Gene Regulation and Systems Biology

ORIGINAL RESEARCH

Application of Structure Equation Modeling for Inferring a Serial Transcriptional Regulation in Yeast

Sachiyo Aburatani

Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan. Corresponding author email: s.aburatani@aist.go.jp

Abstract: Revealing the gene regulatory systems among DNA and proteins in living cells is one of the central aims of systems biology. In this study, I used Structural Equation Modeling (SEM) in combination with stepwise factor analysis to infer the protein-DNA interactions for gene expression control from only gene expression profiles, in the absence of protein information. I applied my approach to infer the causalities within the well-studied serial transcriptional regulation composed of GAL-related genes in yeast. This allowed me to reveal the hierarchy of serial transcriptional regulation, including previously unclear protein-DNA interactions. The validity of the constructed model was demonstrated by comparing the results with previous reports describing the regulation of the transcription factors. Furthermore, the model revealed combinatory regulation by Gal4p and Gal80p. In this study, the target genes were divided into three types: those regulated by one factor and those controlled by a combination of two factors.

Keywords: Structural equation modeling, transcriptional regulation, gene regulatory network, expression profile

Gene Regulation and Systems Biology 2011:5 75–88

doi: 10.4137/GRSB.S7569

This article is available from http://www.la-press.com.

Introduction

Transcriptional regulation of gene expression is necessary for living cells to function, and therefore revealing the mechanisms behind this regulation is a central aim of systems biology. Such regulatory mechanisms are known to function via complex relationships among DNA, RNA and proteins, and to respond to intracellular signals and extracellular conditions to ensure proper gene expression.[1] One of the most common regulatory mechanisms involves protein-DNA interactions. The proteins that bind to a DNA sequence are known as transcription factors, and they are crucially involved in regulating their target genes. Detailed knowledge about how these transcription factors conduct this regulation is essential when inferring a gene regulatory network.

Investigations of gene regulatory systems or complex functional networks among DNA, RNA, proteins and other cellular components in a living cell conventionally follow a standard protocol. After a DNA sequence is completed, the mRNA level is measured by a cDNA microarray, to reveal the gene expression profiles under various conditions. With this information, the regulatory networks can be inferred by employing a number of approaches, including Boolean and Bayesian networks.[2,3] I previously developed an approach based on a Graphical Gaussian Model (GGM) in combination with hierarchical clustering.[4,5] Among the graphical models, GGM is the simplest in a mathematical sense, as the conditional independence between variables is estimated from the partial correlation coefficients. However, GGM infers only the undirected graph, whereas Boolean and Bayesian models infer the directed graph, which shows causality. Although all of these approaches are suitable for establishing relationships among genes, it is difficult to reveal the concrete interactions between proteins and genes, because of the insufficient information about proteins in the gene expression profiles. Thus, an alternative approach is needed when investigating transcriptional regulation that involves protein-DNA interactions.

One possible technique is Structural Equation Modeling (SEM).[6] SEM has been successfully used to elucidate causal relationships in disparate fields such as econometrics, sociology and psychology.[7–9] In addition, it has been applied to Quantify Trait Loci (QTLs) for association and linkage mapping in biology,[10,11] as well as to identify genetic networks from microarray data or SNP data.[12–14] The significant features of SEM are the inclusion of latent variables into the constructed model and the ability to infer the network, including the cycle structure. Additionally, the linear relationships between the latent variables and the observed variables are assumed to minimize the differences between the fitted covariance matrix and the calculated sample covariance matrix.

To clarify the effects of latent variables in a transcriptional regulatory model, I selected a serial transcriptional regulatory system composed of GAL-related genes. This regulatory system in *S. cerevisiae* has a hierarchical structure comprising three transcription factors, Mig1p, Gal4p, and Gal80p.[15–18] Mig1p is a multicopy inhibitor for gene expression and is the initiating factor of this regulation.[19,20] Gal4p, the gene of which is a target for Mig1p, represents the second stage.[21–23] Gal4p and its target genes then regulate the next stage in the system, including Gal80p, which represents the third and final stage. In this serial transcriptional regulatory system, the expression of one transcription factor in the former stage leads to the subsequent expression of transcription factors in the latter stage.

Previous investigations have revealed that transcription factors recognize and bind to upstream DNA motif sequences to regulate the target genes.[24,25] Some target genes have been empirically confirmed, and others were estimated by computational analyses.[26,27] Analyses based on gene expression levels and motif sequences have indicated that various gene expression regulator mechanisms involve Mig1p and Gal4p. Although many genes have been estimated as the targets of those regulators, the regulation of other target genes remained unclear. Therefore, to characterize the entire mechanism of the serial transcriptional regulation, a network model that includes information on the relevant genes and proteins is necessary.

Here, I applied SEM to reveal a serial transcriptional regulation system that is mediated by transcription factors, by using information from numerous gene expression profiles. Since microarray data do not reflect protein expression, a network model that includes not only genes but also proteins should provide new information on the genetic regulatory architecture, assuming that the transcriptional regulation by the transcription factors involves physical interactions between the factors and the corresponding DNA. In

this study, I used SEM to describe the transcription factors as latent variables and their effects. This method estimates not only the number of variables in the model, but also any significant interactions between them. The resulting gene expression profiles have clarified the complicated gene expression mechanism.

Methods

Outlier removal

The Grubb's test was used to identify and delete outlier gene expression data derived from the Gene Expression Omnibus (GEO). The Grubb's test was performed on a gene-by-gene basis for each experimental condition, to detect and delete the abnormal numerical data that are included as flags of experimental error. To quantitatively identify the outlier data, a Z value was calculated.

$$z = \frac{|mean - value|}{SD} \quad (1)$$

where SD is the standard deviation. If Z was large, then the datum was considered to be an outlier. The SD was calculated from all data including outliers. A two-tailed probability for the Student t-distribution was obtained as follows:

$$T = \sqrt{\frac{N(N-2)Z^2}{(N-1)^2 - NZ^2}} \quad (2)$$

Here, N is the number of values in the sample. The probability that a datum is an outlier can be obtained from the Student t-distribution for T and $N-2$ degrees of freedom. The approximate P value for the outlier test, which was calculated by multiplying the obtained probability by N, is represented as the probability of observing an outlier, assuming the data were sampled from a Gaussian distribution. If P was significant, then the outlier was excluded and the Grubb's test was performed on the next suspected outlier. The procedure was repeated until a P-value <0.05 was achieved.

Data selection

A multilevel hierarchical regulation system between proteins and DNA is considered to be a good model for inferring protein-DNA interactions. To detect the hierarchical regulation in yeast, I utilized the TRANSFAC Database (http://biobase-international.com/index.php?id=transfac). TRNSFAC is one of the useful databases that include information about transcriptional regulation.

