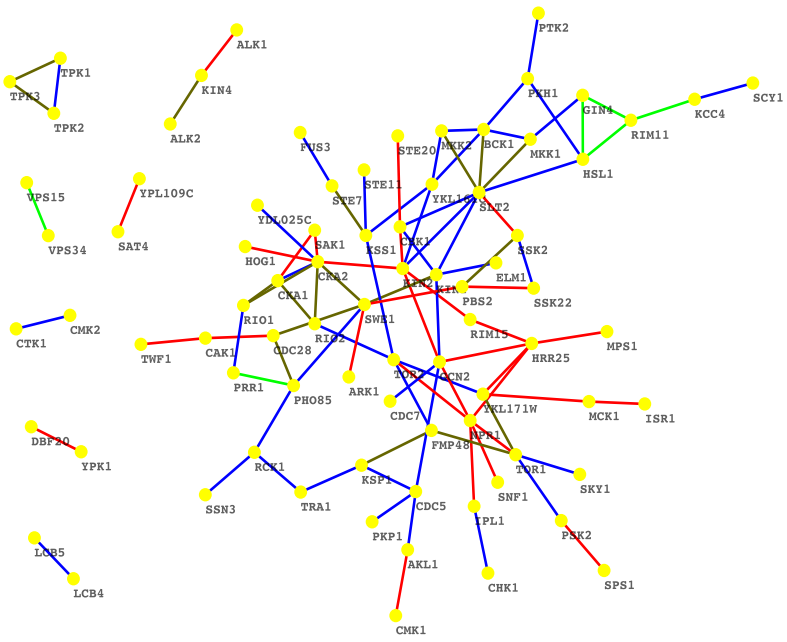


Figure 5.5: Core yeast kinome. Shown above is the yeast kinase network consisting of only kinases that have real interactions with at least one other kinase. Color coding remains the same.

5.4.4 Experimental Validation

In order to evaluate the quality of selected interactions, we have used benchmarked them against all in-vitro interactions thoroughly validated from the experimental sources such as co-crystallization and far-western analysis collected in Biogrid database. These low-throughput experiments are more accurate sources of individual interactions than high-throughput surveys since the protocols purify both interaction partners to target local molecular complexes with specific biological functions.

Figure 5.6 shows that, considering the in-vitro interactions as the true positives, interactions with high SAINT probability contain more true positives than randomly selected interactions of an equal size. Green and blue curves are the number of true positives in twenty different bins defined by the probability range. The plot shows that more true positives are enriched in high probability bins, i.e. probability

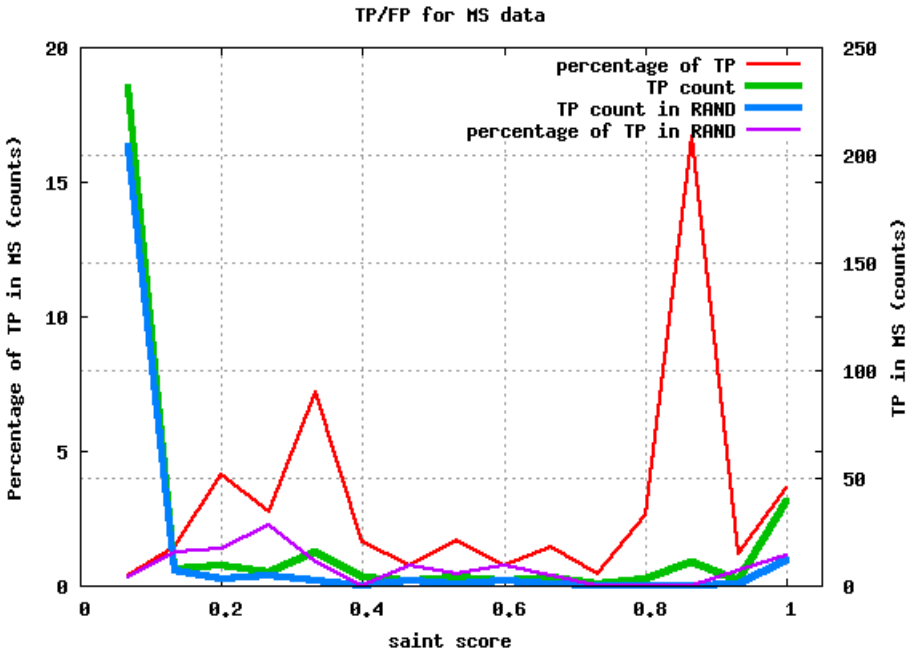


Figure 5.6: Experimental validation of SAINT interactions. High scoring interactions in SAINT was overlapped with the interactions derived from in-vitro interactions in the BioGrid database.

around 0.8 and above. Since each bin contains a different number of interactions with more interactions in the low probability range, the enrichment of true positives in the high probability interactions and the random set is not well presented in absolute counts. In order to see the enrichment of true positives more clearly, the proportion of true positives in each set was calculated, which is shown in red (high probability set) and purple (random set) lines. The percentage is significantly higher in high probability interactions than low probability interactions, although there remain some good percentage of true positives in the range of $0.3 \sim 0.4$. This shows that the quantitative measure corroborates with the confidence of interactions in affinity purification experiments at least indirectly, which also highlights the need for rigorous statistical filtering method such as SAINT.

