ABSTRACT

Motivation: Data from DNA microarrays and ChIP-chip binding assays often form the basis of transcriptional regulatory analyses. However, experimental noise in both data types combined with environmental dependence and uncorrelation between binding and regulation in ChIP-chip binding data complicate analyses that utilize these complimentary data sources. Therefore, to minimize the impact of these inaccuracies on transcription analyses it is desirable to identify instances of gene expression-ChIP-chip agreement, under the premise that inaccuracies are less likely to be present when separate data sources corroborate each other. Current methods for such identification either make key assumptions that limit their applicability, and/or yield high false positive and false negative rates. The goal of this work was to develop a method with a minimal amount of assumptions, and thus widely applicable, that can identify agreement between gene expression and ChIP-chip data at a higher confidence level than current methods.

Results: We demonstrate in Saccharomyces cerevisiae that currently available ChIP-chip binding data explains microarray data from a variety of environments only as well as randomized networks with the same connectivity density. This suggests a high degree of inconsistency between the two data types, and illustrates the need for a method that can identify consistency between the two data sources. Here we have developed a Gibbs sampling technique to identify genes whose expression and ChIP-chip binding data are mutually consistent. Compared to current methods that could perform the same task, the Gibbs sampling method developed here exceeds their ability at high levels (>50%) of transcription network and gene expression error, while performing similarly at lower levels. Using this technique, we show that on average 73% more gene expression features can be captured per gene as compared to the unfiltered use of gene expression and ChIP-chip derived network connectivity data.

Availability: Our algorithm is available on request from the authors, and soon to be posted on the web. See author’s homepage for details, http://www.seas.ucla.edu/~liaoj/

Contact: liaoj@ucla.edu

Supplementary Information: at Bioinformatics online.

1 INTRODUCTION

Gene expression data and ChIP-chip binding data are often used in conjunction to deduce transcriptional regulation (Banerjee and Zhang, 2003; Bar-Joseph et al., 2003; Liao et al., 2003; Gao et al., 2004; Luscombe et al., 2004; Boulesteix and Strimmer, 2005; Kao et al., 2005; Ruan and Zhang, 2005; Tran et al., 2005; Yang and Liao, 2005; Yang et al., 2005; Galbraith et al., 2006; Lemmens et al., 2006; Sun et al., 2006). Ideally, these complimentary data sources would provide sufficient approximations of gene transcription (expression data) and the transcriptional regulatory network (binding data). However the ability of these data types to provide such functionality is often adversely affected by a number of factors. For gene expression data, experimental noise associated with the DNA microarray technique detracts from its ability to properly portray transcription. While for ChIP-chip data, in addition to experimental noise (Bar-Joseph et al., 2003), environmental dependence in binding (Harbison et al., 2004; Luscombe et al., 2004; Boulesteix and Strimmer, 2005) and uncorrelation between binding and regulation (Bar-Joseph et al., 2003) cause difficulties. The uncertainty these issues may impart to transcription analyses could be significant. For instance, using a common combinatorial model of transcription regulation we demonstrate that currently available ChIP-chip data performs as well as random networks with the same connectivity density when attempting to describe gene expression from 10 different environments. This result illustrates that the approximations of transcription as expression data and the transcription network as ChIP-chip binding data are in certain instances insufficient, and may generate misleading inferences from transcription analyses. One way to ensure the accuracy of these approximations is to identify agreement between gene expression and ChIP-chip binding data under the premise that inaccuracies are less likely to be present when separate data sources point to the same conclusion.

