

# Prediction of Transcription Factor Gene Expression in *Saccharomyces cerevisiae*

Daniel F. Simola  
Laboratory of Junhyong Kim  
University of Pennsylvania  
simola@mail.med.upenn.edu

September, 2004

## Introduction

Towards the lofty goal of understanding the complex behavior of a cell, it has become common practice to divide a cell's functionality into components, or layers, such as the metabolic layer, the genomic layer, and the proteomic layer. This reductionistic, divide-and-conquer approach facilitates learning about cells by restricting complexity and the scope of a given question, with the realization that it is not likely to attain a full understanding by investigating one layer alone. In this regard it is worthwhile to determine the kind and quality of predictive models that can be developed using only the genomic layer of a cell. For example, is it possible to predict gene expression levels using information from the genomic layer, such as gene-gene interactions projected from protein-gene interactions and mRNA expression measurements?

Specifically, the present goal is to predict the gene expression levels of known transcription factors in *Saccharomyces cerevisiae*. Towards this end, regression is used to evaluate linear mathematical models based on the yeast transcriptional regulatory network and expression data from the cell cycle. Each model's explanatory variables are taken from a basis subset of transcription factor genes, whose proteins were found to regulate every downstream transcription factor gene. Research is ongoing, but current results and methods are detailed.

## Data Considerations

A plethora of microarray data is available, with gene expression levels of thousands of genes in numerous conditions for several organisms now measured. In addition researchers are working to determine the complete set of regulatory interactions manifest in various genomic regulatory networks. Both types of data are important to understand transcriptional regulation, but the methods used to obtain such measurements are not without problems. Microarray measurements are inherently noisy and few instances of high resolution data collections exist, especially for time-series measurements. It is also difficult to identify the set of all transcription factors that bind to a given gene's promoter region. Some proteins bind weakly to DNA, and most bind during different cellular conditions. Despite these limitations it is feasible to accurately predict the transcript concentrations of sampled genes.

Perhaps the most complete and accurate expression data and genome wide transcription factor (TF) binding data are the Spellman yeast cell cycle time-series expression data [1] and the Lee/Young genome-wide location data [2], for which there is expression data for 6179 genes, and binding data for 2403 genes.

It has been proposed that the yeast network of transcription factor regulatory interactions generates the oscillatory dynamics of its cell cycle [3, 4, 5], a hypothesis we may be able to address using a cell cycle data set.

To investigate the ability of transcription factors to operate and wholly control the cell cycle, focus fell on the interaction network formed by transcription factors operating during the cell cycle. As a result, any and all downstream or non-TF-related effects are masked from analysis. One likely conclusion is that cell cycle operation depends on more than just the set of transcription factor genes (i.e. control extends beyond the genomic layer). Perhaps a group of non-TF-related proteins is responsible for its cyclic nature.

In light of this, we wish to investigate the possibility that the cell cycle is (largely) causally dependent on a small subset of primary or “controlling” transcription factors, whose interactions with other transcription factors generate oscillations in gene expression. At the time of writing, 106 yeast proteins have been experimentally verified as DNA-binding transcription factors, so the initial data set is composed of the 106 TF-encoding genes, their binding interactions with each other, and expression data for these genes measured during the yeast cell cycle. Thus an attempt is made to trace the hierarchical regulation of these 106 transcription factors to the expression patterns of a basis subset of controlling transcription factors. Each member of this basis is presumably independently regulated, and hypothetically serves as a key input to the cell cycle process.

## Procedure Overview

It is possible to formulate the notion of cis-acting transcriptional regulation into a dependency statement, or equation, which relates the expression of a particular target gene  $g$  to a weighted sum of the expression levels of the genes which encode proteins that bind and regulate  $g$  (see Methods). When considering the set of equations describing the regulatory dependencies of all 106 transcription factor genes, two primary questions arise. Topologically, is there a basis subset of transcription factors which can be said to control the response of the system of equations at hand? Is it possible to derive accurate predictions for each gene using the interaction network? The latter question realistically reduces to a more statistical question: what are the weight values assigned to each TF (i.e. the coefficient matrix) that allow a gene’s expression to be predicted?

To answer the latter question, it is useful to consider the following view of how regulatory networks might be constructed over time. Perhaps for a given cell, existing transcription factors are expressed according to certain patterns, and to incorporate into the regulatory network an additional TF generating a novel expression pattern, a subset of the existing TFs was selected to regulate the new TF, effecting the novel pattern. This view implicitly reflects concepts from Fourier analysis, where any (periodic) wave function may be decomposed to arbitrary resolution using a weighted linear combination of sine and cosine waves. Following this interpretation, one expects that each downstream target gene might have a generally periodic expression profile, possibly phase-delayed by several minutes (time for the Central Dogma to run its course) and amplitude-shifted, with respect to the expression profiles of its regulators. In fact recent work suggests that, at least during the yeast cell cycle, expression levels of most genes do oscillate temporally [3, 4, 5]. These encouraging conclusions suggest that it is possible to accurately model gene expression regulation.

