AN IMPROVED METHOD FOR ANALYTICAL MODELING AND ANTICIPATION OF GENE EXPRESSION PATTERNS

M. U. Akhmet, H. Öktem, G.-W. Weber
Institute of Applied Mathematics
Middle East Technical University, Ankara, Turkey
e-mail: marat@metu.edu.tr, hoktem@math.metu.edu.tr,
gweber@metu.edu.tr
J. Gebert
e-mail: gebert@zpr.uni-koeln.de
Institute of Mathematics, Center for Applied Computer Science
University of Cologne, Germany
S. W. Pickl
University of German Defence Munich, Germany
e-mail: stefan.pickl@unibw.de

Вычислительная биология является одной из наиболее популярных областей науки. Статья посвящена обзору и дальнейшему улучшению методов математического моделирования восстановления структурно-функциональных связей в генных сетях по данным анализа динамики экспрессии генов. Сделан учет неустойчивости на стадии временной дискретизации системы, в частности, учтен смешанный дискретно-непрерывный спектр изменения состояний. Реализована техника статистического доверительного интервала и генных сетей для описания приближения к изменению состояния (разрыву) в уровнях экспрессии генов и правых частях системы, описывающей данные процессы. Алгоритмический аспект задачи также рассматривается.

Introduction

In this work, we study the problem of predicting and anticipating gene expression patterns with various objectives. We extend and improve the former contributions [16, 18, 19]. There, we continuously approximated the behaviour of time-series of gene expression patterns by a system of ordinary differential equations, which we analytically and algorithmically investigated under the parametrical aspect of stability or instability. In this article, additionally possible *jumps* are accounted for. Our algorithm [16, 18, 19] strongly exploited combinatorial information. Reducing the possible locally unstable behaviour to the jumps of individual variables, we need

an algorithm to detect those jumps and determine the temporal ranges where the system is locally stable. A trivial solution to this problem an estimation of the transition matrices locally within an adaptively sized running window for both describing the locally stable behaviour and detecting the locally unstable regions which correspond to transitions (jumps) between the stable states. The availability of genome sequence information facilitates large-scale gene expression analysis. DNA-microarray technology enables us to monitor changes in gene expression (mRNA-concentrations) on a whole genome scale.

With the current techniques in hand we are able to get insights into one of the great miracles of life: its organization and adaptation capacity [28, 42]. This miracle does not only manifest itself in an ecological and evolutionary scale but is also reflected in individual, tissue, cell, molecular, and sub-molecular scales. In our focus, there is a basis of the organization of cellular life, i. e., the concerted expression of genes.

Genes control cellular processes by initiating protein synthesis through mRNA transcription, while transcription is controlled through the binding of proteins or protein complexes to the appropriate promoter region. Hence, gene transcription regulates transcription factors, which in turn regulate gene transcription. This cycle defines a dynamical regulatory network involving highly nonlinear feedback mechanisms [3, 21, 23, 43].

The analysis of gene expression by DNA-microarrays leads to the information whether a gene is expressed or not and, if two different expressions states of the same organism are compared, to what extent the expression of particular genes differs in both states (fold-change).

In this work, first we introduce an analytically solvable model for a stable (in terms of Lyapunov stability, pointwise stability) and linear feedback system. Then, we will treat the genomic network as a locally stable and piecewise-linear system to include well known multistationarity phenomena requiring coexistence of both stable and unstable ranges of behaviour with highly nonlinear transitions [9, 30, 44, 45]. We will enrich our linear and stable model by introducing a self-testing statistical method to determine the locally stable and linear portions of the data. Herewith, we can perform the model estimation locally. Later on, we will formulate the transitions between the stable ranges by using a constant derivative approach, investigate and discuss the asymptotic stability.

In this article, possible jumps are accounted for. We introduce a method for obtaining a unique least mean square estimate for the matrix M representing the stationary behaviour. Inference of the possible jump conditions for a particular gene is even simpler and will be presented. We need to search for a Boolean representation of possible jump conditions only for the limited number of genes which may be subject to positive feedback under some circumstances. Additionally, we extend that approach and present first numerical results and a unique optimum model.

1. Prediction and Anticipation for a Linear System without Delays

The pointwise stability of a system is identified by Lyapunov exponents. A state of a dynamical system is stable if a differential deviation in the initial states is diminished in time, neutrally balanced if a differential deviation in the initial states is preserved in time and unstable if a differential deviation in the initial states is amplified in time. A linear time-invariant system is stable (or unstable) for all values if it is stable (or unstable) at any point of its state space. In this work, we mention Lyapunov stability by *stability* and we will especially indicate when

any other type (e.g. asymptotic) of stability is mentioned.

1.1. Stability, Predicting and Anticipating

The problem of stability is one of the most discussed in the literature to difference equations as Corduneanu mentioned in [5]. He refers to the interesting work of [12, 20] and the PhD Thesis of [47]. There, stability problems are considered using Lyapunov-type functions. In contrast to our polyhedral approach [18], comparison techniques were used there. In [5], stability is investigated by the linearization method.

We give only a short sketch of that interesting method in order to briefly introduce into our discrete procedure.

A necessary and sufficient condition for the exponential stability of the zero-solution of the linear time-invariant system in biology according to [5] is

$$x_{k+1} = \sum_{j=k_0}^{k} M_{kj} x_j, \quad k \in \mathbb{N}_0,\ k \geq k_0 \geq 0, \tag{1.1}$$

where $x_j \in \mathbb{R}^m$ and $M_{kj} \in L(\mathbb{R}^m, \mathbb{R}^m)$ are given. If $x(k, k_0, x_0)$ is the solution of (1.1) with

$$x(k_0; k_0, x_0) = x_0,$$

then the exponential stability is defined by

$$\parallel x(k; k_0, x_0) \parallel \leq C \parallel x_0 \parallel \gamma^{k-k_0},$$

where $C > 0$ and $\gamma \in (0, 1)$ are fixed numbers. In [5], necessary and sufficient conditions are formulated by introducing the following sequence spaces:

$$L^\gamma := \left\{ f; f = (f_k)_{k \in \mathbb{N}_0},\ f_k \in \mathbb{R}^m\ (k \in \mathbb{N}_0), \quad \sum_{k=0}^{\infty} |f_k| \gamma^{-k} < \infty \right\},$$

$$C^\gamma := \left\{ f; f = (f_k)_{k \in \mathbb{N}_0},\ f_k \in \mathbb{R}^m\ (k \in \mathbb{N}_0), \quad \sup |f_k| \gamma^{-k} < \infty \right\}.$$