It is also important to note that there are hundreds of other in-vitro interactions missed as low probability interactions in SAINT (left side of the figure). This can

be a result of two possible reasons. For one, the kinome data was generated using a high-throughput affinity purification method that target a proteome-wide coverage of interactions, and thus low abundance interactions (low spectral counts) are more likely missed here than in data generated from local quantification method such as far-western blotting. On the other hand, proteome coverage of the affinity purification experiments is often very limited, which was also confirmed in a data mining study of Krogan and Gavin data mentioned earlier [17]. This implies that the high-throughput experiment in the kinome study may have discrepancies in the low abundance interactions despite the good overlap in high abundance interactions. Moreover, since the kinome data was generated from a pool of a small number of yeast strains, this issue of limited coverage may be further aggravated.

5.4.5 Comparison with Affinity Scores

To assess the performance of SAINT from a comparative standpoint, SAINT was compared to another existing scoring method based on affinity measures. The Purification Enrichment (PE) score by Collins et al [17] can be considered as the most widely accepted method at present, and thus PE scores were calculated for this data. For the comparison, kinase-kinase interactions that are entries in BioGrid biochemical interaction were used as the positive, or high confidence validation set. Only kinase-kinase interactions were used because PE score requires that the data was generated accounting for reciprocal interactions, while this does not affect SAINT. Meanwhile, all interactions with ribosomal proteins were used as a negative set, or a low confidence validation set. In this dataset, there were 27 positive and 3147 negative interactions in HA data and 29 positive and 2605 negative interactions in FLAG data.

Figure 5.7 shows the comparison. In both tags, it is clearly illustrated that the

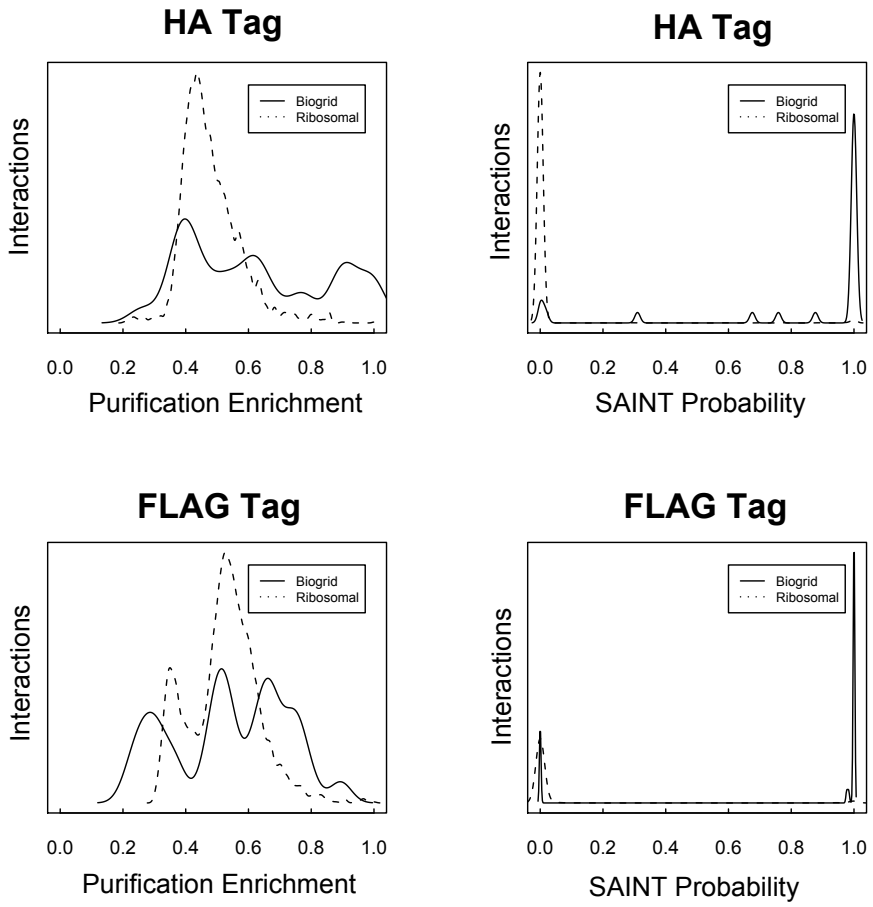


Figure 5.7: Purification Enrichment score versus SAINT probability. Kinase-kinase interactions included in Biogrid were used as a positive set, while all interactions with ribosomal proteins were used as a negative set.

probabilities from SAINT gives a good separation of the positive and the negative interactions, while the PE score fails to distinguish the two groups. This result demonstrates that the application of the PE score is conditional on the wide coverage of baits so that most preys have also been used as baits, while it also shows the advantage of direct measure of protein abundance in assigning confidence measures to PPIs. With no conditional assumption on the topology of the interactome, SAINT capitalizes on this predictive power of the quantitative information and performs a reliable confidence assessment.

5.5 Discussion

In this chapter, a novel model-based method addressing confidence assignment in AP/MS experiments was demonstrated. With the ability to pool the statistical information across the preys and the baits, SAINT normalizes spectral counts by normalization factors that have implications in the differentiation of the background contaminants and the non-contaminants, and directly translates them into the confidence measures of PPI. The reproducibility of spectral counts for interactions can also be validated over biological replicates and across different tags for a selected set of baits using the model-based method QSPEC.

Particularly, SAINT has shown a superior performance to the affinity score-based approach in distinguishing high confidence interactions from low confidence ones in the dataset analyzed in this study. This is because the coverage of baits in the kinase network data is relatively localized to a specific class of proteins compared to other large scale datasets, where reciprocal interactions provide key information to an elevated confidence on the true positive interactions. With an increasing interest in the study of local interactome in the literature, the utility of SAINT has far-reaching

practical implication.