Among all registered yeast transcription factors in the TRANSFAC database, 108 genes were detected as encoding transcription factors, which regulate 252 other genes. Some of these regulated genes had empirical confirmation for their transcriptional regulation, but other regulated genes were estimated by a computational analysis of their 5′UTR sequences. To compose the hierarchical regulation among the genes, I compiled 864 binomial relationships between the 108 transcription factor genes and the 252 regulated genes. The regulatory network with hierarchy was constructed from these binomial relationships. Among the constructed hierarchical structures, a 3 layered structure composed of 15 genes was the most complicated hierarchy. In this 3 layered hierarchical structure, many parts have been experimentally confirmed, but some parts still remain uncertain. The details of the components in this hierarchical structure are shown in Table 1.

In the TRANSFAC database, the target genes of each transcription factor have been estimated from the binding site motif sequences. In addition, the database provides quality scores ranging from 1 to 6, which reflect the experimental reliability of a particular protein-DNA interaction. Table 1 lists the regulated genes and their respective quality scores for each transcription factor. Ten genes are known to be targets of Mig1p, but three of them have not been empirically confirmed as targets. Furthermore, for 5 genes among the ten genes, their binding locations and binding site sequences have not been identified. On the other hand, the five genes known to be targeted by Gal4p have experimentally confirmed binding sites. In the last stage of this serial transcriptional regulation, the transcription factor Gal80p targets two genes. Among the 14 regulated genes in Table 1, the quality scores of the two genes were 1, which represents experimentally confirmed transcriptional regulation. The quality scores of the putatively regulated genes were 2, meaning that the transcription factor protein has been confirmed to bind to the motif sequences of the regulated genes, but the transcriptional regulation has not been confirmed. Thus, the inference of the transcriptional regulation of these

Table 1. Known binding sites and regulated genes of selected genes.

Transcription factor	Regulated genes (coding protein)	Location of binding sites	Quality score
Miglp	YLL043C (FPS1)		2
	YBR020W (GAL1)	–140 to –120, –210 to –195	2
	YDR009W (GAL3)		2
	YPL248C (GAL4)	–72 to –51, –50 to –30	2
	YKL109W (HAP4)		2
	YMR011W (HXT2)	–504 to –493, –427 to –416	2
	YIL162W (SUC2)	–505 to –483, –451 to –426	2
	YDR040C (ENA1)		6
	YLR377C (FBP1)		6
	YLR044C (PDC1)	–505 to –483, –451 to –426	6
Ga14p	YBR020W (GAL1)	367 to 384, 386 to 405, 404 to 421, 465 to 487	1
		–391 to –374, –373 to –357, –356 to –339, –291 to –275	5
	YLR081W (GAL2)	–383 to –366	2
	YBR018W (GAL7)	–292 to –270, –205 to –183	2
	YML051W (GAL80)	Around –95	2
	YNL239W (LAP3)	–150 to –134	2
Ga180p	YBR020W (GAL1)	386 to 405	1
	YBR018W (GAL7)	–205 to –183	2

Notes: Quality scores ranging from 1 to 6 reflect the experimental reliability of a particular protein-DNA interaction by TRANSFAC. Quality Score = 1, functionally confirmed transcription factor binding site = 2, binding of pure protein (purified or recombinant) = 5, binding of an uncharacterized extract protein to an element = 6, no quality assigned.

genes is informative for the functional confirmation of transcription factor binding.

I utilized an abundance of gene expression data for the 15 genes to reconstruct and clarify the 3-layered hierarchical structure composed of them. Since the genes within each layer are controlled by one transcription factor, this 3-layered structure is considered to represent serial transcriptional regulation passing through 3 stages, connected by transcription factors.

All gene expression profiles were obtained from *S. cerevisiae* and were downloaded from the GEO Database (http://www.ncbi.nlm.nih.gov/geo/). The expression profiles from 11,923 experiments were obtained as series matrix files, which describe the expression levels as a log2-ratio of the raw expression signals. The data from these matrix files were transformed into Z-scores and compared. The Grubb's test was performed on these 11,923 profiles, resulting in the analysis of 4,013 *S. cerevisiae* expression profiles.

Factor analysis

A skeleton network structure of serial transcriptional regulations was constructed by using SEM. The framework of the network structure defined transcription factors as latent variables, and their coding genes and their target genes as observed variables. In this study, 15 genes were described as observed variables, but the number of latent variables was unknown. To find this number, factor analyses were used in each of the three stages of the serial transcriptional regulation.

Factor analysis is a statistical method for describing the variability among observed variables in terms of a potentially lower number of latent variables.[28] The initial assumption is that any observed variables may be related with any latent variables. Let us assume that there are p latent variables (proteins) and q observed variables (genes) $x_1, x_2, \ldots x_q$, with means $u_1, u_2, \ldots u_q$. Note that the number p of latent variables is always smaller than the number q of observed variables. Each observed variable is expressed as linear combinations of p latent variables, as follows.

$$x_i - u_i = \alpha_{i1}F_1 + \alpha_{i2}F_2 + \cdots + \alpha_{ip}F_p + \varepsilon_i \qquad (3)$$

where x_i is the vector of the expression levels of the gene i, α_{ij} is the partial regression weight of the latent variable F_j, and ε_i is an independently distributed error term with zero mean and finite variance. In matrix form, equation (3) is expressed as

$$X - U = \Lambda F + Q. \qquad (4)$$

If there are n samples in each of the observed variables, then X and U are the $(q \times n)$ matrices composed of the observed data and their means, respectively. The partial regression coefficients of each latent variable are indicated as elements of Λ, the $(q \times p)$ latent interaction matrix. In the matrix Λ, each column corresponds to a factor and each row corresponds to an observed variable, and thus each element of Λ indicates the strength of the regulation from each protein to each gene. The matrix F is the $(p \times n)$ latent variable matrix, and Q is the $(q \times n)$ error matrix.

In the factor analysis model, the error terms ε are independent and multivariate normally distributed with a mean of zero, $\varepsilon_i \sim N(0, \Psi_i)$. If we let $Var(\varepsilon_i) = \Psi_i^2$, then the covariance matrix of ε is expressed as

$$Cov(\varepsilon) = diag(\Psi_1^2, \Psi_2^2, \ldots \Psi_p^2) = \Psi^2. \qquad (5)$$

The following assumptions are imposed on F and ε:

1. F and ε are independent.
2. $E(F) = 0$, $Var(F) = \Phi$.
3. $E(\varepsilon) = 0$, $Var(\varepsilon) = \Psi^2$.

The variance-covariance matrix between the observed variables Σ is given by

$$Var[X] = E\left[(X-U)(X-U)^t\right] = \Sigma = \Lambda\Phi\Lambda^t + \Psi^2. \qquad (6)$$

The covariance matrix of the observation variables is structurized by the parameters. From this structurized matrix, the values of the partial regression weight matrix Λ and the variances of the "errors" ε are estimated.