One important assumption used to model this idea is that the expression level of a gene is wholly dependent on which transcription factors bind to its promoter. Naturally this assumption is too stringent, as there are obvious structural and kinetic constraints that contribute to transcriptional activation. There are also two important considerations. The data used to construct predictive models should be accurate,

lest the results become meaningless. This constraint is somewhat flexible with respect to microarray data, but current models rely completely on a correct and complete interaction network. (Future work may relax this dependency). Another consideration is the functional form used to describe transcriptional regulation. It is believed that gene expression is most often controlled in a nonlinear fashion. In many cases, however, quite reasonable linear approximations may be achieved. In this procedure, a log-linear equation form is used, using log-transformed expression data.

## Previous Work

Previous work [6] has been done using the same biological system and data set, as well as in *E. coli* [7], where the goal was to decompose a matrix of gene expression levels from varying conditions into a weighted connectivity matrix and a protein activation level matrix. In this way the authors tried to derive both the weights assigned to transcription factor regulation events, as well as the activation levels of the proteins, given an initial binary-valued network and expression data. This approach, called Network Component Analysis, assumes protein activation levels may be different from transcript levels (which is true in many cases). In addition weights for a given TF are assigned per equation.

Our goal is to develop an approach which predicts the expression level of a set of genes in a network, based upon the expression levels of a small subset of basis genes for the network. This approach emphasizes the topology and dependency relations of the network. This approach, unlike NCA, estimates a set of coefficients for each regulatory gene per equation, where the size of the set is determined by the number of genes separating a basis gene from the given target gene. In other words, the number of coefficients weighting a basis gene depends on the depth of the path connecting the basis gene and the target gene. The NCA approach estimates one coefficient per regulatory gene in a given equation. That said, our approach is not yet complete, and currently only one coefficient value is estimated per basis gene. The approach is nevertheless unique, because of its emphasis on topological dependencies. The coefficients used are weighting basis genes that often indirectly regulate target genes, instead of weighting regulators that directly activate a target gene.

## Methods

### Choice of Data Set

To attempt to control the variability in transcription factor regulation, we focus on gene expression changes within a single condition over time. Within a single condition it is reasonable to assume that the connectivity matrix will remain constant. It is important to work with a large amount of data, because each time point supplies an additional degree of freedom for parameter estimation, which is used to increase the possible predictive resolution. The number of data samples must exceed the number of parameters to estimate for a given equation, lest the estimation problem become over-determined (estimation is trivial with more DOFs than data points). In addition larger data sets allow for more robust estimates. In particular temporal data is used to test our ability to predict changes in expression levels over a time interval.

It is necessary to use connectivity data collected for the same set of genes for which expression data exists, although no current experiments have been performed in which large-scale expression measurements and genome-wide binding assays are taken together in condition-synchronized cells. Thus we assume binding data corresponds to the gene regulatory activities of yeast progressing normally through the cell cycle.

Data sets based on both yeast and *E. coli* were promising, but the yeast cell cycle microarray data set from Spellman [1] and genome-wide location binding data from Lee [2] were chosen. The microarray data features 24 time points of 6179 genes, and Young’s data has connectivity information for 106 transcription factors. To summarize, the initial data set includes expression measurements for 106 genes over 24 time points (106x24 log-valued matrix), and DNA-binding measurements for all 106 genes against themselves (106x106 binary-valued matrix).

## Expression Data Transformation

The Spellman expression data set originally had been log-normalized and Fourier transformed [1]. As measurements using microarrays are inherently noisy, important patterns in the data, specifically oscillatory patterns of transcript levels, may remain hidden unless certain transformations are applied. In order to reveal any patent oscillatory signals in the data [3], a kernel smoothing function was applied over each gene’s temporal expression data. A discrete binomial approximation to the normal distribution was used with a window size of 5, so that the transformed expression value at a single time point is a sum of at most 5 values from surrounding time points, weighted by a binomial distribution.

## Model Form

Realistically, the factors composing the promoter complex form dependent non-linear interactions in such a way that the removal of individual factors may result in drastic changes in the target gene’s transcript level. Thus a representative model used to predict the expression level of gene  $y_i$  is the following:

$$y_i = \prod_{j: x_j \rightsquigarrow y_i, x_j \in B} x_j(t)^{c_j}, \tag{1}$$

where  $x_j(t)$  is the raw expression level of gene  $j$  at time  $t$ ,  $c_j$  is the coefficient that weights the influence  $x_j(t)$  has on  $y_i$ , and  $B$  is the basis set for  $y_i$ . This model becomes linear in log space, where the log is applied to all  $x_j(t)$ :

$$y_i = \sum_{j: x_j \rightsquigarrow y_i, x_j \in B} c_j \log(x_j(t)), \tag{2}$$

or as a system of linear equations:

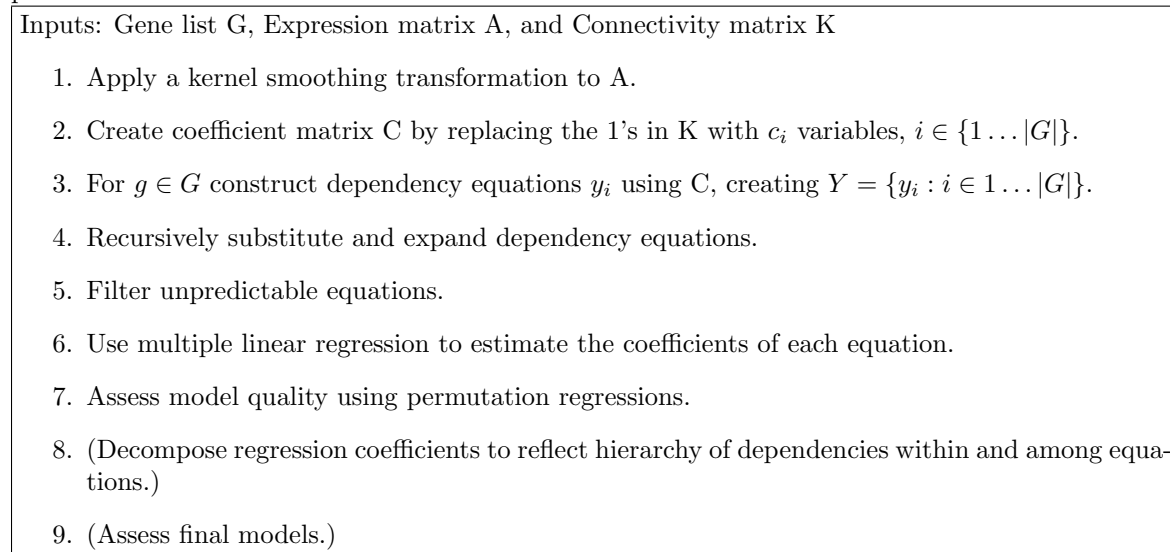
$$Y = C \log(X), \tag{3}$$

if one considers  $y_i$  to be dependent on every  $x$ , with non-regulating genes having  $c_j = 0$ . Each gene  $y_i$  is equivalent to  $x_j$ , when  $i = j$ . A different symbol is used simply to distinguish regulators from target in each equation.

This formulation was chosen because it offers a simple representation of transcriptional gene regulation, where the concentration of each regulator mRNA positively or negatively influences the concentration of the target gene’s transcript level, independent of other regulators. Note that in this model, mRNA transcript levels are assumed to be identical to the corresponding transcripts’ protein levels. This assumption is not always reasonable, as many genes are regulated post-transcriptionally.

The actual equation for each gene is determined by the connectivity matrix  $K$ , where  $K_{i,j} = 1$  indicates that transcription factor  $j$  binds and regulates the expression of transcription factor  $i$ , and  $K_{i,j} = 0$  indicates no regulation. Thus the number of regulators of a given gene  $i$  is given by the row sum of the connectivity

Figure 1: Flowchart of expression prediction procedure. Steps 4 and 5 are analogous to performing a row reduction on the coefficient matrix of the system of equations. Items within parentheses have not yet been performed.



matrix  $\sum_i K_{i,:}$ .  $K$  was built from the Lee binding data matrix  $P$ , which contains p-values for each binding experiment.  $K_{ij} = 1$  if  $P_{ij} \leq \text{cutoff}$ , where the cutoff is 0.001, used in the author's original analysis [2].

## Equation Expansion and Simplification

The most important aspect of the set of equations is that it should reflect the regulation of genes, not in terms of their direct cis-acting TFs, rather in terms of the set of TFs most upstream to the entire network. In order to transform the initial set of dependency equations  $Y$ , equation substitution and simplification are performed recursively, until a basis variable is inserted. That is, for each equation  $y_i = c_1x_1 + c_2x_2 + \dots + c_kx_k$ , each gene variable  $x_j$  is replaced by gene  $j$ 's dependency equation. By recursively substituting variables for their dependency equations (essentially replacing a variable by its definition), each  $y_i$  is rewritten in terms of some basis set of transcription factors whose expression levels alone influence  $y_i$ . Each basis TF is weighted by the coefficients of each of its downstream genes. e.g., the dependency equation of *sko1* is written as follows:

$$y_{sko1} = c_{gts1}c_{dal82}x_{dal82} + c_{gts1}c_{reb1}c_{cbf1}x_{cbf1} \quad (4)$$

In this example, the expression levels of two upstream genes (*dal82* and *cbf1*) are needed to describe the expression of *sko1*. The amount of influence these genes have is modulated by the nested set of coefficients reflecting the hierarchy in the interaction network (see Figure 2). The union of per-gene basis sets constitutes the basis set of transcription factors for the entire yeast TF interaction network.

## Regression and Coefficient Estimation

To estimate the values of the coefficients for each equation, multiple linear regression is used with the log-linear model described in Equation 3. As a preliminary step, we attempted to estimate only one coefficient for each basis gene in an equation. Specifically, in the *sko1* example from Equation 4, three coefficients were needed to weight the effect of *cbf1*. For this regression model, an approximation is made of the form

$C_{cbf1} \approx c_{gts1}c_{reb1}c_{cbf1}$ . Thus initially the coefficients for each basis variable were combined into a single parameter in the regression model. Note that regression estimates each model’s parameters independently, supporting the idea that TFs may have a different influence on each gene.

## Assessment of Regression

To assess the significance of the regression models, a procedure of random sampling was used in which, given a model in the form of Equation 2,  $k$  TFs are randomly chosen from the total set of  $n$  and replace the existing  $x$  variables. Then multiple regression is performed on this permuted model. Significance of true and random models was assessed using the  $R^2$  metric (square of correlation coefficient). The  $R^2$  metric was chosen because it has a monotonic distribution over successively better fits in the closed range of  $[0,1]$ . This makes it suitable for comparing different models. Since it is often interpreted as the amount of variation explained by the model, one expects that the  $R^2$  value computed from the true model should be greater than that of the random model, if the given network topology is correct.

