A Linear Discrete Dynamic System Model for Temporal Gene Interaction and Regulatory Network Influence in Response to Bioethanol Conversion Inhibitor HMF for Ethanologenic Yeast

Mingzhou (Joe) Song[1] and Z. Lewis Liu[2]

[1] Department of Computer Science, New Mexico State University
P.O. Box 30001, MSC CS, Las Cruces NM 88003, U.S.A
[2] National Center for Agricultural Utilization Research
U.S. Department of Agriculture, Agriculture Research Service
1815 N University Street, Peoria, Illinois 61604, U.S.A

Abstract. A linear discrete dynamic system model is constructed to represent the temporal interactions among significantly expressed genes in response to bioethanol conversion inhibitor 5-hydroxymethylfurfural for ethanologenic yeast *Saccharomyces cerevisiae*. This study identifies the most significant linear difference equations for each gene in a network. A log-time domain interpolation addresses the non-uniform sampling issue typically observed in a time course experimental design. This system model also insures its power stability under the normal condition in the absence of the inhibitor. The statistically significant system model, estimated from time course gene expression measurements during the earlier exposure to 5-hydroxymethylfurfural, reveals known transcriptional regulations as well as potential significant genes involved in detoxification for bioethanol conversion by yeast.

1 Introduction

Computational modeling of gene regulatory networks (GRNs) is a central focus in systems biology. By far, few approaches are capable of describing the information flow over time in a large network. Only a dynamic system model of a GRN can empower biologists to fully understand the interactions among entities in a network. Verhulst equation, a discrete dynamic system model of one variable, is an example that is widely used in mathematical biology (Edelstein-Keshet, 2004) to study population dynamics in evolution. Although early work that utilizes difference equations to model GRNs exists (D'haeseleer et al., 1999), which estimates system coefficients by least squares, the potential of discrete dynamic systems in modeling GRNs has remained largely unrecognized until recent endeavors by systems biology researchers such as Bonneau et al. (2006) and Schlitt and Brazma (2006), who characterize gene interactions by discrete dynamic system models composed of linear difference equations or finite state linear

T. Ideker and V. Bafna (Eds.): Syst. Biol. and Comput. Proteomics Ws, LNBI 4532, pp. 77–95, 2007.

equations. Our work moves along with three innovations. The first is to perform log-time domain interpolation to reposition non-uniformly spaced samples to equally spaced time locations. The second is to assess statistical significance of all possible linear difference equations for a given gene node and to choose the most significant one, as well as to assess the statistical significance of the entire system. The third is to enforce power stability on the discrete dynamic system model so that it does not exhibit chaotic or unstable behaviors under a normal condition. A discrete dynamic system is power stable if variables in the system stay bounded as time goes to infinity given a bounded initial state.

A major motivation of our work originates from the investigation of genetic mechanisms for bioethanol conversions in yeast in pursuit of renewable sources of energy. As interest in alternative energy sources rises, the concept of agriculture as an energy producer has become increasingly attractive. Renewable biomass, including lignocellulosic materials and agricultural residues, has become attractive low cost materials for bioethanol production. One major barrier of biomass conversion to ethanol is inhibitory compounds generated during biomass pretreatment, which interfere with microbial growth and subsequent fermentation. For economic reasons, dilute acid hydrolysis is commonly used to prepare the biomass degradation for enzymatic saccharification and fermentation (Bothast and Saha, 1997; Saha, 2003). However, numerous side-products are generated by this pre-treatment, many of which inhibit microbial metabolism. More than 100 compounds have been detected to have potential inhibitory effects on microbial fermentation (Luo et al., 2002). Among these compounds, 5-hydroxymethylfurfural (HMF) and furfural are the most potent and representative inhibitors derived from biomass pretreatment (Taherzadeh et al., 2000; Martin and Jonsson, 2003). Other commonly recognized inhibitors include acetic acid, cinnamic acid, coniferyl aldehyde, ethanol, ferulic acid, formic acid, levulinc acid, and phenolics. Few yeast strains tolerant to inhibitors are available due to a lack of understanding of mechanisms involved in the stress tolerance for bioethanol fermentation. Based on functional genomic studies, a concept of genomic adaptation to the biomass conversion inhibitors by the ethanologenic yeast is proposed (Liu and Slininger, 2006a; Liu, 2006). However, a great deal of detailed knowledge of GRNs involved remains unknown.

In the computational and biological context described above, we have developed discrete dynamic system models to study the genetic basis underlying metabolic pathway of the ethanologenic yeast. As initiated in this study, we have delineated through discrete dynamic system models how a biological system behaves in response to inhibitor HMF during the earlier exposure to the inhibitor for ethanol production. In this model, the change in expression level of a target gene at a discrete time point is a linear function of the expression levels of influential genes at previous discrete time points. This model facilitates the characterization of gene interactions in efficient production of ethanol in yeast under both control and stress conditions, allowing one to introduce specific perturbations into a system and predict the effects on biomass conversion under various

stress conditions. Furthermore, the model enables one to identify relevant genes and gene interactions for optimal genetic manipulations that will guide the engineering of more robust yeast strains for economic ethanol production.

Although other alternative modeling methodologies have been developed, discrete dynamic system models are advantageous given the increased availability of experimental designs that collect time-course gene expressions at the whole system scale. *Dynamic Bayesian networks* (DBNs) extend the static Bayesian networks by introducing the time aspect. Both models have been used for modeling GRNs: the former used by (Ong et al., 2002) and the latter used by (Imoto et al., 2003; Friedman, 2004). A DBN describes statistical dependencies among genes temporally, by extending Bayesian networks to incorporate time transitions between the Bayesian networks at consecutive time points. Since a DBN does not describe functional relations among genes, it is not a suitable tool to understand the dynamics of a GRN, though there is no doubt that Bayesian networks and DBNs are indeed successful in extracting probabilistic dependencies among genes. The *Boolean networks* (Liang et al., 1998; Akutsu et al., 2003; Pal et al., 2005) have gained momentum recently. Shmulevich et al. (2002) introduce stochastic components for GRNs by creating probabilistic Boolean networks. Since a Boolean network represents gene expression level in two states: on and off, this qualitative abstraction limits its capacity in discriminating quantitative changes in gene expression levels under perturbed situations. Our primary goal is to establish a gene interaction network model inferring regulatory mechanisms in biomass conversion to ethanol, especially the quantitative shift of biotransformation and detoxification of the inhibitors, which requires information beyond the presence or absence of genes. Thus, Boolean networks are not the best dynamic strategy to describe accurately the amount of ethanol product as a function of the concentration of glucose substrate. *Differential equations* in both deterministic (Meir et al., 2002) and stochastic (van Kampen, 1997) formulations have been used to model interactions among entities in a GRN in continuous time. The E-CELL Project (Tomita et al., 1999; Takahashi, 2004; Takahashi et al., 2005) targets at reproducing *in silico* intracellular biochemical and molecular interactions within a single cell with the differential equation model. The stochastic differential equations (Master equations) represent the dynamics of probabilities of states by differential equations, which is impractical for GRNs involving more than a handful of genes because the amount of data needed to characterize stochastic behaviors is subject to curse of dimension, to be encountered in probability density estimation. However, almost all differential equations reduce to difference equations in practical applications. Direct discrete dynamic modeling overrides this intermediate step and speaks the native discrete time language of a computer. We believe it is more effective to go without the intermediate mode of differential equations. In addition, the time interval between discrete points in difference equations can be adjusted to the sparsity of data, making it more flexible to model the dynamics at different resolutions.