To clarify the possible number of latent variables in the network, Exploratory Factor Analysis (EFA) was performed. EFA is utilized to reveal the latent structure by assuming that the observational data are a synthetic amount of a lower number of latent variables. The number of latent variables was suggested by a principal factor method with varimax rotation, which is a general method for rotating factors to fit a hypothesized structure of latent variables. To extract the number of latent variables as factors of observed variables, an eigen value >1 and a scree plot were applied, as usual.[28]

The number of latent variables estimated by EFA may not correctly result in the network with latent variables. To determine the number of latent variables in each stage, I performed a Confirmatory Factor Analysis (CFA) for all numbers smaller than or equal to the number of estimated p latent variables. CFA seeks to determine whether the number of factors and associations between the factors and observed variables follow the assumption. Suggested latent variables were arranged within a genetic network as indicators, according to the initial assumption.[29] As the initial assumption, I utilized the model that includes all possible associations between the factors and the observed variables. The details of the model assumption are described in the following section.

Model assumptions

Regulatory network analysis by SEM consists of two parts: parameter fitting and structure fitting. Before parameter fitting, a network structure must be assumed. The framework of the network structure defined transcription factors as latent variables, and their coding genes and their target genes as observed variables. Based on their relationships, the transcription factors were considered as the predictor variables for the target genes, which in turn were represented as the criterion variables. Causality between transcription factors and their coding genes was treated inversely, with coding genes as the predictor variables and transcription factors as the criterion variables. Therefore, this SEM is a particular implementation of the Multiple Indicator Multiple Cause (MIMIC) model, with the latent variables and observed variables arranged alternately throughout the three transcription factor-regulated stages of the network.

Each gene was arranged as a variable in the network, according to the information registered in the TRANSFAC database. The three targeted transcription factors, Mig1p, Gal4p, and Gal80p, are known to be regulators of gene expression in the hierarchical transcription. I defined these proteins as latent variables. However, the number of latent variables in each stage and the details of the regulatory relationships between latent variables and observed variables were not initially known. Thus, I applied a modified four-step procedure, based on work by Mulaik and Millsap, which is known as a method for constructing models when there is no information about latent variables (Fig. 1).[29]

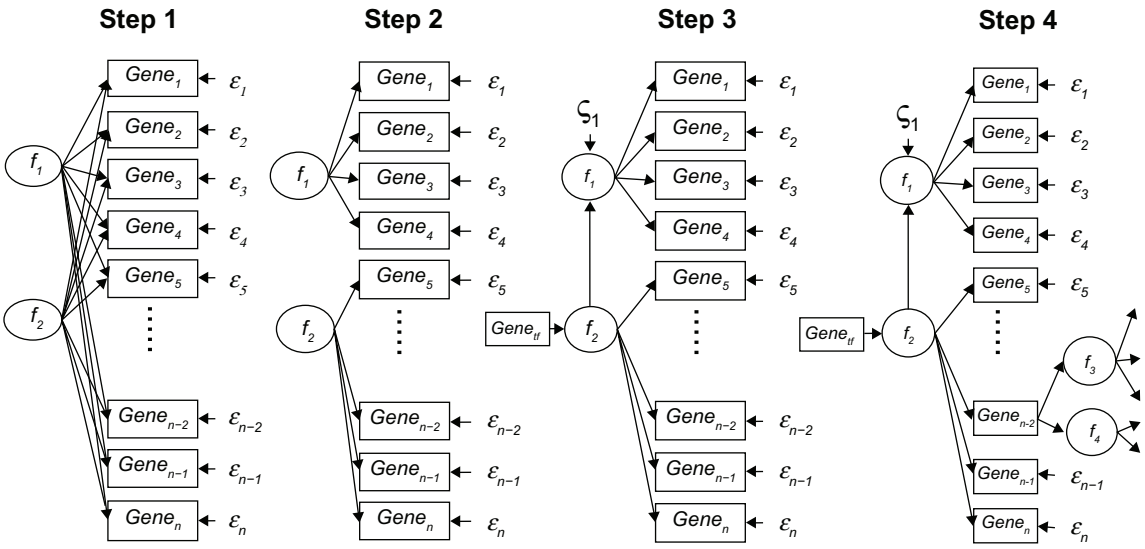


Figure 1. Modified four-step procedure. Step 1, construction of unrestricted models for each stage of the serial transcription; Step 2, construction of measurement models for each stage; Step 3, construction of structural equation models for each stage; Step 4, stepwise modeling to connect the different structural equation models.

Step 1: *Construction of the unrestricted model.* By the application of EFA to the observed variables, the number of latent variables was determined. EFA is a technique for discovering the base structure of the variable. It is assumed that a relationship exists between the regulators and the regulated variables. The factor loading value is used to find the factor structure of the data by intuition.[28–30]

Step 2: *Construction of the measurement model.* The regulatory relationships between the observed and latent variables were solved by CFA in this step. Varimax rotation, which is the most common rotation strategy, was used to seek the factor loadings across variables for each factor. All paths from latent variables to observed variables were tested for their effectiveness in factor loading. Paths with low loading scores were deleted from the models; those with high loading scores were retained. This was done for each stage.

Step 3: *Construction of the structural equation model.* The relationships between latent variables were established in this step. This step also determined the causal relationships between observed variables and latent variables. The details of this step are described in the following "Structural equation modeling" section.

Step 4: *Stepwise modeling connecting different stages of the network.* The models for each stage were connected to create one model of the entire network. I tested all network structure possibilities for one observed variable and two latent variables, to identify the most likely ones. The stages were then connected sequentially by SEM. Finally, significant relationships between error terms, as estimated by the Modification Index (MI) defined by AMOS,[31] were determined. These relationships among the error terms were used for the calculations but not incorporated into the network, and thus they have been excluded from the figures.

Structural Equation Modeling (SEM)

SEM is a comprehensive statistical model that includes two types of variables: observed and latent. These variables constitute the structural models that consider relationships between latent variables and the measurement models that consider relationships between the observed variables and the latent variables. These relationships can be presented both algebraically, as a system of equations, and graphically, as path diagrams.

In this study, the target genes and the coding genes of the transcription factors in the transcriptional regulation were defined as observed variables, since their expression data were obtained from the GEO. Meanwhile, the transcription factors were defined as

latent variables, owing to the lack of measured data in the expression profiles. All variables were classified as one of two types: exogenous variables and endogenous variables. Exogenous variables are those that are not regulated by other variables in the system. Endogenous variables, on the other hand, are. In my model, the mig1 gene is defined as an exogenous variable, while all other genes are defined as endogenous variables. All latent variables are endogenous variables, since they are regulated by coding genes. There are three possible relationships between the variables: relationships between latent variables, causal relationships between observed variables and latent variables, and regulatory relationships between latent variables and observed variables. The model is defined as follows:

$$\eta = B\eta + \Gamma y + \varsigma \tag{7}$$

$$y = \Lambda\eta + \varepsilon. \tag{8}$$

Here, η is a vector of p latent variables (number of transcription factors); y is a vector of q observed variables (measured gene expression patterns); B is a $p \times p$ matrix representing relationships between latent variables; Γ is a $p \times q$ matrix representing causal relationships between observed and latent variables; and Λ is a $q \times p$ matrix representing regulatory relationships between latent and observed variables. Errors that affect the observed and latent endogenous variables are denoted by ς and ε, respectively.

