Reverse Engineering of the Stress Response during Expression of a Recombinant Protein

W. Schmidt-Heck[1], R. Guthke[1], S. Toepfer[2], H. Reischer[3], K. Dürrschmid[3], K. Bayer[3]
[1] Hans Knoell Institute for Natural Products Research (HKI),
Beutenbergstr. 11a, D-07745 Jena, Germany,
Phone: +49-3641-656820, Fax: +49-656825,
email: reinhard.guthke@hki-jena.de
[2] BioControl Jena GmbH, Wildenbruchstr. 15, D-07745, Jena, Germany,
Phone: +49-3641-675511, Fax: +49-3641-675512, email: BioControl@t-online.de
[3] Institute of Applied Microbiology, University of Agricultural Sciences (BOKU),
Muthgasse 18, A-1190 Vienna, Austria, Phone: +43 1 36006 6220, Fax: +43 1 3697615,
email: bayer@mail.boku.ac.at

ABSTRACT: The overexpression of recombinant proteins in microorganisms may lead to a metabolic depression or collapse of the cell factory. In order to understand this process and to optimize the cellular productivity the stress response was investigated. The expression of the recombinant human superoxide dismutase (SOD) was induced under steady state conditions and the expression of all 4289 protein coding genes of the microorganism *Escherichia coli* was monitored using microarrays. After normalization by the LOWESS method 102 differentially expressed genes were selected by a novel criterion that includes the measurement error. These differentially expressed genes were clustered using the *EcoCyc* database and the fuzzy-c-means clustering method. The results from clustering were interpreted in terms of dynamic models, which have been constructed either via Singular Value Decomposition (SVD) or a novel heuristic algorithm for dynamic model structure optimization.

KEYWORDS: Systems Biology; Gene Expression; Stress Response; Model Structure Identification

INTRODUCTION

In the life sciences with the availability of "post-genomic" technologies to generate huge amounts of experimental data, transcriptomics, proteomics, metabolomics and cytomics as well as the holistic interdisciplinary concept of "systems biology" [1] to discover the structure and dynamics of complex bioprocesses have emerged. In order to discover and to understand the complex molecular and cellular interactions during the stress response after induction of recombinant protein synthesis gene regulatory networks and signaling pathways have to be reconstructed by reverse engineering methods [2-5]. This approach is both, data- and model- (i.e. hypothesis-) driven. In order to determine the main interactions between the genes and proteins and to construct continuous dynamic linear models for instance the Singular Value Decomposition method (SVD) was applied recently [6]. This algorithm works fast and guarantees an optimal model fit to the gene expression profiles and/or better to the interpolation of the gene expression profiles. A disadvantage is the non sparseness of the matrix of coupling constants. The number of parameters must be reduced later by suitable algorithms (robust regression [6]).
During stress response the cell must react with an appropriate adjustment by the conversion of its gene expression . This takes place very fast (from seconds to minutes after induction) by so-called alarmones, e.g. ppGpp and cAMP. This early response is communicated via signal conversion cascades and by response modulators, e.g. by sigma factors, and results in the change of the expression of stress proteins (after some hours).
This leads to the synthesis of proteins, which are involved in defensive strategies such as the degradation of foreign proteins or the repair and protection of cell structures. At the same time genes, whose products are unnecessary for the actual cellular state, are repressed.

EXPERIMENTAL DATA

Escherichia coli was grown in a chemostate culture. Under steady state conditions the expression of the recombinant protein SOD (human superoxide dismutase) was induced by IPTG (isopropyl-beta-D-thiogalactopyranoside) dosage. (IPTG is a nonmetabolizable analog of the normal substrate, lactose. Both, lactose and IPTG are inducers of the *lac*-promoter that is used in the genetically engineered *E. coli* to control the recombinant protein synthesis.) Samples were analyzed 8, 15, 22, 45, 68 90, 150 and 180 minutes after induction and compared with the control from pooled samples before induction. To monitor changes of the transcription levels whole genome microarrays (MWG Pan® *E. coli*) with probes for all 4289 protein coding genes of *E. coli* were used for hybridization The gene expression was analyzed by dye swap experiments, i.e. the gene expression was analyzed twice in two hybridization experiments using Cy3 versus Cy5 and Cy5 versus Cy3 for labeling. The data were pre-processed by LOWESS normalization.

