# Neocybernetic Modeling of a Biological Cell

Olli Haavisto*
*Helsinki University of Technology
Control Engineering Laboratory,
P.O.Box 5500, FIN-02015 TKK, Finland
olli.haavisto@tkk.fi

Heikki Hyötyniemi†
†Helsinki University of Technology
Control Engineering Laboratory,
P.O.Box 5500, FIN-02015 TKK, Finland
heikki.hyotyniemi@tkk.fi

**Abstract**

A biological cell forms an extremely complex system with a complicated metabolic reaction network. The functionality of the cell is controlled by the genome consisting of thousands of individual genes. Novel measurement technologies provide us with huge amounts of biological data in a quantitative form, but the suitable modeling methods to deal with the high dimensionality and large systems are still missing. In this article, a novel approach for complex system analysis, *neocybernetics*, is applied to modeling a biological cell. It is shown that under certain assumptions both the metabolic network and the gene expression network can be interpreted as a combined linear system. The analysis of this system from the neocybernetic starting points is then elaborated. As an example, the dynamics of yeast gene expression network are modeled according to the presented principles.

## 1 Introduction

The challenge of complex system analysis is the high number of internal variables and complexity of the connections between them. In the following, we concentrate on a special case of complex systems designated as *elastic systems* (Hyötyniemi, 2006a). A characteristic property of these systems is their ability to adapt to and compensate external disturbances by modifying their internal state. For example, an elastic plate can be deformed by applying external forces to it, and when the forces disappear, the original state of the plate is recovered. It is also assumed that variables in elastic systems are highly connected to each other and there exist internal feedback loops which stabilize the behaviour.

Elastic systems can be analysed using the *neocybernetic* approach proposed in (Hyötyniemi, 2006b), which emphasizes steady states or *balances* of the system instead of the transients leading to these states. It is proposed here that complexity of systems should be attacked by leaving the dynamical transients aside and concentrating on their tendency to remain stable and reach a steady state. Data-based multivariate methods should be utilized to reveal the degrees of freedom of the system variables rather than sticking to the numerous constraints which connect the variables together.

In complex chemical systems elasticity is also manifested: If there are pressures, the system yields. This phenomenon is known as *Le Chatelier principle:* If the environment changes, a new chemical balance is found so that the environmental changes are compensated, at least to some degree. Specially in biochemical systems, this behavior is very dominant, all processes being strongly buffered.

In this article, the vast field of *systems biology* is attacked from a truly systemic viewpoint. Biological cells are interpreted as elastic systems and the usage of neocybernetic modeling principles to model their properties is elaborated. First, a short introduction to cell biology is presented. Then the modeling approach is discussed and a mathematical framework for it is derived. Finally, as an example case, a study of modeling the yeast gene expression dynamics is introduced.

## 2 Biological cell

The biological cell is the principal unit of all living organisms. Seen from outside, an individual cell can be analyzed as a system that is capable of transferring molecules in and out through its cell membrane. Inside the cell the raw materials are transformed into energy or cell components by a complex machinery involving metabolic pathways, which are strictly controlled by the genetic regulatory network.

Structurally cells can be divided into two main classes, *procaryotes* and *eucaryotes*. Procaryotic cells are typically simpler than eucaryotic ones, since they do not contain membrane separated nucleus and

have less intercellular organelles and structures. The cells of one-cellular organisms are usually procaryotic (e.g. bacteria), whereas the multicellular organisms like animals and plants consist of eucaryotic cells.

In the following analysis we are mainly focusing on the properties and functionalities of eucaryotic cells. More specifically, as probably the simplest case of an eucaryotic cell and a typical model organism the brewer's and baker's yeast *Saccharomyces cerevisiae* is considered.

## 2.1 Genetic regulatory network

Hereditary material of cells is stored in genes, which are parts of the *deoxyribonucleic acid* (DNA) of the *chromosomes* located in the nucleus. Together all the genes form the genome of the organism. A rough estimate of the complexity of an organism is the number of its genes: Simple one-cellular organisms typically contain a few thousand genes, whereas invertebrates have over 10000 and mammals about 30000 genes (see Table 1).