From the above equations, I have $\begin{bmatrix}\eta\\ y\end{bmatrix} = \begin{bmatrix}\mathrm{B} & \Gamma\\ \Lambda & \mathrm{O}\end{bmatrix}\begin{bmatrix}\eta\\ y\end{bmatrix} + \begin{bmatrix}\varsigma\\ \varepsilon\end{bmatrix}$ to represent the model. The structural equation modeling is based on a covariance analysis defined as $S = \Sigma(\theta)$, where S is the covariance matrix calculated from the observed data and $\Sigma(\theta)$ is a matrix-valued function of the parameter θ. Let Φ denote the covariance matrices of the error terms ς and ε, and G denote a $q \times (p + q)$ combined matrix of the $q \times p$ zero matrix and the $q \times q$ identity matrix. The covariance matrix of model $\Sigma(\theta)$ is given by

$$\Sigma(\theta) = G\begin{bmatrix}I - \mathrm{B} & -\Gamma\\ -\Lambda & -I\end{bmatrix}^{-1}\Phi\begin{bmatrix}I - \mathrm{B} & -\Gamma\\ -\Lambda & -I\end{bmatrix}^{-1'}G'. \tag{9}$$

Each element of the covariance matrix model $\Sigma(\theta)$ is expressed as a function of the parameters that appear in the model. The unknown parameters were estimated, in order to minimize the difference between the model covariance matrix $\Sigma(\theta)$ and the sample covariance S.

The SEM software package SPSS AMOS 17.0 (IBM, USA) was used to fit the model to the data. The quality of the fit was estimated by four different model fitting scores: GFI, AGFI, CFI and RMSEA. These scores were considered to be useful to clarify the degree of model fitting in this study, since the model can be evaluated by a general threshold, rather than a huge experimental number.

Parameter estimation

To make the model covariance matrix $\Sigma(\theta)$ closer to the sample covariance matrix S, the parameters within the model were estimated by the maximum likelihood (ML) method with the fitting function. Since the ML estimators are known to be consistent and asymptotically unbiased, ML is commonly used as a fitting function to estimate SEM parameters:

$$F_{ML}(S, \Sigma(\theta)) = \log|\Sigma(\theta)| - \log|S| + tr(\Sigma(\theta)^{-1}S) - q. \tag{10}$$

Here, $\Sigma(\theta)$ is the estimated covariance matrix; S is the sample covariance matrix; $|\Sigma|$ is the determinant of matrix Σ; $tr(\Sigma)$ is the trace of matrix Σ; and q is the number of observed variables. The fitting function $F_{ML}(S, \Sigma(\theta))$ is a discrepancy function that compares the difference between the estimated covariance matrix $\Sigma(\theta)$ and the sample covariance matrix S. The principal objective of SEM is to minimize $F_{ML}(S, \Sigma(\theta))$, which is the objective function that is used to obtain the maximum likelihood. Generally, $F_{ML}(S, \Sigma(\theta))$ is a nonlinear function. Therefore, iterative optimization is required to minimize $F_{ML}(S, \Sigma(\theta))$ and to find the solutions by Fisher's scoring method in the AMOS software, since the Fisher method is highly effective, and may even converge in a single iteration.[32] To avoid a local optimization, I executed iterative optimization procedures with different initial values, and chose the global optimum solutions of the parameters. In the optimization process, the range of parameters was not chosen, but

one parameter of the relationships from the latent variable to the observed variable was fixed to one, as a restriction of the model. Furthermore, I used 1E-5 as the convergence criteria. As the limit on the number of iterations, I chose 100 iterations to be performed by AMOS. When this limit was reached and the convergence criteria had not been met, the model was rejected.

Results

Factor analysis in each stage

In the hypothetical network structure, 15 genes were described as observed variables, but the number of latent variables was unknown. To find this number, EFA was used in each of the three stages of the serial transcriptional regulation. After the number of transcription factors was determined, CFA was applied with a severer check. The check by CFA is required before the SEM analysis.[28,29]

EFA was applied to the ten genes regulated by Mig1p in the first stage, the five genes regulated by Gal4p in the second stage, and the two genes regulated by Gal80p in the last stage. In the first stage, four latent variables were extracted by EFA as the effective factors of ten regulated genes, and the parameters of the two models composed of one or two latent variables could be calculated by CFA. The remaining models composed of three or four latent variables cannot be calculated by the partial regression coefficients, even though the number of parameters was smaller than the number of equations. Thus, the limitation of the latent variables in the first stage was considered to be two. In the second stage, two latent variables were extracted by EFA, and the parameters in the model with all extracted latent variables were calculated. Hence, the limitation of the latent variables of the second stage was regarded as two. In contrast, only two regulated genes were included in the third stage. According to the restriction that the number of latent variables is smaller than the number of observed variables, the number of latent variables in the third stage was considered to be one. Actually, only one latent variable was extracted by EFA, and the parameters of the model composed of one latent variable could be calculated by CFA.

By a combination of EFA and CFA, the maximum number of latent variables in each stage was assumed to be two. Furthermore, the regulatory factors were expected to be encoded by only one gene in each stage of this hypothetical network structure. Thereby, the two factors were considered to be simple substances or complexes of multiple molecules. Although it is possible that the target genes are regulated by other factors, those factors are considered to be encoded by other genes that are outside of this hypothetical network in this study.

After the number of transcription factors was determined, the modified four-step procedure was applied in a stepwise manner. The structure of the resulting model was not simple. For example, the latent variables clearly had high loading for some observed variables, but low loading for others. Similarly, some observed variables were strongly affected by both latent variables, but others were influenced by only one.

Results of SEM

SEM was used to determine the relationships among the transcription factors and the corresponding coding genes in this serial transcriptional regulation. The target genes were considered to be physically bound by the transcription factors represented by latent variables. The regulatory and causal relationships between the latent and observed variables, and the relationships between latent variables were considered in each stage, but the relationships between observed variables were not. By connecting the observed variables, representing the coding genes, to the latent variables, representing the corresponding translated transcription factors, the stage separated models were combined to construct a model for the whole transcriptional regulation in the cell. To connect the different stages, all possible connections between one observed variable and the estimated latent variables were considered. This was done sequentially, with the first two stages being connected by SEM, followed by the third stage. Figure 2 shows the various network structure models between observed variables and latent variables, which molecularly describe a coding gene (gene) being transcribed and translated into one or two effective proteins (F1 and F2).

All possible models were evaluated in terms of their goodness-of-fit scores by using the goodness-of-fit index (GFI), which measures the relative discrepancy between the empirical data and the modeled network, and the adjusted GFI (AGFI), which is a GFI modified according to the degrees of freedom.