2 Results and Discussion

Using first order linear difference equations, we build discrete dynamic system models for the transcriptional interactions among genes in yeast during the earlier exposure to the inhibitor HMF for ethanol production. In a discrete dynamic system model, the expression change rate of a gene is a linear function of the concentrations of potential regulator genes – one equation is used for each gene. A network is derived from a discrete dyamic system model by creating an edge from every potential regulator to each gene it regulates. These models were developed based on mRNA abundance over five time points in the presence or absence of HMF. Data were collected with two biological replications each with two technical replications.

An inferred interaction network with a subset of 46 gene nodes plus an HMF node is depicted in Fig. 1. Based on ANOVA and cluster analysis, 46 significantly induced expressed genes by the HMF treatment were selected and used for the prototype computation modeling development. This network model captured temporal dependencies among the 46 genes and HMF during the earlier exposure to the inhibitor in yeast fermentation process. The system model underlying the network is an optimal solution after searching all possible directed graphs with 47 nodes, except that the HMF node is not allowed to have incoming edges and the maximum number of incoming edges for a gene node is at most 5. Existence of an edge from YAP1 to DDI1 indicates a temporal dependency of the rate of change in DDI1 expression on the mRNA level of YAP1. The number 1.2e-07, positioned next to the edge, is the p-value of this temporal dependency. The original system matrix was stabilized by scaling all eigenvalues by the spectral norm 3.09. The overall p-value, 1.6e-5, of the entire system model indicates that the model is statistically significant. The p-value is based on a stringent standard and the resulting model has high levels of consistency with biological observations because the probability of the model arising by chance is as low as 1.2e-07.

Among three known transcription factors, PDR1, PDR3, and YAP1, in this subset of genes, YAP1 was shown as one of the most influential regulators as demonstrated by this model in earlier response to the HMF stress for ethanol production (Fig. 1). This is strongly supported by current knowledge and documented experimental observations (Teixeira et al., 2006). For example, the following edges have been reported as transcriptional regulations including YAP1 to DDI1 (Haugen et al., 2004), YAP1 to ATM1 (Haugen et al., 2004), YAP1 to GRE2 (Lee et al., 1999), YAP1 to SNQ2 (Lee et al., 2002; Lucau-Danila et al., 2005), and YAP1 to TPO1 (Lucau-Danila et al., 2005). Four more edges from YAP1 demonstrated enhancement to SCS7, PDR1, PDR11, and HIS3, suggesting regulatory rules of YAP1 to these genes (Fig. 1). According to YEASTRACT, SCS7, PDR1, PDR11, and HIS3 are considered potential transcriptional regulatees of YAP1 based on sequence motifs (YEA, 2006). In addition, transcriptional factor PDR3 showed regulatory rule to RSB1 as demonstrated in this model, which is in agreement with and supported by previous documented observations (Devaux et al., 2002). It also showed enhancement to SAM3, ATM1,

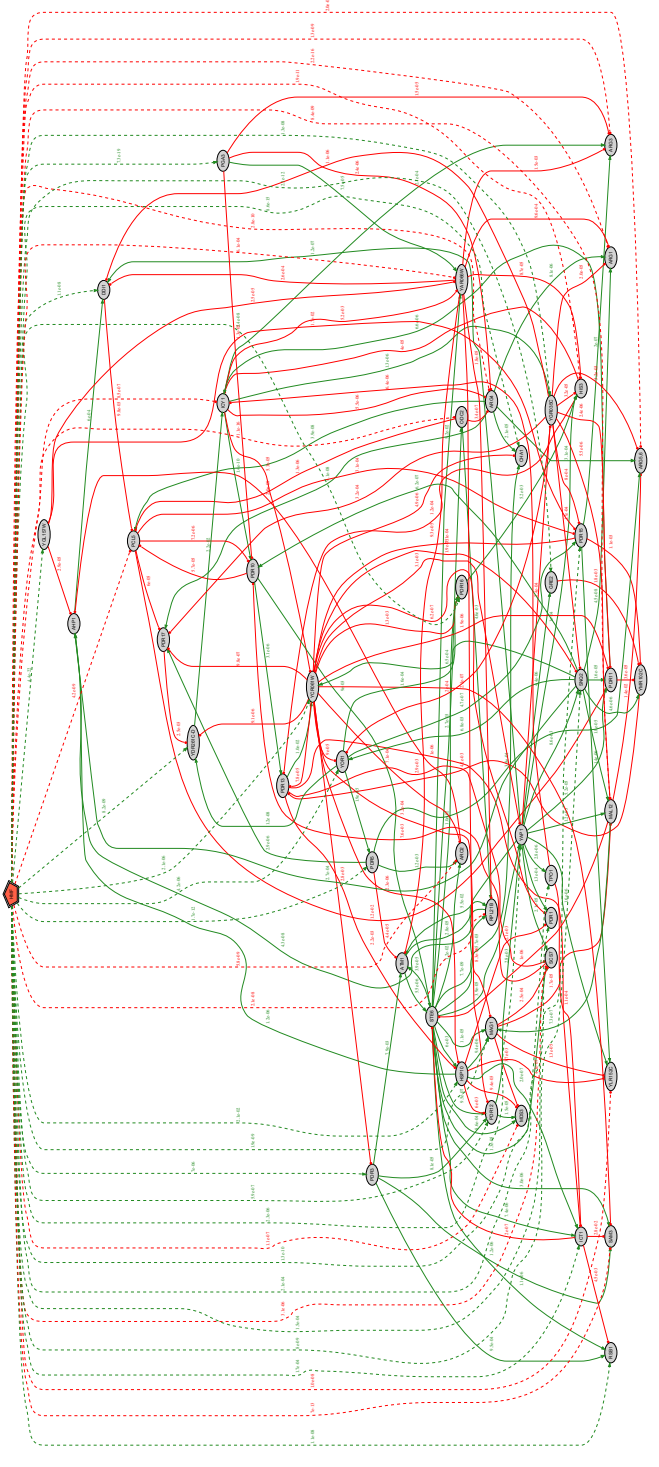

Fig. 1. Temporal interactions for a subset of 46 genes in response to HMF for biomass conversion to ethanol by ethanologenic yeast. The p-values of each edge are displayed. A solid directed edge in green from the first gene node to the second gene node with an arrowhead indicates enhancement of the second gene by the first gene; An edge in red from the first gene node to the second gene node with a solid dot indicates repression of the second gene by the first gene. The dashed edges represent the external influence from HMF to each gene: red for repressing and green for enhancing. The graph is rendered by the software GraphViz (available from www.graphviz.org).

and PDR12. It is very encouraging that the gene regulatory network model developed in this study is highly consistent with the current knowledge including documented experimental observation and sequence motif based analysis. More significantly, the model demonstrated in this study showed statistical significance on the temporal dependencies.