Genes control the cell functions by regulating the amount of proteins inside the cell. When activated, a gene is able to transfer the code it contains into a protein in the process called *gene expression*, which has two parts: *transcription* and *translation*. In transcription, the information stored in the sequence of the nucletides of the gene are read to a *ribonucleic acid* (RNA) molecule, which then moves from the nucleus to the cell organelles called *ribosomes*. There translation takes place: According to the sequence of the RNA molecule, an amino acid chain is constructed. As the chain folds to a stable three dimensional structure, it forms the ready functional protein.

Expression of an individual gene is controlled during all the steps of the expression process. To begin with, transcription is activated only if a correct combination of proteins is attached to the DNA strand. Additionally, by binding to the DNA strand certain gene specific protein complexes called *activators* and *repressors* can either increase or decrease the rate of transcription, respectively. Also the density of the DNA packing, that is, the amount of folding of the DNA double helix affects how easily the gene can be read. After translation, the RNA molecule goes through a series of controlled processes which determine e.g. the final nucleotide sequence and the lifetime of the ready RNA chain. The longer the RNA molecule lasts, the more times it can be used in translation to create proteins. Finally, even the protein folding is a controlled process which determines the usage of the ready protein molecule.

Generally the proteins that control gene expression by binding to the DNA strand are referred to as *transcription factors* and they themselves can be products of other genes. Thus the whole genome forms a connected regulation network with strong feedbacks. As the number of genes is typically high and each transcription factor may affect more than one gene, this network is very complex and hard to analyse.

## 2.2 Metabolism

The purpose of cell metabolism is to process nutrients, produce energy and keep the cell alive and growing. The complete metabolism of a cell can be divided into many individual and well controlled *metabolic pathways*, which in turn consist of series of metabolic reactions. These reactions are typically catalysed by *enzymes*, a group of proteins produced in gene expression. When a certain gene is activated, more catalysing enzymes for a certain reaction are produced and the rate of the reaction increases resulting to higher concentrations of the corresponding end products. This in turn may affect the activity of some other gene, thus creating a feedback connection between the gene regulation network and the network of metabolic reactions of the cell.

An example of a typical metabolic pathway is shown in Figure 1. In the figure, only the main reactions and metabolites are shown. Each reaction is catalysed by one or more enzymes, shown as boxes with numbers.

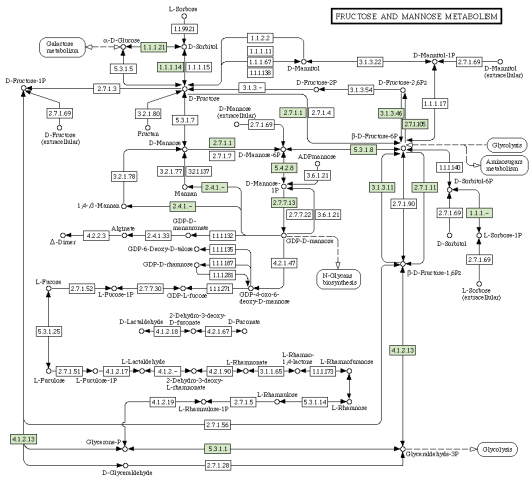


Figure 1: Yeast fructose and mannose metabolism. Each arrow represents a metabolic reaction or a set of reactions catalyzed by an enzyme. Original image from KEGG (www.genome.jp/kegg/)

Table 1: The number of genes of some organisms (Ewing and Green, 2000; Mouse Genome Sequencing Consortium, 2002)

| Organism | Group | #genes |
|---|---|---|
| *Escherichia coli* | bacteria | 4300 |
| *Saccharomyces cerevisiae* (yeast) | one-cellular eucaryote | 6000 |
| *Drosophila melanogaster* (fruit fly) | invertebrate | 13600 |
| *Caenorhabditis elegans* (roundworm) | ” | 19000 |
| mouse | mammal | 30000 |
| human | ” | 35000 |

For more detailed information on cell biology, see for example the book by Alberts et al. (2002).

# 3 Modeling approach

Traditionally, investigating the connections and functions in the cells has been concentrated on studying individual genes or chemical reactions for example by disabling a certain gene in some cell cultivation. Behaviour of this cultivation has then be compared to cells having a normal genome. However, using this kind of reductionistic approach leads to difficulties when the combined behaviour of all the small partial reactions and connections should be analyzed, especially as the dimensionality of the problem is large. On the other hand, when using the neocybernetic approach where only the net effects of the whole system are analysed, one is able to form a more holistic picture of the system.