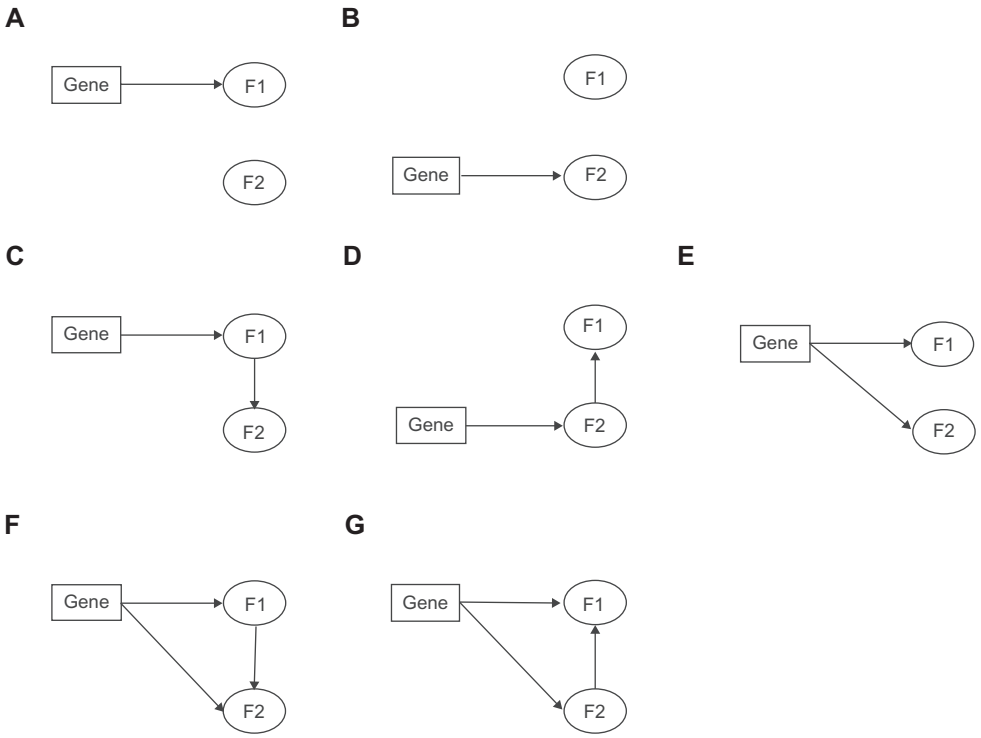
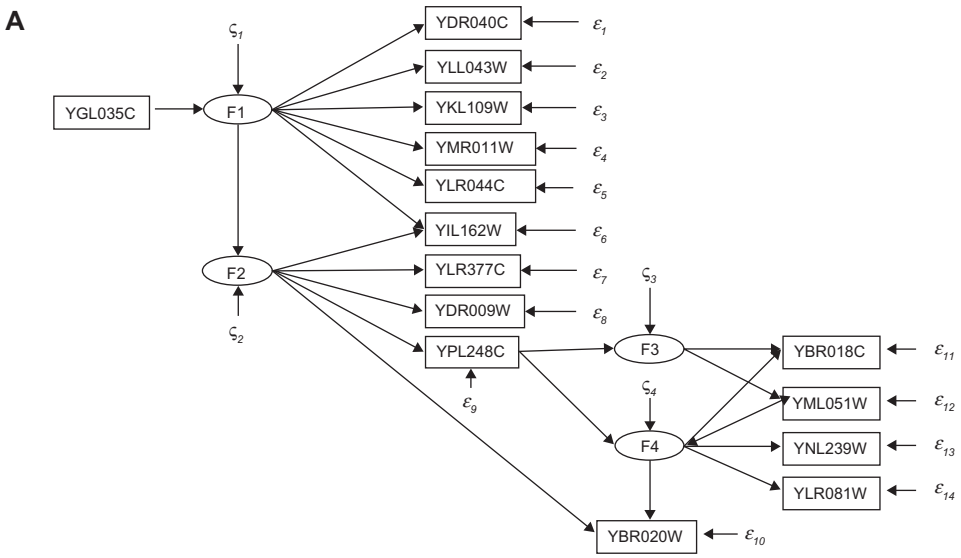


Figure 2. Possible relationships between variables. The relationships between one observed variable (gene) and two estimated latent variables (F1 and F2). Based on empirical studies, two possible causal relationship directions exist: the observed variable to the latent variables, and one latent variable to the other latent variable. (**A**), (**B**) Possibilities for one causal relationship between the observed variable and one latent variable. (**C**), (**D**), (**E**) Possibilities for two causal relationships between the observed variable and the latent variables. (**F**), (**G**) Possibilities for three causal relationships between the observed variable and the latent variables.

Furthermore, I used CFI and RMSEA as fitting scores to evaluate the model fitting. Since these indices have threshold values as criteria to decide whether the model is suited to obtain data independent of a huge sample number, they are considered to be useful to clarify the degree of model fitting in this study. By evaluating the attached models in terms of their goodness-of-fit scores and expanding the network model in a stepwise manner, I constructed a complete model of the serial transcriptional regulation composed of the GAL-related genes.

In the first and second stages, two latent variables were chosen as the regulators of the target genes. The associations from those latent variables to the observed variable had been estimated in the first two steps of the four-step-procedure, and thus the models for inferring the connected part were constructed while maintaining the estimated associations. According to Figure 2, I constructed 7 models while maintaining the association from the latent variables to the observed variables. There was no difference in the goodness-of-fit scores of the 7 constructed models in the first step, and all of the goodness-of-fit scores were low, such as GFI < 0.9 and RMSEA > 0.5. To select the best model, I modified the models as follows: (1) Deletion of the associations between the variables with $P > 0.05$, and (2) Addition of the associations between a latent variable and an observed variable according to the MI scores. All of the modifications were performed step by step to infer the best structure across the two different stages.

Figure 3 shows the estimated relationships between the genes and their corresponding transcription factors. The regression weights for each path are shown in Table 2. Table 2 shows all of the significant regression weights ($P < 0.05$). The MI was used to estimate the relationships among the error terms after all of the relationships among the variables were estimated. The MI measures how much the chi-square statistic is expected to decrease if a particular parameter setting is constrained. Different covariance structures were constructed for the error terms, and thus the relationships between the error terms were tested last. The MI showed thirty relationships between the error terms. Adding these relationships between the error terms to the model improved the goodness-of-fit (Fig. 3b).



B

GFI	AGFI	CFI	RMSEA
0.992	0.977	0.922	0.035

Figure 3. Inferred network model of the GAL regulatory system. (**A**) Estimated main structure of the transcriptional regulation. Arrows show causal relationships between genes (rectangles) and transcription factors (circles). Error terms are indicated by ε. Relationships between errors are considered to represent other regulatory systems in the cell. For simplicity, these relationships are not shown. (**B**) Goodness-of-fit scores. The calculations for these scores included relationships between errors. Four criteria were used: GFI $>$ 0.95, AGFI $>$ 0.95, CFI $>$ 0.90 and RMSEA $<$ 0.05.
Note: All four scores indicate that the model fit the measured data well.