This system model also presented numerous interesting network interactions among genes with potential significance. For example, STE6, SNQ2, ARG4 and YOR1 significantly enhanced directly or indirectly 15, 8, 5, and 4 other genes, respectively. These genes have been observed to be core stress response genes and many related genes are observed to be interested to cope with the HMF stress for survival. Resolution of such interactions could have a significant impact to understand the mechanism of detoxification and the stress tolerance caused by HMF. Although they have not been reported, such statistically significant gene interactions presented by this model could be potentially biologically significant to predict unknown gene interaction networks. With the high consistency between the model network on YAP1 presented in this study and current knowledge, it is reasonable to assume potential relationships presented in this model with significant p-values. However, a common transcription factor PDR1 did not show significant regulatory rule to the selected subset genes in this model. We need to examine it further using biological experiment. Although it is highly homologous with PDR3, PDR1 does not always respond the same with PDR3.

Another impact of the system model is to prescribe desired system behaviors by applying perturbation to the system. A perturbation can be changing the concentration level of the inhibitor HMF, silencing of a subset of genes in the network, or mutating of a subset of genes. To increase the tolerance to the inhibitor HMF, one can consider adjusting the influential genes to achieve an effect similar to the transcriptome profile observed in the absence of HMF. In Fig. 1, the following genes were identified as potential significant elements in gene interaction networks for detoxification and HMF stress tolerance: STE6 (15/46), YCR061W (14/46), YAP1 (12/46), YGR035C (10/46), SNQ2 (8/46), HSP10 (7/46), and YAR066W (7/46). By perturbing these major regulators, one will exert the most control over expressions of other genes, which might be economically desirable.

Another strategy to genetic engineering for wild type yeast to become tolerant to the inhibitor HMF is to study the system model of HMF resistant yeast strains. Preliminary tolerant strains for *in situ* detoxification of the inhibitors have been developed (Liu et al., 2004; Liu and Slininger, 2005; Liu et al., 2005; Palmqvist et al., 1999; Wahbom and Hahn-Hägerdal, 2002). By comparing the system models for the wild type and the tolerant yeast strain, one can identify those genes that behave differently between the two strains. Those different genes can be the targets of genetic engineering for the wild type strains to become HMF tolerant.

Figures 2 to 3 show how well the model fits the observed trajectory data from the 46 genes. The model is able to capture trends in the data precisely such as ARG1, ICY1, MDS3, TPO1, and YCR061W.

Figure 4 demonstrates the prediction made by the model how the time courses evolve differently when the same sample is subject to different experimental conditions.

Figures 2 to 3, each corresponding to a different sample, show how well the model fits the observed data from the 46 genes. In these figures, the original time course sample, the log-time interpolated data, and the fitted time course by the model are illustrated. The model captured the trend in the data precisely for genes such as ARG1, ICY1, MDS3, TPO1, and YCR061W, given the large sample variation present in most microarray experiments. We are primarily interested in detecting significant interactions that can be captured by the capability of linear discrete dynamic system model. The poor fits suggest that there might be nonlinear interactions in addition to the linear interactions, which we plan to address in the future work.

Based on the estimated coefficients in above tables, simulations are performed to evaluate the effectiveness of the estimated difference equations. Figure 4 demonstrates the prediction made by the model how the time courses evolve differently when the same sample is subject to different experimental conditions. It can be observed that the influential gene nodes in Fig. 1 evidently exhibit sharper transitions in the time course than the non-influential genes. The presence of HMF has significantly influenced all the selected genes. However, the effect takes on different courses. Some genes have been enhanced such as ICT1, while staying on similar curvatures; some genes are repressed severely such as TPO1; other genes show opposite transitions such as HIS3.

A strong temporal dependency of gene X on gene Y can indicate a transcriptional regulation from Y to X. However, a real transcriptional regulation from transcription factor W to Z may not show up as a temporal dependency of Z on W due to other factors involved in the expression of W. It is possible that W can have a high concentration of mRNA, but somehow the translation of W mRNA to its protein product is blocked by the presence of other regulatory proteins during translation. Therefore, the mRNA concentration of Z will be low due to the scarcity of the protein product of its transcription factor W. No temporal dependencies of Z on W can be possibly established in such a scenario. It is also plausible that a temporal dependency does not equate to a real transcriptional regulation: Two genes S and T can co-express in similar patterns and only one of them S is a real transcription factor of a third gene R. Although it is unlikely that two genes have identical expression patterns as the nature of biology tends to go parsimony, the measurement may not discern a difference that is below the noise level, which can be high in current microarray technology. These limitations can be overcome when proteome measurements are available and uncertainty in measurements is reduced. They are not inherent problems of discrete dynamic system models.

The methodology presented in this paper can be applied to the analysis of a network from data sets that contain both transcriptome and proteome measured simultaneously on the same sample. With such data sets that encapsulate complete snapshots of molecular processes during bioethanol conversion, the

