Journal of Bioinformatics and Computational Biology
Vol. 5, No. 1 (2007) 31–46

Imperial College Press
www.icpress.co.uk

STATE SPACE MODELING OF YEAST GENE EXPRESSION DYNAMICS

OLLI HAAVISTO* and HEIKKI HYÖTYNIEMI†

Control Engineering Laboratory, Helsinki University of Technology
PO Box 5500, FI-02015 TKK, Finland
**olli.haavisto@tkk.fi*
†*heikki.hyotyniemi@tkk.fi*

CHRISTOPHE ROOS

Medicel Ltd, Huopalahdentie 24
FI-00350 Helsinki, Finland
christophe.roos@medicel.com

Received 20 February 2006
Revised 2 June 2006
Accepted 11 October 2006

Combined interaction of all the genes forms a central part of the functional system of a cell. Thus, especially the data-based modeling of the gene expression network is currently one of the main challenges in the field of systems biology. However, the problem is an extremely high-dimensional and complex one, so that normal identification methods are usually not applicable specially if aiming at dynamic models. We propose in this paper a subspace identification approach, which is well suited for high-dimensional system modeling and the presented modified version can also handle the underdetermined case with less data samples than variables (genes). The algorithm is applied to two public stress-response data sets collected from yeast *Saccharomyces cerevisiae*. The obtained dynamic state space model is tested by comparing the simulation results with the measured data. It is shown that the identified model can relatively well describe the dynamics of the general stress-related changes in the expression of the complete yeast genome. However, it seems inevitable that more precise modeling of the dynamics of the whole genome would require experiments especially designed for systemic modeling.

Keywords: Gene expression; microarray data; dynamic modeling; subspace identification; state space model.

1. Introduction

Activities of genes in a biological cell can be analyzed using microarray measurements, which actually give a snapshot of the current mRNA concentrations in the cell. It can be assumed that these concentrations are directly proportional to

*Corresponding author.

the activity levels of the corresponding genes. The amount of microarray data is continuously increasing and there is a constant need for more powerful data analyzing tools. Consequently, many studies have been published that typically aim either at classification, that is, detection of groups of similarly behaving genes, or at derivation of regulation networks showing the connections between different genes or gene groups. More generally the different modeling approaches can be divided into static[1–3] and dynamic[4–10] ones depending on the purpose of the analysis.[11]

When considering all relevant genes of an organism, the main challenge of dynamic modeling is that the identification problem is highly underdetermined. The number of genes is much higher than the length of typical time series, hence no unique solution to the problem exists. As listed by Guthke *et al.*,[12] several alternative approaches deal with this dimensionality problem. The most interesting and well-justified of these approaches assumes that the actual gene expression machinery is driven by a small group of latent variables. Holter *et al.*[13] call these variables "characteristic modes" and calculate them as principal component vectors of the original data. They show that different gene expression time series data sets can be reproduced using only a couple of the most important principal components. Also the dynamics of the principal components can then be studied.[14] Wu *et al.*[15] further refine this approach by forming the internal variables using an extension of the principal component analysis. These "eigengenes" are interpreted as the state variables of a modified version of the basic linear state space model. However, it turns out that the estimated model is not stable.

A typical feature for previously used dynamic models is that the purpose of the modeling is more or less to create a (static) connectivity matrix, which should tell the regulative connections between different genes. Our approach presented in this study has a slightly different goal: Instead of deriving a regulation network graph with more or less probable connections for a small subprocess of the cell, we concentrate on analyzing and modeling the dynamics of the whole genome on a more abstract level. That is, we derive a state space model which represents the stress response dynamics of all the yeast genes. The aim of the work is to enable simulation and prediction of the gene expression response so that the results would qualitatively match the real system.

In this study we are focusing on stress experiments concerning yeast *Saccharomyces cerevisiae*. In a stress experiment, the environmental conditions of a yeast cultivation are suddenly changed and the response in the gene expression is measured. Changes can be made for example in the temperature or by modifying the concentration of some chemical in the growth media.

Our approach for model identification applies a rather new modeling method called *subspace identification*,[16] which enables the state space model identification even when all the genes in the genome are included in the calculations. The model also contains an input signal which corresponds to the physical change or stress factor applied to the yeast cultivation.

2. Approach

We are considering in this study biological cells which live in a relatively constant environment, so that the state of the cells can be assumed to remain near a nominal or normal state. External perturbations may force the cells to change their state, but they still can reach a new balance and carry on living. When the original balance of a cultivation is disturbed in a stress experiment, the internal state of the cells is assumed to change so that the effect of the disturbance is canceled out as well as possible. A new balance is obtained after a transient phase during which typically more genes are active than in the final state. This means that in a stress experiment the expression level of each gene is assumed to remain (approximately) constant before the environmental change as well as after the effect of the change has died away. That is, the system remains in a (possibly different) steady state before and after the transient phase, and the transition from the nominal state to the final state is caused by the environmental change.