Then, the operator is defined as follows. For each $f \in L^\gamma$, let y be the sequence

$$y_k = \sum_{j=0}^{k-1} F_{k,j+1} f_j, \quad k \geq 1,\ y_0 = \theta.$$

Here, $F_{k,j+1}$ denotes the fundamental matrix of our system (1.1). The main issue of such an approach is that necessary and sufficient conditions now can be expressed by an inclusion

$$FL^\gamma \in C^\gamma, \quad \text{for some} \quad \gamma \in (0, 1).$$

1.2. State of the Art

The dynamic nature of gene expression time-series is captured by different models, including linear models [11], dynamical Bayesian networks [29] and others.

Chen *et al.* [4] proposed in 1999 to use a system of differential equations $\dot{E} = ME$ to model gene expression data with $E(t)$ being the vector of mRNA and protein concentrations at time t and M being a constant matrix. Due to the fact that protein concentrations are not as often available as mRNA-concentrations, many researchers build their models solely on the basis of mRNA-data. Models of the kind $\dot{E} = ME + B$ are called *linear additive regulatory models*, because they assume that the effect of the regulating genes on a regulated gene could be accumulated. D'haeseleer [11] and Mjolsness *et al.* [27] investigated such models to show that several important known regulatory interactions could be found, although these linear models are only an extreme simplification of the biological system.

The matrix M has the following property: the rows and columns of the matrix stand for genes and the entry m_{ij} represents the level of influence of the expression level of gene j on the change of expression of gene i. By calculating the model's parameters one therefore achieves a gene regulatory network, represented by the matrix M, see also Section 2.

Chen *et al.* proposed two algorithms to construct the model based on time series data of mRNA and protein concentrations. The first one uses minimum weight solutions to linear equations (MWSLE), but the number of genes influencing the expression of the target gene has to be set. The other algorithm is called the Fourier Transform for Stable System (FTSS). It assumes that the system is stable and that genes repeat their expression patterns at cell cycle periods.

De Hoon *et al.* [7, 8] applied their linear model already on mRNA-data of Bacillus subtilis estimating the matrix M with the help of *Akaikes Information Criterion* [1, 22]. Their approach maximizes a likelihood function under the constraint that the matrix M is *sparse*. This is done by estimating the number of parents for each gene in the graph with the help of an information criterion.

In a more flexible approach, Sakamoto and Iba [38] chose the model $\dot{E}_i = f_i(E_1, ..., E_n) \, \forall i = 1, ..., n$ with n being the number of genes and f_i being a function in $E_1, ..., E_n$. Sakamoto and Iba found the functions f_i with the help of genetic programming combined with the least mean square method. Their method worked very well for small samples. For a large-sized network one could use methods of pre-processing to reduce the given networks to small-sized sub-networks.

1.3. Mathematical Modeling Based on DNA Experiments

In [16], we extend the approach of de Hoon *et al.* and Chen *et al.* by letting the matrix M depend on E. We obtain the following time-continuous differential equation

$$(\mathcal{CE}) \quad \dot{E} = M(E)E.$$

Experimentators expect that $M(E)$ better fits the data than a constant matrix M. Furthermore, we intend to calculate the matrix $M(E)$ not only for a small number of genes. This is an extension of the fruitful approach of Sakamoto and Iba. At the end of Section 2, we present a numerical example which is also containded in [16]. This example is restricted to the case where M does not depend on E. The algorithm from [19] gives us information whether our system is stable or unstable according to the corresponding time-discretization.

Our model also extends an approach of Chen *et al.* Therefore, we regard $E = E(t)$ as expression patterns at different times t. The equation $\dot{E} = M(E)E$ describes the *continuous* process of the gene expression. Such biological processes should be stable from an energetic point of view. At the moment we consider 10–50 time-steps of a gene expression experiment.

We simulate the dynamical behaviour with a set of suitable matrices (candidates) which occur if we apply the Eulerian discretization principle in the following way:

According to [16] we get for all $k \in \mathbb{N}_0$:

$$\frac{E_{k+1} - E_k}{h_k} = M(E_k)E_k,$$

delivering the sequence

$$E_{k+1} = (I + h_k M(E_k))E_k,$$

where $h_k = t_{k+1} - t_k$ and t_k is the k-th time step.

Now, we define

$$M_k = I + h_k M(E_k),$$

such that we obtain the following *time-discrete equation* and dynamics:

$$(\mathcal{DE}) \quad E_{k+1} = M_k E_k, \quad k \in \mathbb{N}_0.$$

So we have obtained various candidates for our gene expression process. We can now analyze whether a finite set of these system matrices (or: an a finite approximation of all of them), $\mathcal{M} = \{M_0, ..., M_{m-1}\}$ consisting of m distinct matrices is stable or not with respect to its parametrical entries. Often, the M_j are generations of a matrix group.

We note that the right-hand side of our time-discrete system is of multiplicative form, that is of a great advantage of the modelling of expression data in time by our continuous system $(\mathcal{CE})$. Indeed, the multiplicative recursive definition of $(\mathcal{DE})$ can easily be calculated, and it allows a natural stability analysis of the time-discrete dynamics.

As step by step a multiplication of matrices from $\mathcal{M}$ is performed, stability of the system will mean boundedness of these matrix products, i. e., of the linear mappings defined by them. Thus boundedness can be studied by the matrices' eigenvalues.

Our aim is to analyze gene expression processes with the aid of microarray experiments. The gene expression process has some continuous character (we shall specify this below), whereas each microarray experiment is a discrete *spot*.