It is important to distinguish the goal of SAINT from that of other common algorithms performing protein complex formulation. While most of the protein-complex weaving algorithms operate on a given set of interaction pairs, SAINT performs a more rudimentary operation of validating individual interactions that are constituents of protein complexes, which therefore precedes the stage of contructing complexes. The value of this significance analysis should be appreciated especially when the motivation for a large-scale experiment is the study of *connectivity* between proteins, rather than modularity of them. The goal of the data analyzed in this study was to investigate the inter-links of protein kinases that consist of a large signaling cascade of phosphorylation. Another potential example of this kind is a set of purification of proteins that form the skeleton of a metabolic pathway, in which case denoising of biologically spurious protein interactions should be helpful for identifying the core sequence of reactions in the pathway.

There remain a few important issues in applying SAINT. First, SAINT performs a frequency-based filtering of the background contaminants and thus takes away the need to run negative control experiments. In this process, special care must be paid to the relationship between the frequency-based filter and the overall topology of the network. It was earlier explained that SAINT partly relies on a user-specified threshold for filtering the background contaminants in addition to the model-based classification. Specifying a particular threshold directly influences the statistical model in SAINT because the mean counts of contaminants and non-contaminants are dynamically altered as the assignments of preys into these two groups are changed. If a threshold is too stringent, some of the important *bona fide* interactions, including hub proteins, can be assigned to the group of background contaminants. Thus the

frequency-based filter has an important implication for the large hub proteins, and these need to be cautiously examined in the list of contaminants reported by the model. On the contrary, if a threshold is too liberal, then the overall topology will be dominated by these contaminants and the key structure of the network will be overshadowed by the noise.

Second, the choice of tags may lead to differences in the composition of the network. If the pool of all preys identified in at least one of the purifications is considered, the source of this difference comprises a very small fraction of the entire set of interactions. However, if one focuses on the positive identifications and all baits that were purified using both HA and FLAG tags (498 in HA and 648 in FLAG), only 19% of the real interactions are common (179 interactions), indicating there could be a considerable amount of influence in the composition of the network by the choice of the tag. This also implies that different tags may play a complementary role, unless a precise recipe is written for choosing an optimal tag for individual baits in large-scale experiments.

Third, the spectral counts have so far been regarded as the only given features of the data. However, biased spectral counting may pose a danger in modeling the data, especially when homologous proteins that share common amino acid sequence are co-purified. In the kinase network data, these candidates were manually scanned and they were forcefully classified as contaminants so that they will not be assigned high probabilities. More specifically, the spectral count distribution of a few members of SSAx and SSBx proteins resembled that of key proteins such as CDC14, so that the model would have made a large number of significance calls that would have masked the core topology of the network, had they not been manually tagged to be unsure targets. A potential cure for this problem would be modifying the manner in

which spectral counts are computed, e.g. utilizing additionally available information such as the number of unique peptides represented by the spectra and the sequence coverage, to come up with an extended measure of protein abundance.

Finally, it must be pointed out that the interpretation of the final composition of the network is always conditional on the choice of baits, or at least successfully purified baits. Although it is often argued that the reciprocity provides a stronger evidence for the presence of interaction, the assertion does not uniformly apply to all proteins, due to the unique enzyme-substrate relationships between baits and preys, e.g. in the signaling pathways. Since signaling cascades require a complex balance of active kinases, phosphotases, and their inhibitors and competing enzymes, enriching one of the interaction partners may not necessarily harvest the identification of proteins in the counterpart. Therefore, it is important to validate individual PPIs based on the quantitative measure such as spectral counting that does not use *a priori* information derived from the topology of the network, and the interactive nature of statistical modeling in SAINT is expected to illustrate the full advantage in exploiting all the information sources shared by a large number of interactions.

CHAPTER VI

Conclusion

This thesis presented Bayesian hierarchical models for the analysis of high-throughput experiments in contemporary molecular biology. In the presence of modeling complexity and limited sample size, hierarchical Bayes can be a powerful model framework. The collection clearly demonstrates that, with the current level of computing power and the efficient sampling algorithms, Bayesian inferential methods originally developed for datasets of a much smaller size remain useful for the new datasets. In addition to hierarchical Bayes, the proposed methods incorporate a variety of existing inferential topics, including change point problem (Chapter 2), model selection perspective (Chapter 4), and mixture models (Chapters 2-3 and 5) for distinguishing biological signals from experimental noise.

In Chapter 2, a hierarchical model called Double-Layered Mixture Model (DLMM) was developed for analyzing a combined set of DNA copy number and gene expression data to obtain a genome-wide mapping for copy number-associated expression changes. The change-point estimation procedure based on the reversible jump Markov chain Monte Carlo algorithm provides a solution to the segmentation problem. The posterior probability calculated from DLMM allows one to select the target genes based on a unified scoring scheme across the genome and the samples, which

would have been difficult otherwise due to lack of control of global error rates or unfitting assumptions made in other existing methods. With the mixture hierarchical prior specification, local copy number changes are probabilistically identified based on the distributional information attained from the entire chromosome, where the confidence of observing a real breakpoint is determined by the overlap of the mixture components in the hierarchical prior.

In Chapter 3, a hierarchical hidden Markov model (HHMM) was presented as a tool that combines the statistical evidence of transcription factor binding from multiple platforms of ChIP experiments. By modeling separate hidden Markov models as emission from a master hidden Markov model, the binding site identification using the posterior probability estimated from HHMM improves the receiver operating characteristic over other competing methods. Moreover, the application of HHMM extends beyond the analysis of ChIP-seq and ChIP-chip. The method has a natural extension to analysis of combining data from k sources, with the simple adjustment of a few model parameters in the master-level hidden Markov model. This implies that data from any new technology mapping can be incorporated in the existing analysis by combining posterior probabilities estimated from each data source. With the increasing number of new technologies to profile transcription factor binding sites, HHMM will become a very powerful analysis framework for integrative genomic data analysis.