In other words, incorporating the relationships between the error terms led to a better model. Since many regulatory systems in a cell act simultaneously, the relationships between the error terms can be considered to represent the presence of other regulatory networks. In this study, the network structure among the latent variables and the observed variables focuses only on the hierarchical transcription initiated by Mig1p. Therefore, for simplicity, these other regulatory systems are not displayed in Figure 3A.

The relative strength of each association is shown as a standardized regression weight in Table 2. Interestingly, the standardized regression weight of the association from F1 to F2 was negative, even though both F1 and F2 were considered as proteins. This is one of the features of the SEM analysis. The negative interactions between the observed variables are summarized as the existence of a negative interaction between an observed variable and a latent variable. Thus, the negative relationship between F1 and F2 does not indicate the negative association from F1 to F2, but the negative relationships between YGL035C and the target genes of F2.

In the first stage, 6 target genes were regulated by F1 and 5 other genes were regulated by F2, although YIL162W was regulated by both F1 and F2. Among the 6 genes that were regulated by F1, 3 genes have identified Mig1p binding sites in Table 1; the binding sites of YMR011W are −504 to −493, and −427 to −416; the binding sites of YLR044C are −505 to −483, and −451 to −426; and the binding sites of YIL162W are −505 to −483, and −451 to −426. In contrast, the known binding sites of the F2 regulated genes were closer to the transcription start points.

The latent variable F4 in Figure 3(A) shows that the latent variable in the third stage was combined with one latent variable in the second stage. The association from YML051W to one latent variable in the second stage was revealed by the MI scores during the model modification step. Furthermore, the probabilities of the associations between the third latent variable and the target genes were not significant ($P > 0.05$). Therefore, the associations between the third latent variable and the target genes were deleted from the inferred model, instead of adding the association from YML051W to F4. This model

Table 2. Estimated regression weights ($P < 0.05$).

Regulated variable	Direction	Regulator variable	Regression weight	Standardized regression weight	Level of significance probability
F1	<---	YGL035C	0.075	0.345	***
F2	<---	F1	–2.248	–0.467	***
YPL248C	<---	ε9	0.949	0.997	***
YPL248C	<---	F2	0.083	0.084	***
F3	<---	YPL248C	0.285	0.263	***
YML051W	<---	F3	0.209	0.301	***
YML051W	<---	ε12	0.685	0.954	***
F4	<---	YPL248C	0.376	0.331	***
F4	<---	YML051W	0.259	0.172	***
YDR040C	<---	F1	1	0.181	
YDR040C	<---	ε1	1.085	0.984	***
YLR377C	<---	ε7	1.336	0.977	***
YLL043W	<---	F1	0.721	0.202	***
YLL043W	<---	ε2	0.696	0.979	***
YDR009W	<---	ε8	0.962	0.975	***
YKL109W	<---	F1	1.991	0.344	***
YKL109W	<---	ε3	1.083	0.939	***
YMR011W	<---	F1	3.223	0.432	***
YMR011W	<---	ε4	1.344	0.902	***
YLR044C	<---	F1	–0.445	–0.068	0.027
YLR044C	<---	ε6	1.306	0.998	***
YIL162W	<---	F1	1.513	0.24	0.002
YBR018C	<---	ε11	0.733	0.654	***
YNL239W	<---	ε13	1.021	0.994	***
YLR081W	<---	ε14	1.126	0.989	***
YIL162W	<---	ε5	0.805	0.641	***
YLR377C	<---	F2	0.3	0.211	***
YIL162W	<---	F2	1	0.765	
YBR020W	<---	ε10	0.826	0.923	***
YBR020W	<---	F4	0.313	0.379	***
YBR018C	<---	F3	–0.637	–0.586	***
YDR009W	<---	F2	0.229	0.223	***
YNL239W	<---	F4	–0.133	–0.14	***
YBR020W	<---	F2	0.059	0.063	0.006
YBR018C	<---	F4	0.581	0.562	***
YLR081W	<---	F4	0.148	0.14	***

Notes: The value in the Regression weight column indicates the degree of influence from "Regulator variable" to "Regulated variable". The "Regression weight" is estimated with the restriction for the discrimination of the model. Thus, the Regression weight from F1 to YDR040C is fixed to one. The standardization regression weights were calculated after the standardization coefficient first calculated by the standardization of all original data. The result is often interpreted as being comparable by using it in general in SEM, relative to the influence of two or more passing coefficients. For example, when YGL035C increases by 1 standard deviation, F1 increases by 0.345 standard deviation.

shows that the transcriptional product of YML051W (Gal80p) may interact with the transcriptional product of YPL248C (Gal4p). Actually, the assumed transcription factor Gal80p has been experimentally confirmed to form a complex with the transcription factor Gal4p. Thus, the stepwise method using both the MI scores and the probabilities of the estimated parameters seems to be useful to expand the network model.

Discussion

The scheme of the entire transcriptional regulation composed of GAL-related genes is shown in Figure 4. In the first two stages, the transcription factors are determined by two latent variables. The transcription factor in the last stage, Gal80p, appears as one of the two latent variables of Gal4p. This suggests that some latent variables represent protein complexes. This is no surprise, as Gal80p is known to form a complex