Previous methods have been devised to identify agreement between ChIP-chip and gene expression data (Bar-Joseph et al., 2003; Gao et al., 2004; Boulesteix and Strimmer, 2005; Ruan and Zhang, 2005). For instance, Bar-Joseph et al. 2003 identified regulatory modules as those genes that were bound by the same set of transcription factors in ChIP-chip data that had highly correlated expression. By using gene expression data under the assumption that regulator activity should correlate with target gene expression, Gao et al. 2004 concluded that on average 58% of the binding targets identified by ChIP-chip data in Saccharomyces cerevisiae are true regulatory targets. Following the same assumption, Boulesteix and Strimmer 2005 documented an environmental dependence in ChIP-chip accuracy. Ruan and Zhang 2005 used decision
trees to investigate how well ChIP-chip data predicts the up/down/unchanged expression of genes under stress and cell cycle conditions. While all these approaches have had success detecting instances of ChIP-chip and gene expression data agreement, they all make key assumptions that may not be valid under all circumstances. For instance, Bar-Joseph et al. 2003 assumed that genes bound by the same set of regulators must be controlled proportionally. This ignores the possibility of independently acting transcription regulators. While justifiable when searching for genes controlled by transcription complexes, this assumption cannot account for instances of combinatorial regulation. For Gao et al. 2004 and Boulesteix and Strimmer 2005, the assumption of correlation between transcription factor activity and target gene expression may be valid for singly regulated genes, but may not hold true for genes controlled by multiple regulators. For Ruan and Zhang 2005, the implicit assumption that transcriptional regulation is an on/off event from a basal state ignores more complicated regulation, such as a spectrum of induced and repressed states.

Here we have developed a method that can identify instances of ChIP-chip and gene expression data agreement, allows for the function of transcription complexes, combinatorial regulation, and the independent action of transcription factors, accounts for uncorrelation between transcription factor activity and target gene expression when a gene is controlled by more than one regulator, and allows differential gene expression to be more than an on/off event from a basal state. Our method is based on concepts and principles from (Brynildsen et al., 2006) concerning the behavior of networked systems, and a model for gene expression from (Liao et al., 2003). By employing a Gibbs sampler and Bayesian statistics to deal with inaccuracies in ChIP-chip and gene expression data we are able to identify genes whose binding and expression data corroborate one another more effectively than current approaches.

To demonstrate the utility of our technique we analyzed S. cerevisiae gene expression data from 10 different environments, and S. cerevisiae ChIP-chip data from a variety of conditions (Gasch et al., 2000; Lyons et al., 2000; Gasch et al., 2001; Lee et al., 2002; Yoshimoto et al., 2002; Harbison et al., 2004). In each environment our method identified genes whose expression and binding data were in agreement (consistent genes). To illustrate how such screenings for enhanced accuracy improve transcription analyses, Network Component Analysis (NCA) was performed on those genes identified by our method to have accurate expression and binding data, and performed separately on the datasets as a whole. We found that when datasets were analyzed as a whole, without concern for which genes had data that provided an accurate approximation, random networks could describe approximately the same amount of variation in the data. However, when only consistent genes were analyzed, significantly more variation could be described. On average 73% more data variation could be described per gene as compared to use of the whole dataset. This clearly shows that the accuracy of gene expression and ChIP-chip binding data representations of transcription and the transcription network are important for transcription analyses. This suggests that a pre-processing step, such as our Gibbs technique, should be used to determine what portions of the system can be analyzed with confidence.

Lastly, we looked for enrichment of functional annotation of consistent gene sets in each condition (http://db.yeastgenome.org/cgi-bin/GO/goTermFinder). For every environment there was an enrichment of genes associated with the ribosome and metabolism. This is not surprising since ribosomal function and metabolism are core processes and portions of them are most likely regulated in a similar fashion over a variety of environments.

2 METHODS

2.1 Background

Figure 1 illustrates the problem we are addressing. DNA microarray measured gene expression and ChIP-chip derived network connectivity under certain circumstances may not represent transcription and the transcriptional regulatory network well. We seek to identify genes with both accurate expression and accurate connectivity (consistent). Intuitively, consistent genes could be those genes that best fit a transcriptional regulatory model. Following this line of reasoning, gene expression and ChIP-chip binding data could be fit to commonly used transcriptional regulatory models in order to identify consistent genes. Typically, gene expression data are decomposed using the following model:

\[ E = AP + \Gamma \]

where \( E \) is the expression data \((N \times M)\), \( A \) is the transcription network \((N \times L)\) which can be constructed from ChIP-chip binding data, \( P \) is the collection of transcription factor activities \((L \times M)\), \( \Gamma \) is the residual variation term that accounts for differences between the model and data \((N \times M)\), \( N \) is the number of genes, \( L \) is the number of transcription factors, and \( M \) is the number of experiments or time points.