In Chapter 4, a hierarchical Bayes method was introduced for the analysis of mass spectrometry-based quantitative proteomics data analysis. QSpec addresses the limited sample size issue in the typical experimental data with hierarchical Bayes, where protein specific model parameters are linked through the proteome-level prior distribution, whose posterior distribution gives the characterization of the abundance and

the differential expression patterns in the proteome covered by mass spectrometry experiments. QSpec utilizes Bayes Factors in order to sort proteins in the order of significance of differential expression, incorporating a model selection perspective.

In Chapter 5, Significance Analysis of Interactome (SAINT), a mixture model-based method, was devised for assigning confidence scores to protein-protein interactions. SAINT calculates the probability of real interaction based on the quantitative measures from large-scale affinity purification-mass spectrometry experiments, with a probabilistic filtering of background contaminants. Through the connections across interactions involving the same baits or preys and hierarchical prior, all relevant model parameters share statistical information. The analysis of yeast kinome shows that SAINT has the ability to distinguish both strong and weak interactions with biological relevance from experimental noise.

In all chapters, the hierarchical prior serves as the channel for sharing distributional information across the genome or the proteome, and fast and efficient implementation of Markov chain Monte Carlo allowed parameter estimation despite the presence of a substantial number of parameters. The model parameters were estimated as a combination of the prior information and the data, where informative priors were given only when it is necessary to establish the identifiability of parameters or the stability of sampling algorithms. In congruence with the theme of the thesis, i.e. the classification of relevant signal and experimental noise, mixture models using latent variables have been extensively used throughout the work. In sum, these examples show that the combination of hierarchical Bayes and mixture models, along with other complementary inferential techniques, has promising grounds to remain as a model framework for incorporating future innovations across many basic science disciplines.

Bibliography

[1] P.O. Brown and D. Botstein. Exploring the new world of the genome with dna microarrays. *Nat. Genet.*, 20:33–37, 1999.

[2] A. Schulze and J. Downward. Navigating gene expression using microarrays - a technology review. *Nat. Cell Biol.*, 3:E190–E195, 2001.

[3] R. Aebersold and M. Mann. Mass-spectrometry-based proteomics. *Nature*, 422:198–207, 2003.

[4] H. Steen and M. Mann. The abc's (and xyz's) of peptide sequencing. *Nat. Rev. Mol. Cell. Biol.*, 5:699–711, 2004.

[5] R.G. Sadygov, D. Cociorva, and J.R. Yates. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat. Methods*, 1:195–202, 2004.

[6] A.I. Nesvizhskii, O. Vitek, and R. Aebersold. Mass-spectrometry-based proteomics. *Nat. Methods*, 4:787–797, 2007.

[7] D.C. Ward and D.C. White. The new 'omics era. *Curr. Opin. Biotech.*, 13:11–13, 2002.

[8] A.R. Joyce and B. Palsson. The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell. Biol.*, 7:198–210, 2006.

[9] C.D. Brown, D.S. Johnson, and A. Sidow. A functional architecture and evolution of transcriptional elements that drive coexpression. *Science*, 317:1557–1560, 2007.

[10] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, 2:166–176, 2003.

[11] A.L. Barabasi and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5:101–113, 2004.

[12] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, 3:78, 2007.

[13] I. Yanai, L.R. Baugh, J.J. Smith, C. Roehrig, S.S. Shen-Orr, J.M. Claggett, A.A. Hill, D.K. Slonim, and C.P. Hunter. Pairing of competitive and topologically distinct regulatory modules enhances patterned gene expression. *Mol. Syst. Biol.*, 4:163, 2008.

[14] Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A.R. Willems, H. Sassi, P.A. Nielsen, K.J. Rasmussen, J.R. Andersen, L.E. Johansen, L.H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B.D. Sorensen, J. Matthiesen, R.C. Hendrickson, F. Gleeson, T. Pawson, M.F. Moran, D. Durocher, M. Mann, C.W.V. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–183, 2002.

[15] A. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L.J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drews, G. Neubauer, J.M. Rick, B. Kuster, P. Bork, R.B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, 2006.

[16] N.J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, T. Punna, J.M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M.D. Robinson, A. Paccanaro, J.E. Bray, A. Sheung, B. Beattie, D.P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M.M. Canete, J. Vlasblom, S. Wu, C. Orsi, S.R. Collins, S. Chandran, R. Haw, J.J. Rilstone, K. Gandi, N.J. Thompson, G. Musso, P. St Onge, S. Ghanny, M.H.Y. Lam, G. Butland, A.M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J.S. Weissman, C.J. Ingles, T.R. Hughes, J. Parkinson, M. Gerstein, S.J. Wodak, A. Emili, and J.F. Greenblatt. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440:637–643, 2006.

[17] S.R. Collins, P. Kemmeren, X. Zhao, J.F. Greenblatt, F. Spencer, F.C.P. Holstege, J.S. Weissman, and N.J. Krogan. Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Mol. Cell. Proteomics*, 6:439–450, 2007.

[18] A-C. Gingras, M. Gstaiger, B. Raught, and R. Aebersold. Analysis of protein complexes using mass spectrometry. *Nat. Rev.*, 8:645–654, 2007.

[19] B. Efron. Robbins, empirical bayes and microarrays. *Ann. Statist.*, 31(2):366–378, 2003.

[20] B.P. Carlin and T.A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, Boca Raton, FL, 2000.

[21] J.F.C. Kingman. Uses of exchangeability. *Ann. Prob.*, 6:183–197, 1978.

