Time Series Analysis of Gene Expression and Location Data

Chen-Hsiang Yeang
Artificial Intelligence Lab, MIT.
Cambirdge, MA 02139, USA.
chyeang@ai.mit.edu

Tommi Jaakkola
Artificial Intelligence Lab, MIT.
Cambridge, MA 02139, USA.
tommi@ai.mit.edu

Abstract

We develop a method for integrating time series expression profiles and factor-gene binding data to quantify dynamic aspects of gene regulation. We estimate latencies for transcription activation by explaining time correlations between gene expression profiles through available factor-gene binding information. The resulting aligned expression profiles are subsequently clustered and again combined with binding information to determine groups or subgroups of co-regulated genes. The predictions derived from this approach are consistent with existing results ([11], [8]). Our analysis also provides several hypotheses not implicated in previous studies.

1 Introduction

Gene regulation is a complicated process that needs to be understood at multiple levels of description. We consider here two levels, *physical* level involving molecular interactions and *functional* level dealing with the computational behavior of the system. A single high-throughput data source such as expression profiling is geared toward one or the other level of description (functional in this case) and is insufficient for obtaining a full understanding of the regulatory process. It is necessary to constrain the set of possible models of gene regulation from multiple complementary data sources.

Time course expression profiles (e.g., [6, 2]) and factor-gene binding (a.k.a. *location*) data [9, 11] provide two complementary sources. Time course profiles are advantageous over typical expression profiles as time can be used to disambiguate causal interactions. Location data, on the other hand, provides high-throughput quantitative information about in-vivo binding of transcriptional activators to intergenic (regulatory) regions of the DNA. These two data sources – both causally unambiguous – can be readily combined.

A number of previous papers address both time course expression analysis [6, 2, 1] and the combination of such data with location or other data sources such as sequence motifs [5, 11, 12, 13]. However, systematic approaches of combining properties of time course profiles with other genomic data are relatively scarce in previous studies. While, for example, [11, 12] certainly fall in this category, their focus were more on using one data source as a means of validating the other. To fully use time course profiles it seems necessary to temporally align individual gene profiles.

The contribution of this paper is twofold. First, we propose an algorithm to determine the time lag of transcription activation. Second, we introduce a clustering method for expression profiles that involves both the inferred lag times (temporal alignment) and factor-binding constraints. We validate the methods on the cell cycle expression data [12] and location data of 113 transcription factors ([11]; [8]).

The paper is organized as follows. Section 2 provides a brief introduction of data sources. In section 3 we explicate our major assumptions about gene regulation. Section 4 describes the algorithm for delay optimization via constrained temporal alignment. Section 5 subsequently provides the new clustering method for time series expression data. Experimental validation is described in Section 6.

2 Location and time course expression analysis

Location analysis is a genomic scale assay ([9]) measuring the in-vivo abundance of transcription factors that bind to intergenic regions of the DNA. Unlike expression or knock-out experiments, location analysis provides direct evidence about the physical processes underlying gene regulation.

Each factor (transcriptional activator) profiled in the location analysis is associated with a set of p-values computed from an error model such as the one described in [7]. The p-value of each factor-gene pair represents the confidence that the factor binds to the corresponding intergenic region. By thresholding the confidence values, we can view the location data as a directed graph $G = (V, E)$, where the ver-

Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)
0-7695-1907-5/03 $17.00 © 2003 IEEE

tices V represent the factors and genes, and the presence of a directed edge in E represents a significant binding event. We use this simple graph representation in the remainder of the paper.

Gene expression profiling represents a more established high-throughput data source. Models such as Bayesian networks [4] or probabilistic relational models [10] have been used to capture interactions among the measured expression levels (random variables in the model).

Factor analysis provides a very simple class of models involving continuous normally distributed variables. The statistical dependencies between the observable random variables (in this case expression measurements) are explained in this model by a small number of latent factors [3]. The model can be expressed as

$$\mathbf{y} = \mu + \mathbf{\Lambda x} + \mathbf{e}. \quad (1)$$

where $\mathbf{x} \sim N(0, I)$ are independent Gaussian latent variables, $\mathbf{e} \sim N(0, \mathbf{\Phi})$ are independent Gaussian residuals ($\mathbf{\Phi}$ is diagonal), Λ are unknown coefficients (factor loadings), and $\mathbf{y}$ are the observable variables. This model is simple enough to warrant efficient estimation of the associated parameters.

In the context of expression profiling we use latent factor values as proxies for (active) protein levels associated with the transcription factors. This is useful since the expression measurements of activators are themselves often poor indicators of the corresponding protein activities. The interpretation of the latent factors as transcriptional activators also facilitates the combination of these models with constraints from location data (next section). The inherent linearity of the factor analysis models is a major restriction, however, and will be removed in later work.