RESULTS

ERROR CORRECTED FOLD ANALYSIS

For the recognition of differentially expressed genes the fold analysis was used [7]. The aim is to distinguish between significant changes of gene expression and the intrinsic biological and technical variability. Several studies considered a 2-fold ratio $F = I_T/I_R$ of the signal intensity I_T of the treated sample (e.g. after induction) and the signal intensity I_R the reference sample (prior to induction) as significant. The general application of such a formula, however, leads to numerous false-positive results, in particular for low spot intensity. To avoid these effects we introduced the corrected fold change by $F_{corr} = F \cdot 2^{-E(I_G)}$ using an error function

$E(I_G) = p_1 + p_2 \cdot e^{p_3 \cdot I_G}$ with the geometric mean of both intensities $I_G = \sqrt{I_T \cdot I_R}$.

The parameters p_1, p_2 and p_3 of the error function were identified by model fit to the expression data of control experiments before induction. With the selection criterion $|(\mathrm{ld}(F_{corr})| > 1$ we identified 102 genes as differentially expressed in one or more samples after induction.

CLUSTER ANALYSIS AND SELECTION OF REPRESENATIVE GENES

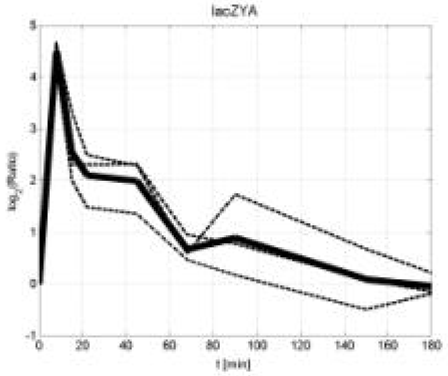

Figure 1: Expression profile of the genes *lacZ, lacY* and *lacA* (dashed lines) and the averaged profile of the transcription unit *lacZYA*

In a first step the 102 gene expression profiles were merged to 69 transcription profiles using the available knowledge: This is done by averaging of expression profiles of genes belonging to a transcription unit as described in the database *Ecocyc* (http://BioCyc.org). 72 of the 102 genes belong to 39 transcription units. For instance Figure 1 shows the transcription profile of the *lac*-operon obtained by averaging of the expression profiles of the genes *lacZ, lacY* and *lacA*.

In the second step the time series of the 69 transcription profiles were clustered into 3 clusters by the fuzzy c-means algorithm. 66 profiles were assigned to one of these classes with a membership degree greater than 0.6 (Figure 2). 19 profiles were assigned to the cluster 1, 12 profiles were assigned to the cluster 2 and 35 profiles were assigned to the cluster 3. The remaining three transcription units (*lacZYA*, *cspA* and *artPIQMJ*) could not be assigned to one of these three classes with a membership degree greater than 0.6. The transcription unit *lacZYA* (Fig. 1) plays an important role for the recombinant product formation due to the use of a *lac*-promoter regulated plasmid for the recombinant protein synthesis. Therefore, the transcription profile of the *lac*-operon, shown in Figure 1, was considered for modeling. The kinetics of the two other transcription units are to complicate (i.e. oscillatory and are considered to be artificial) and were not considered for the modeling of stress response.

For each class one representative gene was selected that i) has a high degree of membership, ii) is annotated, i.e. has a known physiological function, and iii) has a function that is typical for more genes belonging to the cluster. Typical members of the first cluster (Figure 2 left) are the σ^{32}-regulated transcription units *dnaKJ, ibpAB, grpE* and *ftsJ-hflB* genes, i.e. chaperones and proteases. Thus, the transcription unit *ibpAB* coding for the inclusion body protein (a chaperone which mediates protein folding) was selected as representative for the first cluster. As representative for the second cluster (see Figure 2 middle part) the gene *cchB* coding for the detox protein was selected. In the third cluster

(Figure 2 right part) are many σ^{70}-regulated transcription units whose products are involved in the energy metabolisms and the carbon source uptake. As representative of the third cluster we selected the transcription unit *nuoABCEFGHIJKLMN* coding for the NADH dehydrogenase I.