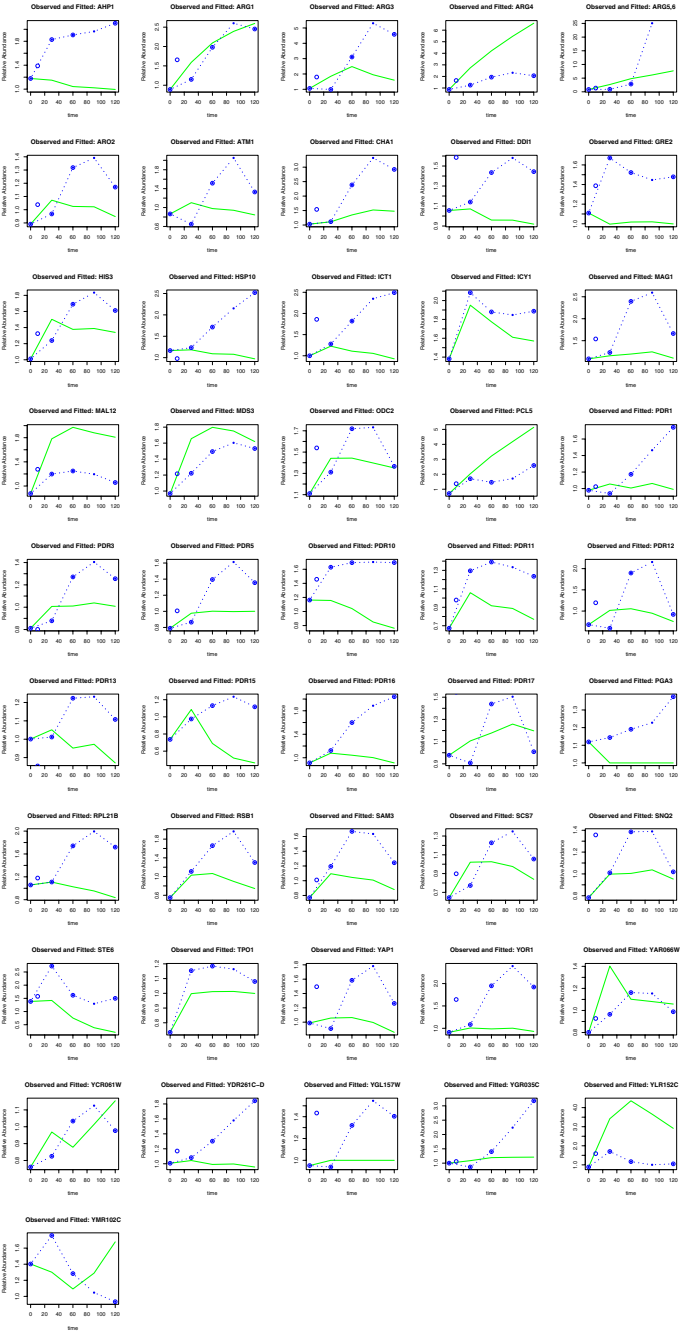

Fig. 2. Sample 1. Control: Not exposed to HMF. Fitted gene expression time courses (green solid lines) from the model versus the observed ones (blue dotted lines). The big open blue circles represent the original values; the small solid blue circles are interpolated values actually used.

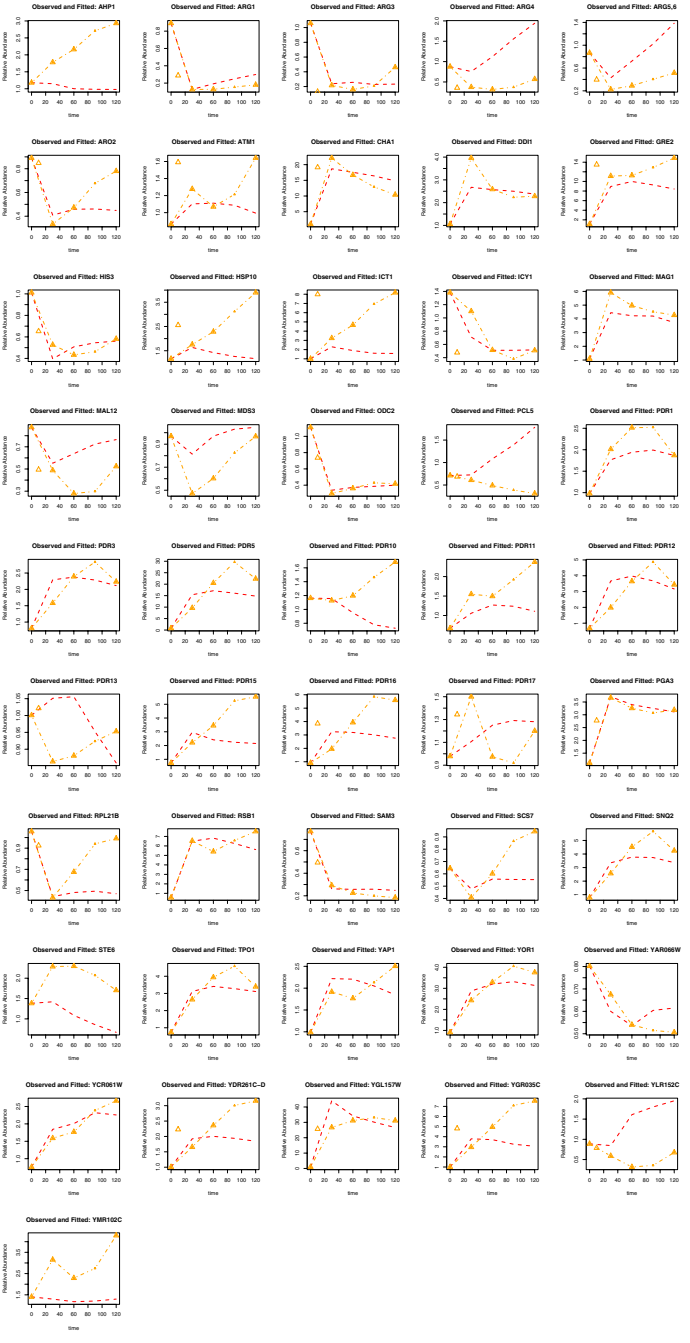

Fig. 3. Sample 2. Treatment: Exposed to HMF. Fitted gene expression time courses (red dashed lines) from the model versus the observed ones (yellow dash-dotted lines). The big open yellow triangles represent the original values; the small filled yellow triangles are interpolated values for model estimation.