3 Assumptions of data integration

The semantics of the two types of data discussed above – location and expression – are very different: one characterizes physical interactions and the other characterizes functional interactions. These two types of interactions need not agree in the sense that the expression profile of a gene encoding a factor needs not be correlated with the transcript levels of genes that the factor binds to. We avoid this ambiguity by using latent variables to describe the activity levels of transcription factors rather than relying on their observed transcript levels.

We are interested in explaining the observed dependencies among expression profiles with the physical mechanisms revealed by the location data. For example, we assume that the observed correlations between profiles of genes sharing a common factor can be explained by the latent activity of this factor.

The framework we use in this paper for integrating location and expression data has the following structure. First we build a location graph $G = (V, E)$ by thresholding the p-values of the location data. Since we assume that protein-DNA binding is necessary for gene regulation, each subgraph of G yields a valid dependency model for explaining expression data. Such an expression model takes the form of a factor analysis model. Specifically, the transcription factors in the location graph denote latent variables of protein activities and the genes bound by the factors are replaced by variables encoding the observed transcript levels.

To account for time-series expression data, the lag time of transcription activation/repression needs to be incorporated. In our factor analysis model each edge has a distinct delay, and the expression level of a gene at current time is affected by the *activities* of the latent factors at previous times in accordance with the edge delays (τ_{ij}). The expression profile of gene i can be modeled as

$$y_i(t) = \mu_i + \sum_j \lambda_{ij} x_j(t - \tau_{ij}) + e_i(t). \quad (2)$$

The model construction will be described in section 5.

4 Estimating edge delays

Finding edge delays in the location graph is a nontrivial task. They cannot be determined simply by correlating the expression profiles of factor-gene pairs, since factor expression profiles are poor indicators of their activities. A more reasonable approach would be to build a factor analysis model according to the location graph, write down the likelihood function for the observed expression levels $\mathbf{y}(t)$ in equation 2, and estimate Λ, Φ, μ and the edge delays τ simultaneously. The resulting likelihood function would be difficult to optimize, however, and we resort to a simpler approximation in this paper.

For the purpose of optimizing the edge delays, we start with a second order approximation to the log-likelihood function corresponding to a complete Gauss-Markov model. The resulting log-likelihood function, evaluated at the maximum likelihood setting of the parameters, can be written approximately as a sum of square delayed covariances (derivations are in Appendix A). This property provides us with a simple criterion for adjusting the edge delays, namely maximizing the sum of square delayed covariances.

The delays associated with the edges in the location graph still need to be tied with the observed correlations between the expression profiles. We associate each pair of genes with a set of common (not too distant) ancestors in the location graph that could in principle explain the observed delayed correlation between the genes. In other words, the observed correlation can be attributed to a common cause in

Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)
0-7695-1907-5/03 $17.00 © 2003 IEEE

terms of the cascading effect of protein-DNA interactions. The selection of which common ancestor should explain any particular correlation is carried out jointly with the estimation of the edge delays. We assume here that a delayed correlation between expression profiles arises from the differences between the edge delayes accrued along the paths from the common ancestor to the genes.

Suppose two genes have normalized expression profiles $x_i(t)$ and $x_j(t)$, and there is a non-convergent path (a path without colliding edges, i.e., a path which is not $\rightarrow o \leftarrow$) π from gene i to gene j in the location graph. The empirical delayed covariance function $R_{ij}(\tau)$ can be evaluated relative to a fixed-sized window. The goal is to find the edge delays which maximize the square covariance functions:

$$\Delta_\pi^* = \arg\max_{\Delta_\pi} R_{ij}^2(\sum_{e\in\pi} \chi_\pi(e)\delta_e). \quad (3)$$

where δ_e is the delay associated with edge e in the location graph, Δ_π denote edge delays along path π, and $\chi_\pi(e)$ is the sign of the direction of the path relative to the edge. In other words, $\chi_\pi(e) = +1$ when the path traverses along the edge, and $\chi_\pi(e) = -1$ when the path goes against the edge. Since this is an under-determined problem (unless π has only one edge), there are multiple optimal solutions.