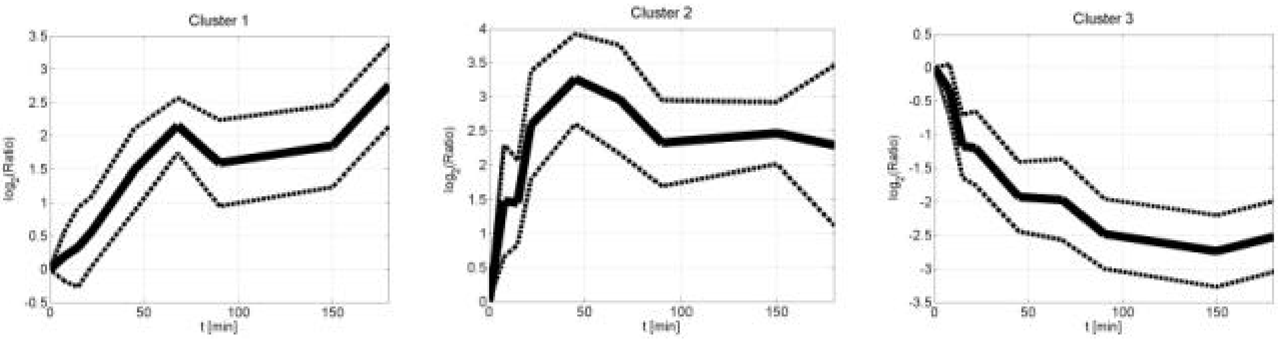

Figure 2: Result of fuzzy-c-means clustering with 3 classes: mean normalized gene expression profiles averaged over the number of 19, 12 and 35 genes with membership degrees >60% to the clusters 1, 2 or 3, respectively. The dashed lines show the standard deviation.

REVERSE ENGINEERING OF GENE NETWORKS

The reverse engineering approach aims at the reconstruction of gene regulatory network structures based on transcription data. In the system of differential equations (1) the network structure is represented by a matrix of coupling constants *W*.

$$\dot{X}(t)=\sum_{j=1}^{N}W_{ij}X_{j}(t)+b_{i}(t)+x_{i}(t) \qquad (1)$$

$\dot{X}$ derivative of *X* with respect to time *t*

W matrix of coupling constants; here: with the dimension 5 x 5

X measured data; here: expression profiles of the 5 representative transcription units

b external stimuli; here: induction by IPTG

x noise

For the modeling of the stress response we used the expression profiles *X(t)* of the transcription units *ibpA*B, *nuoABCEFGHIJKLMN* and *cchB* selected as representatives of the three clusters as well as the profile of *lacZYA* (Figure 1) and the profile of the product *cp* (SOD – superoxide dismutase).

Singular Value Decomposition

To compute the matrix *W* by Singular Value Decomposition (SVD) from the experimental data *X* we used the LAPACK routines [8] implemented in Matlab as formulated by eqs. (2) and (3):

$$[U,S,V]=svd(X) \qquad (2)$$

U left singular vectors

V right singular vectors

S singular value

For the determination of the unknown matrix *W* it is necessary calculate the first temporal derivative of the matrix *X*. The computation of the derivative results from spline interpolation of the transcription profiles. The matrix of coupling constant W results from equation (3).

$$W=(\dot{X}-b)\cdot U\cdot S^{-1}\cdot V^{T} \qquad (3)$$

The matrix *W* calculated by eqs. (2) and (3) is shown in Table I. We checked by model simulation if parameter reduction is possible without great losses in approximation quality. We found that the parameters smaller than 1 were not relevant for the model fit. The model simulation with 21 relevant parameters (>1) is shown in Figure 3.

	lac	*cp*	*ibp*	*nuo*	*cch*
lac	-7.115	338.5	-785.6	39.57	-84.55
cp	(-0.0091)	15.0673	-34.91	2.5865	-3.4267
ibp	(0.0033)	11.41	-26.4	1.826	-2.520
nuo	(0.1297)	44.81	-101.34	5.999	-7.908
cch	(0.004)	-16.44	39.47	-2.714	4.448

Table I: Matrix of coupling constants *W* as computed by the SVD method, i.e. by eqs. (2) and (3) from the measured and pre-processed expression profiles of the transcription unites *ibpAB, nuoABCEFGHIJKLMN* and *cchB* selected as representatives of the three clusters (Figure 2) as well as the transcription profile of *lacZYA* (Figure 1) and the concentration profile *cp* of the recombinant product SOD. The parameters in parentheses that are smaller than one, as well as the vector *b* in eq. (1) were set to zero for the simulation shown in Figure 3.