In general, two types of model are used. The first type (type I) uses static magnitudes to describe regulator-promoter affinity (Bussemaker et al., 2001; Wang et al., 2002; Roven and Bussemaker, 2003; Gao et al., 2004). These values often come from the number of regulator motifs found within a promoter region, as in the case of REDUCE (Bussemaker et al., 2001; Roven and Bussemaker, 2003), or ChIP-chip derived binding strengths, as in the case of MA-Networker (Gao et al., 2004). The second type of model (type II) allows some of the regulator-promoter strengths to

![Figure 1: Possible disparity between A) ChIP-chip derived connectivity and true network B) DNA microarray measured gene expression and true expression.](image)
float, while others are held constant at zero (Liao et al., 2003; Tran et al., 2005). In general, the type of mathematical procedure where certain values are deduced and others held constant, is termed Confirmatory Factor Analysis (CFA) (Thurstone, 1947; Koopman and Reiersol, 1950; Anderson and Rubin, 1956; Anderson, 1984). For transcription regulation, a model based form of CFA was developed and termed Network Component Analysis (NCA) (Liao et al., 2003). Network Component Analysis was derived as a power-law based transcription regulatory model from mRNA generation and degradation kinetics to relate gene expression ratios (E) to regulator activities (P) through the transcription network (A) in Eq. (1). In NCA connectivity data (ChIP-chip, sequence analysis, etc) directs which regulator-promoter interactions are allowed to float and which remain fixed. If connectivity data suggests an interaction the magnitude of that interaction is allowed to float, while if no such evidence exists the interaction is constrained to zero. This more closely follows experimental observations about environmental dependence in regulator-promoter binding, since it allows regulator-promoter strengths to vary when analyzing different environments.

While either type I or type II models could be used to identify consistent genes, both models tend to favor highly connected genes due to overfitting, and struggle when a high degree of uncertainty is present. This is evidenced in Figure 2, where it shows that when a large degree of error in expression and network connectivity data is present false positive rates for fitting procedures become substantial.

![Figure 2: False positive and false negative rates for synthetic data comparison between type II model (Huber, Cauchy, Fair, OLS) and the Gibbs sampler described here. Specific values and type I model comparison can be found in Table S1 of Supplemental Information.](image)

### 2.2 Gibbs Overview

In contrast to the model fitting mindset, our Gibbs technique views the problem from a network perspective. In bipartite networked systems, such as transcriptional regulation, numerous relationships exist between network components that result from the network topology (Brynildsen et al., 2006). In transcription systems, these relationships are often obscured by error and uncertainty in gene expression and ChIP-chip data. Consistent genes would be those genes that satisfy all network relationships amongst each other at a given tolerance level, due to noise (see Supplemental Information section 3). By utilizing concepts from (Brynildsen et al., 2006), we recognized that properly chosen sub-networks of L genes (L = TFs in the network) can be used as seeds to interrogate network relationships (see section 2.4) and thus identify consistent genes. To uncover the largest number of consistent genes in the network, these subnetworks should be populated with consistent genes. Since we do not know which genes are consistent a priori, a Gibbs sampler was devised to populate these sub-networks with consistent genes and thus maximize the number of consistent genes identified.

![Gibbs Sampler](image)

Fig.2: False positive and false negative rates for synthetic data comparison between type II model (Huber, Cauchy, Fair, OLS) and the Gibbs sampler described here. Specific values and type I model comparison can be found in Table S1 of Supplemental Information.

Fig.3: Gibbs sampling technique. Blue nodes are current seeds, pink nodes are updated seeds, and green nodes are genes deemed consistent by the Gibbs sampler.