[22] D. Aldous. *Lecture Notes in Mathematics 1117*, chapter Exchangeability and related topics, pages 1–198. Springer-Verlag, 1985.

[23] P. Diaconis and D. Freedman. Finite exchangeable sequences. *Ann. Prob.*, 8:745–764, 1980.

[24] B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academica Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale*, 4:251–299, 1931.

[25] E. Hewitt and L.J. Savage. Symmetric measures on cartesian products. *Trans. Am. Math. Soc.*, 80:470–501, 1955.

[26] J. Besag, J. York, and A. Mollie. Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, 43:1–59, 1991.

[27] P. Diggle, P.J. Heagerty, K. Y. Liang, and S.L. Zeger. *Analysis of Longitudinal Data*. Oxford Univ. Press, 2002.

[28] P. Baldi and A. Long. A bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, 2001.

[29] J. Ibrahim, M.H. Chen, and R. Gray. Bayesian models for gene expression with dna microarray data. *J. Am. Statist. Assoc.*, 97:88–99, 2002.

[30] M. Tadesse, J. Ibrahim, and G. Mutter. Identification of differentially expressed genes in high-density oligonucleotide arrays accounting for the quantification limits of the technology. *Biometrics*, 59:542–554, 2003.

[31] G.K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3:3, 2004.

[32] H. Ishwaran and J.S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Ann. Statist.*, 33:730–773, 2005.

[33] Y.C. Tai and T.P. Speed. A multivariate empirical bayes statistic for replicate microarray time course data. *Ann. Statist.*, 34:2387–2412, 2006.

[34] K.Y. Yeung, C. Fraley, and A. Marua. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977–987, 2001.

[35] J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West, editors. *Bayesian Statistics*, chapter Bayesian Clustering with Variable and Transformation Selections, pages 249–275. Oxford University Press, 2003.

[36] M. Medvedovic, K.Y. Yeung, and R.E. Brmgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20:1222–1232, 2004.

[37] A.E. Raftery and N. Dean. Variable selection for model-based clustering. *J. Am. Statist. Assoc.*, 101:168–178, 2006.

[38] B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96:1151–1160, 2001.

[39] C. Genovese and L. Wasserman. Bayesian and frequentist multiple testing. *Technical Report, Carnegie Mellon University*, 2002.

[40] M. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5:155–176, 2004.

[41] P. Sebastiani, M. Abad, and M.F. Ramoni. Bayesian networks for genomic analysis. *Genomic Sig. Proc. Statist.*, pages 281–320, 2004.

[42] P. Salzman and A. Almudevar. Using complexity for the estimation of bayesian networks. *Stat. Appl. Gen. Mol. Biol.*, 5:21, 2006.

[43] J.S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *J. Am. Statist. Assoc.*, 89:958–966, 1994.

[44] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[45] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. Pat. Anal. Mach. Intel.*, 6:721–741, 1984.

[46] A.F.M. Smith and G.O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *J. R. Statist. Soc. B*, 55:3–23, 1993.

[47] W.R. Gilks, D.G. Clayton, D.J. Spiegelhalter, N.G. Best, and A.J. McNeil. Modelling complexity: Applications of gibbs sampling in medicine. *J. R. Statist. Soc. B*, 55:39–52, 1993.

[48] F. Liang and W.H. Wong. Evolutionary monte carlo: Applications to cp model sampling and change point problem. *Stat. Sinica*, 10:317–342, 2000.

[49] D.J. Sargent, J.S. Hodges, and B.P. Carlin. Structured markov chain monte carlo. *J. Comput. Graph. Statist.*, 9:217–234, 2000.

[50] S.C. Kou, Q. Zhou, and W.H. Wong. Equi-energy sampler with applications in statistical inference and statistical mechanics. *Ann. Statist.*, 34:1581–1619, 2006.

[51] B.P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *J. R. Statist. Soc. B*, 57:473–484, 1995.

[52] P. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[53] S.J. Godsill. On the relationship between markov chain monte carlo methods for model uncertainty. *J. Comput. Graph. Stat.*, 10:230–248, 2001.

[54] G. Parmigiani, E.S. Garrett, R. Anbazhaghan, and E. Gabrielson. A statistical framework for expression-based molecular classification in cancer. *J. Roy. Statist. Soc. B*, 64:717–736, 2002.

[55] A.E. Raftery and V.E. Akman. Bayesian analysis of a poisson process with a change point. *Biometrika*, 73:85–89, 1986.

[56] B.P. Carlin, A.E. Gelfand, and A.F.M. Smith. Hierarchical bayesian analysis of changepoint problems. *Appl. Statist.*, 41:389–405, 1992.

[57] A.E. Gelfand and A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.*, 85:398–409, 1990.

[58] R.E. Kass and A.E. Raftery. Bayes factors. *J. Am. Statist. Assoc.*, 90:773–795, 1995.

[59] S. Chib. Marginal likelihood from the gibbs output. *J. Am. Statist. Assoc.*, 90:1313–1321, 1995.

[60] C. Han and B.P. Carlin. Markov chain monte carlo methods for computing bayes factors: A comparative review. *J. Am. Statist. Assoc.*, 96:1122–1132, 2001.

[61] B.E. Stranger, M.S. Forrest, M. Dunning, C.E. Ingle, C. Beazley, N. Thorne, R. Redon, C.P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S.W. Scherer, S. Tavare, P. Deloukas, M.E. Hurles, and E.T. Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, 2007.