It is important to note that the assumption of cell steady states does not exclude the possibility of oscillations in the genome. Since the focus is on a cultivation consisting of a group of cells that is not synchronized, the oscillatory effects (e.g., cell cycle) are averaged away. Thus the measured net genomic activation is mainly related to the stress response common to all the cells.

Our main assumption is that the gene expression network in a stress experiment actually has few possible degrees of freedom, that is, all the relevant changes in the gene activities can be described by only a small number of latent variables. This enables the use of a low-dimensional state vector to comprise the core of the system dynamics, whereas the actual gene expression values are produced as a linear combination of these state variables or "functional modes". It has already been shown in a few publications that this assumption is quite well justified; the latent variables can be calculated for example using principal component analysis (or singular value decomposition).[15,17,14]

Despite the dimension reduction obtained by the latent structure, the modeling problem still remains extremely high-dimensional, at least when compared with the typical number of available data vectors. This aspect is dealt with model structure selection; only linear or slightly nonlinear structures are utilized, which is the only way to preserve the analyzability of the models. Furthermore, if we assume that the changes in the gene expressions remain quite small, a linear model can be a sufficient approximation of the phenomena.

3. Data

Source data for this study were obtained from the two yeast stress gene expression experiment series by Gasch *et al.*[18] and Causton *et al.*[19] In Ref. 18, the measurements are performed using cDNA technology, where the gene activities are measured with respect to some reference pool. Typically the time zero values of a time series

experiment are used as the reference pool so that the actual time series measurements can be obtained as proportional (ratio) changes with respect to the original state of activity. These normalized and background corrected data were loaded from the publication web page (`http://genome-www.stanford.edu/yeast_stress`).

In Ref. 19, on the other hand, the Affymetrix technology is used which provides the absolute values of the gene activities. In this study, normalized data available on the web page (`http://web.wi.mit.edu/young/environment`) were used. The normalization had been made according to the Affymetrix recommendations and we used the data as such.

To combine the two datasets, the absolute values of the Causton dataset were transformed into ratio values by dividing each time series measurement with the time zero absolute measurement. These ratio values were then combined with the ratio values in the Gasch dataset. Two open reading frames (ORF) in the datasets were considered the same if the given systematic names were identical. All the Causton dataset experiments were used, whereas only the time series experiments were included from the Gasch dataset.

Combining microarray data from different platforms and laboratories is not straightforward; it has been shown that both the platform type and laboratory may cause differences in the results. However, some recent studies also show that the activation levels of the set of most active genes can be in a rather good agreement.[20] Thus the combination of the different data sets is justified in the sense that in the modeling the high activation values are emphasized and the small variations remain less significant. Furthermore, the abstraction level of our modeling approach is quite high, so that it was decided that it is more important to gather as much data as possible instead of confining to a more compatible but insufficiently small data set. This also lead to the assumption that the time series of the different experiments are comparable as well.

To increase the number of usable measurements, missing values in the data were estimated using the k nearest neighbor method found to be effective by Troyanskaya *et al.*[21] In this study the value $k = 10$ was used and the imputation was performed using $\log_2$ transformed ratio data. For dynamic modeling purposes, the data were resampled to have a constant sample time. Ordinary linear interpolation was used here, individually for each ORF and time series. The selected new sample time was 15 min, which was the smallest reasonable choice based on the original data. Long time series with sparse samples after 120 min were truncated so that the average of the last samples was treated as the final sample at time 135 min.

As a result of the data collection and preprocessing, a set of 21 time series containing altogether 158 whole-genome expression data points was obtained. The included stress experiments are listed in Table 1. In each series, a stepwise change in the cultivation conditions has been made at time instant zero which corresponds to the measurement time of the first sample in the series.

The gene expression values were considered as the outputs of the system whereas the changes in the environmental conditions and growth media concentrations were

Table 1. Experiments included in the study and the number of samples in each series after data pre-processing. The data were gathered from the two public data sets.[18,19]