## 3.1 Dynamics vs. balances

The neocybernetic modeling approach is based on the concept of dynamical balance; it is assumed that the system remains stable and that the system variables are stationary, meaning that their statistical properties remain constant. A biological cell can easily be seen as an neocybernetic system, at least when it is living in a relatively constant environment and remaining healthy and fully functional. In such a case the cell is able to compensate (minor) external disturbances introduced in the form of temperature, pH or chemical concentration variations. When reacting to an external disturbance, the internal state of the cell may change, i.e. balances of the chemical reactions may shift a little and some genes may activate more and some less than before. This means that the cell can adapt to the new environmental situation by changing its internal state and carry on the processes it requires to survive.

The idea of the neocybernetic model is to describe the set of dynamical balances, i.e. the final states of the dynamic system in different homeostatic environments. It is assumed that the inputs of the system (e.g. environmental conditions) are changing slowly when compared with the speed of the internal dynamics of the system. Thus the system state reaches quickly a new balance as the inputs change, and the system mostly stays in balance.

Strictly speaking, biological systems are not in the state of homeostasis, as the living cells continuously exhaust nutrients and produce metabolic products. Here it is assumed that the balanced variables also contain *rates of change,* so that the dissipative processes can also be modeled: It is assumed that in certain conditions the chemical conversion rates remain constant. Inclusion of such derivatives does not ruin the linearity of the model.

## 3.2 Constraints vs. degrees of freedom

When dealing with dynamical systems containing a large number of variables, the traditional modeling approaches have been based on the *constraints* which determine the relations of the system variables. For example, *differential equations* are a typical choice when modeling the dynamics: Each individual equation covers one relation or constraint between the investigated variables. To unambiguously describe the whole system, as many differential equations are required as there are system variables. However, even in the case of a “simple” biological cell like yeast, for example the gene regulation network includes a huge number of genes (over 6000) which can be connected to each others. Even though the gene regulation networks are typically assumed to be sparse instead of being completely connected, this means that a large number of constraints is required to define the system dynamics. When the system dimension is high, this approach is not feasible anymore.

The opposite way to analyse a multivariable system is to collect data from the system and use them to find out the main directions of variation or *degrees of freedom* present in the system. If the system variables are highly connected and the dynamics are restricted by many constraints, the actual number of meaningful degrees of freedom remains low. Accordingly, it may turn out that the dynamics can be described with a much lower number of variables. That is because each individual constraint reduces the number of possible variation directions by one, thus hopefully resulting to a low dimensional space of degrees of freedom.

### 3.3 Linearity

Linearity is a strong assumption which is usually not globally valid for real world systems. However, if the analysis is restricted to a small region in the vicinity of a nominal operating point, it may be possible to describe the system behaviour using a linear model. After all, every smooth nonlinearity can be locally approximated to be linear.

There are clear advantages when sticking to linear models: the analyzability and scalability of the models can be preserved even when the model dimension is increased, and there exists a well justified theory for linear systems. On the other hand, if even a minor nonlinearity is allowed, the theoretical analyses become much harder and no general theory exists.

## 4 Modeling framework

The common factor for different modeling applications in the neocybernetic approach is the assumption of a underlying network structure. Indeed, in both gene expression and metabolism it seems natural to use a network structure as a starting point. In the first case, genes are the nodes of the net, whereas in the latter case the individual molecule concentrations together with environmental factors like temperature and pH value form the nodes. Traditionally when network models are created, e.g. graph theory is used and in order to reduce the complexity of the model, the connectivity of the network is limited. When applying the neocybernetic approach, however, the model is simpler if the network is assumed to be fully connected or *pancausal*. This is because instead of concentrating on the individual connections and their strengths, the aim is to detect the few emerging degrees of freedom.

A linear model structure with multiple variables and parameters can be presented in the form

$$0 = \Gamma^T z, \tag{1}$$

where column vector $z$ contains the system variables and matrix $\Gamma$ the parameters. For example, the ordinary $d$'th order single input, single output (SISO) dynamic system with input $u$ and output $y$

$$y(k) = \sum_{i=1}^{d} a_i y(k-i) + \sum_{j=0}^{d} b_j u(k-j), \tag{2}$$

can be defined as

$$\Gamma = \begin{pmatrix} -1 \\ a_1 \\ \vdots \\ a_d \\ b_0 \\ b_1 \\ \vdots \\ b_d \end{pmatrix} \quad \text{and} \quad z(k) = \begin{pmatrix} y(k) \\ y(k-1) \\ \vdots \\ y(k-d) \\ u(k) \\ u(k-1) \\ \vdots \\ u(k-d) \end{pmatrix}. \tag{3}$$

If there exist more than one constraint between the variables, more columns can be added to matrix $\Gamma$ and the same structure still remains valid.