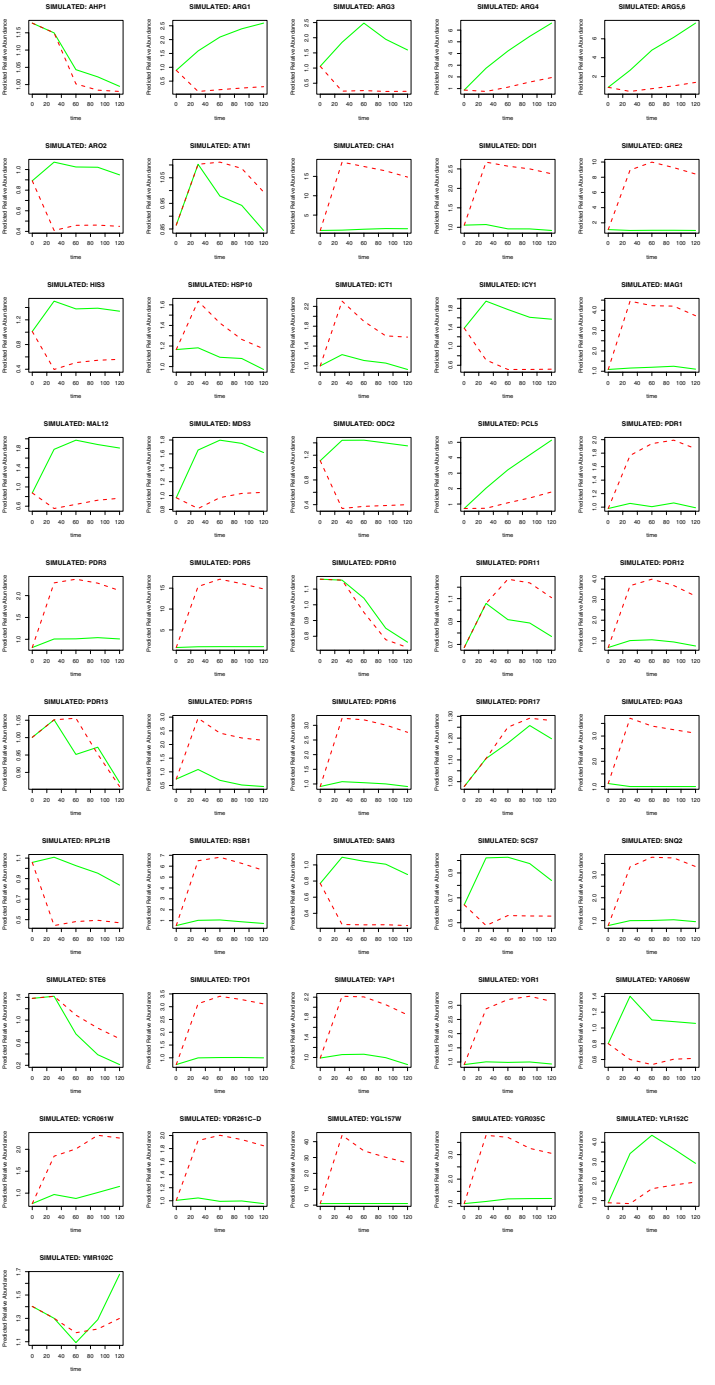

Fig. 4. Predictions of mRNA expression time courses of the 46 genes: HMF in absence (green solid lines) versus HMF in presence (red dashed lines)

temporal dependencies depicted by our approach will be able to provide a more accurate account of the genomic mechanism on inhibitor detoxification and tolerance in ethanologenic yeast.

3 Materials and Methods

3.1 Microarray Experiments, Data Preprocessing, and Gene Expression Analysis

Target genome microarray of *Saccharomyces cerevisiae* was fabricated using GeneMachine OmniGrid 300 microarrayer robot. A recent version of 70-mer oligo set representing 6,388 genes was applied and Codelink activated slides were used. DNA oligo samples were resuspended in 150 mM of sodium phosphate printing buffer (pH 8.5) at a final concentration of 20 M probes for printing. Each genome microarray was designed with two replications on one slide. Each microarray slide consisted of 13,000 elements including target genes and spiking-in quality controls for linear dynamic calibration, ratio reference, DNA sequence background, and slide background controls. The first developed universal external RNA control was applied in microarray experiments (Liu and Slininger, 2006b). The universal quality control consisted of six unique RNA transcripts that can be applied to different assay platforms of microarray and real time quantitative RT-PCR, including SYBR Green and TaqMan probe-based chemistry. It was demonstrated that the signal intensity detected from these controls are independent from cell treatment of stress or environmental conditions in a host RNA background. Highly fitted linearity and dynamic ranges provided a basis for estimation of mRNA abundance in gene expression analysis. Such external RNA controls provide an unbiased normalization reference, valid dynamic range of linearity, and estimate of variations of microarray experiments. It guards reliability and reproducibility of expression data and also makes it possible to compare data derived from different experiments and different assay platforms for data verification and confirmation. Each of the control elements in each array had 48 replications and was distributed evenly in each block of the microarray. A mini array consisting of the controls and two background controls was designed on top of a target genome array with 16 replications. This mini array served as a reference to adjust PMT Gain balance of GenePix 4000B scanner for two dye channels prior to a scanning of the entire target array for data acquisition.

Ethanologenic yeast *S. cerevisiae* NRRL Y-12632 was used and maintained lyophilized in the ARS Culture Collection, National Center for Agricultural Utilization Research, USDA, Peoria, IL. Yeast cultures were incubated on a synthetic complete medium for 6h prior to a treatment by HMF (30 mM) as previously described (Liu et al., 2004). Briefly, HMF was added to the medium in a fleaker fermentation system at 30℃. A set of gene expression profiles derived from a yeast culture grown under the same conditions without the HMF treatment served as a control. The time point of inhibitors added was designated as hour 0. Yeast cells were harvested periodically starting from 0h, 10 min, 30 min, 2h and 4h. Cell samples were harvested by centrifugation at 25℃ and immediately

frozen and stored at −80℃ until use. Total RNA was isolated using a protocol based on Schmitt et al. (1990) with modifications. The RNA was further purified using a nucleic acid purification column. RNA probe was labeled using an indirect dUTP Cy3 or Cy5 labeling procedure based on Hegde et al. (2000) with modifications. Microarray slide was scanned and data acquisition obtained using GenePix 4000B scanner and GenePix Pro software. Pre-scan control was used to adjust PMT Gain against Cy3 and Cy5 channels and ratios of signal intensities between Cy3 and Cy5 were balanced to 1 using the calibration controls determined using the mini-array. Microarray data were analyzed using GeneSpring program. Control gene CtrlGm_5 was used as normalization reference for each gene. Median of foreground signal intensity subtracted by background for each dye channel was used. Data were filtered between each dye channel and among multiple microarray experiments. A gene list shared by all microarray experiments was generated and used for data analysis. ANOVA was performed to identify genes significantly expressed in comparison with the control. Based on expression patterns, subsets of gene lists were generated by self-organizing map and cluster analysis.

3.2 Log-Time Interpolation

Non-uniform time sampling is often used in a time course experimental design, such that high frequency components in the original continuous signal can be preserved. Conversely, interpolation in the original time domain over non-uniform samples tends to distort high frequency components in the original signal. To save sharp transitions at densely sampled time locations, we apply a logarithm transform on time by

$$t' = \log(t + t_0)$$

where t' is the time variable in the log-time domain. Selection of the constant t_0 is determined by how well it equalizes the distance between each consecutive pair of time points after the log-time transform. The observed samples are then interpolated by cubic splines in the log-time domain, by assuming that the sampling times are designed sufficiently well to capture major change of the signals; or equivalently, the change of gene expression levels between two consecutive time points can be captured by the cubic splines. Let $x = f(t')$ be the interpolated cubic spline. One can obtain values at equally spaced time points 0, h, $2h$, …, kh, …, in the original time domain by

$$x_k = f(\log(kh + t_0))$$

where h is the sampling interval. We pick the same number of interpolated points as the number of points in the original data set. So the interpolation solely serves to equalize the non-uniform time points in the log-time domain. If more points were interpolated, the p-value must be adjusted to that effect, otherwise, faulty significance might arise. The discrete dynamic system model will be fitted to the interpolated values in the original time domain.