Number	Publication	Experiment	Data Points
1.	Gasch *et al.*	Heat shock (hs-1)	6
2.	”	Heat shock (hs-2)	5
3.	”	Heat shock 37 → 25°C	7
4.	”	Heat shock 29 → 33°C	3
5.	”	Heat shock 29 → 33°C (in sorbitol)	3
6.	”	Heat shock 29 → 33°C (sorbitol 1 M → 0 M)	3
7.	”	H_2O_2 treatment	10
8.	”	Menadione exposure	10
9.	”	DTT exposure (dtt-1)	10
10.	”	DTT exposure (dtt-2)	10
11.	”	Diamide treatment	7
12.	”	Hyper-osmotic shock (Sorbitol)	9
13.	”	Hypo-osmotic shock (Sorbitol)	5
14.	”	Amino acid starvation	10
15.	”	Nitrogen depletion	10
16.	Causton *et al.*	Heat shock	9
17.	”	Acid	7
18.	”	Alkali	7
19.	”	H_2O_2 treatment	9
20.	”	Salt	9
21.	”	Sorbitol	9

collected in the corresponding input vectors. The input quantities were the following: Temperature, pH and the concentrations of H_2O_2, menadione bisulfate, diamide, sorbitol, amino acids, ammonium sulfate, dithiothrietol (dtt) and sodium chlorine in the growth medium. To match the output values, the input variables were calculated as $\log_2$ transformations of the relative changes in each of the input quantities. Additionally, the input vectors were augmented to include the squared values of the changes, so that the final input dimension of the model was $m = 20$. This augmentation was done in order to improve the modeling results of those general stress response genes which activate positively (or negatively) in any disturbance. Otherwise, the linear model structure would not be able to handle these genes properly.

4. Model

4.1. *Model structure*

The standard discrete state space model form was selected to describe the gene expression system:

$$\begin{cases} x(k+1) = Ax(k) + Bu(k) + w(k) \\ y(k) = Cx(k) + Du(k) + v(k). \end{cases} \tag{1}$$

Here the input column vector $u \in \mathbb{R}^m$ includes the changes in the environmental conditions and chemical concentrations, $y \in \mathbb{R}^l$ is a high-dimensional column vector

describing the gene activities and state $x \in \mathbb{R}^n$ is the low-dimensional latent variable vector. The model has also a stochastic part, so both the state equation and the measurement equation have an additional unknown noise term (w and v). k is the discrete time index referring to the time instant kT, where T is the sample time. Finally, the compatible time-independent matrices A, B, C and D contain the free parameters of the model.

The state space model nicely corresponds with the assumptions made earlier: it contains a user-defined number n of latent variables (states) to collect the dynamics in terms of degrees of freedom and it is purely linear. Since the state space structure is a well-studied model in systems theory, selecting it also enables the utilization of all the powerful results of linear systems theory, e.g., the Kalman filter.

4.2. *Model identification*

To estimate the parameters of the discrete state space model (Eq. (1)) using the interpolated microarray data points and the known environmental changes the *deterministic–stochastic subspace identification* algorithm[16] was applied. The main idea of the method is to transform the originally dynamic identification problem into a static regression calculation. This is done so that the original (interpolated) data points are collected into interlaced windows of consecutive data points, where the length of the window is i data points. For example for the output data one such window would be:

$$\mathbf{y}^T(\kappa) = \big(y^T(\kappa - i) \big| y^T(\kappa - i + 1) \big| \cdots \big| y^T(\kappa - 1) \big) \in \mathbb{R}^{il}. \tag{2}$$

As the index κ goes through all the values for which the required original data points exist, a set of windows or "static data points" $\mathbf{y}$ is obtained. The corresponding static data points for the input data are defined identically. For the identification algorithm, all the static data points are collected to block Hankel matrices, one for input data and one for output data. Static estimation methods are then applied to these matrices to form the compressed state sequence, from which the actual system matrices of the model (1) can be calculated. Thus, as distinct from the ordinary static regression approaches, subspace identification method is dealing with dynamical data and leads to a model which describes the dynamics present in the data.

It is normally assumed that all the N data points are from a single long and continuous time series. However, in this study our data contained multiple (21) short time series, one from each environmental condition experiment, which all described the same dynamical system that was to be modeled. To be able to use all the data points efficiently in the identification, two additional operations were performed: (1) *padding* of the short time series and (2) combination of the static data points from different time series.

In padding each short time series (jth series containing originally N_j data points) was extrapolated using the assumptions of the system steady states: Assuming that *before* the environmental change the gene expression values remain

constant, we can say that for all negative time indices the values of the output are equal to the time zero value, i.e., $y(k) = y(0)$, when $k < 0$. Correspondingly, if the system has reached the final state at the end of the time series ($k = N_j - 1$), gene expression remains constant *after* that, i.e., $y(k) = y(N_j - 1)$, when $k > N_j - 1$. This way it is possible to include in the identification the static data points also which contain some original data points outside the range $k = 0, 1, \ldots, N_j - 1$ and thus increase the number of static data points.

The combination of the static data points from different time series was straightforward since the identification algorithm ignores the temporal order of the static data points; the dynamics are embedded inside the static data points, but not between them. So the complete block Hankel matrices for both the input and the output were obtained by combining all the static data points from all time series into columns of two large block Hankel matrices.