When there are multiple non-convergent paths connecting x_i and x_j, we do not know a priori which paths can be applied to explain their pairwise correlation. This uncertainty is expressed as the probability of choosing a particular path to explain pairwise correlations. Suppose $\{\pi_{ijk}\}_{k=1}^{|\pi_{ij}|}$ are the paths connecting x_i and x_j, and p_{ijk} is the probability of assigning path π_{ijk} to explain the delayed correlation between $x_i(t)$ and $x_j(t)$. The expected value of the square pairwise delayed correlation is:

$$E\{R_{ij}^2(\Delta, P)\} = \sum_{k=1}^{|\pi_{ij}|} p_{ijk} R_{ij}^2(\sum_{e\in\pi_{ijk}} \chi_{\pi_{ijk}}(e)\delta_e) \quad (4)$$

Here both edge delays Δ and path assignments P are unknown. The overall objective function is the sum of expected square delayed correlations over all pairs of genes which share common ancestors in the location graph:

$$R^2(\Delta, P) = \sum_{i\sim j}\{\sum_{k=1}^{|\pi_{ij}|} p_{ijk} R_{ij}^2(\sum_{e\in\pi_{ijk}} \chi_{\pi_{ijk}}(e)\delta_e)\} \quad (5)$$

where $i \sim j$ denotes i and j share common ancestors in the location graph. Here we are interested only in these pairs since we use edge delays to explain correlated pairs. The problem of finding edge delays amounts to maximizing equation 5 subject to constraints

$$\delta_e \geq 0, 0 \leq p_{ijk} \leq 1, \sum_{k=1}^{|\pi_{ij}|} p_{ijk} = 1.$$

Equation 5 specifies the global constraint on edge delays for explaining the time-series data. While the constraint from one pairwise correlation cannot uniquely determine an edge delay (unless the genes are connected by a one-edge path), a reasonable solution of Δ can be obtained by combining all constraints regarding pairwise correlations.

The approximate log-likelihood function in equation 5, though legitimate, has two disadvantages in real datasets. First, the log-likelihood function of a Gauss-Markov model is insensitive to the signs of pairwise correlations. This property leads to an ambiguity for periodic datasets such as cell cycle data: two periodic signals $x_i(t)$ and $x_j(t)$ can be viewed as correlated with a phase shift τ_0 or anti-correlated with a phase shift $\tau_0 + \frac{T}{2}$. The model used in this paper cannot distinguish this ambiguity without external information. Furthermore, the noise in the data might mislead the model to favor the wrong delay (e.g., the actual delay is τ_0 but the noise makes the log-likelihood function of $\tau_0 + \frac{T}{2}$ slightly higher). Second, the $R_{ij}^2(.)$ in equation 5 are empirical square covariance functions. This makes the optimization problem difficult. To simplify the problem it is better to approximate the empirical functions with analytic functions.

For these two reasons we replace $R_{ij}^2(.)$ terms in equation 5 with simple, periodic approximations $\hat{R}_{ij}(\tau - \tau_{ij}^*)$ of covariance functions (rather than square covariance functions): $\hat{R}_{ij}(\tau) = c_{ij} - k_{ij}\tau^2$ within $[-T/2, T/2]$, repeating with period T. τ_{ij}^* is the delay which achieves the maximum delayed correlation between $x_i(t)$ and $x_j(t)$. It is a reasonable approximation for the covariance of two signals with the same frequency: the covarianec has a single peak within each period. This is appropriate for the cell cycle data since all the cell cycle genes (in the same dataset) tend to have the same frequency.

The objective function now reduces to:

$$\hat{R}(\Delta, P) = \sum_{i\sim j}\{\sum_{k=1}^{|\pi_{ij}|} p_{ijk}\hat{R}_{ij}(\sum_{e\in\pi_{ijk}} \chi_{\pi_{ijk}}(e)\delta_e - \tau_{ij}^*)\} \quad (6)$$

This function is a concave function of P and Δ but not jointly concave. We solve it via alternative maximization with respect to P and Δ. Initially we set all paths equally likely to explain pairwise correlations, i.e., $p_{ijk} = \frac{1}{|\pi_{ij}|}$ for each path connecting x_i and x_j.

When Δ is fixed, optimizing P reduces to linear programming:

$$\max_P \sum_{k=1}^{\pi_{ij}} p_{ijk}\hat{R}_{ijk}, \quad (7)$$

where $\hat{R}_{ijk} \equiv \hat{R}_{ij}(\sum_{e\in\pi_{ijk}}(\chi_\pi(e)\delta_e) - \tau_{ij}^*)$, subject to $0 \leq p_{ijk} \leq 1, \sum_{k=1}^{\pi_{ij}} p_{ijk} = 1$. The problem is decoupled

Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)
0-7695-1907-5/03 $17.00 © 2003 IEEE

for each pair of genes, thus the solution is straightforward:

$$p_{ijk} = \begin{cases} \frac{1}{N^*_{ij}} & \text{when } k = \arg\max_l \hat{R}_{ijl} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $N^*_{ij} = \#(k = \arg\max_l \hat{R}_{ijl})$.