[62] J.L. Freeman, G.H. Perry, L. Fuek, R. Redon, S.A. McCarroll, D.M. Altshuler, H. Aburatani, K.W. Jones, C. Tyler-Smith, M.E. Hurles, N.P. Carter, S.W. Scherer, and C. Lee. Copy number variation: new insights in genome diversity. *Genome Res.*, 16:949–961, 2006.

[63] R. Redon. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, 2006.

[64] L. Feuk, A.R. Carson, and S.W. Scherer. Structural variation in the human genome. *Nat. Reviews*, 7:85–97, 2006.

[65] D. Pinkel, R. Segraves, D. Sudar, S. Clark, C.I. Poole, D. Kowbel, C. Collins, W. Kuo, C. Chen, Y. Zhai, S. Dairkee, B. Ljung, J. Gray, and D. Albert-

son. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, 20:207–211, 1998.

[66] J.R. Pollack, T. Sorlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.L. Borrensen-Dale, and P.O. Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *PNAS*, 99:12963–12968, 2002.

[67] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringner, G. Sauter, O. Monni, A. Elkahloun, O.P. Kallioniemi, and A. Kallioniemi. Impact of dna amplification on gene expression patterns in breast cancer. *Cancer Res.*, 62:6240–6245, 2002.

[68] G. Tonon, K.K. Wong, G. Maulik, C. Brennan, B. Feng, Y. Zhang, D.B. Khatry, A. Protopopov, M.J. You, A.J. Aguirre, E.S. Martin, Z. Yang, H. Ji, L. Chin, and R.A. DePinho. High-resolution genomic profiles of human lung cancer. *PNAS*, 102(27):9625–9630, 2005.

[69] A. Olshen, E. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5:557–572, 2004.

[70] J. Fridlyand, A.M. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain. Hidden markov models approach to the analysis of array cgh data. *J. Mult. Anal.*, 90:132, 2004.

[71] J. Marioni, N. Thorne, and S. Tavare. Biohmm: a heterogeneous hidden markov model for segmenting array cgh data. *Bioinformatics*, 22:1144–1146, 2006.

[72] S. Stjernqvist, T. Ryden, M. Skold, and J. Staaf. Continuous-index hidden

markov modelling of array cgh copy number data. *Bioinformatics*, 23:1006–1014, 2007.

[73] O. Rueda and R. Diaz-Uriarter. Flexible and accurate detection of genomic copy-number changes from acgh. *PLoS Comp. Biol.*, page e122, 2007.

[74] P. Wang, Y. Kim, J. Pollack, B. Narasimhan, and R. Tibshirani. A method for calling gains and losses in array cgh data. *Biostatistics*, 6(1):45–58, 2005.

[75] N.R. Zhang and D.O. Siegmund. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63:22–32, 2007.

[76] F. Picard, S. Robin, E. Lebarbier, and J.J. Daudin. A segmentation/clustering model for the analysis of array cgh data. *Biometrics*, 63:758–766, 2007.

[77] H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21:4084–4091, 2005.

[78] W.R. Lai, M.D. Johnson, R. Kucherlaptai, and P.J. Park. Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics*, 21:3763–3770, 2005.

[79] R. Chari, W.W. Lockwood, and W. Lam. Computational methods for the analysis of array cgh. *Cancer Informatics*, 2:48–58, 2006.

[80] J.H. Kim, S.M. Dhanasekaran, R. Mehra, S.A. Tomlins, W. Gu, J. Yu, C. Kumar-Sinha, X. Cao, A. Dash, L. Wang, D. Ghosh, K. Shedden, J.E. Montie, M.A. Rubin, K.J. Pienta, R.B. Shah, and A.M. Chinnaiyan. Integrative

analysis of genomic aberrations associated with prostate cancer progression. *Cancer Res.*, 67:8229–8239, 2007.

[81] D. Lipson, A. Ben-Dor, E. Dehan, and Z. Yakhini. Joint analysis of dna copy numbers and gene expression levels. *Proceedings of Algorithms in Bioinfomatics*, 3240:135–146, 2004.

[82] W.N. van Wieringen and M.A. van de Viel. Nonparametric testing for dna copy number induced differential mrna gene expression. *Biometrics*, pages Epub May 12, 2008, 2008.

[83] J.G. Ibrahim, M.H. Chen, and D. Sinha. Criterion based methods for bayesian model assessment. *Statistica Sinica*, 11:419–443, 2001.

[84] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nature Reviews Genetics*, 4(3):177–183, 2004.

[85] M. Baudis and M.L. Cleary. Progentix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics*, 12:1128–1129, 2001.

[86] M.A. van de Wiel, K.I. Kim, S.J. Vosse, W.N. van Wieringen, S.M. Wilting, and B. Ylstra. Cghcall: Calling aberrations for array cgh tumor profiles. *Bioinformatics*, 23:892–894, 2007.

[87] G Jr. Dennis, B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, and R.A. Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biol.*, 4(5):P3, 2003.

[88] M.J. Solomon, P.L. Larsen, and A. varshavsky. Mapping protein-dna inter-

actions in vivo with formaldehyde: evidence that histone h4 is retained on a highly transcribed gene. *Cell*, 53(6):937–947, 1988.

[89] V. Orlando and R. Paro. Mapping polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell*, 75(6):1187–1198, 1993.

[90] B. Ren, F. Robert, J.W. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C. Wilson, S.P. Bell, and R.A. Young. Genome-wide location and function of dna-associated proteins. *Science*, 290:2306–2309, 2000.

[91] V.R. Iyer, C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, and P.O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409:533–538, 2001.

[92] D.R. Bentley et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59, 2008.

[93] D.S. Johnson, A. Mortazavi, R.M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316:1497–1502, 2007.

[94] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O.L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4:651–657, 2007.