The subspace identification algorithm was applied to the source data with the parameter value $i = 5$. That is, each static data point contained five consecutive original data points collected together meaning that the dynamic range of the model was assumed to be equal to or less than five. This choice was made because the shortest data series only contained three time points, thus limiting the length of the dynamic range. Based on the analysis of the singular values of the oblique projection and the estimation results, the system order n was selected to be four, which is in line with the results obtained in the previous publications of the topic. For example, despite the different modeling approaches presented in the publication by Raychaudhuri *et al.*,[17] the study confirms that the gene expression system can be quite accurately reduced to a small number of latent variables. Additionally, the system order selection is also limited by the properties of the subspace identification algorithm, since the order is restricted to stay below the assumed maximum dynamic range.

In practice, the calculations of the algorithm were implemented in Matlab. Some care had to be taken with the matrix operations involving high dimensional matrices, and the fact that there were less samples than the data dimension affected the calculations as well. For example, the economical QR-decomposition-based subspace identification algorithm presented by Van Overschee and De Moor[16] had to be replaced by a version which used directly the definitions of the projections. Additionally, the covariance matrices of the noise sequences $w(k)$ and $v(k)$ could not be determined due to the low number of samples. This means that if the identified state space model is to be used for Kalman filter simulations, some additional assumptions of the noise have to be made.

5. Results and Discussion

5.1. *Identified model*

As a result of the subspace identification calculation, the state space model (Eq. (1)) parameter matrices A, B, C, and D were obtained. Some analysis of the identified

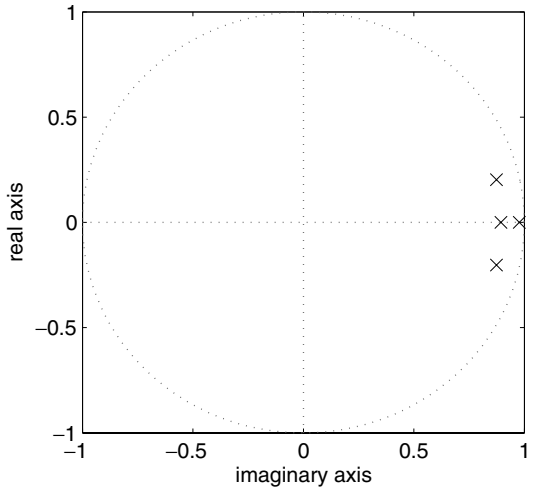


Fig. 1. Poles of the estimated model. All four poles lie inside the unit circle, which means that the system is stable. Additionally, the complex pole pair indicates oscillating behavior.

model can be done by simply investigating these parameters. Figure 1 shows the location of the estimated system poles (eigenvalues of matrix A). Obviously the model turned out to be stable (all the poles are inside the unit circle) and to have two real poles and one complex pole pair. All the poles are located near the point $(1, 0)$ in the complex plane thus indicating that the system emphasizes low frequencies with respect to higher ones. Additionally, the complex pole pair refers to oscillating behavior which is also present in the source data; the stress responses of individual genes tend to have some overshoot before they reach a new balance after the stepwise disturbance. This behavior can already be seen in the average stress responses illustrated by Gasch *et al.*[18]

5.2. *Simulation results*

The identified model could easily be used to calculate the response of all the genes to any known change in the input variables. In this case we considered only the *simulation* of the model, where the model is used without information of the real system output, that is, the real gene expression values. Starting from a certain initial state $x(0)$ (here chosen to be zero) the two equations of the model (1) are consecutively calculated with the input series for which the gene expression values are to be simulated. As a result, this approach enables also the simulation of the responses of new input sequences for which no real experiments have been performed.

To evaluate the model, all the experiments present in the source data were simulated and compared with the measured expression series. Figure 2 shows the scatter plots of the experiments with measured values on the horizontal axis and

