A Review on Data Mining and Continuous Optimization Applications in Computational Biology and Medicine

G.-W. Weber [a,*], S. Özöğür-Akyüz [b], E. Kropat [c]

[a] *Institute of Applied Mathematics, Middle East Technical University, 06531 Ankara, Turkey*

[b] *Department of Electronics, Faculty of Engineering and Natural Sciences, Sabancı University, Orhanlı, Tuzla, 34956 Istanbul, Turkey*

[c] *Department of Mathematics, University Erlangen-Nuremberg, 91058 Erlangen, Germany*

Abstract

An emerging research area in computational biology and biotechnology is devoted to mathematical modeling and prediction of gene-expression patterns; it nowadays requests mathematics to deeply understand its foundations. This paper surveys data mining and machine learning methods for an analysis of complex systems in computational biology. It mathematically deepens recent advances in modeling and prediction by rigorously introducing the environment and aspects of errors and uncertainty into the genetic context within the framework of matrix and interval arithmetics.

Given the data from DNA microarray experiments and environmental measurements we extract nonlinear ordinary differential equations which contain parameters that are to be determined. This is done by a generalized Chebychev approximation and generalized semi-infinite optimization. Then, time-discretized dynamical systems are studied. By a combinatorial algorithm which constructs and follows polyhedra sequences, the region of parametric stability is detected. In addition, we analyze the topological landscape of gene-environment networks in terms of structural stability.

As a second strategy, we will review recent model selection and kernel learning methods for binary classification which can be used to classify microarray data for cancerous cells or for discrimination of other kind of diseases.

This review is practically motivated and theoretically elaborated; it is devoted for a contribution to better health care, progress in medicine, a better education and more healthy living conditions.

Key words: Gene-Environment Networks, Computational Biology, Diseases, Birth Defects, Classification, Support Vector Machines, Machine Learning, Model

Selection, Generalized Semi-Infinite Programming, Errors, Uncertainty, Infinity, Modeling, Dynamical System, Intervals, Matrix, Structural Stability, Conic Programming, Continuous, Discrete, Hybrid, Medicine, Health Care

1 Introduction

"*Can mathematics under the limitations of modern technology model the complexity of nature?*", "Yes", but still within the margins of our developing understanding, the margins of approximation in modeling only. Any new improvement to the model gives a chance for a deeper insight into the nature and a hope for a continuously supported service to the people. Likewise, the complexity of the environment which also includes psychological or societal phenomena, and its relation to nature and life of humankind are not an easy modeling task [53]. This paper bases on four foundations: *(i)* contemporary advances in modeling and prediction of gene-expression patterns, *(ii)* recent inclusions of the interactions of biological life with the environment and of *(iii)* errors in measurement by modern DNA microarray technology or in the quantification of the environment and various mutual influences, *(iv)* recent development in classification techniques and relation with gene networks and diseases. We aim at a deepening contribution to scientific progress and services in medicine, health care, food production, industry and education.

There are two quantities coupled for modeling and prediction of gene-expression patterns: the levels (concentrations, *states*) of gene-expressions and their rates of change (*dynamics*); both of them are of a "primal" importance. For the environmental effects, a "dual" role can be identified, such that we speak of some "duality" [53,61] which entirely characterizes our learning problem, represented by a bilevel problem of optimization and decision [61,62]. Indeed, one class of variables contains parameters under perturbation whose response is observed by the other remaining variables that constitute the second class. For a deep understanding about the states and the variation of genetic and environmental patterns we use *matrices*, representing duality and obtained via least-squares (or maximum likelihood) estimation, and an interpretation of their algebra.