The vector-valued (time-continuous) differential equations given in the system $(\mathcal{CE})$ $\dot{E} = M(E)E$ are regarded as being *given*. In [16], we obtained the right-hand side by means of an *ansatz* about the matrix-valued function $M(E)$ and, then, by evaluating a time-series of finitely many measurements (DNA-microarray experiments) $\widehat{E}_j$ at times $\hat{t}_j$ by a *discrete approximation* (called method of "least squares" [10, 35, 36]). We can use different kinds of mostly elemenentary functions in $M(E)$, such as polynomials, trigonometric, exponential or logarithmic functions, connected with parameters where, then, the discrete least-squares approximation refers to [16]. The data at times $\hat{t}_j$, to be approximated, do not primarily consist of the states (measurements) $\widehat{E}_k$, but of (approximate) increase or decrease (mimicking a derivative). These tendencies can be given by the difference quotients

$$\dot{\widehat{E}}_k := \frac{\widehat{E}_k - \widehat{E}_{k-1}}{\hat{t}_k - \hat{t}_{k-1}}, \quad k \in \{1, \ldots, l\}.$$

The choice of how to define $\dot{\widehat{E}}_k$ should depend on the lengths of the time-intervals $\hat{h}_k := \hat{t}_k - \hat{t}_{k-1}$. In the equidistant case $\hat{h}_k \equiv c$, for some small constant $c > 0$, $\dot{\widehat{E}}_k := \frac{1}{2c}(\widehat{E}_{k+1} - \widehat{E}_{k-1})$ $(k \notin \{1, 2\})$ is a common choice.

As often it is very difficult to obtain the right-hand side of $(\mathcal{CE})$ and to resolve the time-continuous system, that discrete matrix calculus is an advantage.

1.4. Perturbation and Noise

Usually, the range of values of the $\widehat{E}_k$, coming from counting realizations of the brightness of the fluorescence signals from the DNA-microarray experiment at time $\hat{t}_k$, is large, e.g., $\widehat{E}_k \in \{0, 1, 2, \ldots, 255\}^n$. In the following, we regard the stochastic aspect of such measurements.

For that reason we begin with a very large number $l + 1$ of measurements and, possibly, some repetition of the sequence of experimental treatments (varying the time intervals hereby). Then, however, the turning to a time-*continuous* system can serve as a good *approximation* of the underlying biochemical reactions. This possible repetition rules out some "noise" firstly and statistical learning begins to take place [22]. Differentiability is understood as the (continuous) differentiability of the solutions (trajectories) of $(\mathcal{CE})$.

Nevertheless, there is already one central condition in our research by which we also take account of random phenomena. In fact, as we study the *stability* (and *instability*) of the system $(\mathcal{CE})$ under slight perturbations [18, 19], there may also be random perturbations considered. Hereby, the perturbations are small stochastical errors. One main advantage of our approach is the evaluation of a certain degree of uncertainty of prediction, i. e., stochasticity in modelling of the open time-horizon.

Our approach based which is on least squares method is to some extent, robust to measurement noise and the accuracy of the method increases with the increasing number of measurements.

2. Determination of the Matrix M

In this section we will recall and deepen our mathematical modeling [16], distinguishing for $M(E)$ the constant and the general case.

2.1. Gene Regulatory Networks

As we mentioned, linear models can already discover some important regulatory interactions. Thus, as a first step, we will focus on the case that M is independent from E, i. e., $M(E)$ is *constant*. Therefore, we have to solve the least square, optimization problem from Section 1.3, which now has the form

$$\min_{M=(m_{ij})} \sum_{k=1}^{l} ||M\widehat{E}_k - \dot{\widehat{E}}_k||^2 ,$$

with $\widehat{E}_k$, $\dot{\widehat{E}}_k$ being the measurements and difference quotients of the gene expressions of genes $1, ..., n$ at time $k = 1, ..., l$, and $||.||$ being the Euclidian norm.

This minimization problem can be restated in vector form as

$$\min_{m=(m_i)} ||\widehat{E}m - \dot{\widehat{\mathrm{e}}}||^2.$$

The vector $m = (m_i)_{i=1,\ldots,n^2}$ consists of the entries of the matrix $M = (m_{ij})$ as follows:

$$m_{(i-1)n+j} := m_{ij} \quad i, j = 1, 2, \ldots, n.$$

The vector $\dot{\widehat{\mathrm{e}}} = (\dot{\widehat{\mathrm{e}}}_j)_{j=1,\ldots,nl}$ is built in a similar way by

$$\dot{\widehat{\mathrm{e}}}_{(i-1)l+k} := \dot{\widehat{E}}_{k,i}$$

for $i = 1, ..., n$ and $k = 1, ..., l$, if we interpret $\widehat{E}_{k,i}$ as the measurement of the expression of gene i at time; k, $k = 1, ..., l$, and $i = 1, ..., n$.

It remains to define our system matrix $\widehat{E}$:

$$\widehat{E} = \begin{pmatrix} \tilde{E} & & \mathrm{O} \\ & \tilde{E} & \\ & & \ddots \\ & \mathrm{O} & & \tilde{E} \end{pmatrix},$$

where $\tilde{E} = (\tilde{e}_{ki})_{k=1,...,l;i=1,...,n}$ with

$$\tilde{e}_{ik} := \widehat{E}_{k,i}.$$

This special *block diagonal* structure of $\widehat{E}$ in our optimization problem

$$\min_{m=(m_i)} ||\widehat{E}m - \dot{\hat{\mathrm{e}}}||^2$$

allows us to decompose the problem into n subproblems, which can be written as

$$\min_{m_{i1},...,m_{in}} \left|\left| \begin{pmatrix} \widehat{E}_{1,1} & \cdots & \widehat{E}_{1,n} \\ \widehat{E}_{2,1} & \cdots & \widehat{E}_{2,n} \\ & \vdots & \\ \widehat{E}_{l,1} & \cdots & \widehat{E}_{l,n} \end{pmatrix} \begin{pmatrix} m_{i1} \\ m_{i2} \\ \vdots \\ m_{in} \end{pmatrix} - \begin{pmatrix} \dot{\widehat{E}}_{1,i} \\ \dot{\widehat{E}}_{2,i} \\ \vdots \\ \dot{\widehat{E}}_{l,i} \end{pmatrix} \right|\right|^2$$

for $i = 1, ..., n$.