For each equation describing the expression of gene  $y_i$ , let  $k$  be the number of basis genes used to describe the regulation of  $y_i$  ( $k = \text{indegree}(y_i)$ ). Then there are  $\binom{n}{k}$  ways to randomly create a model. Since enumerating all  $\binom{n}{k}$  possibilities becomes intractable for certain values of  $n$  and  $k$ , a fixed number of samples,  $s$  is chosen, so that  $s$  random permutations are generated for each equation, and thus  $s$   $R^2$  values sampled per equation. One expects that the true model has a  $R^2$  value in a top percentile of the distribution of such values, indicating that the model represents one of the best models for the expression of gene  $y_i$ . If  $s$  is large, it is possible that some random models will share many of the true variables, and so the  $R^2$  values of some random models may be as high as the true model.

There is one caveat to the random gene selection procedure described above. For each sample, if the selected gene is identical to the model’s target gene (if  $x$  is chosen where  $x = y_i$ ), that gene is disregarded and another is chosen. If this were allowed, the regression procedure would consistently estimate the parameter for the  $x$  variable as 1, and assign 0 values to all other variables, guaranteeing  $R^2 = 1$  (see Results for further discussion).

## Results

### Self-consistent Subgraphs and Autoregulation

Within a graph, different local mappings between subsets of nodes can be classified into categories of motifs, commonly called network motifs. Cyclic motifs comprise an interesting category, as nodes in the motif may receive feedback via autoregulation, which modulates the behavior of the nodes within the motif. Such a motif may be simple, having one node which regulates itself, or complex, incorporating all nodes of a graph into some form of cycle where every node is regulated by a combination of other nodes, forming a closed, self-consistent loop. Such a motif may be called a self-consistent motif (SCM), whose feedback mechanism may provide a way to sustain oscillatory flows through its nodes [4, 3].

A self-consistent motif  $M$  is defined as follows:  $M = \{u \rightarrow v : v \rightsquigarrow u, \forall u, v \in M\}$ . That is, an SCM is a set of connected nodes  $u, v$  such that if there exists an edge from  $u$  to  $v$ , there also exists a path from  $v$  to  $u$  in the motif. This definition assumes that SCMs should be strongly connected, that every node in a motif should be directly or indirectly regulated by every other node. Under this definition, no self-consistent motifs larger than one node (simple autoregulatory SCM) exist in the yeast TF interaction network used

Table 1: Three classes of nodes in the yeast transcription factor regulatory network. Expression models are formulated using members of the controller class. The prediction of controllers is trivial, since there are no known regulators, and thus no parameters to estimate. The intermediate class includes autoregulatory nodes, because these nodes are known to be regulated.

<ol style="list-style-type: none"> <li>1. Controllers: ADR1, AZF1, CAD1, CBF1, CIN5, DAL82, DIG1, FKH1, FZF1, GCR2, GLN3, HAL9, HAP3, INO2, INO4, LEU3, MAC1, MBP1, MCM1, PHD1, PHO4, RAP1, RME1, RTG3, SKN7, STB1, SWI6</li> <li>2. Intermediates: ABF1, ACE2, ARG80, ARG81, ARO80, BAS1, CHA4, CUP9, DOT6, FHL1, FKH2, GAL4, GAT1, GCN4, GCR1, GTS1, HAA1, HAP2, HAP4, HIR2, HMS1, HSF1, IME4, IXR1, MAL13, MAL33, MET31, MIG1, MSN2, MSS11, NDD1, NRG1, PDR1, PHO2, RCS1, REB1, RFX1, RGT1, RLM1, ROX1, RPH1, RTG1, SFP1, SKO1, SMP1, STE12, STP2, SUM1, SWI4, SWI5, THI2, YAP1, YAP3, YAP5, YAP6, YAP7, YFL044C, YJL206C, ZAP1</li> <li>3. Effectors: A1, ASH1, CRZ1, DAL81, GAT3, HAP5, HIR1, MET4, MOT3, MSN1, MSN4, PUT3, RGM1, RIM101, SFL1, SIP4, SOK2, STP1, UGA3, ZMS1</li> </ol>
--

(see Figure 2). Prediction of autoregulatory nodes is trivial as the only parameter to estimate necessarily takes the value 1.

Instead of finding a self-consistent, cyclic network structure, a predominantly hierarchical, and unidirectional, topology exists, composed of three classes of nodes: a controller class that have no regulators (indegree 0, positive outdegree), an intermediate class which receives connections from the controller class but also from the intermediate class itself (positive indegree and outdegree), and a terminal, or effector class, whose nodes are regulated by either of the above classes, and elicit a response not known to regulate any other TFs (positive indegree, outdegree 0) (see Table 1). Thus no instances of SCMs having edges that span more than one class were found.

One may define a self-consistent motif in other ways. In some cases it may be too strict to require that a SCM be strongly connected. Instead it may be reasonable to require that each node have positive outdegree, with edges to nodes in the motif. That is,  $M' = \{u \rightarrow v, v \rightarrow u', u, v, u' \in M'\}$ . Under this definition a node may be included in a SCM even if it has indegree 0. This is useful both because connectivity information is most likely incomplete (there are missing edges), and because it allows for the formulation of a network’s dynamics in terms of the controlling basis of nodes. However this definition is more suitable when investigating network dynamics [8], rather than expression prediction, because it effectively eliminates the effector class.