In the following it is shown that under certain assumptions both the gene regulation network and the metabolic system of a cell can be written in the form (1).

### 4.1 Gene regulation network

It has been suggested (Bunde and Havlin, 1994; Barabasi, 2002) that many distributions in self-organized complex networks statistically follow the *power law*

$$z_j = c z_i^D, \tag{4}$$

where $z_i$ is the free variable, $z_j$ is some emergent phenomenon related to the probability distribution of $z_i$ and $c$ and $D$ are constants. The law (4) can e.g. be applied to relate the popularity of Internet web pages; if $z_i$ is the "ranking of an Internet page" and $z_j$ is the "number of visits per time instant", the dependency of these variables follows the power law.

The single-variable formula (4) can be augmented to include multiple variables:

$$1 = c' z_1^{D_1} \cdot \ldots \cdot z_\mu^{D_\mu}, \tag{5}$$

where $\mu$ is the number of variables. Assuming further that there are multiple dependency structures, the fol-

lowing set of equations is obtained:

$$\begin{cases} 1 &= c_1 z_1^{D_{11}} \cdot \cdots \cdot z_\mu^{D_{1\mu}} \\ &\vdots \\ 1 &= c_\nu z_1^{D_{\nu 1}} \cdot \cdots \cdot z_\mu^{D_{\nu\mu}}, \end{cases} \tag{6}$$

where $\nu$ is the number of dependency structures connecting the variables. Taking a logarithm on both sides of the equations the multiplicative expressions can be transformed to linear structures:

$$\begin{cases} 0 = \log c_1 + D_{11} \log z_1 + \cdots + D_{1\mu} \log z_\mu \\ \quad \vdots \\ 0 = \log c_\nu + D_{\nu 1} \log z_1 + \cdots + D_{\nu\mu} \log z_\mu. \end{cases} \tag{7}$$

Even simpler model structure can be obtained by differentiating the equations around the nominal values $\bar{z}_i$:

$$\begin{cases} 0 &= D_{11} \frac{\Delta z_1}{\bar{z}_1} + \cdots + D_{1\mu} \frac{\Delta z_\mu}{\bar{z}_\mu} \\ &\vdots \\ 0 &= D_{\nu 1} \frac{\Delta z_1}{\bar{z}_1} + \cdots + D_{\nu\mu} \frac{\Delta z_\mu}{\bar{z}_\mu}, \end{cases} \tag{8}$$

where variables $\frac{\Delta z_i}{\bar{z}_i}$ are relative deviations from the nominal state. Since the original system (7) is nonlinear with respect to the original variables, the differentiated version (8) is only valid locally. However, if we assume only small changes near the nominal values, this model is accurate enough and it is easy to see that it actually is of the form (1).

## 4.2 Metabolism

The metabolic system of a cell can be characterized as a group of metabolic reactions as explained in Section 2.2. Let us now write one of these equations in the form

$$a_1 A_1 + \cdots + a_\alpha A_\alpha \underset{k_A}{\overset{k_B}{\Leftrightarrow}} b_1 B_1 + \cdots + b_\beta B_\beta, \Delta H, \tag{9}$$

where there are $\alpha$ reactants $A_i$ and $\beta$ products $B_j$. Additionally, $k_A$ is the reaction speed in forward direction and $k_B$ in backward direction and $\Delta H$ is the change in enthalpy as the reaction proceeds. For each molecule concentration, the rate equation defining the rate of concentration change can be written (see e.g. Atkins (1997)):

$$\frac{dC_{A_1}}{dt} = -k_B C_{A_1}^{a_1} \cdots C_{A_\alpha}^{a_\alpha} + k_A C_{B_1}^{b_1} \cdots C_{B_\beta}^{b_\beta}. \tag{10}$$