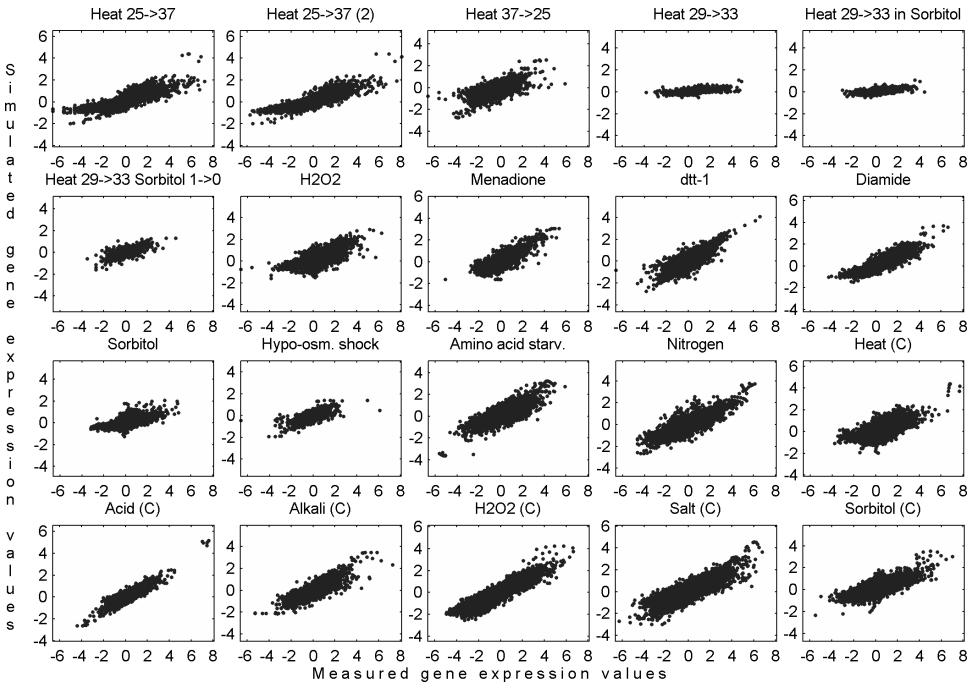


Fig. 2. Scatter plots of the measured and simulated gene expression values for all experiments. Each subfigure shows the comparison of the measured and simulated responses for one experiment.

corresponding simulated values on the vertical axis. Each plot includes all measurements of the corresponding experiment for all genes present in the final data.

Although there is some variation in the figures, the overall performance of the model is good; there clearly is a positive correlation between the measured and simulated values for all the experiments. When calculated, the linear correlation coefficients of the experiments vary from 0.47 to 0.92 (Fig. 3), thus showing that at least most of the experiments can be explained by the model. The correlation measure was selected for evaluation instead of a mean square error (MSE) based criterion because of the complexity and high-dimensionality of the problem: Even though there are errors in the simulations, the most important thing is that the simulator can tell the direction of the gene activation change correctly. This can be better measured by correlation than MSE.

The best modeling results are given by the experiments Causton H_2O_2, Diamide, Causton Alkali and Acid, DTT-1, Nitrogen, Menadione, both of the strong heat shocks (25°C → 37°C) and Causton Salt. On the other hand, the experiments with smallest correlation values are Sorbitol, the two mild heat shocks (29°C → 33°C) and Causton Heat. From Fig. 2 it can be seen that most of the well-modeled experiments contain many high expression values, whereas the expression in the less-accurately modeled experiments is generally lower. This could be explained so that

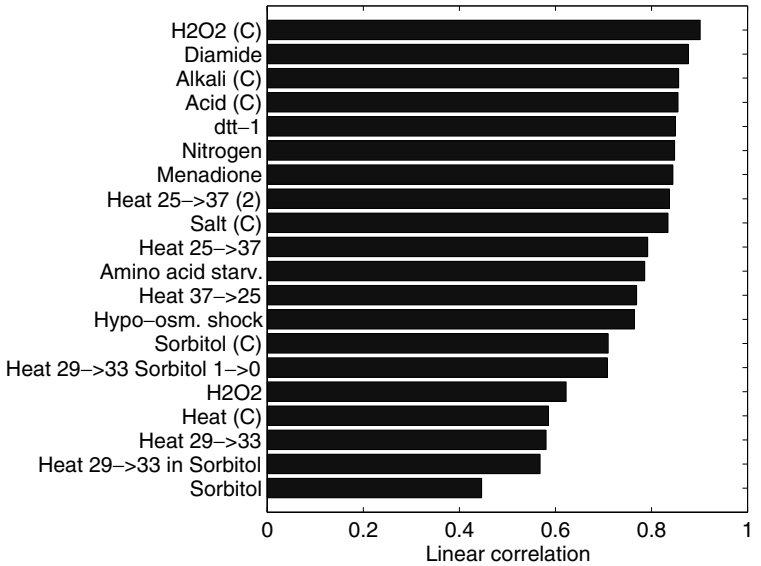


Fig. 3. Linear correlation coefficients between the measured and simulated values. The higher correlation values correspond to the experiments with stronger environmental changes.

larger measurements affect more in the modeling process and are thus repeated better. In addition, the proportion of noise in the measurements with large expression values is most likely smaller than that in the other samples, which naturally improves the modeling performance. One important factor affecting the modeling process is also the length of the individual time series. For example, the mild heat shock time series contain only three original samples each and are thus harder to model accurately.