3.3 First-Order Linear Discrete Dynamic System Model

Although dynamics in molecular processes are largely nonlinear – reflected by various nonlinear kinetics models, the number of observations sufficient to induce a nonlinear model for a biological system is too large to be practical for a system with more than a handful of variables. Instead of nonlinear models, we use the first-order linear discrete dynamic model to capture the linear effect of a system. A system can only be considered linear when the perturbation to the system is sufficiently small. A large perturbation could lead the system out to another state of linearity. In our experiment, the time points we collected reflected the initial response of gene expressions to the inhibitor HMF before major dramatic dynamic effect takes place. We consider the linear discrete dynamic system model can approximate primary expression response to HMF.

In a first-order linear discrete dynamic system model, the transition from one state at discrete time t to the next state at $t+1$ depends linearly on the state of the system at time t. Let h be the constant time span of 1 unit of discrete time. First order refers to the transition from t to $t+1$ does not depend on the state of the system at $t-1$, $t-2$, and so on, except the state at t. Let $\mathbf{g}[t] = [g_1[t], g_2[t], \ldots, g_N[t]]^T$ be a vector of the expression levels of N genes at time t. Let $\mathbf{e}[t] = [e_1[t], e_2[t], \ldots, e_K[t]]^T$ be a vector of the strength of K external signals at time t. A first-order linear discrete dynamic system model can be written as

$$\mathbf{g}[t+1] - \mathbf{g}[t] = h\ \{A\ \mathbf{g}[t] + B\ \mathbf{e}[t]\} + \epsilon[t] \tag{1}$$

where $A = \{a_{i,j}\}$ is an $N \times N$ system matrix and $a_{i,j}$ $(i \neq j)$ is the influence of gene j on gene i, $a_{i,i}$ is the self-control rate, $B = \{b_{i,k}\}$ is an $N \times K$ influence matrix where $b_{i,k}$ is the influence of the k-th signal on gene i, $\epsilon[t] = [\epsilon_1[t], \epsilon_2[t], \ldots, \epsilon_N[t]]^T$ is a vector of noise levels to each gene at time t. The noise is estimated by fitting the linear discrete dynamic system model, and thus is a function of the time interval as well as the observed data. In the modeling process, we assume the noise model Gaussian. We also introduce a possible intercept vector I to the right hand side of the above equation during model selection for each node.

Solving the Linear Difference Equations. From the experiments under different conditions, one can collect M time course observations or trajectories of the system at the discrete time points $0, 1, 2, \ldots, T$. Let $\mathbf{g}^m[0]$, $\mathbf{g}^m[1]$, …, $\mathbf{g}^m[T]$ $(m = 1 \ldots M)$ be all the observed system states, and $\mathbf{e}^m[0], \mathbf{e}^m[1], \ldots, \mathbf{e}^m[T]$ be all the external stimulus applied to the system. We use the least squares to find optimal estimates of system matrix A and influence matrix B. The system model defined in Eq. (1) can be written as a collection of all M observations by

$$\mathbf{g}^m[t+1] - \mathbf{g}^m[t] = h\ \{A\ \mathbf{g}^m[t] + B\ \mathbf{e}^m[t]\} + \epsilon^m[t]$$

where

$$\mathbf{g}^m[t] = \begin{pmatrix} g_1^m[t] \\ g_2^m[t] \\ \vdots \\ g_N^m[t] \end{pmatrix}, \ \mathbf{e}^m[t] = \begin{pmatrix} e_1^m[t] \\ e_2^m[t] \\ \vdots \\ e_K^m[t] \end{pmatrix}, \ \text{Noise: } \epsilon^m[t] = \begin{pmatrix} \epsilon_1^m[t] \\ \epsilon_2^m[t] \\ \vdots \\ \epsilon_N^m[t] \end{pmatrix}$$

The above formulation can be rearranged into a multiple linear regression form

$$\mathbf{g}^m[t+1] = (hA + I)\mathbf{g}^m[t] + hB\mathbf{e}^m[t] + \epsilon^m[t]$$

Equivalently, for each gene node, we have

$$g_i^m[t+1] = \left[\sum_{j=1}^{N}(ha_{ij} + I(i=j))g_j^m[t]\right] + \left[\sum_{k=1}^{K} hb_{ik}e_k^m[t]\right] + \epsilon_i^m[t]$$

Let $\mathbf{a}_i = (a_{i1}, \ldots, a_{ik})^\top$ and $\mathbf{b}_i = (b_{i1}, \ldots, b_{ik})^\top$ be the parameters associated with gene node i. Thus, $\mathbf{a}_i$ and $\mathbf{b}_i$ can be solved independently of the other nodes. By least squares, optimal estimates for $\mathbf{a}_i$ and $\mathbf{b}_i$ are

$$\begin{aligned}\mathbf{b}_i = &\left[2h^2 \sum_{m=1}^{M}\sum_{t=1}^{T-1} \mathbf{e}^m[t]\mathbf{e}^m[t]^\top \right. \\ &- \left(2h^2 \sum_{m=1}^{M}\sum_{t=1}^{T-1} \mathbf{e}^m[t]\mathbf{g}^m[t]^\top\right)\left(2h^2 \sum_{m=1}^{M}\sum_{t=1}^{T-1} \mathbf{g}^m[t]\mathbf{g}^m[t]^\top\right)^{-1}\left(2h^2 \sum_{m=1}^{M}\sum_{t=1}^{T-1} \mathbf{g}^m[t]\mathbf{e}^m[t]^\top\right)\Bigg]^{-1} \\ &\left[\left(2h \sum_{m=1}^{M}\sum_{t=1}^{T-1} (g_i^m[t+1] - g_i^m[t])\mathbf{e}^m[t]\right)\right. \\ &- \left(2h^2 \sum_{m=1}^{M}\sum_{t=1}^{T-1} \mathbf{e}^m[t]\mathbf{g}^m[t]^\top\right)\left(2h^2 \sum_{m=1}^{M}\sum_{t=1}^{T-1} \mathbf{g}^m[t]\mathbf{g}^m[t]^\top\right)^{-1}\left.\left(2h \sum_{m=1}^{M}\sum_{t=1}^{T-1} (g_i^m[t+1] - g_i^m[t])\mathbf{g}^m[t]\right)\right]\end{aligned} \tag{2}$$

$$\mathbf{a}_i = \left(2h^2 \sum_{m=1}^{M}\sum_{t=1}^{T-1} \mathbf{g}^m[t]\mathbf{g}^m[t]^\top\right)^{-1}\left[\left(2h \sum_{m=1}^{M}\sum_{t=1}^{T-1} (g_i^m[t+1] - g_i^m[t])\mathbf{g}^m[t]\right) - \left(2h^2 \sum_{m=1}^{M}\sum_{t=1}^{T-1} \mathbf{g}^m[t]\mathbf{e}^m[t]^\top\right)\mathbf{b}_i\right] \tag{3}$$