In thermodynamical equilibrium all the concentrations remain constant, so that the derivatives are equal to zero. Thus it holds

$$K = \frac{k_B}{k_A} = \frac{C_{B_1}^{b_1} \cdots C_{B_\beta}^{b_\beta}}{C_{A_1}^{a_1} \cdots C_{A_\alpha}^{a_\alpha}}. \tag{11}$$

This equation connects all the system variables together in a multiplicative way. As in the case of gene expression network, the structure can be transformed into a linear one by taking logarithms:

$$\begin{aligned} \log K' = \; & b_1 \log C_{B_1} + b_\beta \log C_{B_\beta} \\ & - a_1 \log C_{A_1} - a_\alpha \log C_{A_\alpha}. \end{aligned} \tag{12}$$

When dealing with local changes in the close vicinity of the nominal operating point $\bar{C}_{B_i}$ and $\bar{C}_{A_j}$, one can further derive the differentiated linear model:

$$\begin{aligned} 0 = \; & b_1 \frac{\Delta C_{B_1}}{\bar{C}_{B_1}} + \cdots + b_\beta \frac{\Delta C_{B_\beta}}{\bar{C}_{B_\beta}} \\ & - a_1 \frac{\Delta C_{A_1}}{\bar{C}_{A_1}} + \cdots + a_\alpha \frac{\Delta C_{A_\alpha}}{\bar{C}_{A_\alpha}}. \end{aligned} \tag{13}$$

In the case of a complete set of metabolic reactions, each reaction leads to its own equation of the form (13). By collecting all the system metabolites into the vector $z$ and adding a column for each reaction in the matrix $\Gamma$ (having zero coefficients for the metabolites not involved in the reaction), one is able to express the complete metabolic system in the form (1).

## 4.3 Combining gene expression and metabolism

We have shown that both gene regulation network and metabolic system of a cell can be assumed locally linear phenomena and described using the same framework. Whereas the metabolic reactions proceed quite fast, the gene regulation network has much slower dynamics; it has been estimated that it takes about 15 min for an activated gene to produce enough transcription factors to activate another (target) gene. However, assuming that we only are interested on the steady state behaviour of the system where the transient dynamics have died away, it is possible to combine the two networks. One can simply collect all the gene activation values with the metabolite concentrations and environmental conditions (temperature, pH) into a high-dimensional state vector, and apply multivariate tools to reveal the low-dimensional latent structure containing the degrees of freedom.

### 4.4 Interpretations of the model structure

When analysing cell functionality it becomes evident that individual components and reactions are not responsible for separate tasks. Instead, each cell function involves multiple reactions and require activity changes in several genes. This means that the behaviour is *redundant* but also *robust*; if one gene or metabolic reaction is disabled, in many cases the cell can overcome the problem and use alternative ways to complete the task. This makes the traditional SISO analysis hard, because strong changes in one variable are required until the effects are seen. However, when analysing all the variables in a neocybernetic manner, the robustness only justifies the assumption of the highly connected network of variables with strong feedbacks.

When the data of the cell are analyzed, some intuitions can be proposed. For example, the degrees of freedom found by multivariate analysis can be interpreted as "functional modes" or cellular functionalities. Each steady state can be described as a linear combination of these directions; some functionalities are going on more than others. It is also possible to analyse the nature of these functional modes; some metabolites and genes are heavily involved in some degree of freedom, thus forming a group of variables connected to that functionality.

## 5 Example case

As a test case, the gene regulation network of yeast *Saccharomyces cerevisiae* was analysed (for a more technical description of the work, see Haavisto et al. (2006)). The yeast is a common model organism and widely studied because it is easy to cultivate and as an eucaryote contains many functional similarities with more advanced organisms. However, the complexity of the metabolic and genetic systems of yeast is very high, and the cells are capable of exploiting multiple different nutrients and live in varying environments (Walker, 2000).

There are two main types of yeast genome time series experiments available, namely stress experiments and cell cycle measurements. In this study we focused on the former type, where the environmental conditions of a non-synchronized yeast cultivation are disturbed, and a time series of the activities of the genes is measured after the shock. Since there are a large number of cells in the cultivation, other genomic activities like cell cycle are averaged away, and in the data only the effects of the shock are seen. These series are also referred to as *stress experiments*.