It can also be noted that the environmental factor changes in the group of well-modeled experiments (i.e., addition of H_2O_2, diamide, alkali, acid, dtt, nitrogen, menadione) typically cause strong metabolic effects in the cells, whereas the stress responses of the other environmental factors are milder. Additionally, in the low-correlating Gasch H_2O_2 experiment the final concentration of hydrogen peroxide (0.3 mM) is lower than that in the corresponding experiment by Causton *et al.* (0.4 mM), which is modeled very well.

The results for a single experiment can also be analyzed by illustrating the measured and estimated gene expression values with respect to time. Figure 4 shows the measured and simulated responses of a group of stress response related genes in the Causton H_2O_2 experiment. The group included all the genes which are annotated to the GO-Slim term "process: response to stress" in the *Saccharomyces* genome database (`http://www.yeastgenome.org`). This group of genes was selected only for visualization purposes to constrain the number of genes to be shown, although the model could predict responses for all the 4176 genes included in the model calculation. Clearly the simulated responses behave very similar to the measurements; predicted changes in the gene expression are almost always to the

Fig. 4. Measured and simulated responses to H_2O_2 addition. The measurements shown are from the Causton hydrogen peroxide experiment and the genes listed are all stress related according to the GO-slim annotation. White corresponds to high and black to low expression values when compared with the time zero values.

right direction, and the dynamical behavior is reproduced quite well. However, the model is a bit too timid, giving in general too small deviations from the zero level. This behavior of the model reflects the high variability of the data.

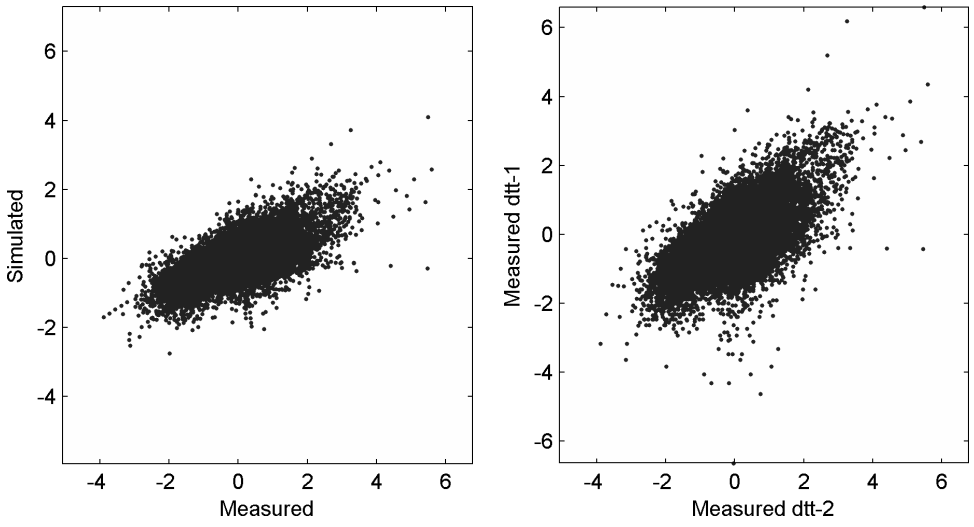
The largest differences between the simulated and measured responses are in quick changes at the beginning of the time series. Apparently the 15 min sample

time is too long that the model could not learn these rapid responses as well as the longer lasting ones. However, with the current measurement data it is not possible to use any shorter sample time. It is also known that the delay from an activated transcription factor to its target gene is about 25 min. This means that the changes in the first sample (15 min) are direct effects of the environmental change in the system input rather than regulative responses mediated by transcription factors.

5.3. *Validation*

For cross validation checking of the modeling accuracy it would have been necessary to leave some data series out of the model estimation and use them as independent data in the validation. However, the model structure was selected so that each component of the input vector u represented one environmental factor, and the data typically contained only one or two experiments for each environmental factor. This lead to the situation where leaving one experiment out of the modeling would have reduced the amount of data drastically and affected the modeling performance.

As a test case, the second DTT addition experiment (dtt-2) by Gasch *et al.* was left out from the parameter identification and then taken into account in validation, so that the model learned the correct DTT response only from the first DTT experiment (dtt-1). Figure 5(a) shows the scatter plot for this data set, where the measured and simulated expression values are compared. Clearly there is a correlation between the measured and simulated data, the linear correlation coefficient being 0.59. This shows that although the correlation is lower than that in most of



(a) (b)

Fig. 5. Scatter plots of the DTT addition experiments. (a). Comparison of the measured and simulated values of the DTT-2 experiment not included in the model estimation. (b). Comparison of the measured values from the identical DTT-1 and DTT-2 experiments.

the data sets included in the identification, the model is still able to give a reasonable simulation of the replicate experiment not used in the teaching of the model.