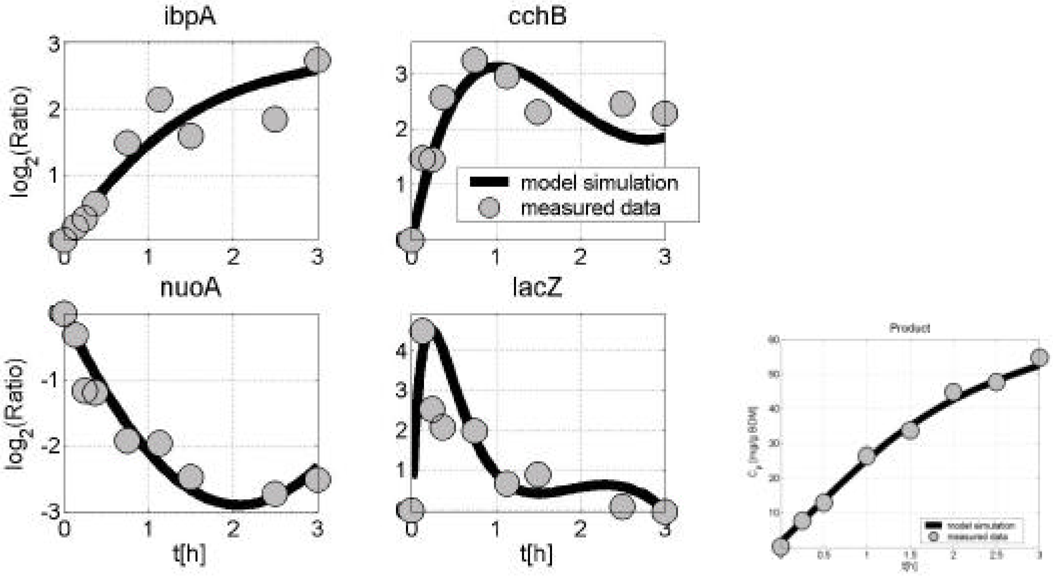

Figure 3: Measured transcription profiles *ibpAB, nuoABCEFGHIJKLMN, cchB* and *lacZYA* as well as the concentration *cp* of the recombinant product SOD and the kinetics simulated by eqs. (1) with the coupling constants *W* calculated by SVD and shown in Table I

Optimization of the Model Structure with the NetGenerator Algorithm

The heuristic *NetGenerator* [9] structure optimization method for systems of differential equations optimizes both the model structure as well as the model parameters. The algorithm aims to find a minimum number of relevant parameters of the system of differential equations required for an adequate identification of the given data. The *NetGenerator* algorithm is able to generate alternative model structures that can be compared using the Generalized Coss-Validation coefficient *GCV* according to eq. (4).

$$GCV = \frac{mse}{(1-\frac{n}{N})^2} \qquad (4)$$

mse mean square error

n number of parameters to be estimated (e.g. *n*=12 for the example shown in Tab. II)

N number of observations (sample size, e.g. *N*=44 in the given example)

The *NetGenerator* algorithm permits also the integration of available expertise of biologists. The following interactions are expected to be valid:

- IPTG → *lac* (IPTG induces the *lac*-promoter)
- *lac* → C_P (the *lac*-promoter induces the synthesis of the recombinant protein SOD)
- C_P → *ibp* (the recombinant product SOD up-regulates the synthesis of the inclusion body protein).

Applying the *NetGenerator* algorithm with this configuration four different model structures as shown in Figure 5 and 6 were found. Figure 4 shows the measured data together with the kinetics simulated by the model structure that is defined by Table II and illustrated in Figure 5.