When P is fixed, solving Δ becomes a nonlinear programming problem. The objective function becomes quadratic if we restrict the longest path delay to be within one period of the data. In this setting, only the first "mode" of each $\hat{R}_{ij}(\tau)$ is considered. Hence we do not need to worry about the degeneracy of different path delays yield the same pairwise correlation values.

5 Dynamic expression models

Once edge delays are obtained, we can build a dynamic model of gene expression profiles. A decent estimation of dynamic factor analysis model parameters often requires considerable number of time-series data points or repeated experiments which are scarce in gene expression analysis. To overcome this limitation, we choose a simplified model. Align the expression time points in $y_i(t) = \mu_i + \lambda_i x(t - \tau_i) + e_i(t)$ such that $y'_i(t) = y_i(t + \tau_i)$ and equation 2 becomes

$$y'_i(t) = \mu_i + \lambda_i x(t) + e_i(t + \tau_i).$$

Since $e_i(t+\tau_i)$ is a white noise, we can treat $y'_i(t)$ as a realization of a Gaussian random variable instead of a time series data, and the model reduces to a factor analysis model. The activation delay τ_i between a transcription factor and a gene is estimated according to the alternating maximization algorithm in section 4. The covariance matrix of aligned random variables $y'_1, \cdots, y'_n$ then becomes

$$\Sigma = \Lambda\Lambda^T + \Phi \quad (9)$$

where $\Sigma_{ij} = E\{(y'_i - \mu_i)(y'_i - \mu_j)\}$ and $\Lambda = (\lambda_1, \cdots, \lambda_n)^T$. This is identical to static factor analysis except the covariance matrix is built from aligned profiles $y'_i(t)$. The log likelihood is:

$$L(\Lambda, \Phi) = m\left(-\frac{1}{2}p\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}tr\Sigma^{-1}S\right) \quad (10)$$

where m is the size of the sample, p is the dimension of observed variables, and S is the sample covariance matrix. Λ and Φ are the unknown parameters to be estimated.

This simple dynamic expression model can be viewed as a generalization of clustering with protein-DNA binding constraints. For each transcription factor f, a factor analysis model is built based on the genes that it binds to. The latent factor corresponds to the regulatory activity of f, and the observables are the expression profiles of these genes aligned by the inferred edge delays.

We hypothesize that protein-DNA bindings are necessary but not sufficient conditions for gene regulation. Therefore, we need a computational method to tell which genes are dependent (correlated or anti-correlated) with the core of the cluster. Since the parametric model (the factor analysis model) is given, we apply nested hypothesis testing to incorporate genes in the cluster.

Suppose factor f binds to genes $g_1, \cdots, g_n$ in the location data, but having a functional impact only on $g_1, \cdots, g_m$ ($m < n$). Then the appropriate factor analysis model is H_0, where $g_1, \cdots, g_m$ are controlled by the latent variable, while $g_{m+1}, \cdots, g_n$ are independent. In contrast, H_1 assumes that all $g_1, \cdots, g_n$ are linked to the latent variable. The log likelihood ratio (deviance score) between H_1 and H_0 is:

$$\begin{aligned} R &= 2\log(\tfrac{L(H_1,D)}{L(H_0,D)}) \\ &= m\{tr((\Sigma_0^{-1} - \Sigma_1^{-1})S) - \log|(\Sigma_0^{-1} - \Sigma_1^{-1})S|\} \end{aligned} \quad (11)$$

which has an asymptotic χ^2 distribution ([3]) if the data is generated by H_0. The degrees of freedom equal $\nu = 2p$ [3], where p is the number of edges in $H_1 - H_0$.

The p-value of nested hypothesis testing is used to determine the core cluster whose members are co-regulated by the factor. To avoid evaluating all possible submodels, we follow a top-down greedy method Start with H_1, incrementally remove an edge at each step which minimizes the deviance score with respect to the previous model until the p-value of edge removal drops below a threshold. Figure 1 describes this algorithm.

6 Experimental results and discussions

We test our methods on two datasets: Spellman's cell cycle expression data [12] and the location analysis data from [8] and [11].

6.1 Dataset

The expression dataset of *S. cerevisiae* published in [12] contains 5 time-series expression data synchronized by different methods. In the same paper, the authors identified 800 genes as cell cycle related and labeled the active phases of these genes. We use three of these – CDC15, α factor, and CDC28 – due to the size and quality of the data.