One may ask how are the controller nodes regulated. An obvious possibility is that the cell cycle TF network truly is a large-scale SCM, a fact masked due to an incomplete data set. Missing edges may connect controller nodes, establishing cyclic regulation. Additional connections might arise from nodes not included in the verified set of 106 TFs. Estimates suggest that more than 450 yeast genes act as transcription factors, 140 of which have been experimentally verified.

Controller regulation may also result from mechanisms not explicitly involved in transcriptional regulation, such as those which affect the structural dynamics of a cell’s DNA. Such regulation may be a causal side-effect of the conformational changes of chromatin, a phenomenon which itself is caused by transcriptional regulation, but which is not represented in our model. Genes in open chromatin are accessible to polymerase, and as such, basal activation of controller TFs may be sufficient to “breathe life” into the TF network, successively resulting in production of proteins encoded by terminal genes which promote chromatin condensation, thereby damping the overall expression levels of transcription factors.

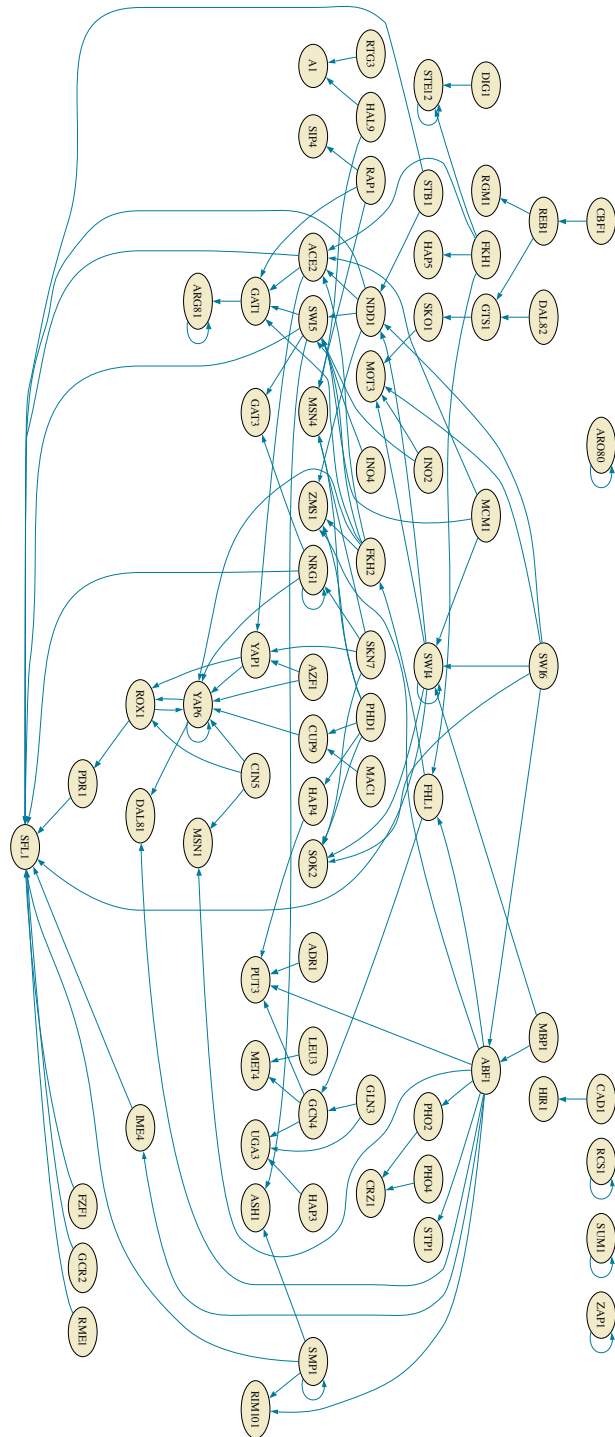


Figure 2: Yeast Transcription Factor Network

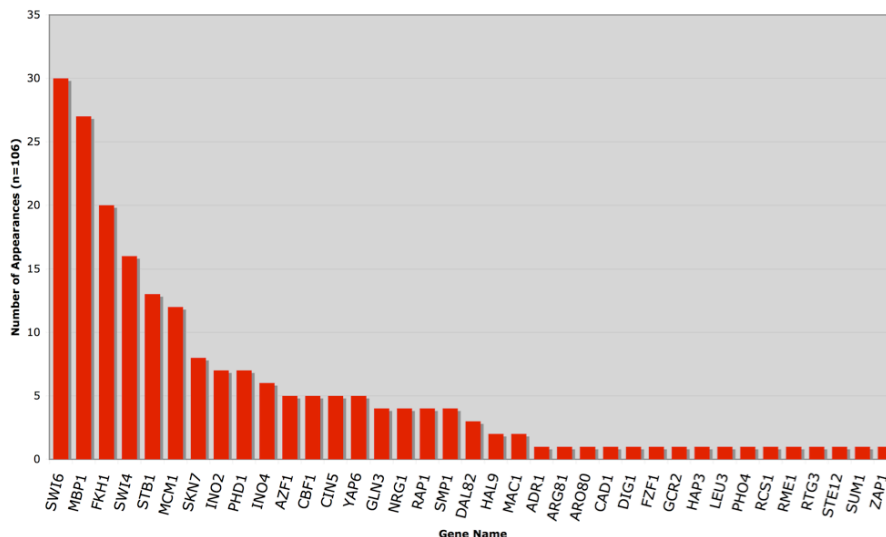


Figure 3: Counts of Basis Genes. The x-axis lists the names of 37 genes, and the y-axis represents the number of equations in which a particular gene is used as a transcription factor. The remaining 69 genes are not used as basis variables in any predictive models.