Matrices include our gene-environment networks by specifying the concrete dynamical systems on which a testing of the *goodness of data fitting* and

* Corresponding author
Email addresses: gweber@metu.edu.tr (G.-W. Weber), sozogur@sabanciuniv.edu (S. Özöğür-Akyüz), kropat@am.uni-erlangen.de (E. Kropat).

prediction base. They represent linear mappings which determine the time-discrete or time-continuous changes of the states (levels). Their common effect can be expressed in terms of equilibrium, expansion, contraction, cyclicity or mixed asymptotic properties; these behaviours contribute to *stability* or *instability*. Differently from the time-discrete dynamics which can be called a *forward problem*, there is the underlying *inverse problem* of parameter estimation. Those discrete "forward" orbits are resulting from the matrix multiplication stepwise performed, and we analyze them by the combinatorial algorithm of Brayton and Tong [8,52]. This procedure generates and observes a sequence of compact neighbourhoods of the origin. Choosing these neighbourhoods as polytopes allows a translation into the combinatorial language of their vertices; on them the construction principle iteratively applies finitely many matrix multiplications.

Another way of understanding the behaviour or structure of the microarray data can be finding patterns by outlier detection or classifying genes which are cancerous or not, or diseased or not. In this survey, we will also give recent developments in classification tasks to solve these kinds of problems. In [42], a new model selection tool called "confidence level based model selection" is developed for pattern analysis of aminoacid sequence data. Likewise, the same approach can be useful for gene expression or DNA sequence data. Based on the intuiton of "confidence level" approach in [42], *model selection via test margin* is developed in [43] which can be applied on all kind of binary classification problems. Both of these methods have smaller running time than existing model selection methods such as cross validation, and have comparable accuracies.

Classically, e.g., in classically science, technology and medicine, stability has a positive interpretation in terms of some local order, a coming to a rest (recovering) or as the robustness of system against small perturbations such as infections or attacks [26]. In contrast, there is also the negative meaning. An organism, a living being or biosystem which is unable to adapt to a changing environment is in a serious danger caused by bacteria, viruses, radiation and other kinds of attacks. What is more, a stability analysis can also serve for the acceptance or rejection of a mathematical model, i.e., to a testing of the goodness of data fitting and, if needed, by a model improvement. In fact, if any state dimension of the model behaves unbounded under slight parametric variations, then this contradicts the natural-technical limitation of the genetic of environmental levels by bounded intervals.

Genetic network is an established and yet exciting subject of modern science. It means a weighted directed graph composed of nodes representing genes, and of arcs with functional weights standing for the influences between the genes; but also each node can be equipped with a (level) function of the other genes' combined effects on it. For each gene we wish to predict how it

influences the other genes. Various analytic and numerical tools have been developed for the construction and understanding of such networks [1,11,13,19–22,24,30,40,41,47,49,52,60–62,64,65]. Genetic networks are nowadays widely used in computational biology and they have gained significant importance since the human genome project started. As an example to such applications: In [25], a genetic network of a mouse is analyzed to further characterize the differential response to alcohol, and in [66], *GenePath* is developed which is a computer-based system that supports the inference of genetic networks from a set of genetic experiments. GenePath uses abductive inference to explain network constraints based on background knowledge and experimental results.

In [52,53,60–62], we firstly extended genetic networks to *gene-environment networks*. A simple additive shift included on the right-hand side of differential equations served to appropriately extend the model space; then, with our coauthors, we interpreted the shift by the relevant environmental factors. Now, the new nodes are environmental items such as poison in soil, groundwater, in air or food, radiation, but also the welfare and living conditions, temperature (concerning, e.g., global warming), but also education and campaigns for a healthy lifestyle.

For a large number of genes the expression levels can easily be monitored by *DNA-microarray technology* [10]. Despite the fast advances in this high technology, it is nevertheless affected with different uncertainties and measurement ambiguities. Therefore, we included these errors into our model [53,62]. Likewise for the environmental levels and concentrations, we are facing measurement and reliability problems, such that we represent them in error terms, too. As introduced in [53,62], we will represent various kinds of errors by *intervals*.