In the *general* case

$$M = M(E),$$

where M depends on E, we can easily modify the optimization problem by adding entries to the vector m according to some functional expression. More precisely, if we model the changes of gene expression instead of using a linear function

$$\dot{\widehat{E}}_{k,i} = \sum_{j=1}^{n} m_{ij} \widehat{E}_{k,j},$$

now, with a second order polynomial function

$$\dot{\widehat{E}}_{k,i} = \sum_{j=1}^{n} m_{ij} \widehat{E}_{k,j} + \sum_{j=1}^{n} \sum_{m=1}^{n} m_{jn+m} \widehat{E}_{k,i} \widehat{E}_{k,j},$$

then we expand our system matrix E with entries representing the quadratic term.

Analogously, we expand the vector m. This procedure can be regarded as addition artificial genes to our model. Note, that the number of decomposed optimization problems remains the same. This more general approach has the drawback that we have to consider a huge amount of unknown variables. Even in the case where M is constant we have a high degree of freedom, because the number of genes typically by far exceeds the number of time points for which data are available. Therefore, it is worthwhile to restrict the solution space.

2.2. Restriction of the Solution Space

According to our underlying biological motivation, we want to restrict our solution space, because otherwise we would need a huge amount of expression data to solve our minimization problem.

One aspect is to consider the existence of *sibling genes*, which are genes with very similar expression scores. That is why we can not decide which of these sibling genes regulates another gene; so we summarize these genes to one gene. We would like to detect them before the optimization step. Thus give rise to a reduction of the number of free variables. For detecting such genes we can use clustering algorithms [46].

Moreover, we assume that the values m_{ij} for $i \neq j$ are not less than zero, because no gene consumes another gene. If two genes i and j do not interact at all, the entry m_{ij} is zero. We restrict the decrease of the transcript concentration by a constant vector $\lambda \in \mathbb{R}^n$ assuming that the time steps are small enough. We can add such a condition to our model as follows:

$$m_{ij} \geq \begin{cases} -\lambda(i) & \text{if} \quad i = j, \\ 0 & \text{if} \quad i \neq j. \end{cases}$$

As, in biology, gene regulatory networks are known to be sparse, it is useful to limit the maximum outdegree and indegree of each node [17]. Thus, we get a relatively sparse network, where only the highly significant relations between two genes i and j are exposed as non-zero elements $m_{ij} \neq 0$. To preserve the decomposition property of the minimization problem, we bound only the indegree of each gene by $deg_{\max}$. This means that the maximum number of genes which regulates the expression of this gene, is restricted to $deg_{\max}$ genes. Mathematically, we bound the indegree of each node by imposing binary variables y_i on our optimization problem, defined as follows:

$$y_i = \begin{cases} 0 & \text{if} \quad m_i = 0, \\ 1 & \text{if} \quad m_i \neq 0, \end{cases}$$

reminding that $m_{(i-1)n+j} := m_{ij}$ for $i, j = 1, ..., n$. Thus, our minimization problem becomes:

$$\min_m \left|\left|\widehat{E}m - \hat{\dot{\mathrm{e}}}\right|\right|^2,$$

with

$$\begin{aligned} \mathrm{m_i} \quad &\geq \quad \begin{cases} -\lambda(i) & \text{if} \quad i = kn + k, \ k \in \{1, \ldots, n\}, \\ 0 & \text{if} \quad \text{otherwise}, \end{cases} \\ \sum_{j=(k-1)n+1}^{kn} y_i \quad &\leq \quad deg_{\max} \qquad \text{for} \quad k = 1, \ldots, n+1. \end{aligned}$$

This problem is a mixed-integer programming problem, which can be treated computationally by a branch and cut algorithm.

With this method, it is even feasible to compute a matrix $M(E)$ which depends on E.

The choice of the maximum indegree of each node is implemented in a static manner. Therefore, we plan also to use other methods to obtain a *sparse* network. Good results have been obtained by simultaneously trying to minimize the norm of the matrix M. Perrin *et al.* [33] used the L_1-norm $||M||_1 = \sum_{ij} |m_{ij}|$ in their approach with a dynamical Bayesian network. They stated that instead of just getting a lot of weak connections, it turns out that some coefficients decrease more slowly than others, such that one indeed gets a sparse matrix.

2.3. Numerical Example

We conclude this section by a small numerical example, dealing with four different genes and their expression levels at four different time points. The first gene A does not change its expression level in time. The expression levels of gene B and C decrease or increase, respectively, whereas the expression score of gene D is alternating; see Table 1.

We want to calculate a constant M, which is independent from E, and set the time step equal to one, i. e., $\hat{h}_k = \hat{t}_k - \hat{t}_{k-1} := 1 \ \forall \ k = 1, \ldots, 3$, such that the difference quotients become: $\dot{\widehat{E}}_k = \widehat{E}_k - \widehat{E}_{k-1}$, where $\dot{\widehat{E}}_1 = (0, -50, 50, -255)^T$, $\dot{\widehat{E}}_2 = (0, -20, 20, 255)^T$, $\dot{\widehat{E}}_3 = (0, -10, 10, -255)^T$.

For the calculation we choose $\lambda(i) = 2 \quad$ for $i = 1, \ldots, 4$ and $deg_{\max} = 2$, i. e., the expression of each gene depends on a maximum of two genes. We have to solve the following the mixed-integer minimization problem:

$$\min_{M=(m_{ij})} \sum_{k=1}^{3} ||M\widehat{E}_k - \dot{\widehat{E}}_k||^2$$

subject to

$$\begin{aligned} m_{ij} &\geq \begin{cases} -2 & \text{if} \quad i = j, \\ 0 & \text{if} \quad i \neq j, \end{cases} \\ \sum_{i=(k-1)n+1}^{kn} y_i &\leq 2 \quad \text{for} \quad k \in \{1, \ldots, 4\}, \end{aligned}$$

where y_i is defined as in Section 2.2. After solving the minimization problem we get the matrix M:

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.4 & -0.61 & 0 & 0 \\ 0 & 0.2 & -0.39 & 0 \\ 1 & 0 & 0 & -2 \end{pmatrix}.$$

T a b l e 1
Expression scores of the genes A, B, C and D at four time points

time \ genes	A	B	C	D	
1	[255	250	0	255]	$= \widehat{E}_1^T$
2	[255	200	50	0]	$= \widehat{E}_2^T$
3	[255	180	70	255]	$= \widehat{E}_3^T$
4	[255	170	80	0]	$= \widehat{E}_4^T$

T a b l e 2
Approximation and extrapolation of gene expression using the matrix M

time \ genes	A	B	C	D
1	255	250	0	255
2	255	199.5	50	0
3	255	179	70.4	255
4	255	172	78.9	0
5	255	169	82.6	255
6	255	167	84.2	0
7	255	167	84.9	255
⋮	⋮	⋮	⋮	⋮
100	255	167	85.8	255
101	255	167	85.8	0

By calculating $E_{k+1} = (I + M)E_k$, we get the approximation and extrapolation of the gene expression data as shown in Table 2.