## Model Statistics

106 predictive models were derived, one model per target gene. It is not possible, however, to perform regression on every model. For the 27 controller genes, regression is not possible because there is currently no connectivity information (see Figure 1).

31 models are unpredictable because they form no known connections with any nodes (indegree and outdegree are 0). There are no cases where a model is unpredictable because the indegree of the target gene exceeds the available number of data points/DOF (24). Thus 48 out of 106 (45%) of the models are predictable, and the corresponding models are shown in Figure 2. 5 of these 48 models (*smp1*, *zap1*, *rsc1*, *sum1*, *aro80*), however, are trivially predictable, because the only known input is the gene itself; thus regression estimates a parameter value of 1 for these equations. Such simple autoregulatory genes exist independently of the hierarchical structure of the TF network.

Figure 3 illustrates the number of times each transcription factor is used as a basis variable. 37 genes appear in at least one model; these 37 comprise the basis set of the network. 6 of these genes (*swi6*, *mbp1*, *fkh1*, *swi4*, *stb1*, *mcm1*) each appear in more than 10 equations. It may not be an extreme generalization to posit that the yeast cell cycle may be more or less completely controlled by adjusting the expression levels of these 6 factors.

The basis set is not the same as the controller set. Notice that so-called internal autoregulatory genes are currently allowed to exist in the basis set, such as *swi4*. This brings into question the use of the term basis, since the expression of *swi4* is predicted by other genes in the basis set. However since SWI4 protein regulates the *swi4* gene, it may be useful to use this value in predicting downstream expression levels. As it turns out this situation becomes problematic when regression methods are used (see Regression Estimates).

Table 2: Five Example Predictive Models. The  $y_i$  variable represents the index of the gene whose expression we are trying to predict.  $x_j$ 's are the basis variables, and the  $c_k$ 's are the scalar coefficients of all TFs which weight each basis variable. The form of each model is derived from Equation 2. Regression estimates are shown in Figure 3.

1. $c_{34}x_{34} + c_{80}x_{80} + y_1$
2. $c_{51}x_{51} + c_{95}x_{95} + y_2$
3. $c_{22}x_{22} + c_{21}c_{22}c_{23}x_{22} + c_2c_{21}c_{23}c_{51}x_{51} + c_{52}x_{52} + c_{61}c_{88}x_{88} + c_{61}c_{93}x_{93} + c_2c_{21}c_{23}c_{95}x_{95} + c_{61}c_{95}x_{95} + y_3$
4. $c_6x_6 + c_3c_{22}c_{26}x_{22} + c_{21}c_{22}c_{23}c_{26}x_{22} + c_3c_{21}c_{22}c_{23}c_{26}x_{22} + c_{21}c_{22}c_{23}c_{26}c_{94}x_{22} + c_{26}c_{44}c_{94}x_{44} + c_{26}c_{45}c_{94}x_{45} + c_2c_{21}c_{23}c_{26}c_{51}x_{51} + c_2c_3c_{21}c_{23}c_{26}c_{51}x_{51} + c_2c_{21}c_{23}c_{26}c_{51}c_{94}x_{51} + c_3c_{26}c_{52}x_{52} + c_{26}c_{52}c_{94}x_{52} + c_{26}c_{68}x_{68} + c_3c_{26}c_{61}c_{88}x_{88} + c_{26}c_{61}c_{88}c_{94}x_{88} + c_3c_{26}c_{61}c_{93}x_{93} + c_{26}c_{61}c_{93}c_{94}x_{93} + c_2c_{21}c_{23}c_{26}c_{95}x_{95} + c_2c_3c_{21}c_{23}c_{26}c_{95}x_{95} + c_3c_{26}c_{61}c_{95}x_{95} + c_2c_{21}c_{23}c_{26}c_{94}c_{95}x_{95} + c_{26}c_{61}c_{94}c_{95}x_{95} + y_6$
5. $c_7x_7 + y_7$

## Regression Estimates

Linear regression was performed on each predictable model, and the  $R^2$  metric was used to assess the quality of the model's fit to the data (see Methods). Figure 4 illustrates the distribution of  $R^2$  scores for the 48 predictable models. Ideally one would observe a histogram resembling a steep exponential curve whose largest values lay near 1, indicating the models fit the data very well. Instead there are no extreme peaks, an observation exacerbated by the small sample size. Just under half of the models have an  $R^2$  value greater than .8, but this indicates that the majority of the models fit the data poorly. To assess the significance of the models and thus the TF network, a set of 500 randomly connected models was built for each true model (see Methods). The distribution of values for all randomly sampled models is shown in Figure 5. As expected, the vast majority of random models fit the data extremely poorly, with  $R^2$  values below 0.1. However a several models (2314/24000) were able to generate scores above 0.95.