[95] A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei,

I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837, 2007.

[96] T.S. Mikkelsen, M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.K. Kim, and R.P. Koche. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448:553–560, 2007.

[97] S. Shivaswamy, A. Bhinge, Y. Zhao, S. Jones, M. Hirst, and V.R. Iyer. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biology*, 6(3):e65, 2008.

[98] D.E. Schones, K. Cui, S. Cudapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, 2008.

[99] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, E. Wong, Y.L. Orlov, W. Zhang, and J. Jiang et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, 2008.

[100] P. Park. Epigenetics meets next-generation sequencing. *Epigenetics*, 3:6:318–321, 2008.

[101] G.M. Euskirchen, J.S. Rozowsky, C. Wei, W.H. Lee, Z.D. Zhang, S. Hartman, O. Emanuelsson, V. Stolc, S. Weissman, M.B. Gerstein, Y. Ruan, and M. Snyder. Mapping of transcription factor binding regions in mammalian cells by chip: Comparison of array- and sequencing-based technologies. *Genome Res.*, 17:898–909, 2007.

[102] Y.H. Loh, Q. Wu, J.L. Chew, V.B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, and J. Liu et al. The oct4 and nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, 38(4):431–440, 2006.

[103] C.L. Wei, Q. Wu, V.B. Vega, K.P. Chiu, P. Ng, T. Zhang, A. Shahab, H.C. Yong, Y. Fu, and Z. Weng et al. A global map of p53 transcription factor binding sites in the human genome. *Cell*, 124(1):207–219, 2006.

[104] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32:41–62, 1998.

[105] H. Bui, D. Phung, and S. Venkatesh. Hierarchical hidden markov models with general state hierarchy. *Proceedings of AAAI*, 2004.

[106] S.P. Shah, W.L. Lam, R.T. Ng, and K.P. Murphy. Modeling recurrent dna copy number alterations in array cgh data. *Bioinformatics*, 23:i450–i458, 2007.

[107] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[108] S. Scott. Bayesian methods for hidden markov models: Recursive computing in the 21st century. *J. Am. Statist. Assoc.*, 97:337–351, 2002.

[109] P.C. Consul. *Generalized Poisson Distributions*. Marcel Dekker, New York, 1989.

[110] N.L. Johnson, S. Kotz, and A.W. kemp. *Univariate discrete distributions*. John Wiley & Sons, New York, 1992.

[111] M. Stephens. Dealing with label switching in mixture models. *J. R. Statist. Soc. B*, 62:795–809, 2000.

[112] K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner. Matind and matinspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, 23:4878–4884, 1995.

[113] K. Cartharius, K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein, and T. Werner. Matinspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 21:2933–2942, 2005.

[114] T.H. Kim, Z.K. Abdullaev, A.D. Smith, K.A. Ching, D.I. Loukinov, R.D. Green, M.Q. Zhang, V.V. Lobanenkov, and B. Ren. Analysis of the vertebrate insulator protein ctcf-binding sites in the human genome. *Cell*, 128:1231–1245, 2007.

[115] C.J. Schoenherr and D.J. Anderson. The neuron-restrictive silencer factor (nrsf): a coordinate repressor of multiple neuron-specific genes. *Science*, 267:1360–1363, 1995.

[116] C.J. Schoenherr, A.J. Paquette, and D.J. Anderson. Identification of potential target genes for the neuron-restrictive silencer factor. *Proc. Natl. Acad. Sci. USA*, 93:9881–9886, 1996.

[117] R. Ohlsson, R. Renkawitz, and V. Lobanenkov. Ctcf is uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.*, 17:520–527, 2001.

[118] K.L. Dunn and J.R. Davie. The many roles of the transcriptional regulator ctcf. *Biochem. Cell Biol.*, 81:161–167, 2003.

[119] B. Domon and R. Aebersold. Mass spectrometry and protein analysis. *Science*, 312:212–217, 2006.

[120] A.I. Nesvizhskii and R. Aebersold. Interpretation of shotgun proteomic data. *Mol. Cell. Proteomics*, 4:1419–1440, 2005.

[121] A.I. Nesvizhskii, O. Vitek, and R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods*, 4:787–797, 2007.

[122] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.*, 17:994–999, 1999.

[123] W.A. Tao and R. Aebersold. Advances in quantitative proteomics via stable isotope tagging and mass spectrometry. *Curr. Opin. Biotechnol.*, 14:110–118, 2003.

[124] W. Wang, H. Zhou, S. Roy, T. Shaler, L. Hill, S. Norton, P. Kumar, M. Anderle, and C. Becker. Quantitation of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.*, 75:4818–4826, 2003.

[125] H. Liu, R.G. Sadygov, and J.R. III Yates. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.*, 76:4193–4201, 2004.

[126] P.L. Ross, Y.N. Huang, J.N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D.J. Pappin. Multiplexed

protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics*, 3:1154–1169, 2004.

[127] J. Listgarten and A. Emili. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics*, 4:419–434, 2005.

[128] X. Li, E.C. Yi, H. Zhang, and R. Aebersold. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics*, 4:1328–1340, 2005.

[129] J.D. Jaffe, D.R. Mani, K.C. Leptos, G.M. Church, M.A. Gillette, and S.A. Carr. Pepper, a platform for experimental proteomic pattern recognition. *Mol. Cell. Proteomics*, 5:1927–1941, 2006.

[130] W.M. Old, K. Meyer-Arendt, L. Aveline-Wolf, K.G. Pierce, A. Mendoza, J.R. Sevinsky, K.A. Resing, and N.G. Ahn. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics*, 4:1487–1502, 2005.