Here, we see the necessity and possibility of prediction and anticipation methods in the challenging field of approximation and extrapolation of gene expression data.

3. Qualitative Behaviour of Gene Regulatory Networks and Local Approach

3.1. Multistationarity

Up to here, we considered a system under the stability aspect, described an algorithm to infer the model parameters for a linear system, and we expanded the algorithms for including some non-linear behaviour. This approach is adequate for the mechanisms involved in homeostasis which maintain the concentrations of some metabolites near a supposed optimal level [44]. However, there are also mechanisms involved in cell specialisation, differentiation, adoption and memorization.

Delbrück [9] suggested that epigenetic differences including those involved in differentiation might reflect *multistationarity*; or existence of more than one stationary steady-state in a biological system (1949). A *stationary steady-state* stands for a property that the mean, variance and autocorrelation structure do not change over time, while time is going to infinity. Multistationarity strictly requires existence of unstable fixed points or ranges of state space exhibiting sensitive dependence to the initial states. This means, there exists $M(E_0) = I + M(E_0)$ such that at least one of its eigenvalues $\lambda_k(M(E_0)) > 1$. Furthermore, those unstable ranges coexist with a multiple number of possible steady-states which imply non-linearity. Various examples of such non-trivial behaviour were demonstrated with Boolean Delay Equations in [30], with Boolean Networks and continuous variable systems in [45].

In fact, a stationary steady-state does not strictly imply stability, nor does instability strictly imply sensitive dependence of the future steady state to the initial conditions for all cases. However, for a piecewise linear system which we will consider here, some types of chaotic attractors corresponding to unstable but almost steady behaviour do not exist.

Genomic regulation can be treated as a multistationary system which may exhibit more than one possible stable state and can perform a transition from one stable state to another. Some expected qualitative features of genomic regulation, their biological meaning and mechanism are summarized in the following Table 3. Stable and locally stable states are expected to appear

T a b l e 3

Aspects of genomic regulation

System features	Biological relevance	Mechanism
Stable (locally stable) Behaviour	Homeostasis	Negative feedback
Multistationarity	Epigenetic differentiation Cell specialization	
Transitions from a possible stable state to another	Adoption Learning Memorization	Activation of a positive feedback circuit

in genomic activity which corresponds to *homeostasis*. Homeostatis stands for maintaining the cell functions at constant or periodically oscillating desired levels. Homeostasis is controlled by negative feedback mechanisms. Multistationarity corresponds to epigenetic differentiation and cell specialization. Epigenetic differences are the differences observed between the cells having the same genome. Existence of various different cells and cell types in the same organism can be explained by multistationarity, i. e., the same system may have various different steady states. On the other hand, *transitions* (*jumps*) from a possible stable state to another also frequently occur in cells. Those may be adoption (for example to the nutrition conditions, responses to drugs, etc.), learning (e.g., plastic changes in neuron connectivities), memorization (e.g., immunity). Such transitions can take place by activation of a positive feedback circuit [44]. Positive feedback can take place only within a range of values. Otherwise, the variables may tend to go to infinity which is not possible.

3.2. Model Class Selection

Modelling of gene regulatory dynamics is a type of *inverse problem*: The available information is the time-evolution of mRNA-concentrations, by which we wish to infer the governing equations for being able to predict the future behaviour, diagnostic state, etc. Let us consider the following time-discrete system:

$$E_{k+1} = f(E_k), \qquad E_0 \text{ is given.} \tag{3.1}$$

When the relation between the next and previous states of the variables can be taken as any function, a very wide range of dynamics can be represented and functions representing the real biological relation can be found within this set. However, in this case, the inverse problem does not have a unique solution, i. e., infinitely many different functions fitting an observation can exist, hence, such an approximation does not have any predictive value. When f is taken as a linear function, the state transition function will be a matrix and (3.1) will turn into $E_{k+1} = ME_k$ form. In this case, a unique optimum solution of M providing the minimum least mean square deviation between the model and observation can be determined. However, as it is mentioned before, linear systems are not capable to exhibit a multistationary behaviour, hence, the dynamic capabilities of a linear approach are beyond the requirements of representing gene regulatory dynamics. As an option, f can be taken as a piecewise linear function. This means, f can be represented by a matrix which is constant within some values and the elements of that matrix may switch to other values when one or some of the variables exceed a *threshold*. A system which may exhibit more than one possible stable state and can jump from one stable

T a b l e 4

Aspects and special properties of difference equations

Consideration of f	Dynamic capabilities	Inverse problem	Biological relevance
Can be any function	Very high	No unique solution	Very high
Linear	Poor	optimal solution	Insufficient
Piecewise linear	Sufficient	Possible	Sufficient

state to another under some circumstances, can be represented by *piecewise linear functions*. Furthermore, Jacob and Monod [24] found out that activation or inhibition of a gene takes place when the concentration of a promoter or inhibitor protein exceeds a *threshold* (1961). This is a very strong motivation to consider piecewise linear class of functions to express the genomic interactions. We can presume that constant rate reaction kinetics (which is linear) determines the expression levels until the concentration of a transcription factor reaches its threshold value. Unless the thresholds are exceeded, no new genes are activated or inhibited to interrupt the locally stable behaviour.