The Gibbs sampler maximizes the total number of consistent genes which is a function of all network components and is considered a joint distribution. To maximize this high dimensional joint distribution we iteratively sample from a series of conditional distributions, and select updates probabilistically. The method is outlined in Figure 3. The algorithm is initialized with a random selection of seeds, and proceeds through a series of iterations that execute the following steps:

1. Select $\eta_i$ for updating: From the current $L$ seeds (blue nodes) one is chosen for updating, $\eta_i$, either at random or in a specified order.
2. Construct conditional distribution: The conditional distribution, $\eta_i = |\eta_{i|S_i}|, Z_A,E$ is constructed by examining all $\eta_i$ candidates as $\eta_i$. The set of $\eta_i$ candidates, $S_i$, consists of all genes regulated by TF $i$ that are not current seeds for other transcription factors. Every gene candidate for $\eta_i$ ($gene \in S_i$) is evaluated as $\eta_i$ for its ability to identify consistent genes, holding $\eta_{j|S_i}$ constant. The number of consistent genes identified by different $\eta_i$ is taken to be proportional to the likelihood that it is itself a consistent gene. This constructs the conditional distribution, $\eta_i = |\eta_{i|S_i}|, Z_A,E$ (see Section 2.5). Note that as the algorithm progresses $\eta_{j|S_i}$ can
change, thus the conditional distribution needs to be constructed for any new seed sets, \( \eta_j = \).

3) Sampling step: The conditional distribution is normalized, and an update for \( \eta_j \) (pink node) is selected probabilistically.

This procedure preferentially selects consistent genes to occupy \( \eta_j \), since erred genes would be unlikely to identify consistent genes. In a single iteration of the Gibbs sampler an update is performed for every \( \eta_j \). Seed sets that identify a comparatively large number of consistent genes are logged, and convergence is tested after 1000 of these sets have been catalogued. After a convergence criterion based on ranking has been met (see Supplemental Information section 3.4), the method outputs a set of consistent genes.

Both model based fitting and our Gibbs sampler are justifiable procedures. In general, model fitting procedures assume that connectivity and expression are correct, and attempt to minimize the residual, \( \Gamma \), based on these assumptions. Gibbs sampling on the other hand allows for errors in connectivity and expression by not enabling them to impact the rest of the analysis. This is done by preferentially disallowing erred genes to become seeds. A detailed comparison of the ability of these methods to identify consistent genes in synthetic data is presented in the Results section.

2.3 Gene Expression Model

Our approach uses the transcriptional regulation model from (Liao et al., 2003):

\[
mRNA_i(t) = \prod_{j=1}^{L} \left( \frac{TFA_j(t)}{TFA_j(0)} \right)^{CS_{ij}}
\]

where mRNAs, \( \frac{mRNAs(t)}{mRNAs(0)} \) is the relative abundance of mRNA from promoter \( i \), \( TFA_j(t)/TFA_j(0) \) is the relative activity of TF \( j \), and \( CS_{ij} \) is the strength with which TF \( j \) controls expression from gene promoter \( i \). To investigate genome-wide and over multiple experiments we can use the matrix form of Eq. (2):

\[
E = AP
\]

Further, \( A \) is characterized by the connectivity network \( Z_A \).

\[
Z_A = \{A \in \mathbb{R}^{NL} | a_{ij} = 0, \text{for a given set of } (i, j) \}
\]

where \( a_{ij} = 0 \) indicates a lack of evidence to support binding of TF \( j \) to the promoter of gene \( i \). When noise, error, or uncertainty is present in connectivity and gene expression data, it transforms Eq. (3) into Eq. (1).

\[
E = AP + \Gamma
\]

2.4 Consistency and Network Relationships

Based on the premise that in networked systems certain relationships should exist between network components, we partition Eq. (3) such that:

\[
E_1 = A_1P
\]

\[
E_2 = A_2P
\]

\[
E_1 = A_1A_2^T E_2
\]

where \( A_2 \) is an invertible subnetwork of \( A_1 \). \( E_2 \) is the expression data of genes whose connectivity comprise \( A_2 \), \( A_1 \) is the network of the remaining genes not within \( A_2 \), and \( E_1 \) is the expression data of genes whose connectivity comprise \( A_1 \). Note that \( A_2 \) must be full rank for an invertible \( A_2 \) to be found. Eq. (8) states that the expression of genes in \( E_2 \) can be related to those in \( E_1 \), through the network term \( A_2^T \) (see Supplemental Information section 3.1 and (Brynildsen et al., 2006)). However, if error or uncertainty exists, Eq. (3) transforms into Eq. (1), and relationships between the genes break down.