[131] Y. Ishihama, Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann. Exponentially modified protein abundance index for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mo. Cell. Proteomics*, 4:1265–1272, 2005.

[132] J. Colinge, D. Chiappe, S. Lagache, M. Moniatte, and L. Bougueleret. Differential proteomics via probabilistic peptide identification scores. *Anal. Chem.*, 77:596–606, 2005.

[133] B. Zybailov, A.I. Mosley, M.E. Sardiu, M.K. Coleman, L. Florens, and M.P.

Washburn. Statistical analysis of membrane proteome expression changes in saccharomyces cerevisiae. *J. Proteome Res.*, 5:2339–2347, 2006.

[134] P. Lu, C. Vogel, R. Wang, X. Yao, and E.M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.*, 25:117–124, 2007.

[135] X. Fu, S.A. Gharib, P.S. Green, M.L. Aitken, D.A. Frazer, D.R. Park, T. Vaisar, and J.W. Heinecke. Spectral index for assessment of differential protein expression in shotgun proteomics. *J. Proteome Res.*, 7, 2008.

[136] B. Zhang, N.C. VerBerkmoes, M.A. Langston, E. Uberbacher, R.I. Hettich, and N.F. Samatova. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.*, 5:2909–2918, 2006.

[137] V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98:5116–5121, 2001.

[138] G. Parmigiani, E.S. Garrett, R.A. Irizarry, and S.L. Zeger. *The analysis of gene expression data*. Springer-Verlag, New York, 2003.

[139] K.A. Do, P. Muller, and M. Vannucci. *Bayesian inference for gene expression and proteomics*. Cambridge University Press, New York, 2006.

[140] E. Segal, N. Friedman, N. Kaminski, A. Regev, and D. Koller. From signatures to models: understanding cancer using microarrays. *Nat. Genet.*, 37:S38–S45, 2005.

[141] N.M. Pavelka, M.L. Fournier, S.K. Swanson, M. Pelizzola, P. Ricciardi-Castagnoli, L. Florens, and M.P. Washburn. Statistical similarities between

transcriptomics and quantitative shotgun proteomics data. *Mol. Cell. Proteomics*, Advanced E-publication, 2007.

[142] S.L. Zeger and M.R. Karim. Generalized linear models with random effects; a gibbs sampling approach. *J. Amer. Stat. Assoc.*, 86:79–86, 1991.

[143] H. Jeffreys. *The Theory of Probability*. Oxford University Press, Oxford, 1961.

[144] G.D. Bader, D. Betel, and C.W.V. Hogue. Bind - the biomolecular interaction network database. *Nucleic Acids Res.*, 29:242–245, 2001.

[145] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S. Kim, and D. Eisenberg. Dip: the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, 30:303–305, 2002.

[146] H.W. Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman. Mips: a database for genomes and protein sequences. *Nucleic Acids Res.*, 30:31–34, 2006.

[147] B.J. Breitkreutz, C. Stark, and M. Tyers. The grid: the general repository for interaction datasets. *Genome Biol.*, 3:R0013, 2002.

[148] P. Jorgensen, J.L. Nishikawa, B. Breitkreutz, and M. Tyers. Systematic identification of pathways that couple cell growth and division in yeast. *Science*, 297:395–400, 2002.

[149] W. Kolch, M. Calder, and D. Gilbert. When kinases meet mathematics: the systems biology of mapk signalling. *FEBS Lett.*, 579:1891–1895, 2005.

[150] I. Amit, R. Wides, and Y. Yarden. Evolvable signaling networks of receptor tyrosine kinases: relevance of robustness to malignancy and to cancer therapy. *Mol. Syst. Biol.*, 3:151, 2007.

[151] M.R. Birtwistle, M. Hatakeyama, N. Yumoto, B.A. Ogunnaike, J.B. Hoek, and B.N. Kholodenko. Ligand-dependent response of the erbb signaling network: experimental and modeling analyses. *Mol. Syst. Biol.*, 3:144, 2007.

[152] H. Pham, R. Ferrari, S.J. Cokus, S.K. Kurdistani, and M. Pellegrini. Modeling the regulatory network of histone acetylation in saccharomyces cerevisiae. *Mol. Syst. Biol.*, 3:153, 2007.

[153] N.N. Batada, L.D. Hurst, and M. Tyers. Evolutionary and physiological importance of hub proteins. *PLoS comp. biol.*, 2:e88, 2006.

[154] G.D. Bader, A. Heilbut, B. Andrews, M. Tyers, T. Hughes, and C. Boone. Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends. Cell. Biol.*, 13:344–356, 2003.

[155] A. Breitkreutz and M. Tyers. A sophisticated scaffold wields a new trick. *Science*, 311:789–790, 2006.

[156] D. Scholtens and R. Gentleman. Making sens of high-throughput protein-protein interaction data. *Stat. Appl. Genet. Mol. Biol.*, 3:39, 2004.

[157] D. Scholtens, M. Vidal, and R. Gentleman. Local modeling of global interactome networks. *Bioinformatics*, 21:3548–3557, 2005.

[158] M.E. Sardiu, Y. Cai, J. Jin, S.K. Swanson, R.C. Conaway, J.W. Conaway, L. Florens, and M.P. Washburn. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc. Natl. Acad. Sci. U.S.A.*, 105:1454–1459, 2008.

[159] F. Desiere, E.W. Deutsch, N.L. King, A.I. Nesvizhskii, P. Millick, J. Eng,

S. Chen, J. Eddes, S.N. Loevenich, and R. Aebersold. The peptideatlas project. *Nucleic Acids Res.*, 34:D655–658, 2006.

[160] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, 2003.