In general, genetic and gene-environment networks are too large to be easily investigated. Therefore, we imply bounds into the parameter estimation problem which force the number of edges to diminish and make the parameter estimation become a *mixed continuous-discrete programming problem*. Relaxing the inequality constraints to become continuous and depending on the environmental items, maybe also on time intervals and, what is more, on errors and uncertainties located in intervals, the problem becomes a one from *semi-infinite programming* (*SIP*). In addition, by allowing dependence of the domain of combined external effects on the unknown environmental parameters, we obtain a *generalized semi-infinite programming* (*GSIP*) problem. By this, we permit regulation of the network's edge density in a more refined way and we can more confidently guarantee existence and tractability of genetic and metabolic processes.

In [53,60–62] we connected the discrete mathematics of networks with GSIP, by this introducing a new and pioneering scientific approach into computational biology. GSIP is an advancing wide problem class with many moti-

vations, results, future challenges and many practical applications even today [44,46,57]. In computational biology, a sound *modeling*, *prediction* and *process optimization* are very important for a well-understanding of *genetic processes*, of the *optimization of cell metabolism*, and for their applications in medicine, health care, food production, in industry and energy supply. Today, in a time of globalization, of rapid information exchange, of mobility and multicausalities in all kinds of biosystems, communities and societies, the ways how the *environment* expresses itself and exercises effects – often in mutually catalyzing or multiplicative ways, are becoming more and more important. This paper acknowledges this situation and tries to give a scientific service in it.

2 Gene-Expression and Environmental Data, Modeling and Dynamics

2.1 The Interval-Valued Model

At early stages of modeling, gene-environment networks were represented by time-continuous systems of ordinary differential equations (ODEs):

$$\dot{\mathbb{E}} = \mathbb{F}(\mathbb{E}).$$

Here, the d-vector $\mathbb{E} = (\mathbb{E}_1, \mathbb{E}_2, \ldots, \mathbb{E}_d)^T$ comprises the positive concentration levels of proteins (or mRNAs, or small components) and certain levels of the environmental factors, while $\dot{\mathbb{E}}\,(= \frac{d\mathbb{E}}{dt})$ represents a continuous change in the gene-expression data, and $\mathbb{F} : \mathbb{R}^d \to \mathbb{R}^d$ is composed of nonlinear coordinate functions $\mathbb{F}_i : \mathbb{R}^d \to \mathbb{R}$ that determine the rate of change of each data item (cf. [11,29,45,52] for different dimensions). In this paper, we offer a parameter estimation on unknowns implied into the definition of $\mathbb{F}$, established on experimental data vectors $\overline{\mathbb{E}}$ of those levels. Since the vectors $\overline{\mathbb{E}}$ obtained from microarray experiments and from environmental measurements in a widest sense are merely approximating the actual states $\mathbb{E}$ at the sample times of the experiments, we have the following relations at these times [53]

$$\mathbb{E}_i = \overline{\mathbb{E}}_i \pm \mathrm{err}_i \quad (i = 1, 2, \ldots, d);$$

here, $\mathrm{err}_i \geq 0$ is an error likely to be made at the experimental measurements of the gene- or environmental expression level $\mathbb{E}_i$. For a closed representation of all cases, we use intervals determined by some maximal measurement error $\mathrm{Err}_i > 0$ which leads us to consider the state $\mathbb{E}_i$ just to be the interval

$$[\overline{\mathbb{E}}_i - \mathrm{Err}_i, \overline{\mathbb{E}}_i + \mathrm{Err}_i]$$

and, hence, $\mathbb{E} = (\mathbb{E}_1, \mathbb{E}_2, \ldots, \mathbb{E}_d)^T$ to be in the d-dimensional parallelpipe

$$\bigtimes_{i=1}^{d} [\overline{\mathbb{E}}_i - \mathrm{Err}_i, \overline{\mathbb{E}}_i + \mathrm{Err}_i].$$

For this approach we suppose that there is no functional dependence among any two of the errors made in the measurements of the gene-expression levels $\mathbb{E}_i$. We obtain confidence intervals and a confidence parallelpipe here, when taking into account dependence in some stochastic or statistical sense [6]. In general, there are confidence regions, e.g., given by confidence levels (yielding ellipsoids and other kinds of level sets; cf. Subsection 6.2). For further details and definitions on interval analysis we refer to [59].