Consider a set of genes that contain accurate gene expression and accurate connectivity data (consistent). We can partition our matrices such that:

\[
\left[ \begin{array}{c} E_1 \\ E_2 \end{array} \right] = \left[ \begin{array}{c} A_1 \\ A_2 \end{array} \right] P + \left[ \begin{array}{c} \Gamma_1' \\ \Gamma_2' \end{array} \right]
\]

where the \( e \) subscript designates erred genes and the \( c \) subscript designates consistent genes. Consistent genes would have a negligible \( \Gamma_1' \) and reduce to:

\[
E_1 = A_1 P
\]

If an invertible subnetwork could be found within \( A_1 \), an equation analogous to Eq. (8) could be formed, and all relationships dictated by \( A_1 A_2^T \) would be satisfied. Further, if erred genes were not partitioned away from consistent genes, and \( A_2 \) was still populated with consistent genes, the relationships between the consistent genes would be satisfied, since this would simply produce the following:

\[
\left[ \begin{array}{c} E_1' \\ E_2' \end{array} \right] = \left[ \begin{array}{c} A_1 \\ A_2 \end{array} \right] P + \left[ \begin{array}{c} 0 \\ \Gamma_2' \end{array} \right]
\]

where \( c' \) designates the remaining consistent genes once \( L \) have been partitioned into \( A_2 \), and \( \Gamma_2' \) is the error associated with \( E_2 \) being approximated as \( A_2^T E_2 \). However, if erred genes populated \( E_2 \), Eq. (11) would be:

\[
\left[ \begin{array}{c} E_1' \\ E_2' \end{array} \right] = \left[ \begin{array}{c} A_1 \\ A_2 \end{array} \right] P + \left[ \begin{array}{c} \Gamma_1' \\ \Gamma_2' \end{array} \right]
\]

and neither \( \Gamma_1' \) nor \( \Gamma_2' \) would be negligible, and the relationships dictated by \( A_1 A_2^T \) would be violated.

Interestingly, if \( E_2 \) were populated with both consistent and erred genes consistent genes can still be identified, although not as well as if \( E_2 \) were solely populated with consistent genes. Following Appendix C of (Brynildsen et al., 2006), we realize that \( A_1 A_2^T \) and \( A_1 A_2^T \) can have zero patterns. Therefore, if a consistent gene in \( E_1 \) is solely related to consistent genes in \( E_2 \) through \( A_1 A_2^T \), the network relationships between those genes will be satisfied. However, if a consistent gene in \( E_1 \) is related at all to an erred gene in \( E_2 \), it will be deemed erred. Therefore, to identify the largest number of consistent genes we need to populate \( E_2 \) with consistent genes. Since we do not know consistent genes a priori, we rely on Gibbs sampling and Bayesian statistics to identify the best candidates to populate \( E_2 \).

It is important to note that in many datasets an invertible \( A_2 \) does not exist, thereby requiring \( E_2 \) to be populated with both consistent and erred genes. Our Gibbs sampling method is designed to find those genes that when placed as seeds \( (E_2) \) maximize the number of consistent genes. If an invertible \( A_2 \) exists, our method is designed to find it, however, if it one...
does not exist, our method identifies the seeds \( \mathbf{E}_2 \) that maximize the number of consistent genes. Thus, our method populates seeds with consistent genes when possible, and does not require an invertible \( \mathbf{A}_2 \) to exist. The single requirement of our method is that \( \mathbf{A} \) be of full rank.

2.5 Gibbs Strategy
By sampling we look to identify the correct list of consistent genes from a dataset. We begin by recognizing that if the consistent genes are selected for use in Eq. (8) the number of consistent genes will be maximized. This is a valid assumption since erred genes will violate network relationships needed to be intact for genes to be deemed consistent. We define \( r \) as a vector of length \( N \) \( (N = \# \text{ genes}) \), populated by 0’s and 1’s:

\[
r = [1, 0, \ldots, 1]
\]

where a 1 designates a consistent gene whose expression and connectivity data are consistent with each other, and 0 designates an inconsistent/erred gene. Note that the decision for consistency is dependent on the choice of \( L \) genes in \( \mathbf{E}_2 \) (seed genes), given the set of expression data \( \mathbf{E} \) and topology \( \mathbf{Z}_A \):

\[
r = f(\eta_1, \eta_2, \ldots, \eta_L; \mathbf{Z}_A, \mathbf{E})
\]