Location analysis experiments of 113 yeast transcription factors were undertaken by Lee et al. [8]. This collection comprises about half of the total transcription factors in yeast genome. Here we also use the location data of the 9 cell cycle transcription factors published earlier [11]. The latter is subsumed by the former dataset. The p-values for

Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)
0-7695-1907-5/03 $17.00 © 2003 IEEE

Figure 1. Algorithm of clustering genes bound by the same transcription factor

Input: location graph G, time series expression data D, location graph edge delays Δ, threshold on expression model p-value p_t.
Output: clusters of genes under each transcription factor.
Procedures:

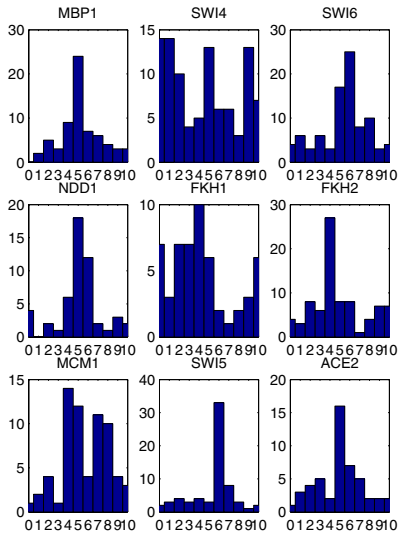
1. For each transcription factor f, find $V_f = \{g : (f, g) \in G\}$.
2. For each V_f, align the expression profiles of its members according to edge delays: $y'_g(t) = y_g(t + \delta_{fg}), \forall g \in V_f$.
3. For each V_f:
   (a) $M =$ the factor analysis model of $\{y'_g(t) : g \in V_f\}, p = 1$.
   (b) while $p > p_t$:
      i. $M' = \arg\max_{K \in re(M)} R(K, M; D)$, where $R(.)$ is the log likelihood ratio and $re(M) = \{K \subset M : |\text{edges}(K)| = |\text{edges}(M)| - 1\}$.
      ii. $\nu =$ the degree of freedom $= 2$.
      iii. $p =$ p-value of $R(M, M'; D)$, $M \leftarrow M'$.
4. Report $G'_f =$ the non-singular cluster in M.

protein-DNA bindings were obtained from the empirical error model developed in [7]. We set the p-value threshold to be 0.006 since it yields a graph consistent with the functional interactions of the 9 cell cycle transcription factors. The results are robust against small changes in the p-value threshold.

We are interested in the genes which are both cell cycle related in the expression data and bound by transcription factors in the location data. 399 genes are in this intersection when the location p-value cutoff is 0.006. There are 240 cell cycle genes bound by 9 cell cycle transcription factors.

Only 15 of out the 113 transcription factors are labeled cell cycle related according to Spellman et al. Among the 9 cell cycle regulators only 5 transcription factors are labeled. The lack of periodicity in the observed expression profile of some of these cell cycle regulators can be attributed to either experimental errors (for example, FKH2 data points are missing in CDC15 dataset) or the nature of physical regulation. In either case, we cannot identify the functional roles of these factors by relying on their expression profiles.

Figure 2. Histograms of factor-gene edge delays under single factors, x: time, y: counts



6.2 Relevant transcription factors for cell cycle regulation

Apparently not all transcription factors are involved in cell cycle regulation. Since mRNA levels are poor indicators for transcription factors' activities, we often cannot determine relevance to cell cycle by inspecting the expression profiles of transcription factors. A more sensible approach is to investigate the expression profiles of the genes they bind to. Simon et al. discovered that all the 9 cell cycle transcription factors bind primarily to genes at one or two phases [11]. This finding is quantitatively verified by our method. We align the expression profiles of the genes bound by each cell cycle regulator. The distributions of edge delays of CDC15 dataset are plotted in figure 2. Notice the histogram is wrapped around the cell cycle period (10 sample intervals), hence $t = 0$ is adjacent to $t = 10$. This plot clearly indicates that these factors bind to genes primarily at one or two phases. Moreover, genes bound by some factors are active not only at the same phase, but also at the same time within the accuracy of the sampling rate.

Based on this observation we establish a criterion for the relevance to cell cycle regulation. We argue that a factor is involved in cell cycle regulation if the transcript levels of cell cycle genes it binds to are concentrated at one or two narrow intervals during the cell cycle. To quantify this statement we derive a p-value for the time preference. The derivation of the corresponding p-value is described in Appendix B.

Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)
0-7695-1907-5/03 $17.00 © 2003 IEEE

Table 1. Cell cycle factors and their functions

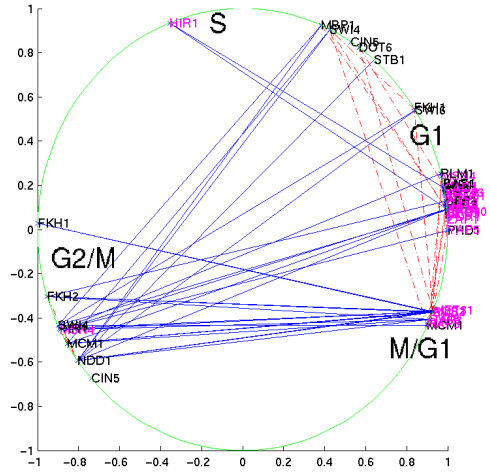
Factor	Function	Factor	Function
MCM1	cell cycle control	MBP1	cell cycle control
SWI4	cell cycle control	SWI6	cell cycle control
FKH1	cell cycle control	FKH2	cell cycle control
NDD1	cell cycle control	ACE2	cell cycle control
SWI5	cell cycle control	PHD1	filament growth
CIN5	cell cycle control	CRZ1	calcineurin response
DOT6	telomeric silencing	HAP3	CCAAT binding factor
IXR1	oxygen regulation	MAC1	metal binding
MAL13	maltose pathway	MSN4	zinc finger
PDR1	drug resistance	RGM1	represor
SMP1	MADS-box	STB1	cell cycle control
ZMS1	acid tolerance	SFP1	cell cycle control
GAL4	galactose pathway	CUP9	copper homeostasis
GLN3	nitrogen regulation	ARG81	arginine metabolism
HAP4	CCAAT binding factor	HAP5	CCAAT binding factor
LEU3	amino acid metabolism	MET31	amino acid metabolism
GAT3	nitrogen regulation	HAA1	regulator
RLM1	MADS-box	ZAP1	zinc response
IME4	sporulation	RCS1	iron repressor
HIR1	histone regulator	ARO80	amino acid metabolism
BAS1	amino acid metabolism	CAD1	drug resistance
YAP5	leucine zipper		

Based on this criterion, we select the factors among the 104 transcription factors (excluding the 9 known transcription factors) whose time-preference p-values are below 0.002 and which bind to more than 10 cell cycle genes. There are 34 transcription factors which satisfy this criterion. Table 1 enlists the names and functions of these factors and the 9 cell cycle transcription factors. Among the 34 putative factors, 8 of them are relevant to cell cycle regulation according to previous studies: CIN5, DOT6, STB1, GLN3, SFP1, RCS1, RLM1 and HIR1.

The fact that some cell cycle relevant factors bind to genes at two phases suggests that they may carry multiple functional roles in cell cycle regulation. In figure 2, MCM1, FKH1 and SWI4 have this property. MCM1 regulates both G2/M and M/G1 genes according to previous studies [11].

Distinguishing multiple functional roles of a transcription factor is crucial for delay optimization and clustering. Since we estimate factor-factor edge delays by correlating clusters of genes bound by the same factors, errors occur if we correlate one cluster with a coherent phase to another cluster with mixed phases. We introduce the notion of *physical* and *functional* factors in the model. A physical factor is an actual transcription factor, while a functional factor denotes a cluster of genes with a coherent cell cycle phase and bound by a factor. One physical transcription factor may bind to several clusters of genes which are expressed at different cell cycle phases. In this case one physical factor corresponds to multiple functional factors. We separate genes bound by the same transcription factor into two clusters if their edge delay distribution is bimodal (for example, SWI4 in figure 2). The following transcription factors are splitted to double functional factors: MCM1, SWI4, FKH1, CIN5, MSN4, and HIR1.

Figure 3. Location graph of putative cell cycle relevant factors



6.3 Edge delays and scores in the location graph

We apply the delay optimization algorithm on the 43 cell cycle factors and the 356 cell cycle genes they bind to. A set of edge delays consistent with the expression data are chosen by the delay optimization algorithm. Figure 3 visualizes the delays of factor-factor edges in the location graph. A point represents the average expression profile of the gene cluster bound by a particular functional factor. We apply multi-dimensional scaling to project average profiles on a circle. Lines in the graph reflect the connectivity of the location subgraph of transcription factors. A solid line denotes a longer edge delay ($\geq$ 20 minutes), and a dash line denotes a shorter edge delay ($<$ 20 minutes). An enlargement of figure 3 is available in the supplementary webpage of this paper [1]. There are several interesting properties in figure 3. First, most of the factors bind to genes at G1 phase. Second, M/G1 and G1 phases are very close compared to the delays between M/G1 and G2/M genes. Edges of long estimated delays are inter-phase edges which span larger angles on the circle. Edges of short estimated delays are intra-phase edges which span smaller angles on the circle.