It was assumed in the analysis that the stress reaction of yeast cultivation proceeds as follows. Originally, the cultivation is grown in a static environment, so that it reaches some density and the average of the cell internal states remains in the nominal level. At the time instant zero, the environmental conditions are suddenly changed, for example the temperature of the cultivation medium is increased from $25°\mathrm{C}$ to $30°\mathrm{C}$. As a response to this, the activities of several genes are either increased or decreased as the cells are trying to adapt to the new situation. This *transient phase* ends when the gene activities stabilize to a new steady state, where the levels of the activities typically differ from the nominal state. In this case, it could be defined that only the transient phase is actually 'stress', whereas the steady states are interpreted as normal function states of the cultivation.

### 5.1 Data

The gene activities of an organism with known genome can be measured using the *microarray* technology (see e.g. Schena (2003)), which gives a "snapshot" of the activities of all the genes at a certain moment. This rather new technology provides huge amounts of biological data and has made possible the data-based analysis of biological cells using the increased computational capacity and data mining tools. However, to model the dynamic behaviour of the genome still remains a nontrivial task, since a typical time series of gene activities contains about 10 time points. As the system dimension e.g. in the case of yeast is over $6000$, the problem is highly underdetermined.

In this study two publicly available data sets were utilized, which originally are published by Gasch et al. (2000) and Causton et al. (2001). These stress response experiments both include responses to temperature and pH changes as well as addition of some chemicals (e.g. salt, sorbitol, etc.). Each time series contains measurements of the transient phase, and it is assumed that in the end of the series new balance is reached. There were altogether $21$ time series with a total of $152$ measurement points. After preprocessing the data and discarding the incomplete genes, there remained about $4000$ genes.

### 5.2 Modeling

As discussed in (Hyötyniemi, 2006b), the framework of elastic systems makes it possible to make hypotheses concerning *goals of evolution.* It can be assumed that individual actors try to maximize their

coupling with the environment, simultaneously maximizing the intake of resources. If this is the case in the real cell, and if there has been enough time for the evolutionary mechanisms to find the optimum, strong mathematical intuitions and tools are available. It is not whatever degrees of freedom that are manifested in data, they are axes of *principal components* (see e.g. Basilevsky (1994)). And when such principal component analysis is carried out for the observation data, it is not just some data analysis; it can be claimed that it is *system analysis,* the mathematical structures corresponding to real physiological structures.

Instead of modeling only the degrees of freedom present in the final steady states of the time series, also the dynamics of the genome were included in the model. This violates the assumption of steady states with stationary statistical properties, since during the transient phase additional degrees of freedom activate and thus the latent dimension of the system state may remain high. However, the results obtained also contain the initial and final states where the assumptions hold.

For creating a dynamic model of gene regulation network we utilized a fairly new modeling method, *subspace identification* (Van Overschee and De Moor, 1996). The method is especially suitable for multidimensional system analysis. However, due to the small number of available data points and their high dimensionality, some modifications to the steps of the basic algorithm had to be made. The method produced a discrete state space model with the environmental conditions (pH, temperature and cultivation medium molecule concentrations) as inputs and gene activity levels as outputs. This kind of model is suitable for example in development of a *Kalman filter*, which can optimally estimate the system state (see e.g. Grewal and Andrews (1993)).

In the modeling phase, the system internal dimension was selected to be four based on the singular value inspection during the subspace identification algorithm calculation. As a result, a drastic model complexity reduction takes place when compared to the original number of genes. This means that all the relevant functional modes present in the stress responses can be coded using these four degrees of freedom, and even some of the transient phase variations are allowed.

## 5.3 Results

The obtained dynamical model could be used to estimate or simulate the yeast gene expression when an environmental change is introduced. To test the quality of the model, responses to the environmental changes present in the original data were simulated by the model, and the results were compared to the measured values. Generally the modeling results were good; the correlation coefficient between the measured and simulated values for individual experiments varied mainly around the value $0.8$, even though a couple of lower values were also detected.

Figure 2 shows the measured and simulated responses of the yeast genome to a step addition of salt. On the left, the measured values are shown, whereas on the right are the simulations. Each row of the figure corresponds to one gene, and each column represents one time point. White color corresponds to high and dark color to low activity values. The number of genes was limited for visualization reasons to the group of $255$ stress-related genes. However, the model was able to simulate the activities of all the $4000$ original genes.