Selecting the Most Significant Linear Difference Equation for Each Gene Node. For each gene node, the more variables involved in the difference equation for that node, the better the fit. However, statistical significance starts to drop once a maximal complexity has reached to a point that the sample does not support more variables to be involved. Thus, we select the best subset of potential regulators for each gene node such that the corresponding linear difference equation yields the most statistically significant fit. The statistical significance is determined by the F-test. For N genes and K external signals, there are 2^{N+K-1} possible subsets to consider, which is computationally feasible only for a network with less than a dozen of nodes. We limit the number of possible incoming edges or potential regulators for each node to some computational

doable number. Although this lead to an incomplete exploration of the system search space, our experience indicates that major influential gene nodes can be identified even when the number of regulator nodes explored is small.

Stabilization. Although solutions to the linear difference equations constitute an optimal fit to the observed data, the resulted system can be unstable, meaning that the log expression levels of some genes increase to infinity or decrease to negative infinity as time goes on when the initial state of the system is finite. Thus we stabilize the system model when no external stimuli are present.

Now we derive the stabilization formula. Equivalently, Eq. (1) can be written as

$$\mathbf{g}[t+1] = (hA + I)\mathbf{g}[t] + hB\ \mathbf{e}[t] + \epsilon[t] \tag{4}$$

When the system is not subject to external stimulus or noise, it becomes

$$\mathbf{g}[t+1] = (hA + I)\mathbf{g}[t] \tag{5}$$

In the bioethanol conversion process, this system equation describes the ideal behavior of the yeast gene expression without the inhibitor HMF in a zero-noise environment. In such a system, one does not expect the expression of any gene becomes unstable during the experiment since otherwise the subject perishes. An optimal solution found for A by Eq. (3) might lead to an unstable system in Eq. (5). Let $W = hA + I$. A necessary and sufficient condition for the system described by Eq. (5) to be stable is to require W to be power stable – all eigenvalues of W must be located within or on the unit circle; or the spectral norm must be no greater than one. Let $\lambda(W)$ be the sequence of eigenvalues of W. The spectral norm $\rho(W)$ is defined by (Golub and van Loan, 1996)

$$\rho(W) = \max\{|\lambda| : \lambda \in \lambda(W)\}$$

Let Λ be a diagonal matrix $\mathrm{diag}(\lambda(W))$ and V be a matrix whose columns are the eigenvectors in an order corresponding to the order of eigenvalues in $\lambda(W)$. It follows that

$$W = V\Lambda V^{-1}$$

We stabilize W to W_s by scaling all its eigenvalues by its spectral norm if the spectral norm is greater than 1, while maintaining the same eigenvectors, that is,

$$W_s = \begin{cases} V\dfrac{\Lambda}{\rho(W)}V^{-1} = \dfrac{1}{\rho(W)}W & \text{if } \rho(W) > 1 \\ W & \text{otherwise} \end{cases} \tag{6}$$

Let A_s be the transformed matrix A after stabilization. Plugging in the definition of W, we obtain

$$A_s = \frac{1}{h}\left[\frac{hA + I}{\rho(hA + I)} - I\right]$$

if the spectral norm of W is greater than 1. Replacing A by A_s in Eq. (1), we obtain

$$\mathbf{g}[t+1] - \mathbf{g}[t] = h \left\{ \frac{1}{h} \left[\frac{hA + I}{\rho(hA + I)} - I \right] \mathbf{g}[t] + B\, \mathbf{e}[t] \right\} + \epsilon[t] \tag{7}$$

There are several theoretical and numerical properties associated with our stabilization strategy. It is evident that any coefficients off the diagonal line in A with a value close to 0 will be closer to 0 after stabilization. This ensures that no new interactions between different genes will be introduced by stabilization. The spectral norm can be found efficiently using the power method without obtaining the entire eigenvalues or eigenvectors of matrix W. In addition, since there is no matrix decomposition involved, the stabilized matrix A_s will be real if A is real, which holds true theoretically but could be violated numerically by other approaches.

3.4 Statistical Significance of a Discrete Dynamic System Model

Let the minimum p-value of fitting a linear difference equation to gene i be p_i. The p-value of an entire fitted discrete dynamic system model is computed by

$$\text{p-value} = 1 - \prod_{i=1}^{N} (1 - p_i)$$

where p_i is computed by the F-tests during the fitting of linear model for gene node i. This defines a conservative p-value since it assumes that the mRNA levels are independent to each other. Nevertheless, the p-value of a network is a statistically effective and computationally efficient measure to determine the chance an estimated system would arise randomly. This p-value is influenced by 1) how well each linear difference equation can be fitted to the data and 2) the number of nodes in the network, which constitute two competing factors. Our algorithm minimizes the p-value by trade-off between both factors.

3.5 Implementation and Modeling Details

The network modeling software is written in the R programming language (R Development Core Team, 2006). For the modeling of the network of 46 genes shown in Fig. 1, it took about 24 hours on 12 networked computers (Sun Java Workstation w1100z, Opteron 150 processor, 2.4 GHz clock frequency 1 GB memory, running 64-bit SuSE Linux [version 10]). The maximum number of potential regulators including the HMF was set to 5 during the system model construction.

Acknowledgement. This study was supported in part by NRI Competitive Grant Program project #ILLR- 2006-02272.

References

YEAst Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT), January 2006. URL http://www.yeastract.com. Last Date of Visit: (September 12, 2006)

Akutsu, T., Kuhara, S., Maruyama, O., Miyano, S.: Identification of genetic networks by strategic gene disruptions and gene overexpressions under a Boolean model. Theoretical Computer Science 298(1), 235–251 (2003)

Bonneau, R., Reiss, D.J, Shannon, P., Facciotti, M., Hood, L., Baliga, N.S, Thorsson, V.: The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biology 7(5), R36 (2006)

Bothast, R., Saha, B.: Ethanol production from agricultural biomass substrate. Adv. App. Microbiol. 44, 261–286 (1997)

Devaux, F., Carvajal, E., Moye-Rowley, S., Jacq, C.: Genome-wide studies on the nuclear PDR3-controlled response to mitochondrial dysfunction in yeast. FEBS Letters 515(1-3), 25–28 (2002)

D'haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R.: Linear modeling of mRNA expression levels during CNS development and injury. In: Pacific Symposium on Biocomputing, pp. 41–52. World Scientific Publishing Co, Singapore (1999)