6.4 Cluster models

Factor analysis models under each factor are constructed on aligned expression data. We apply the algorithm in figure 1 to prune the genes which are not correlated with other

[1] http://www.ai.mit.edu/people/chyeang/circleplot2.ps

Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)
0-7695-1907-5/03 $17.00 © 2003 IEEE

Table 2. Characteristics of removed and retained genes

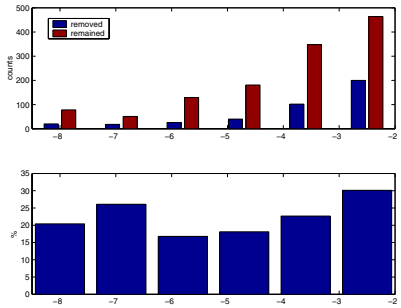
State	Phase	Periodicity	# genes	fraction
Removed	Mismatched	-	215	54.16%
Removed	Matched	Poor	116	29.22%
Removed	Matched	Good	66	16.62%
Retained	Mismatched	-	149	12.00%
Retained	Matched	Poor	696	56.00%
Retained	Matched	Good	398	32.02%

members in the cluster. The hypothesis testing threshold is set at 0.001. As a result, each transcription factor is associated with one or two distinct clusters. Each cluster contains genes which are bound by the same factor, and are correlated in their expression profiles (after being adjusted by edge delays).

Clustering results are not directly shown in the paper due to the page limitation. Instead, they are verified by several means. First, we want to know why some genes are removed from a cluster. There are several possible explanations: the expression profile of a particular gene is not periodic, the occurring phase of the removed gene mismatches the occurring phase of the cluster, and the estimated edge delays are inaccurate. Here we define an expression profile has a clean periodicity if the coefficient of its first harmonic is 1.5-fold greater than coefficients of higher harmonics. The statistics are shown in table 2. About half of the removed genes can be attributed to mismatched phases, whereas only 12% of retained genes have the same property. On the other hand, there are high fractions of genes in both retained and removed sets which do not demonstrate clean periodicity. This suggests the periodicity in terms of Fourier coefficients is not a good indicator for gene removal.

Second, we want to know the relation between expression correlation and binding strength among removed and retained edges. Ideally if physical evidence (protein-DNA bindings) matches functional evidence (co-expression), then the genes of strong binding signals (low p-values) tend to be strongly correlated with other "core" members in the cluster. Figure 4 shows the distributions of location p-values among removed and retained edges. The top figure shows the histograms of location p-values of removed and retained edges, and the bottom figure shows the fractions of removed edges within specified location p-value ranges. In this plot, we observe an increasing trend in the proportion of removed edges as location p-value increases, in spite this relation is not monotonic. The results suggest a link between transcription factor binding affinity and expression dependencies.

Figure 4. Log location p-value histograms of removed and retained genes



7 Conclusion

In this paper, we present a principled framework of integrating location and time series expression data. We propose an alternating optimization algorithm to compute the delays of location graph edges, and a clustering algorithm to identify dependent genes bound by the same transcription factor. These algorithms are applied to cell cycle expression data and location analysis data of 113 transcription factors. By analyzing the results of delay optimization and clustering, we identify the following properties. First, several transcription factors bind to genes occurring at distinct phases, which suggest they play different functional roles in gene regulation. Second, factor-factor edge delays are consistent with the cell cycle phases of their regulated genes across three expression datasets. Third, delay optimization outcomes are robust against location p-value thresholds. Fourth, clustering results are supported by the binding strengths of location analysis and previous studies about protein-DNA interactions.

References

[1] A. Arkin et al. Stochastic kinetic analysis of developmental pathway bifurcation in phase-lambda infected *escherichia coli* cells. *Genetics*, 149(4):1633–1648, 1998.

[2] Z. Bar-Joseph et al. A new approach to analyzing gene expression time series data. In *RECOMB Proceedings*, pages 39–48, April 2002.

[3] D. J. Bartholomew. *Latent variable models and factor analysis*. Oxford University Press, New York, 1987.

[4] N. Friedman et al. Using bayesian networks to analyze expression data. In *RECOMB Proceedings*, April 2000.

Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)
0-7695-1907-5/03 $17.00 © 2003 IEEE

[5] A. J. Hartemink et al. Combining location and expression data for principled discovery of genetic regulatory network models. In *PSB Proceedings*, January 2002.

[6] N. Holter et al. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *PNAS*, 97:8409–8414, 2000.

[7] T. R. Hughes et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[8] T. Lee et al. A transcriptional regulatory network map for saccharomyces cerevisiae. *Science, in press*, 2002.