For comparison, the scatter plot for the measurements of the two identical DTT stress experiments (ddt-1 and dtt-2) is shown (Fig. 5(b)). This reveals well the extremely high variability of the data; the correlation coefficient of the two data sets is only 0.58. Since the variation in the measurements is this high, it cannot be expected that the model could provide any more accurate results either.

5.4. *State sequences*

Given the selected state space model an interesting aspect is the connection of the individual state variables to different biological functions present in the cell. To analyze this we calculated the principal component directions of the state sequence data. Principal component analysis was a justified approach since the state space model is not unambiguous; an arbitrary similarity transformation can be applied to the model so that the input–output behavior remains the same but the state variables change. Since any rotation of the state space is allowed, we used principal component analysis to reveal the mathematically relevant structure of the space spanned by the state variables. It was found that the state sequences of three Causton experiments (Heat, Salt, and Sorbitol) were quite different from the rest of the measurements since they practically spanned a new state space dimension of their own. In Fig. 6, the principal component variances of the state sequence data are

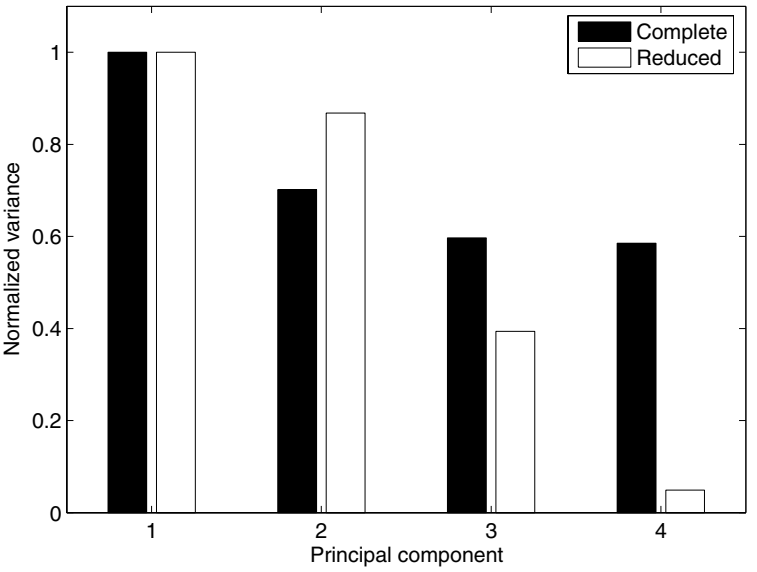


Fig. 6. Normalized variances of the state vector principal components. The complete data set contained all the experiments, whereas from the reduced data set the Causton Heat, Salt, and Sorbitol experiments were excluded.

shown. Clearly the relevant dimension of the state space is four if all the experiments are included, whereas without the three Causton experiments only three main principal components are required. This could be explained by the fact that a different chip type was used in the Causton measurements. Additionally, these results are in line with the not-so-good modeling performance obtained especially for the Causton Sorbitol experiment.

6. Conclusions

Due to the limitations of the source data and simplicity of the model structure it is not realistic to claim that the estimated model could perfectly and globally describe the dynamical behavior of a real yeast cell cultivation. However, the proposed approach is promising since the state space model structure seems to fit quite well to the gene expression modeling. The problems caused by the high-dimensional data vectors are nicely avoided because of the low-dimensional state space, and the model is still able to describe the system dynamics for all the genes. Especially the experiments with environmental shocks causing strong metabolic changes are modeled well. This is most likely because of the larger expression values present in these experiments.

The simulation results are actually surprisingly good, since the data were not well suited for this kind of identification. The input contained only stepwise changes and in one input variable at a time; for better modeling results the input should be richer and contain more variation. There should also be changes in many input variables during one time series, so that the interactions of different stress factors could be modeled. In fact, a change toward more dynamic modeling oriented test planning is required until proper data sets for this kind of whole genome models are achieved.

The actual nature of microarray data is also an interesting subject, since there seems to be quite much variability in the data (e.g., Fig. 5(b)). This may be due to the inaccuracies of the measurement technologies, but the biological variation must also be considered; it seems evident that individual cells simply do not always react exactly the same way even in identical external conditions. It really seems that more precise modeling of these phenomena would require more sophisticated methods and larger data sets with improved test planning. However, in a more abstract level — as presented in this study — the models can already give qualitatively relevant results.

Acknowledgments

This research has been funded by National Technology Agency of Finland (Tekes). Additionally, OH has been supported by the Technology Promotion Foundation (TES). The authors also wish to thank Heli Pykälä, MSc, who implemented the subspace identification routines in Matlab.