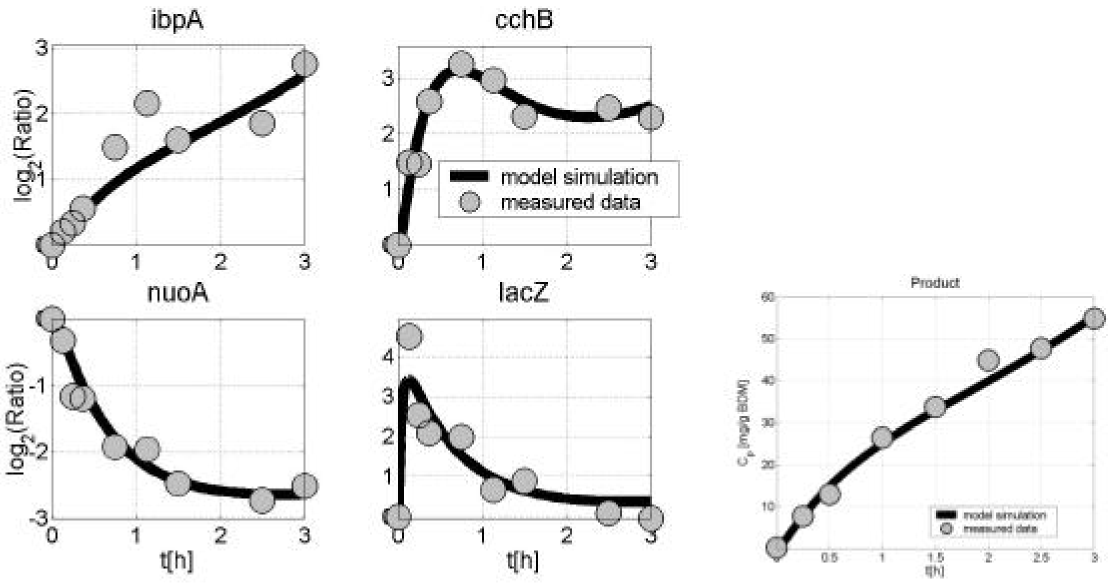

Figure 4: Measured transcription profiles as shown in Figure 3 and kinetics simulated by model eqs. (1) with parameters identified by the *NetGenerator* algorithm and shown in Table II

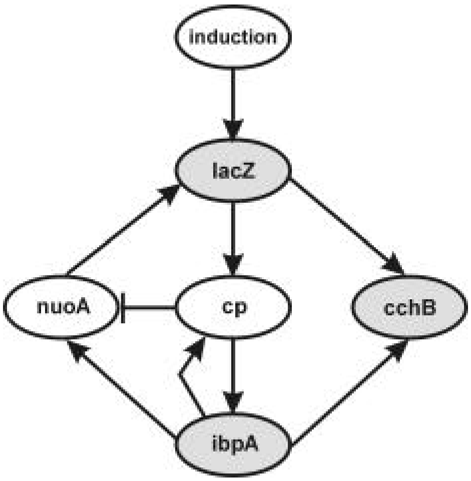

Figure 5: Optimised structure of the dynamic model identified by the *NetGenerator* algorithm and used for the simulation shown in Figure 4. Positive parameters shown in Table II are represented by arrows; negative parameters are represented by T-shaped links (⊥). Shaded circles represent degradation effects by first order kinetics.