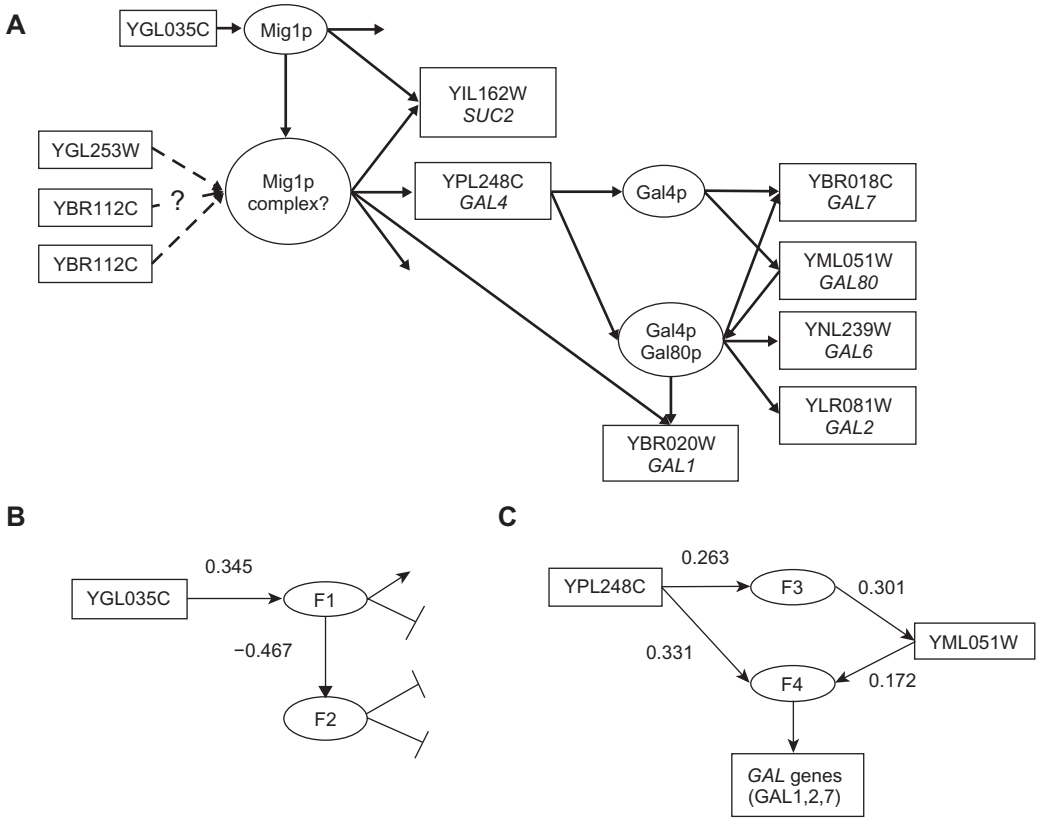


Figure 4. Biological interpretation of the factors. (**A**) Key structures of the inferred network. Genes (rectangles) with two names show the ORF name (upper) and the coding protein name (lower). Circles indicate the transcription factor or a complex of factors. Mig1p and Gal4p form complexes with other proteins. (**B**) Beginning of the serial transcriptional regulation initiated by Mig1p. Numbers indicate the regression weight of the relationships between variables. The edge from YGL035C to F1 was positive, indicating that the expression of YGL035C promoted the translation of the subsequent coding protein. On the other hand, the edge from F1 to F2 is negative, indicating an inhibitory relationship. (**C**) Gal4p and Gal80p. The best fit for the latent variable at the third stage was located at the same location as another latent variable. In other words, Gal80p was inferred to form a complex with Gal4p.

with Gal4p.[33–37] Similarly, Mig1p is also thought to form a complex with Hxk2p,[38,39] although the details of this complex are unclear.

According to my inferred network, one of the two transcription factors in the first stage is Mig1p, which is encoded by the YGL035C gene. The known binding sites of F1's target genes were far from the transcription start point, such as around −500, even though there is no information about the Mig1p binding sites for the remaining 3 genes. In addition, the feature of the Mig1p binding position was not detected for F2's target genes. It is suspected the differences between the latent variables were caused by the Mig1p binding site. Similarly, the unknown binding sites of Mig1p in F1's target genes may be expected to be around −500.

The negative relationship between F1 and F2 can be regarded as the negative relationships between F2 and its target genes. The expression of all of F2's target genes is known to be induced by glucose under poor conditions, even though the relationship with glucose for some of F1's target genes was not revealed. Actually, Mig1p is known to be active and bound to other genes under glucose-rich conditions, to repress its target genes.[40–43] Thus, the associations between F2 and its target genes may represent repressive regulation by Mig1p. Furthermore, the relationship between the two latent variables in this stage appears causal. The second factor is considered to be a complex involving Mig1p, since Mig1p is known to form a complex when binding to F2's target genes. The estimated regression values between YGL035C and the two latent variables are shown in Figure 4B. Only YIL162W was regulated by both F1 and F2. This may be occurred by the difference between the meaning of arrows from F1 and those from F2. The arrows from F1 indicated the position of physical interaction, on the contrast the meaning of arrows from F2 meant binding regulation by Mig1p complex. The expression of YIL162W is known as to be repressed by Mig1p

binding, but its expression may be controlled by the balanced condition in cell.

In the second stage, the observed variable representing the gal4 gene, YPL248C, has a causal relationship with the two latent variables. One of these latent variables is also regulated in the third stage by the gal80 gene, YML051W. The details of this regulation are shown in Figure 4C. The latent variable F4 is regulated by the two genes, and is therefore considered to be a complex of Gal4p and Gal80p. Such a complex has been observed empirically.[33–37,44] The model aptly reflects the fact that the regulation of GAL 1, 2, and 7 is controlled by this complex. In general, the target genes in my model were divided into three types: those regulated by a single transcription factor, those regulated by a transcription factor complex, and those regulated by both. Overall, the model successfully describes the hierarchy of the serial transcriptional regulation among GAL-related genes.

Some relationships between the genes were indicated by the MI scores. The tentative relationships between genes have a reverse direction, such as from a gene in the second stage to a gene in the first stage. In this study, those relationships between genes were not included in the model, since they do not clarify the relationships between proteins and genes in serial transcriptional regulation. However, it is possible that the relationships with a reverse direction indicate feedback regulation. To clarify the feedback regulation in this model, all proteins encoded by all genes should be defined as latent variables in SEM, because it has no previous information about latent variables. This approach contradicts the assumption that the number of latent variables should be smaller than the number of observed variables, and the parameters are not estimated. Since a cyclic model can be analyzed by SEM, the feedback regulation will be clarified with the latent variable information in the future.

Although many factors are suspected to regulate gene expression, since their underlying mechanisms are unclear, these regulators can be viewed as little more than a black box. Here, I have shown that SEM is a powerful approach to estimate the gene expression network controlled by transcription factor binding, based on its gene expression profiles. As biological data accumulate, it is expected that SEM will be applicable to a wide number of gene networks to clarify gene-protein interactions.

Conclusions

In this study, one serial transcriptional regulation was reconstructed by a model that incorporated both genes and proteins from only the gene expression profiles, in the absence of protein information. Since the interactions between proteins and genes can be accurately inferred, this approach should be of great interest to systems biologists. The ability to identify expression profiles and the corresponding biological functions is expected to provide further possibilities for SEM in the inference of regulatory mechanisms in cells. As this approach can be applied to numerous systems and organisms beyond yeast, my findings should be of interest to a wide field of biologists.

Acknowledgements

I thank Dr. Horton and Dr. Toh (CBRC, AIST) for valuable discussions that aided this investigation.

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. Brazhnik P, Fuente A, Mende P. Gene networks: how to put the function in genomics. *Trends in Biotechnology*. 2002;20(11):467–72.
2. Akutsu T, Miyano S, Kuhara S. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J Comput Biol*. 2000;7:331–43.
3. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol*. 2000;7:601–20.
4. Aburatani S, Kuhara S, Toh H, Horimoto K. Deduction of a gene regulatory relationship framework from gene expression data by the application of graphical Gaussian modeling. *Signal Processing*. 2003;83:777–88.
5. Aburatani S, Horimoto K. Elucidation of the Relationships between LexA-Regulated Genes in the SOS response. *Genome Informatics*. 2005;16: 95–105.