3.3. Formulation

Since a positive feedback circuit may cause the corresponding variable to go to a maximum or a minimum in a relatively short period of time, we can treat those transitions as *jumps*. Thus, we can represent them by a ramp of constant derivative to express the time-evolution of gene expression E_k as follows:

$$\begin{aligned} E_{i,k+1} &= \sum_{j=1}^{n} m_{ij} E_{j,k} \quad \text{if} \quad \theta_{i,k} \neq 1, \\ E_{i,k+l} &= E_{i,k} + l_i s_i \quad \text{if} \quad \theta_{i,k-1} \neq 1 \text{ and } \theta_{i,k} = 1, \\ \theta_{i,k} &= F_B(E_{1,k} > T_1, E_{2,k} > T_2, ..., E_{n,k} > T_n). \end{aligned}$$

Here, $\theta_{i,k}$ is a logical variable determined by a Boolean function F_B of the logical variables $E_{m,k} > T_m$ indicating if the expression level of a gene is over or under a corresponding threshold [26], s_i is the slope of the time-evolution of E_i during a jump and l_i is the duration of the jump. In the preceding sections, we introduced a method for obtaining a unique least mean squares estimate for the matrix M representing the stationary behaviour. Inference of the possible jump conditions θ_i for a particular gene is even simpler and it was discussed in Boolean Networks literature [41] in case if such transitions can be detected in the gene expression profiles. Moreover, we need to search for a Boolean representation of possible jump conditions only for the limited number of genes which are subject to positive feedback under some circumstances. This formulation utilizes the quantitative numerical solution and unique optimum model advantages of the linear (piecewise linear) differential equation approach as much as possible. Hereby, it also allows the qualitative dynamical behaviour matching advantages of Boolean Network approach wherever necessary. Combining Boolean representations of gene state relations with piecewise linear differential equations describing the evolution of protein concentrations was also discussed in [13, 14]. There, a different approach was given which is mainly concerned with how mutations affect the asymptotic stability of a model gene network.

3.4. Detecting the Jumps and the Regions of Locally Stable Behaviour

If we reduce the possible locally unstable behaviour to the jumps of individual variables, then we need an algorithm to detect those jumps and determine the temporal ranges where the system is locally stable. A trivial solution approach to this problem is to estimate the transition matrices locally within an adaptively sized running window. Such a method both describes the locally stable behaviour and detects the locally unstable regions, which correspond to transitions (jumps) between the stable states.

When a available microarray measurements are corrupted with various noise components (as it is in real application) the local approach becomes difficult. That is due to the fact that more noise can be eliminated only by using more samples in the state transition matrix estimation. However, this is not impossible.

Here we propose using *confidence interval statistics* and use *intersection of confidence intervals* (ICI) *rule*. This approach has successful applications in robust signal and derivative estimation problems [15, 31, 37].

The *ICI* Rule

Let the observation model for the measurement $X_{i,k}$ be of the form:

$$X_{i,k} = E_{i,k} + N_{i,k}, \tag{3.2}$$

where $E_{i,k}$ and $N_{i,k}$ are a signal and a zero mean noise, respectively.

Let $\hat{E}_{i,k,h}$ be the *estimate* of $E_{i,k}$ obtained by a smoothing estimator with the window size h. Let us introduce a finite set of window sizes: $H = \{h_0 < h_1 < .... < h_M\}$, starting with a small h_0, and determine a sequence of confidence intervals [2] $D(h_l)$ of the biased estimates corresponding to the window size h_k:

$$D(h_l) \;=\; [U_l, L_l], \tag{3.3}$$

where

$$\begin{aligned} U_k &= \hat{E}_{i,k,h_l} + \Gamma\sigma(i,k,h_l), \\ L_i &= \hat{E}_{i,k,h_l} - \Gamma\sigma(i,k,h_l). \end{aligned}$$

Here, Γ is a threshold of the confidence interval, $\hat{E}_{i,k,h_l}$ is the estimate of $E_{i,k}$ using the window h_l, and $\sigma(i,k,h_l)$ is the standard deviation of this estimate.

The ICI rule gives the adaptive window size by the following procedure:

Consider an intersection of the intervals $D(h_l)$, $l = 1, 2, \ldots, M$, with increasing h_l, and let m be the largest of those l for which the intervals $D(h_l)$, $l = 1, 2, \ldots, m$, have a common point [25, 37]. This m defines the *adaptive window size*.

Herewith, the adaptive window size is defined as the largest window size whose confidence interval of the corresponding estimate intersects with the confidence intervals of all smaller window sizes. This ICI window size selection procedure requires knowledge of the estimate $\hat{x}_{N_j}(i)$ and its local variance only. Although the ICI rule is introduced for local polynomials, this rule can be generalized for other bandwidth dependent estimators [15, 31]. Furthermore, it was shown in [15, 31] that this adaptive window estimation procedure works for derivative estimates (thus for the entries $M_{i,j}$ of M), too. However, it is sufficient to apply the bandwidth for the eigenvalues λ_M of M to verify whether there is a change in the local characteristic of the system.

A simple algorithm to identify the regions and detect the jumps can be described by the following steps:

1. Select the minimum window size from the beginning of the data and estimate the state transition matrix M_{kh_1} using only the data within the minimum size window. Set a confidence intervals of the eigenvalues based on the noise estimate of the window size.

2. Then, increase the window size and compute the transition matrix M_{kh_2} for a larger window which provides a more noise-free estimate. Modify the confidence intervals if they are within the confidence intervals of the first estimates.

3. Continue increasing the window size while the estimates are lying within the confidence intervals of the previous estimates.

4. Pass to the next location (move the center of the window, start to the next window) if any eigenvalue of the state transition matrix estimate lies outside of its confidence interval or the largest allowed window size is reached.

If our hypothesis is valid and gene expression patterns exhibit locally stable behaviour with possible transitions, then the following behaviour can be observed ...

• ... within stable range:

— *growing window estimates are likely to lie within the confidence intervals of the previous estimates;*

— *estimates from the successive windows are likely to be close to each other;*

— *state transition matrices are likely to indicate negative feedback.*

• ... within the state transition regions:

— *growing window estimates of the previous locations are likely to fall out of the confidence intervals;*

— *state transition matrices are likely to indicate positive feedback*;

The main motivations of this approach are as follows:

1. The algorithm is based on hypothesis testing. Therefore, we can also test our assumption while proceeding with the estimation task.

2. The algorithm allows usage of not all but as many as possible samples for each range. This allows as much noise elimination as possible.

After detecting the stable regions and jumps, M can be determined using the data lying on the stable regions. Also the variable(s) exhibiting a jump can be identified and included into the formulation by using the data in the corresponding region.