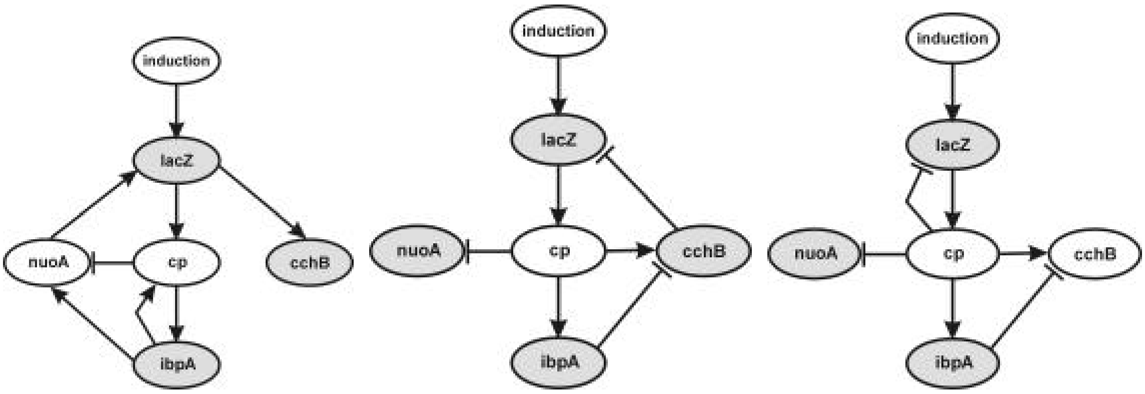

Figure 6: Model structure identified by the *NetGenerator* algorithm from the same experimental data set – alternatives to the structure shown in Figure 5.

	lac	*cp*	*ibp*	*nuo*	*cch*	*b* (Induction)
lac	-26.9055	0	0	34.3966	0	101.1488
cp	0.9998	0	0.5172	0	0	0
ibp	0	15.7149	-33.3333	0	0	0
nuo	0	-34.3753	73.6049	0	0	0
cch	3.3883	0	1.8869	0	-2.2653	0

Table II: Matrix of coupling constants *W* and the vector *b* as computed by the *NetGenerator* algorithm used for simulation by the model eq. (1), the result of which is shown in Figure 4 and illustrated by the model structure shown in Figure 5.

DISCUSSION

Hypotheses on the main interactions of genes and proteins during the recombinant synthesis of the protein SOD in the microorganism *E. coli* were generated by a systems biological approach from genome-wide bioprocess data. The identified model could be applied for the design of optimal control strategies, e.g. the optimal feeding of the inducer IPTG. But, the alternative model structures obtained have to be used for model discriminating experiment design to identify which of the four structures is valid. Common structure elements of the four models are characterized by four interactions:

1. IPTG induces the *lac*-Operon
2. The *lac*-Promoter induces the recombinant product formation SOD with the concentration *cp*.
3. The recombinant product SOD (*cp*) up-regulates the synthesis of chaperones and proteases, e.g. the inclusion body protein (*ibp*)
4. The product SOD with concentration *cp* represses the energy metabolism and carbon source uptake, e.g. represses the *nuo*-operon.

The first three interactions are familiar and need not be proven experimentally once again, whereas the repression of energy metabolism by the recombinant product is a more surprising result and should be the topic of further experimental investigations.

The models were reconstructed from the experimental data by a hybrid, i.e. data- and knowledge driven top-down approach: After the pre-processing and clustering of the time series the model structures and model parameters were identified by fit to cluster representatives using a novel heuristic algorithm implemented in *NetGenerator*. This algorithm identifies the appropriate model structure by growing and pruning of the interaction network. It includes also first and higher order time lag elements and considers available knowledge about interactions. Alternative model structures were compared by a Generalized Cross-Validation coefficient (GCV).

REFERENCES

[1] Kitano H (2002): Systems biology: A brief overview. Science, 295:1662-4.

[2] D'haeseleer P, Liang S, Somogyi R. (2000): Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16, 707-26.

[3] De Jong H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol. 9, 67-103.

[4] Csete ME, Doyle JC (2002). Reverse engineering of biological complexity. Science 295, 1664-9.

[5] Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR (2001): Dynamic modeling of gene expression data. PNAS 98, 1693-8.

[6] Yeung MKS, Tegnér J, Collins JJ (2002): Reverse engineering gene networks using singular value decomposition and robust regression. PNAS 99, 6163-8.

[7] Chen Y, Dougherty ER, Bittner ML (1997): Ratio based decisions and the quantitative analysis of cDNA microarray images, J. of Biomedical Optics 2: 364-374.

[8] Anderson E, Bai Z, Bischof C, Blackford S, Demmel S, Dongarra J, Du Croz J, Greenbaum A, Hammarling S, McKenney A, and Sorensen D (1999); LAPACK User's Guide; (http://www.netlib.org/lapack/lug/lapack_lug.html), Third Edition, SIAM, Philadelphia.

[9] Toepfer S, Driesch D, Woetzel D, Pfaff M, Guthke R (2003): Reconstruction of Gene Regulatory Networks: A Novel Structure Optimization Method for Dynamic Models. German Conference on Bioinformatics GCB2003, München, 12.-14.10.2003