(13)

Here \( \eta_j \) represents the identity of the \( j \)th gene of the \( L \) seeds. We use \( \mathbf{Z}_A \) here to indicate that we do not know the strength with which TFs control their targets, and that we leave these interactions unconstrained within our analysis. For a detailed description of \( f \) please refer to the Supplemental Information section 3.3. We can view \( \sum r_i \) as our joint distribution.

\[
\sum_{i=1}^{N} r_i = f_s(\eta_1, \eta_2, \ldots, \eta_L; \mathbf{Z}_A, \mathbf{E})
\]

(14)

To explore the parameter space of \( \sum r_i \), we have chosen to employ a Gibbs sampler. Our goal is to maximize \( \sum r_i \) by sampling from the conditional distributions:

\[
f_s(\eta_1 | \eta_2, \ldots, \eta_L, \mathbf{Z}_A, \mathbf{E})
\]

\[
f_s(\eta_2 | \eta_1, \eta_3, \ldots, \eta_L, \mathbf{Z}_A, \mathbf{E})
\]

\[
\vdots
\]

\[
f_s(\eta_L | \eta_1, \ldots, \eta_{L-1}, \mathbf{Z}_A, \mathbf{E})
\]

For each \( \eta_j \) there is a list of genes that are possible candidates. This list is denoted as \( S_j \), and we strive to identify the members of \( S_j \) that maximize \( \sum_{i=1}^{N} r_i \) when positioned as \( \eta_j \). Since the set of genes, \( \{\eta_1, \eta_2, \ldots, \eta_L\} \), needs each regulator to partially control at least one gene (full rank of \( \mathbf{A}_2 \) requirement), a candidate for \( \eta_j \) must be regulated by TF \( j \). Hence, every \( S_j \) is composed of all the genes regulated by TF \( j \) that are not currently one of the \( \eta_{h,j} \). This implies that a gene can be a member of multiple \( S_j \)’s if it is regulated by more than one TF, which indeed is true.

For every \( \eta_j \), we set all other variables \( \{\eta_{h,j} \neq j\} \) constant, and construct the conditional distribution, \( f_s(\eta_j | \eta_{h,j}; \mathbf{Z}_A, \mathbf{E}) \), by allowing every member of \( S_j \) to populate \( \eta_j \) during an evaluation of \( \sum r_i \). With the results for the \( sj \) evaluations we construct a probability balanced on a collapsed form of \( f_s(\eta_j | \eta_{h,j}; \mathbf{Z}_A, \mathbf{E}) \) (see Supplemental Information section 3.3). We then select the update for \( \eta_j \) by sampling from this probability distribution. The algorithm is said to converge once a convergence criteria based on general rank invariance of the genes has been attained. Details of the probability distribution and convergence criteria can be found in the Supplemental Information section 3.3 and 3.4, respectively.

3 RESULTS
Transcription analyses often use ChIP-chip data to define transcription network connectivities that are used to analyze gene expression data under the assumption that ChIP-chip derived connectivities are biologically meaningful, and therefore able to produce more biologically pertinent results. Under this assumption, ChIP-chip based network connectivities should explain more of the gene expression data than random networks. Here we demonstrate that this is not necessarily true. Using a common combinatorial transcription model (NCA), S. cerevisiae ChIP-chip data and gene expression data from 10 different environments (calcium treatment, diamide treatment, DTT treatment, γ irradiation, H2O2 treatment, menadione treatment, MMS treatment, nitrogen starvation, salt treatment, zinc deprivation) we show, in Figure 4, that randomized ChIP-chip networks performed comparably to unaltered ChIP-chip derived networks (see Supplemental Information section 1-2). ChIP-chip based networks on average yielded a 58.6% relative residual per gene, while randomized networks with the same edge density on average gave a 60.3% relative residual per gene. This result suggests that the ChIP-chip data explains on average 41.4% of the gene expression variation, which is only 1.7% more than randomized networks. The low gain associated with use of ChIP-chip data over randomized networks presents a major problem for transcription studies involving simultaneous use of both types of data.