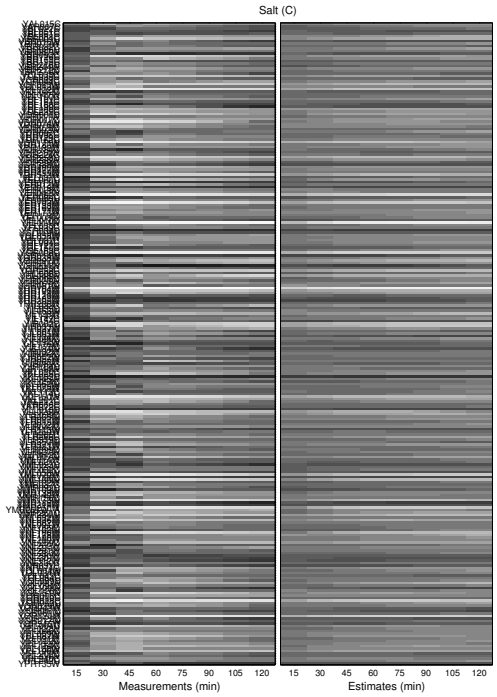


Figure 2: Yeast gene expression response to salt addition, measured and simulated values

When analysing the figure, it can be noted that even though there may be differences between the measured and simulated values on the transient phase, the steady states are modeled quite well. This is because the original neocybernetic assumption of dynamical

balance holds only in the steady state, whereas during the transient phase more complex dynamics and additional degrees of freedom are present in the data.

Because of the lack of data, it was not possible to use a proper validation data set. However, since the model dimensionality is strongly reduced when compared with the original system, it can be assumed that the generalization capability of the model should be quite good. That is, the model is actually able to catch some properties of the real phenomena that produced the data instead of just modeling the given data set.

# 6 Conclusions

A neocybernetic approach for modeling complex systems was discussed in this article. It was shown that both gene expression network and metabolism of a cell can be approximated to follow the proposed linear structure. Utilizing this, data-based principles for analysing the behaviour of cell cultivations were elaborated and a case study of modeling the dynamics of yeast gene expression was presented. The results of the case study encourage the usage of the presented approach for gene expression modeling at least at the high abstraction level.

# Acknowledgements

This work was partly done in the project SyMbolic (Systemic Models for Metabolic Dynamics and Regulation of Gene Expression), funded by the NeoBio program of National Technology Agency of Finland (TEKES) during 2004-2006. OH has also been supported by the Technology Promotion Foundation (TES).

# References

Bruce Alberts, Alexander Johnson, Julian Lewis, Mrtin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell, 4th edition*. Garland Science, 2002.

Peter W. Atkins. *Physical Chemistry*. Oxford University Press, 1997.

Albert-Laszlo Barabasi. *Linked: The New Science of Networks*. Perseus Publishing, 2002.

Alexander Basilevsky. *Statistical Factor Analysis and Related Methods*. John Wiley & Sons, New York, 1994.

Armin Bunde and Shlomo Havlin, editors. *Fractals in Science*. Springer Verlag, 1994.

Helen C. Causton, Bing Ren, Sang Seok Koh, Christopher T. Harbison, Elenita Kanin, Ezra G. Jennings, Tong Ihn Lee, Heather L. True, Eric S. Lander, and Richard A. Young. Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell*, 12:323–337, 2001.

Brent Ewing and Phil Green. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics*, 25:232–234, 2000.

Audrey P. Gasch, Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, and Patrick O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11:4241–4257, 2000.

Mohinder S. Grewal and Angus P. Andrews. *Kalman filtering theory and practice*. Prentice Hall, 1993.

Olli Haavisto, Heikki Hyötyniemi, and Christophe Roos. State space modeling of yeast gene expression dynamics. Submitted to *Journal of Bioinformatics and Computational Biology*, 2006.

Heikki Hyötyniemi. Elastic systems: Another view at complexity. In *Proceedings of SCAI '06*, 2006a.

Heikki Hyötyniemi. Neocybernetics in biological systems. Technical report, Helsinki University of Technology, Control Engineering Laboratory, to be published, 2006b.

Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

Mark Schena. *Microarray Analysis*. Hoboken, NJ, 2003.

Peter Van Overschee and Bart De Moor. *Subspace Identification for Linear Systems*. Kluwer Academic Publisher, Boston, Massachusetts, 1996.

Graeme M. Walker. *Yeast Physiology and Biotechnology*. John Wiley & Sons, 2000.