[9] B. Ren et al. Genome-wide location and function of dna binding proteins. *Science*, 290:2306–2309, 2000.

[10] E. Segal et al. Rich probabilistic models for gene expression. In *ISMB Proceedings*, August 2001.

[11] I. Simon et al. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697–708, 2001.

[12] P. T. Spellman et al. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell*, 9:3273–3297, 1998.

[13] S. Tavazoie et al. Systematic determination of genetic network architecture. *Nature genetics*, 22:281–285, 1999.

A The second order approximation of the log-likelihood function

We will show here that the sum of square covariances is a second order approximation to the log-likelihood function of a complete Gaussian-Markov model. The log-likelihood function of the complete model, evaluated at the maximum likelihood setting of the parameters, is given by

$$\hat{L} = m\Big(-\frac{1}{2}p\log(2\pi) - \frac{1}{2}\log|\hat{\Sigma}| - \frac{1}{2}p\Big), \qquad (12)$$

where m is the sample size and p is the number of variables. We assume for simplicity that the diagonal elements of the sample covariance matrix $S = \hat{\Sigma}$ are one due to specific normalization of the data.

We want to derive a second order Taylor expansion of this log-likelihood with respect to the off-diagonal elements. To this end,

$$\frac{\partial}{\partial\Sigma_{ij}}\left(-\tfrac{1}{2}\log|\Sigma|\right) = \tfrac{1}{2}(\Sigma_{ij}+\Sigma_{ji}) \qquad (13)$$

$$\frac{\partial^2}{\partial\Sigma_{ij}\partial\Sigma_{i'j'}}\left(-\tfrac{1}{2}\log|\Sigma|\right) = \delta_{i,i'}\delta_{j,j'} \qquad (14)$$

where $i \neq j$, $i' \neq j'$, and we have used the fact that the covariance matrix is symmetric. Since the diagonal components are fixed at one, they don't need to be considered in the expansion.

Now, the second order expansion around $\Sigma = I$ has only second order terms and thus

$$\hat{L} \approx \sum_{i\neq j}\hat{\Sigma}_{ij}^2 + \text{constant}$$

where the constant terms do not involve the covariance.

B Time preferential binding p-values

We develop a p-value for measuring over-representation of delays among a group of cell cycle genes. First we align all the (800) cell cycle genes using a heuristic algorithm: find the gene which is the most similar to all other genes (after pairwise alignments) and choose it as the centroid for alignment; other genes are aligned against the centroid according to pairwise delays. Each gene is subsequently labeled by its alignment delay with respect to the centroid gene.

Suppose there are n_i genes of time label i and the entire population is $n = \sum_{i=1}^{T_n} n_i$. Assume a factor f binds to k genes, whose time label composition is $k_1, \cdots, k_{T_n}, \sum_{i=1}^{T_n} k_i = k$. We want to compute the probability that the factor binds predominantly to one time point.

Define $p_i \equiv \frac{n_i}{n}$ and $\hat{q}_i \equiv \frac{k_i}{k}$. $(\hat{q}_i - p_i)$ denotes the deviation of the observed label fraction from the population label fraction of time label i. We are interested in the time label whose observed fraction deviates the most from the population fraction. Denote the dominant label $\hat{i} = \arg\max_i(\hat{q}_i - p_i)$, and the test statistic $\hat{T} = \max_i(\hat{q}_i - p_i)$. The p-value is the probability that the test statistic $T = \max_i(q_i - p_i)$ is greater than or equal to $\hat{T}$, where q_i is the label fraction in a random cluster.

Multiplying both sides by k, the event $T \geq \hat{T}$ can be expressed as

$$\max_i(l_i - kp_i) \geq k\hat{T}. \qquad (15)$$

where l_i is the number of genes with label i in a random cluster. By applying the exclusion-inclusion principle and the hyper-geometric distribution, the p-value is approximated as

$$Pr(T \geq \hat{T}) \approx \sum_i \sum_{l=k_{\hat{i}}}^{k} \frac{\binom{n_i}{l}\binom{n-n_i}{k-l}}{\binom{n}{k}} - \sum_{i,j}\sum_{l_1=k_{\hat{i}}}^{k}\sum_{l_2=k_{\hat{j}}}^{k-l_1} \frac{\binom{n_i}{l_1}\binom{n_j}{l_2}\binom{n-n_i-n_j}{k-l_1-l_2}}{\binom{n}{k}}, \qquad (16)$$

where $k_{\hat{i}} \equiv k(p_i + \hat{T})$. The p-value of preferential bindings to two labels in a cluster can be derived analagously.

Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)
0-7695-1907-5/03 $17.00 © 2003 IEEE