References

1. Eisen MB, Spellman PT, Brown PO, Botstein D, Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci USA* **95**:14863–14868, 1998.
2. De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y, Adaptive quality-based clustering of gene expression profiles, *Bioinformatics* **18**(5):735–746, 2002.
3. Kaski S, Nikkilä J, Sinkkonen J, Lahti L, Knuuttila J, Roos C, Associative clustering for exploring dependencies between functional genomics data sets, *IEEE/ACM Trans Comput Biol Bioinformatics*, Special Issue on Machine Learning for Bioinformatics **2**(3):203–216, 2005.
4. Kauffman SA, *The Origins of Order*, Oxford University Press, 1993.
5. Shmulevich I, Dougherty ER, Kim S, Zhang W, Probabilistic boolean networks: A rule-based uncertainty model for gene regulatory networks, *Bioinformatics* **18**(2):261–274, 2002.
6. Friedman N, Linial M, Nachman I, Pe'er D, Using bayesian networks to analyze expression data, *J Comput Biol* **7**(3/4):601–620, 2000.
7. Friedman N, Murphy K, Russell S, Learning the structure of dynamic probabilistic networks, in Cooper GF, Moral S, (eds.), *Proc 14th Conf Uncertainty in Artificial Intelligence*, Morgan Kaufmann, Madison, WI, pp. 139–147, 1998.
8. Ong IM, Glasner JD, Page D, Modelling regulatory pathways in E. coli from time series expression profiles, *Bioinformatics* **18**(Suppl. 1):S241–S248, 2002.
9. Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran D, Gaiba A, Wild DL, Falciani F, Modeling T-cell activation using gene expression profiling and state-space models, *Bioinformatics* **20**(9):1361–1372, 2004.
10. Perrin BE, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alché-Buc F, Gene networks inference using dynamic Bayesian networks, *Bioinformatics* **19**(Suppl. 3):ii138–ii148, 2003.
11. Bar-Joseph Z, Analyzing time series gene expression data, *Bioinformatics* **20**(16):2493–2503, 2004.
12. Guthke R, Möller U, Hoffmann M, Thies F, Töpfer S, Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection, *Bioinformatics* **21**(8):1626–1634, 2005.
13. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV, Fundamental patterns underlying gene expression profiles: Simplicity from complexity, *Proc Natl Acad Sci USA* **97**(15):8409–8414, 2000.
14. Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR, Dynamic modeling of gene expression data, in *Proc Natl Acad Sci USA* **98**(4):1693–1698, 2001.
15. Wu F-X, Zhang WJ, Kusalik AJ, State-space model with time delays for gene regulatory networks, *J Biol Syst* **12**(4):483–500, 2004.
16. Van Overschee P, De Moor B, *Subspace Identification for Linear Systems*, Kluwer Academic Publisher, Boston, Massachusetts, 1996.
17. Raychaudhuri S, Stuart JM, Altman RB, Principal components analysis to summarize microarray experiments: Application to sporulation time series, in *Proc fifth Pac Symp Biocomput*, Vol. 5, pp. 452–463, 2000.
18. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO, Genomic expression programs in the response of yeast cells to environmental changes, *Mol Biol Cell* **11**:4241–4257, 2000.
19. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA, Remodeling of yeast genome expression in response to environmental changes, *Mol Biol Cell* **12**:323–337, 2001.

20. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JGN, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martínez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W, Multiple-laboratory comparison of microarray platforms, *Nature Methods* **2**(5):1–5, 2005.
21. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB, Missing value estimation methods for DNA microarrays, *Bioinformatics* **17**(6):520–525, 2001.

Olli Haavisto Graduated in Electrical Engineering at Helsinki University of Technology, Finland, in 2004. He is currently a postgraduate student and researcher in the Control Engineering Laboratory, Helsinki University of Technology. His research topics are data-based modeling of dynamical systems and multivariate regression methods.

Heikki Hyötyniemi Graduated in Electrical Engineering at Helsinki University of Technology in 1989 and performed his doctoral dissertation in 1994. Currently he is a professor in the Control Engineering Laboratory at the same university. His research interests include artificial intelligence and modeling of complex systems applying "neocybernetic" principles.

Christophe Roos Graduated in Genetics and Mathematics (1982) at the University of Helsinki, Finland, whereafter he performed a PhD thesis in Molecular Biology (1986) at the University of Strasbourg, France. After having directed a Drosophila Developmental Biology Research group at the University of Helsinki, he joined (2000) Medicel Oy, a company developing a systems biology software platform. His principal scientific interests proceed from the use of Bioinformatics in Developmental Biology.