Edelstein-Keshet, L.: Mathematical Models in Biology. SIAM (2004)

Friedman, N.: Inferring cellular networks using probabilistic graphical models. Science 303, 799–805 (2004)

Golub, G.H., van Loan, C.F.: Matrix Computations, 3rd edn. The Johns Hopkins University Press, Baltimore, MD (1996)

Haugen, A.C., Kelley, R., Collins, J.B., Tucker, C.J., Deng, C., Afshari, C.A., Brown, J.M., Ideker, T., Van Houten, B.: Integrating phenotypic and expression profiles to map arsenic-response networks. Genome Biology 5(12), R95 (2004)

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Earle-Hughes, J., Snesrud, E., Lee, N., Quackenbush, J.: A concise guide to cdna microarray analysis. BioTechniques 29, 548–562 (2000)

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., Miyano, S.: Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. Journal of Bioinformatics and Computational Biology 1(2), 231–252 (2003)

Lee, J., Godon, C., Lagniel, G., Spector, D., Garin, J., Labarre, J., Toledano, M.B.: Yap1 and Skn7 control two specialized oxidative stress response regulons in yeast. J. Biol Chem. 274(23), 16040–16046 (1999)

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A.: Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298(5594), 763–764 (2002)

Liang, S., Fuhrman, S., Somogyi, R.: REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Pacific Symposium on Biocomputing 3, 18–29 (1998)

Liu, Z.L.: Genomic adaptation of ethanologenic yeast to biomass conversion inhibitors. Appl. Microbiol. Biotech. 73, 27–36 (2006)

Liu, Z.L., Slininger, P.J.: Development of genetically engineered stress tolerant ethanologenic yeasts using integrated functional genomics for effective biomass conversion to ethanol, CAB International, Wallingford, UK, pp. 283–294 (2005)

Liu, Z.L., Slininger, P.J.: Transcriptome dynamics of ethanologenic yeast in response to 5-hydroxymethylfurfural stress related to biomass conversion to ethanol. In: Recent Research Developments in Multidisciplinary Applied Microbiology: Understanding and Exploiting Microbes and Their Interactions-Biological, Physical, Chemical and Engineering Aspects, pp. 679–684. Wiley-VCH, Chichester (2006a)

Liu, Z.L., Slininger, P.J.: Universal external RNA controls for microbial gene expression analysis using microarray and qRT-PCR. J. Microbiol. Methods, doi:10.1016/j.mimet.2006.10.014 (2006b)

Liu, Z.L., Slininger, P.J., Dien, B.S., Berhow, M.A., Kurtzman, C.P., Gorsich, S.W.: Adaptive response of yeasts to furfural and 5-hydroxymethylfurfural and new chemical evidence for HMF conversion to 2,5-bis-hydroxymethylfuran. J Ind. Microbiol Biotechnol. 31, 345–352 (2004)

Liu, Z.L., Slininger, P.J., Gorsich, S.W.: Enhanced biotransformation of furfural and 5-hydroxy methylfurfural by newly developed ethanologenic yeast strains. Appl Biochem Biotechnol. 121-124, 451–460 (2005)

Lucau-Danila, A., Lelandais, G., Kozovska, Z., Tanty, V., Delaveau, T., Devaux, F., Jacq, C.: Early expression of yeast genes affected by chemical stress. Mol. Cell Biol. 25(5), 1860–1868 (2005)

Luo, C., Brink, D., Blanch, H.: Identification of potential fermentation inhibitors in conversion of hybrid poplar hydrolyzate to ethanol. Biomass Bioenergy 22, 125–138 (2002)

Martin, C., Jonsson, L.: Comparison of the resistance of industrial and laboratory strains of Saccharomyces and Zygosaccharomyces to lignocellulose-derived fermentation inhibitors. Enzy. Micro. Technol. 32, 386–395 (2003)

Meir, E., Munro, E.M., Odell, G.M., von Dassow, G.: Ingeneue: A versatile tool for reconstituting genetic networks, with examples from the segment polarity network. Journal of Experimental Zoology 294, 216–251 (2002)

Ong, I.M., Glasner, J.D., Page, D.: Modelling regulatory pathways in E. coli from time series expression profiles. Bioinformatics 18, S241–S248 (July 2002)

Pal, R., Ivanov, I., Datta, A., Bittner, M.L., Dougherty, E.R.: Generating Boolean networks with a prescribed attractor structure. Bioinformatics 21, 4021–4025 (November 2005)

Palmqvist, E., Almeida, J., Hahn-Hägerdal, B.: Influence of furfural on anaerobic glycolytic kinetics of Saccharomyces cerevisiae in batch culture. Biotechnol Bioeng 62, 447–454 (1999)

R Development Core Team.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, `http://www.R-project.org` (2006)

Saha, B.: Hemicellulose bioconversion. Journal of Industrial Microbiology and Biotechnology 30, 279–291 (2003)

Schlitt, T., Brazma, A.: Modelling in molecular biology: describing transcription regulatory networks at different scales. Philosophical Transactions of the Royal Society B: Biological Sciences 361(1467), 483–494 (March 2006)

Schmitt, M.E., Brown, T.A., Trumpower, B.L.: A rapid and simple method for preparation of RNA from Saccharomyces cerevisiae. Nucl. Acid Res. 18, 3091–3092 (1990)

Shmulevich, I., Dougherty, E.R., Kim, S., Zhang, W.: Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18, 261–274 (February 2002)

Taherzadeh, M., Gustafsson, L., Niklasson, C.: Physiological effects of 5-Hydroxymethylfurfural on Saccharomyces cerevisiae. App. Microbiol. Biotechnol. 53, 701–708 (2000)

Takahashi, K.: Multi-algorithm and multi-timescale cell biology simulation. PhD thesis, Keio University, Fujisawa, Japan (2004)

Takahashi, K., Arjunan, S.N.V., Tomita, M.: Space in systems biology of signaling pathways – towards intracellular molecular crowding in silico. FEBS Letters 579, 1783–1788 (2005)

Teixeira, M.C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A.R., Mira, N.P., Alenquer, M., Freitas, A.T., Oliveira, A.L., Sá-Correia, I.: The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. Nucl. Acids Res. 34, D446–451 (2006)

Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T.S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., Hutchison III, C.A.: E-CELL: software environment for whole-cell simulation. Bioinformatics 15(1), 72–84 (1999)

van Kampen, N.: Stochastic Processes in Physics and Chemistry. Elsevier, Amsterdam (1997)

Wahbom, C.F., Hahn-Hägerdal, B.: Furfural, 5-hydroxymethylfurfrual, and acetone act as external electron acceptors during anaerobic fermentation of xylose in recombinant Saccharomyces cerevisiae. Biotechnol Bioeng. 78, 172–178 (2002)