6. Bollen KA. *Structural Equations with Latent Variables*. New York: Wiley-Interscience; 1989.
7. Haavelmo T. The statistical implications of a system of simultaneous equations. *Econometrica*. 1943;11:1–12.
8. Duncan OD. *Introduction to Structural Equation Models*. New York: Academic Press; 1975.
9. Pearl J. Causality: Models, Reasoning, and Inference. 2nd ed. Cambridge: Cambridge University Press; 2001.
10. Liu B, Fuente ADL, Hoeshele I. Gene Network Inference via Structural Equation Modeling in Genomics Experiments. *Genetics*. 2008;178:1763–76.
11. Aten JE, Fuller TF, Lusis AJ, Horvath S. Using genetic markers to orient the edges in quantitative trait networks: The NEO software. *BMC Systems Biology*. 2008;2(34):1–21.
12. Shieh GS, Chen CM, Yu CY, Huang J, Wang WF, et al. Inferring transcriptional compensation interactions in yeast via stepwise structure equation modeling. *BMC Bioinformatics*. 2008;9(134):1–10.
13. Xiong M, Li J, Fang X. Identification of Genetic Networks. *Genetics*. 2004; 166:1037–52.
14. Lee S, Jhun M, Lee EK, Park T. Application of structural equation models to construct genetic networks using differentially expressed genes and single nucleotide polymorphisms. *BMC proceedings*. 2007;I(S76):1–5.
15. Westergaard SL, Oliveira AP, Bro C, Olsson L, Nielsen J. A systems biology approach to study glucose repression in the yeast saccharomyces cerevisiae. *Biotechnology and Bioengineering*. 2007;96(1):134–45.
16. Prasad V, Venkatesh KV. Stochastic analysis of the GAL genetic switch in Saccharomyces cerevisiae: Modeling and experiments reveal hierarchy in glucose repression. *BMC Systems Biology*. 2008;2(97):1–17.
17. Klein CJ, Olsson L, Nielsen J. Glucose control in Saccharomyces cerevisiae: the role of Mig1 in metabolic function. *Microbiology*. 1998;144:13–24.
18. Griggs DW, Johnston M. Regulated expression of the GAL4 activator gene in yeast provides a sensitive genetic switch for glucose repression. *Proc Natl Acad Sci U S A*. 1991;88:8597–601.
19. Nehlin JO, Ronne H. Yeast MIG1 repressor is related to the mammalian early growth response and Wilms' tumour finger proteins. *EMBO J*. 1990;9:2891–8.
20. Nehlin JO, Carlberg M, Ronne H. Control of yeast GAL gene by MIG1 repressor: a transcriptional cascade in the glucose response. *EMBO J*. 1991;3373–7.
21. Peng G, Hopper JE. Gene activation by interaction of an inhibitor with a cytoplasmic signaling protein. *Proc Natl Acad Sci U S A*. 2002;99:8548–53.
22. Venkatesh KV, Bhat PJ, Kumar RA, Doshi P. Quantitative model for Gal4p-mediated expression of the galactose/melibiose regulon. *Biotechnol Prog*. 1999;15:51–7.
23. Verma M, Bhat PJ, Venkatesh KV. Expression of GAL genes in a mutant strain of Saccharomyces cerevisiae lacking GAL80: quantitative model and experimental verification. *Biotechnol Appl Biochem*. 2004;39:89–97.
24. Hashimoto H, Kikuchi Y, Nogi Y, Fukasawa T. Regulation of expression of the galactose gene cluster in Saccharomyces cerevisiae. Isolation and characterization of the regulatory gene GAL4. Mol. *Gen Genet*. 1983;191(1): 31–8.
25. Lohr D, Venkov P, Zlatanova J. Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *FASEB J*. 1995;9(9):777–87.
26. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*. 2004; 431(7004):99–104.
27. Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, et al. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res*. 2009;19(4):556–66.
28. Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. 2nd ed. Cambridge: The MIT Press; 2001.
29. Mulaik SA, Millasap RE. Doing the four-step right. *Structural Equation Modeling*. 2000;7(1):36–73.
30. Hochreiter S, Clevert DA, Obermayer K. A new summarization method for affymetrix probe level data. *Bioinformatics*. 2006;22(8):943–9.
31. Joreskog KG, Sorbom D. LISREL-VI: *Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*. Redondo Beach: Doss-Haus Books; 1984.
32. Kendall MG, Stuart A. *The Advanced Theory of Statistics*. 3rd ed. New York: Hafner; 1973.
33. Pilauri V, Bewley M, Diep C, Hopper J. Gal80 dimerization and the yeast GAL gene switch. *Genetics*. 2005;169(4):1903–19.
34. Melcher K. Mutational hypersensitivity of a gene regulatory protein: Saccharomyces cerevisiae Gal80p. *Genetics*. 2005:171(2):469–76.
35. Li Y, Chen G, Liu W. Alterations in the interaction between GAL4 and GAL80 effect regulation of the yeast GAL regulon mediated by the F box protein Dsg1. *Curr Microbiol*. 2010:61(3):210–6.
36. Kumar PR, Yu Y, Sternglanz R, Johnston SA, Joshua-Tor L. NADP regulates the yeast GAL induction system. *Science*. 2008:319(5866):1090–2.
37. Jiang F, Frey BR, Evans ML, Friel JC, Hopper JE. Gene activation by dissociation of an inhibitor from a transcriptional activation domain. *Mol Cell Biol*. 2009;29(20):5604–10.
38. Ahuatzi D, Herrero P, de la Cera T, Moreno F. The glucose-regulated nuclear localization of hexokinase 2 in Saccharomyces cerevisiae is Mig1-dependent. *J Biol Chem*. 2004;279(14):14440–6.
39. Ahuatzi D, Riera A, Pelaez R, Herrero P, Moreno F. Hxk2 regulates the phosphorylation state of Mig1 and therefore its nucleocytoplasmic distribution. *J Biol Chem*. 2007;282(7):4485–93.
40. Lutfiyya LL, Iyer VR, DeRisi J, DeVit MJ, Brown PO, et al. Characterization of three related glucose repressors and genes they regulate in Saccharomyces cerevisiae. *Genetics*. 1998;150(4):1377–91.
41. Sarma NJ, Haley TM, Barbara KE, Buford TD, Willis KA, et al. Glucose-responsive regulators of gene expression in Saccharomyces cerevisiae function at the nuclear periphery via a reverse recruitment mechanism. *Genetics*. 2007;175(3):1127–35.
42. Moreno F, Ahuatzi D, Riera A, Palomino CA, Herrero P. Glucose sensing through the Hxk2-dependent signalling pathway. *Biochem. Soc Trans*. 2005; 33(Pt 1):265–8.
43. Westholm JO, Nordberg N, Muren E, Ameur A, Komorowski J, et al. Combinatorial control of gene expression by the three yeast repressors Mig1, Mig2 and Mig3. *BMC Genomics*. 2008;9(601):1–15.
44. Hertveldt K, Dechassa ML, Robben J, Volckaert G. Identification of Gal80p-interacting proteins by Saccharomyces cerevisiae whole genome phage display. *Gene*. 2003;307:141–9.