This entire wide framework allows us to approximately address the nature of biological, environmental phenomena, and technical phenomena of measurement and modeling as well; it extends the one from [22,24] such that the continuous equation looks as follows [52,53,61]:

$$(\mathcal{CE}) \qquad \dot{\mathbb{E}} = \mathbb{M}(\mathbb{E})\mathbb{E}, \quad \mathbb{E}(t_0) = \mathbb{E}^{(0)}.$$

Here, $\mathbb{M}(\mathbb{E})$ is a $(d \times d)$-matrix whose entries are intervals and defined by a family of functions which include unknown parameters. Now, intervals represent uncertainty with respect to the interactions between the genes and to the effects between the environment and the genes; herewith, they will constitute a dynamics. The point $\mathbb{E}^{(0)} = (\mathbb{E}_1^{(0)}, \mathbb{E}_2^{(0)}, \ldots, \mathbb{E}_d^{(0)})^T$ consists of the interval-valued initial levels, available, e.g., by the first experimental data point $\overline{\mathbb{E}}(t_0) = \overline{\mathbb{E}}^{(0)}$. For finding an approximate model and network, the least-squares optimization problem will finally be restricted by bounds imposed on the number of regulating effects exercised per gene and depending on the effects of the environment onto the genes.

2.2 Two Levels of the Task

Concerning the parameterized entries of the model $(\mathcal{CE})$ we have to examine a *bilevel problem* [21,22,32,46,53,57,61] of two different problem stages, namely, *optimization* and *stability analysis*. The *optimization (approximation) problem* of squared errors bases on the following form:

$$\min_{y} \quad \sum_{\kappa=0}^{l-1} \left\| \mathbb{M}_y(\overline{\mathbb{E}}^{(\kappa)})\overline{\mathbb{E}}^{(\kappa)} - \dot{\overline{\mathbb{E}}}^{(\kappa)} \right\|_{\infty}^{2}.$$

The vector y comprises a subset of all the parameters and the vector $\dot{\overline{\mathbb{E}}}^{(\kappa)}$ comprises interval-valued *difference quotients* based on the κth experimental

data $\overline{\mathbb{E}}^{(\kappa)}$ and on step lengths $\overline{h}_\kappa := \bar{t}_{\kappa+1} - \bar{t}_\kappa$ between neighbouring samplings times [19,24,53]:

$$\dot{\overline{\mathbb{E}}}^{(\kappa)} := \frac{\overline{\mathbb{E}}^{(\kappa+1)} - \overline{\mathbb{E}}^{(\kappa)}}{\overline{h}_\kappa} \quad (\kappa = 0, 1, \ldots, l-1).$$

The *stability of the dynamics* is investigated with respect to the remaining parameters. For this a combinatorial algorithm on polyhedra sequences observed is used to detect the regions of stability. Indeed, the key advantage of $(\mathcal{CE})$ lies in its structure that allows a time-discretization represented by a sequence of matrix multiplications. Based on this recursion, a stability analysis of combinatorial and geometrical type with polytope series is permitted [22], combined by us with our matrix algebra [52,53].

2.3 On the Environment

The interaction between the genes and the environment is frequently characterized as *epigenetic*. This refers to stable changes of gene expression patterns in response to environmental factors without any mutations in the DNA sequence. *DNA methylation* is one of the most common epigenetic factors, but there are also others, such as *acetylation*, *ethylation* and *phosphorylation*, providing important epigenetic regulations. Studies on identical twins showed that although they have the same genomic sequences and genes, but no epigenetic difference during the early stages of life, adult twins exhibited very different epigenetic patterns affecting their gene-expression portrait [18]. Furthermore, nutritional conditions of grandparents can have phenotypic consequences in their grandchildren [17,35]. Life style, nutritional supplementation, and environmental conditions can have a very important impact on inheritance by changing the DNA sequence with mutations and also by affecting epigenetic pattern of DNA through methylation, ethylation, etc., without changing the DNA sequence. Hence, for a better explanation of the complexity of nature, genetic networks cannot be studied solely without taking into consideration the environmental factors which affect epigenetic patterns and, thus, gene expression patterns [61].