This problem can be attributed to two possible causes: 1) noise in gene expression data, 2) uncertainty in network connectivity data (e.g. uncorrelation between binding and regulation, environmental dependence in binding). Our Gibbs sampler was designed to address these issues. Another possible cause is the model used in the analysis. Since Eq. (1) is widely used in transcriptome analysis and more complicated models compromise mathematical tractability, we choose to identify a subset of genes whose expression and
connectivity data are mutually consistent based on the existing model.

The most intuitive approach to identify consistent genes is to use model-fitting based methods. Genes that fit the model well (i.e. small residual variation) could be deemed consistent. Both type I and type II models (see section 2.1) were used as benchmarks for evaluating the performance of our Gibbs sampling technique. To demonstrate the effectiveness of our method we first analyzed synthetic data generated from known networks with both expression data and connectivity data corruption at given levels of noise (see Supplemental Information section 2.3). For evaluation purposes, synthetic data has an advantage over real data because one can easily determine false positive and false negative rates. In addition, with synthetic data it is possible to vary error and noise rates in expression and connectivity data to map out the advantages of different methods. However, synthetic data should approximate real data as closely as possible. For that reason we used the transcription regulation model of (Liao et al., 2003) to create the synthetic data, incorporated overall expression signal to noise ratios of 1 and 2, and kept network connectivity edge densities and size proportions approximately equal to those found in the ChIP-chip derived network connectivity (Supplemental Information section 2.3). The synthetic data was then evaluated with the model-fitting methods and our Gibbs sampling approach.

Both type I and type II models utilize regression to minimize $\Gamma$ (1 step for type I, iterative 2-step for type II). Since regression can be performed with a variety of weighting procedures, multiple schemas (Ordinary Least Squares (OLS); Huber, Cauchy, and Fair weighted robust regression) were tested here (see Supplemental Information section 2.3). The results of the type II model and the Gibbs technique are illustrated in Figure 2, while the exact values for type I, type II, and the Gibbs technique are presented in Table S1 of the Supplemental Information (see Supplemental Information section 2 for method discussion).

Gibbs sampling was also performed on the synthetic data according to the outline in Figure 3, and procedure described in section 2.2. Gibbs sampling had distinct advantages over model-fitting procedure when >50% of the genes had erred connectivity, erred expression or both. This advantage was seen in signal to noise ratios of both 1 and 2, and became more apparent as the percentage of erred genes increased. This is demonstrated by lower false positive and false negative rates (Figure 2). Since gene expression data is considerably noisy, binding is condition dependent, and ChIP-chip data is unavailable in many environmental conditions, our method looks better suited to handle the challenges of gene expression and network connectivity derived from ChIP-chip data than model fitting procedures.

To demonstrate the utility of our Gibbs sampler in transcription systems we analyzed gene expression and ChIP-chip data from S. cerevisiae. Gene expression from 10 different environmental conditions, use of consistent gene sets explained 71.7% of the gene expression variation per gene. This is an increase of 30.3 percentage points over that obtained from unfiltered use of ChIP-chip derived connectivity, which explained 41.4%. Therefore, use of consistent genes on average led to a 73% (30.3/41.4) increase in the amount of gene expression variation explained per gene. It is important to note that there is an environmental dependence associated with consistent gene set performance compared to use of the whole dataset. This can clearly be seen in Figure 5. The best improvement (largest decrease in relative residual per gene) yielded by consistent set use belongs to data from $\gamma$ irradiation, menadione treatment, and hydrogen peroxide treatment, while the poorest improvement belonged to data from salt, MMS, and diamide treatments.

In addition, we looked at whether consistent genes had any distinguishing features. In particular, we searched for enrichment of functional annotation (http://db/yeastgenome.org/cgi-bin/GO/go TermFinder). We found that in every environment analyzed there was enrichment of genes associated with the ribosome and metabolism. This is not surprising since ribosomal function and metabolism are core processes and portions of them are most likely regulated in a similar fashion over a variety of environments. Interestingly, in both starvation environments we looked at, nitrogen and zinc, amino acid metabolism genes were notably enriched in the consistent set. This was not a feature common to stress responses not induced by starvation. Of those genes associated with amino acid metabolism that were deemed consistent under zinc deprivation 79% of them were also deemed consistent under nitrogen starvation. These points taken together suggest that the zinc and nitrogen deprivation responses share some of the same transcriptional wiring, especially in regions related to amino acid metabolism.