It is reasonable to think there is some random model of similar functional form to a corresponding true model, except having a small number of different regulatory variables. Such variables may be yet unknown in reality, but very important factors. In many cases however, the two model types do not appear to share the majority of their variables (data not shown). Nevertheless, when these well-fit random models were compared to their true analogues, it was found that most of the true models also had a good fit, with a similar though often slightly smaller  $R^2$  value. An invariant here is that both model types have the same number of variables, suggesting a relationship between number of variables describing a model, and the quality of fit. Figure 6 does indicate a strong relationship between the number of variables used in a model and the  $R^2$  value for the model. Although there is relatively high variance, the regression estimate indicates a strong positive correlation, suggesting that the number of variables may be a good predictor of model fit.

This result suggests a lack of structure or importance in a genomic regulatory network, and most biological evidence would suggest otherwise. One explanation is that the variables that contribute to a well-fit random model lie just downstream of the true basis variables. Many genes are expressed at similar levels during the cell cycle, and such expression analogues may substitute well or better than their true bases. Future work may incorporate a phase-delayed model, where the expression of gene  $y_i(t)$  at time  $t$  depends on the expression levels of its regulators from time  $t - 1$ ; this reflects the natural delay between a transcription

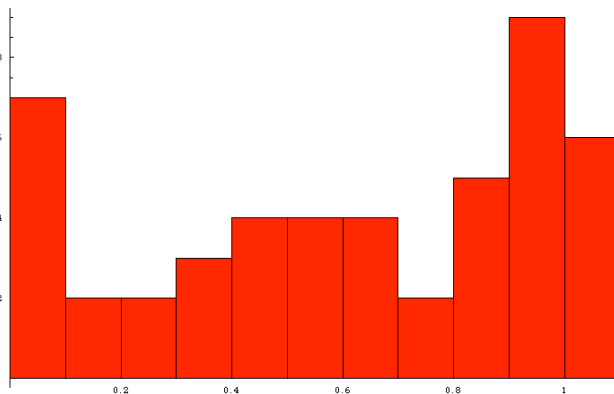


Figure 4: Distribution of  $R^2$  Scores for the 48 Predictable Gene Dependency Models.

event and resulting protein cis-regulation. It may also be the case that some well fitting random models do significantly overlap with corresponding true models; only a small sample of the random models (of 2314) has been analyzed thus far.

In general it would be useful infer the features of models that have very good and very bad fits. To understand the cause of poor fits, all true models with  $R^2 \leq 0.25$  were grouped. Each model was checked to determine if there were any variables which narrowly missed the p-value cutoff of 0.001 [2]. Most models had at least one transcription factor that narrowly missed the cutoff, suggesting that a cutoff of 0.001 establishes a conservatively-connected network. Also all of the poor fitting models have a small number of basis variables (1, 2, or 3), which is consistent with Figure 6. It is then reasonable to expect that with better topology, at least one additional variable would be added to each of these models, improving the quality of fit on average.

To understand the features of very fit models, those which have  $R^2 \geq 0.75$  were grouped. Two subgroups exist, one whose models have an autoregulatory component, and another whose do not. As described above, models with autoregulatory components fit the data perfectly because regression fully weights the autoregulatory variable. Such models are uninformative, and in the future estimates will be determined for these genes having removed the autoregulatory nodes.

The second group of models had scores strictly less than 1. The smallest model has 4 explanatory variables, and the largest has 17, with the majority having fewer than 10. Coefficients for these models were distributed from  $[-1.5, 2]$ , with the majority coming from  $[-0.5, 0.5]$ , indicating that each variable plays a influential role in target gene prediction 7. Perhaps these models fit well because their expression is predicted using several variables, which comprise a complete and accurate set for each gene.

Since Lee’s genome wide-location data was determined using p-values and a cutoff, it will be useful to investigate the effects of adjusting this cutoff, and thus the topology of the network, on model quality. This idea has already been used to look for an ”optimally” connected network [8]. Specifically it would be interesting to see how currently good models are changed (by adding more variables) and whether these new models result in better regression estimates.

## Discussion

This work attempted to provide a method for predicting the gene expression levels of the set of currently verified transcription factors active during the yeast cell cycle. Specifically, a set of predictive models were

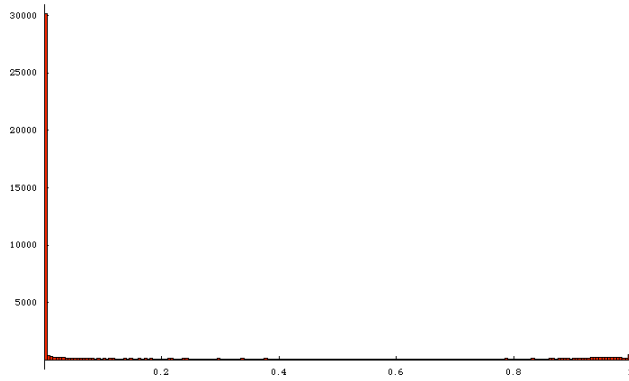


Figure 5: Distribution of  $R^2$  Scores for 48 Random Models, with 500 Samples per Model.