2.4 On Errors as Further Variables

Beyond the extension from n genes to the m environmental factors, in this paper, a further dimensional augmentation is implied by the l errors in data fitting related with the l genetic and environmental sample vectors obtained. These errors can be addressed in a squared or an un-squared form [62], and

they will be detected by minimized upper bounds τ_κ. Then, the entire string of variables is displayed by a vector

$$(E_1, E_2, \ldots, E_n, \check{E}_1, \check{E}_2, \ldots, \check{E}_m, \tau_1, \tau_2, \ldots, \tau_l)^T,$$

but we could further include the cumulative environmental factor as affecting each gene (which would imply $n+m$ rather than n environmental dimensions), or represent the sum of all squares by *one* level τ only (cf. Section 6).

2.5 Example for a Gene-Network

Let us from now on for a while focus on the n genes and their interactions and, then, step by step, return to our general model in dimension $d > n$; actually, $d = m+2n$ as we will see, with m being the number of environmental items. In Section 3, we shall return to the d dimensional (extended) model and mainly add the influence of the environment on the gene.

Example 2.1 *The dynamics of n genes can be determined by the following system of differential equations [23,24]:*

$$\dot{E}_i = -\delta_i E_i + \sum_{\alpha=1}^{\alpha_i} (reg\, f^+)_\alpha + \sum_{\beta=1}^{\beta_i} (reg\, f^-)_\beta + c_i \quad (i = 1, 2, \ldots, n).$$

In this model real- or interval-valued rates of basic synthesis and basic degradation of gene i are represented by $c_i \geq 0$ and $\delta_i \geq 0$ whereas activation or inhibition by other network components are determined by the two sums. The activation and inhibition functions $reg\, f^+$ and $reg\, f^-$ have been shown to possess a sigmoid shape [63]. The resulting $(n \times n)$-matrix $M(E)$, where $E = (E_1, E_2, \ldots, E_n)^T$ consists of the first n components of $\mathbb{E}$, has the entries

$$m_{ii}(E) = \frac{c_i}{E_i} - \delta_i + k_{ii} \frac{E_i^{m_{ii}-1}}{E_i^{m_{ii}} + \theta_{ii}^{m_{ii}}} \quad (i = 1, 2, \ldots, n),$$

$$m_{ij}(E) = k_{ij} \frac{E_j^{m_{ij}-1}}{E_j^{m_{ij}} + \theta_{ij}^{m_{ij}}} \quad (i, j = 1, 2, \ldots, n;\ i \neq j)$$

with k_{ij} and $\theta_{ij}, m_{ij}(E)$ being any or nonnegative reals (or intervals), respectively. Now, some or all of the parameters can be estimated based on data from DNA-microarray experiments.

2.6 Gauss-Chebychev Approximation and Optimization in the Presence of Intervals

In *Chebychev approximation* we refer to infinite data, mostly uncountably many ones in the form of a continuum, and look for a member in a family of functions with less complexity which approximates a given complicated function best, in terms of the "maximal error minimized" [28,31]. If in an approximate sense the solution E (or $\mathbb{E}$) and $\dot{E}$ (or $\dot{\mathbb{E}}$) of initial value problem over some *time* interval, or a *confidence* interval (*error* of *uncertainty*), is given, then the parametric functions can be "uniformly" estimated by finding the matrix $M(E)$ (or $\mathbb{M}(\mathbb{E})$). Chebychev approximation problems can be reformulated as *semi-infinite programming* (*SIP*) *programs* [31]. Sometimes, we are *in between* Gaussian and Chebychev approximation [51]. Then, the functions E and $\dot{E}$ are approximately known by *patterns* in the *piecewise* sense of some subintervals and points which, in this paper, are interval-valued, per coordinate and state. Such a *hybrid* kind of approximation is called by us as *Gauss-Chebychev* (cf. also Section 6).