![Figure 5](http://www.seas.ucla.edu/~liaoj/). On average, over a variety of conditions, use of consistent gene sets explained 71.7% of the gene expression variation per gene. This is an increase of 30.3 percentage points over that obtained from unfiltered use of ChIP-chip derived connectivity, which explained 41.4%. Therefore, use of consistent genes on average led to a 73% (30.3/41.4) increase in the amount of gene expression variation explained per gene. It is important to note that there is an environmental dependence associated with consistent gene set performance compared to use of the whole dataset. This can clearly be seen in Figure 5. The best improvement (largest decrease in relative residual per gene) yielded by consistent set use belongs to data from $\gamma$ irradiation, menadione treatment, and hydrogen peroxide treatment, while the poorest improvement belonged to data from salt, MMS, and diamide treatments.

In addition, we looked at whether consistent genes had any distinguishing features. In particular, we searched for enrichment of functional annotation (http://db/yeastgenome.org/cgi-bin/GO/go TermFinder). We found that in every environment analyzed there was enrichment of genes associated with the ribosome and metabolism. This is not surprising since ribosomal function and metabolism are core processes and portions of them are most likely regulated in a similar fashion over a variety of environments. Interestingly, in both starvation environments we looked at, nitrogen and zinc, amino acid metabolism genes were notably enriched in the consistent set. This was not a feature common to stress responses not induced by starvation. Of those genes associated with amino acid metabolism that were deemed consistent under zinc deprivation 79% of them were also deemed consistent under nitrogen starvation. These points taken together suggest that the zinc and nitrogen deprivation responses share some of the same transcriptional wiring, especially in regions related to amino acid metabolism.

4 DISCUSSION

Consistent genes have expression data that can result from the action of regulators known to bind their promoters. Conceivably, consistency evaluations with our Gibbs sampler can be used to identify when genes are regulatory targets of their ChIP-chip bound regulators. By no means would this be an exhaustive list of
regulatory targets since false negative rates for our algorithm range from 10-45% when tested with synthetic data, however, the identity of a subset of true regulatory targets provides additional means of uncovering more regulatory targets. One approach could be to identify regulator motifs from the consistent set, and use those to aid in identification. Another could be to use the expression of consistent genes or deduced regulator activities to postulate targets, since their expression would be comparatively clean. Indeed, these methods could be used in conjunction to identify putative regulatory targets not found bound in ChIP-chip data.

It is important to note that our approach can be generalized to other types of binding data when ChIP-chip is unavailable. This is of considerable significance since most organisms have not had extensive ChIP-chip assays performed on them. In organisms without ChIP-chip data, sequence analysis or literature compilations could be used to construct network connectivities. Even though one might expect that network connectivities from sequence analysis would be denser with higher connection false positive rates and lower connection false negative rates than ChIP-chip derived connectivities, and those from literature compilations would be sparser with a lower connection false positive rates and higher connection false negative rates, our Gibbs sampler should still be able to provide accurate results, given its performance at varying error levels while evaluating synthetic data.

In addition, our method can be generalized to any type of data that closely follows the structure presented here (ie. bipartite network, combinatorial action). We believe that our Gibbs sampling method can be an integral preprocessing step for transcriptional regulatory analyses. Our method effectively screens through the entire transcriptome and selects out those genes whose gene expression and ChIP-chip binding data are consistent based on the bipartite model. These genes could then become the basis for transcriptional regulatory analyses that provide more significant biological results.

ACKNOWLEDGEMENTS

This work has been supported by the Center for Cell Mimetic Space Exploration (CMISE) a NASA University Research, Engineering and Technology Institute (URETI) under award number NCC 2-1364, NSF-ITR CCF-0326605, and the UCLA-DOE Institute for Genomics and Proteomics.

REFERENCES