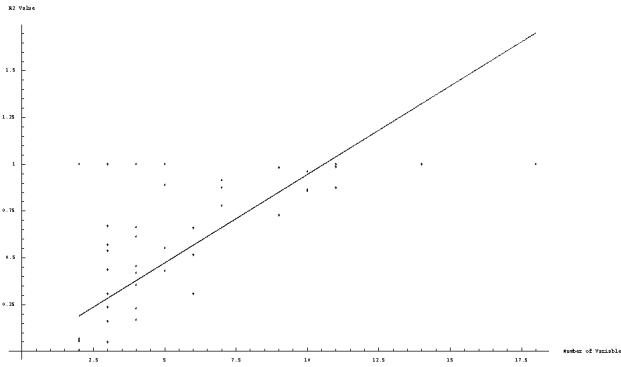


Figure 6: Relationship between Number of Variables per Model and R2 Score. Coordinates are plotted as diamonds, and the least squares estimate plotted as a line. Note the scales of the axes; the y-axis is only valid for  $[0,1]$ . The slope of the line is 0.095.

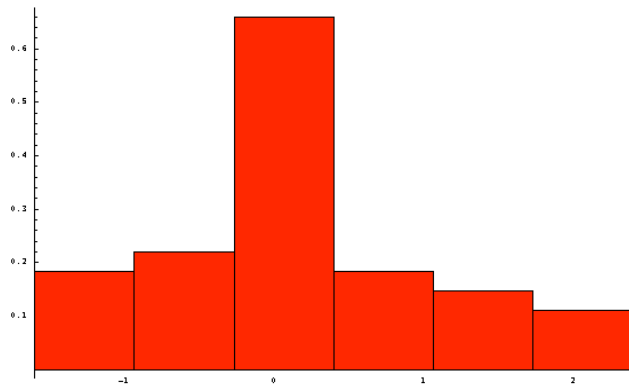


Figure 7: Overall Distribution of Coefficient Assignments for 48 True Models

Table 3: Regression estimates of the models in Figure 2. Note models 4 and 5, whose fits are problematic. One of the basis variables in model 4 is autoregulatory, to which regression assigns a coefficient of 1, and coefficient values near 0 to each other variable, producing an optimal fit. Model 5 is described by a single autoregulatory variable, which regression trivially assigns a coefficient of 1.

1.  $y_1 = -0.024x_{34} - 0.586x_{80}$
2.  $y_2 = 0.384x_{51} + 0.461x_{95}$
3.  $y_3 = 0.318x_{22} - 0.306x_{51} + 0.564x_{52} - 0.677x_{88} - 0.147x_{93} - 1.200x_{95}$
4.  $y_6 = 1.000x_6 - 5.762^{-17}x_{22} + 0.x_{44} + 3.630^{-16}x_{45} - 7.677^{-17}x_{51} + 1.683^{-16}x_{52} - 2.042^{-16}x_{68} - 1.344^{-16}x_{88} + 4.566^{-16}x_{93} + 4.302^{-16}x_{95}$
5.  $y_7 = 1.000x_7$

created, whose explanatory variables are selected from a basis set of transcription factors, which can be used to predict the expression of every downstream transcription factor gene. It is possible that, biologically, the yeast cell cycle is controlled by this basis of genes, adjustments to which alter the behavior of the cell cycle.

Due to inherent limitations in the known network topology, it is possible to predict 48 out of 106 genes. Multiple linear regression was applied to each model, and the  $R^2$  metric was used as an indicator of model significance. One of the simplifications used in model design was to combine a product of coefficients into a single estimable coefficient. Further versions of the model may attempt to relax this restriction.

These initial results revealed many subtle flaws in the model formulation. There are issues when predicting models with autoregulatory components, and possible issues with the network topology used. In addition microarray expression data is inherently noisy, and prediction quality is limited by the number of time points available (24). Future work may incorporate the idea of phase-delayed expression prediction, where the gene expression level of some target gene at a given time point is dependent on the transcription levels of its regulators from a previous time point (see Results).

Since the graph of any genomic network is probably incomplete, an interesting question which results from this work is whether these models can be extended to predict what are the missing edges in the graph. Future work may attempt to generate each predictive model from a distribution of possible basis variables, allowing a variety of topologies to be tested for a given model.

## References

- [1] 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*, vol. 9, pp. 3273–3297, 1998.
- [2] T. I. Lee *et al.*, “Transcriptional regulatory networks in *Saccharomyces cerevisiae*,” *Science*, vol. 298, no. 25, pp. 799–804, 2002.
- [3] R. Klevecz and H. Dowse, “Tuning in the transcriptome: basins of attraction in the yeast cell cycle,” *Cell Proliferation*, vol. 33, pp. 209–218, 200.
- [4] R. Klevecz, J. Bolen, G. Forrest, and D. Murray, “A genomewide oscillation in transcription gates dna replication and cell cycle,” *PNAS*, vol. 101, no. 5, pp. 1200–1205, 2003.

- [5] G. Rustici *et al.*, “Periodic gene expression program of the fission yeast cell cycle,” *Nature genetics*, vol. 36, no. 8, pp. 809–817, 2004.
- [6] J. C. Liao *et al.*, “Network component analysis: Reconstruction of regulatory signals in biological systems,” *PNAS*, vol. 100, no. 26, pp. 15522–15527, 2003.
- [7] K. C. Kao *et al.*, “Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis,” *PNAS*, vol. 101, no. 2, pp. 641–646, 2003.
- [8] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, “Random boolean network models and the yeast transcriptional network,” *PNAS*, vol. 100, no. 25, pp. 14796–14799, 2003.