4. Discussions

As discussed in Section 3.1, inferring the possible jump conditions is not difficult and various methods were discussed in Boolean Networks literature. This inference is not to uniquely solve all possible jump conditions θ_i, but rather to determine the simplest Boolean expression which identifies the jump conditions observed in an experiment. However, the model is applicable for anticipatory prediction, if an asymptotically stable behaviour can be observed as we expect. That is, the variables (gene expression levels) exhibit constant, periodic, almost periodic or quasiperiodic patterns and slight deviations from such patterns converge back to the stationary behaviour [26, 34]. As a multistationary system, a genomic regulatory network (e.g., the model studied here) may also perform transitions from a possible stationary state to another in case of a consistent shift in the environment or a strong stimulus. Naturally, a model identified under some conditions may not give an idea on the epigenetic variants of the same genome. However, the way how epigenetic variations are regulated can convey very valuable information, in case it can be observed (e.g., by changing the environment during the experiment). In [40], possible transitions in steady states were discussed with a perspective of target planning in drug development.

There are two biologically relevant factors important in dynamic behaviour of the system which should also be considered:

1) *delays in the interactions;*

2) *refractory periods.*

The protein synthesis starts with activation of the corresponding gene, continues with mRNA-synthesis, transportation of the mRNA out of the cell nucleus, and synthesis of the proteins [43]. This process definitely takes nonzero time, hence, there exists a *delay* between the concentration of a protein exceeding the threshold and its starting to affect the concentrations of the other proteins. Furthermore, synthesis of different proteins takes time from order of minutes to order of days (for the extreme case). These delays affect a dynamical system in the sense of dynamic variability. When the system can be expressed only by differential (or approximated by difference) equations, the future behaviour of the system merely depends on the values of the variables at a time (initial values). However, when delays are involved, not only the values of the variables but also their whole behaviour within the delay period are distinctive on the future behaviour of the system. Hence, the possible range of initial values and the range of possible stable states are dramatically increased [30]. This dynamical variability is essentially important in genomic regulation. This especially counts if the range of all epigenetic differentiation (both from cell to cell and in time) within an organism is considered. The generability of patterns representing complicated memorization features was demonstrated in [30], and the possibility of generating stable periodic oscillations with negative feedback circuits was shown in [45] when the delays are involved.

Another fact to consider is the *refractory periods*, which stand to some nonzero time before the state of a gene can change again, just after a state change. When a gene is activated or inhibited by binding of a transcription factor to its corresponding promoter region, a reaction requiring nonzero time also must exist to break the bond and another state change can happen. Even though they are not comparable with delays, those refractory periods are very important in the dynamical behaviour of a system too. In fact, they limit the maximum transition frequency. Otherwise, the system may go into a mode of infinitely frequent switchings, called *Zeno effect* in hybrid systems literature [6].

The delays and refractory periods are very difficult to infer in a dynamical system. An approach may include analysis the time differences of the jumps between the locally unstable genes and involving them into the formulation of possible jump conditions. On the other hand, formulating the jumps as constant derivative ramp functions (Section 3.2) somehow involves a refractory period.

Conclusion

In this work, we studied the problem of predicting and anticipating gene expression patterns with various objectives. We started with a solvable formulation based on the system of continuous differential equation:

$$\dot{E} = M(E)E.$$

We paid special attention to the linear case of a constant $M(E) = M$. The matrix-valued function $M(E)$, and the matrix M, were obtained by discrete approximation. Then, we aimed at extending the model to match the qualitative behaviour of the model and the modeled process. On the other hand, we tried to approach most biologically relevant restrictions in the solution space to bring the very general form of the time-discrete formulation

$$E(k+1) = f(E(k))$$

to a solvable form. We formulated a possible locally unstable behaviour as jumps to combine advantages of two popular model classes: discrete difference equations and Boolean networks.

Finally, we introduced a statistical method to detect the possible jumps and we discussed some remaining challenges.

As a future work, formulations by impulsive differential equations [39] and differential equations with discontinuous right side can be studied. Investigating possible algorithms to estimate the possible delays in the interactions is remaining as a further challenge. Possible effects of the perturbations and uncertainities in the gene regulatory networks are also important questions to study.

Acknowledgement: The authors express their graditude to Martin Lätsch, Nicole Radde and Dr. Röbbe Wünschiers for their fruitful collaboration in joint research underlying this article, and to Dr. Igor A. Pestunov and Academician Prof. Yurii Shokin for their kind support.

References

[1] Akaike H. Information theory and an extension of the maximum likelihood principle // 2nd Intern. Symp. on Information Theory. 1971. P. 267–287.

[2] Alder H.L., Roessler E.B. Introduction to Probability and Statistics. W.H. Freeman and Company, 1969.

[3] Chen K.C. et al. Kinetic analysis of a molecular model of the budding yeast cell cycle // Molecular Biology of the Cell. 2000. Vol. 11. P. 369–391.

[4] Chen T., He H.L., Church G.M. Modeling gene expression with differential equations // Proc. Pacific Symp. on Biocomputing. 1999. P. 29–40.

[5] Corduneanu C. Review of some recent results related to discrete-time functional equations // Computing Anticipatory Systems: Proc. Fifth Intern. Conf. CASYS 2001. P. 160–169.

[6] Çamlibel M.K., Schumacher J.M. On the Zeno behaviour of linear complementarity systems // Proc. 40th IEEE Conf. on Decision and Control. 2001. Vol. 1. P. 346–351.

[7] De Hoon M., Imoto S., Kobayashi K. et al. Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations // Proc. Pacific Symp. on Biocomputing. 2003. P. 17–28.

[8] De Hoon M., Imoto S., Miyano S. Inferring gene regulatory networks from time-ordered gene expression data using differential equations // Lecture Notes in Computer Sci. 2534. Berlin: Springer-Verlag, 2002. P. 267–274.

[9] Delbruck M. Discussion. In Unites biologiques douees de continuite genetique. Editions du Centre National de la Recherche Scientifique, Paris. 1949. P. 33–35.

[10] de Trad C.H., Fang Q., Cosic I. An overview of protein sequence comparisons using wavelets // Proc. IEEE-EMBS. 2001. P. 115–119.

[11] D′haeseleer P. Reconstructing Gene Networks from Large Scale Gene Expression Data: Phd thesis, Univ. of New Mexico, 2000.

[12] Driver R.D. Note on a paper of halanay on stability of difference equations // Arch. Rat. Mech. An. 1965. Vol. 18. P. 241–243.

[13] Edwards R., Glass L. Combinatorial explosion in model gene networks // Chaos. Sept. 2000. Vol. 10(3). P. 691–704.

[14] Edwards R., Siegelmann H.T., Aziza K., Glass L. Symbolic dynamics and computation in model gene networks' // Chaos. March 2001. Vol. 11(1). P. 160–169.

[15] Egiazarian K., Katkovnik V., Öktem H., Astola J. Transform based de-noising with window size adaptive to unknown smoothness of the signal // Invited Paper, Proc. SPECLOG 2000, Spectral Transforms and Logic Design for Future Digital Systems. Tampere, Finland, 2000.

[16] Gebert J., Lätsch M., Pickl S.W. et al. Genetic networks and anticipation of gene expression patterns // Proc. CASYS 2003 Computing Anticipatory Systems, Liege, Belgium, 2003.

[17] Gebert J., Lätsch M., Quek E.M.P., Weber G.-W. Analyzing and Optimizing Genetic Network Structure via Path-finding. Preprint, Univ. of Cologne, Germany, Persiaran Institusi, Malaysia, Middle East Technical Univ., Turkey, 2003.

[18] Gebert J., Pickl S.W., Shokina N. et al. Algorithmic analysis of gene expression data with polyhedral structures // 5th Intern. Workshop Similarity Methods, Institute of Statics and Dynamics of Aerospace Structures, Univ. of Stuttgart, 2002. P. 79–87.

[19] Gebert J., Lätsch M., Pickl S.W. et al. An Algorithm to Analyze Stability of Gene-expression Patterns. Preprint ZAIK, Univ. of Cologne, 2002.

[20] Halanay A., Wexler D. Qualitative Theory of Impulsive Systems / Ed. Academiei, Bucharest 4, 1968.

[21] Hartemink A.J., Gifford D.K., Young R.A. Combining location and expression data for principled discovery of genetic regulatory network models // Proc. Pacific Symp. on Biocomputing 2002, Kauai, Jan. 2002.

[22] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. N.Y.: Springer-Verl., 2001.

[23] Holter N.S., Maritan A., Cieplak M. et al. Dynamic modeling of gene expression data // Proc. National Academy of Sci., USA. 2001. Vol. 98(4). P. 1693–1698.

[24] Jacob F., Monod J. Genetic regulatory mechanisms in the synthesis of proteins // J. Molecular Biol. 1961. Vol. 3. P. 318–356.

[25] Katkovnik V. A new method for varying adaptive bandwidth selection // IEEE Trans. on Signal Proc. 1999. Vol. 47, N 9. P. 2567–2571.

[26] Kauffman S.A. Metabolic stability and epigenesis in randomly constructed genetic nets // J. Theoret. Biol. 1969. Vol. 22. P. 437–467.

[27] Mjolsness E., Mann T., Castano R., Wold B. From coexpression to coregulation: an approach to inferring transcriptional regulation among gene classes from large-scale expression data // Advances in Neural Information Proc. Systems. 2000. Vol. 12. P. 928–934.

[28] Monod J. Chance and Necessity. Random House, 1972.

[29] Murphy K., Mian S. Modelling Gene Expression Data Using Dynamic Bayesian Networks. Technical Report, Univ. of California, 1999.

[30] Öktem H., Pearson R., Egiazarian K. An adjustable aperiodic model class of genomic interactions using continuous time boolean networks (Boolean Delay Equations). 2003. Vol. 13. P. 1167–1175.

[31] Öktem H. Transform Domain Algorithms for Biomedical Signal and Image Proceding Problems. PhD Thesis, 2003.

[32] Pearson R. Discrete-Time Dynamic Models. Oxford Univ. Press, 1999.

[33] Perrin B., Ralaivola L., Mazurie A. et al. Gene networks inference using dynamic Bayesian networks // Bioinformatics. 2003. Vol. 19. P. 138–148.

[34] Peleties P., DeCarlo R. Asymptotic stability of m-switched systems using Lyapunov functions // Proc. 31st IEEE Conf. on Decision and Control. 1992. Vol. 4. P. 3438–3439.

[35] Quarteroni A., Sacco R., Saleri F. Numerical Mathematics, Texts in Applied Mathematics. N.Y.: Springer, 1991. Vol. 37.

[36] Riedel K.O. Aspects of Image Processing — Splines, Anisotropic Diffusion, and Biological Models. Mathematical Institute of Univ. of Cologne, 2002.

[37] Ruppert D. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation // JASA. 1997. Vol. 92(439). P. 1049–1062.

[38] Sakamoto E., Iba H. Inferring a system of differential equations for a gene regulatory network by using genetic programming // Proc. Congress on Evolutionary Computation. 2001. P. 720–726.

[39] Samoilenko A.M., Perestyuk N.A. Impulsive Differential Equations. World Scientific, 1995.

[40] Shmulevich I., Dougherty E.R., Kim S., Zhang W. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks // Bioinformatics. 2002. Vol. 18, N 2. P. 261–274.

[41] Shmulevich I., Yli-Harja O., Astola J. Inference of genetic regulatory networks under the best-fit extension paradigm // Proc. 2001 IEEE — EURASIP Workshop on Nonlinear Signal and Image Proc. Baltimore, Maryland USA, Jun. 3–6, 2001.

[42] Schrödinger E. What is life? Cambridge Univ. Press, 1944.

[43] Simon I. et al. Serial regulation of transcriptional regulators in the yeast cell cycle // Cell. 2001. Vol. 106. P. 697–708.

[44] Thomas R. Laws for the dynamics of regulatory networks // Intern. J. Dev. Biol. 1998. Vol. 42. P. 479–485.

[45] Thomas R., Kauffman M. Multistationarity, the basis of cell differentiation and memory. I. Structural conditions of multistationarity and other nontrivial behaviour // Chaos. 2001. Vol. 11. P. 170–179.

[46] Tibshirani R., Hastie T., Eisen M. et al. Clustering Methods for the Analysis of DNA Microarray Data. Technical Report, Division of Biostatistics, Stanford Univ., 1999.

[47] Zouyusefain M.S. Difference Equations of Volterra type. PhD Thesis, 1990.

Received for publication october 4, 2004