3 From the Special to the Extended Dynamics of Gene-Expression and Environmental Patterns

The dynamics of the n genes and their interaction alone can be described as follows:

$$(\mathcal{CE})_{\text{gene}} \qquad \dot{E} = M(E)E.$$

This model shares with $(\mathcal{CE})$ the same multiplicative structure, which is the basis of the recursive iteration idea [22]. Not to loose this recursion property by the shifts proposed in the model extension of $(\mathcal{CE})_{\text{gene}}$ by introducing constant affine linear shifts terms in [64,65], we will reconstruct the form of $(\mathcal{CE})_{\text{gene}}$ by a dimensional model extension. This will even allow to represent our following *affine* continuous equation which includes a variable shift vector [47–49,52,61]:

$$(\mathcal{ACE})_{\text{gene}} \qquad \dot{E} \;=\; M(E)E + C(E).$$

Here, the additional column vector $C(E)$ provides a more accurate data fitting and may represent environmental perturbations or contributions. Differently from $M(E)E$ which exhibits E as a factor explicitly, the shift $C(E)$ does not need to implicitly possess E as a factor. This shift may be, e.g., exponential, logarithmic, trigonometric, but also piecewise polynomial. If the interval entries of $M(E)$ and $C(E)$ are given in a closed or piecewise form by polynomials, then the vector $C(E)$ of various environmental effects should reveal degrees less than the ones in the vector $M(E)E$. An additive decomposition as given by $(\mathcal{ACE})_{\text{gene}}$ can be called a *normal form*, an *unfolding* [5,9,27,32]

or a *(generalized) additive model* [27,50]. In fact, emissions, poison in water or food, dangerous drugs, social stress, changes in the lifestyle, (quantifiable) educational measurements, and other environmental effects are displayed to form the right-hand side of the system $(\mathcal{ACE})_{\text{gene}}$. In this sense, we distinguish and display special effects on each gene examinated by any environmental item itself or cumulatively by all or several items working together or catalyzing each other. This cumulative effect may not be further divisible or quantifiable by the single effects.

With $(\mathcal{ACE})_{\text{gene}}$ we included the disturbances and genetic changes caused by the environment, in long and in short term, but we lost the convenient recursive idea of matrix multiplication first of all. This drawback can be overcome by increasing the dimension of the state space to $d := m + 2n$ such that we reconstruct that product structure. This reconstruction presented in [61] but now modified by interval-valued entries [53], works as follows. We split $C(E)$ of $(\mathcal{ACE})_{\text{gene}}$ into the sum $\mathrm{W}(E)\check{E} + \mathrm{V}(E)$, which yields

$$(\mathcal{ACE}) \qquad \dot{E} \;=\; M(E)E + \mathrm{W}(E)\check{E} + \mathrm{V}(E)$$

with $\check{E}(t) = \Big(\check{E}_1(t), \check{E}_2(t), \ldots, \check{E}_m(t)\Big)^T$ being a specific m-vector (of intervals) which comprises the levels of the m environmental factors that can affect the gene-expression levels and their variation. While some of the coordinates (factors) $\check{E}_\ell$ affect in a short term, the others may affect in a long term. We may think of $\check{E}$ as constant, but also piecewise constant or, generally, time-dependent. In the case of a constant component $\check{E}_j$, we can easily normalize it to unity: $\check{E}_j \equiv 1$.

By the weight matrix $\mathrm{W} = (\mathrm{w}_{i\ell})_{\substack{i=1,\ldots,n \\ \ell=1,\ldots,m}}$, the effects of the factors $\check{E}_\ell$ on the gene-expression data E_i become incorporated into the system, and the n genes and the m environmental factors are individually matched. Differently and complementary to this, the column vector $\mathrm{V}(E) = (\mathrm{v}_i)_{i=1,\ldots,n}$ represents all the cumulative effects of all (or several) environmental items influencing the genes together. This cumulative effect could also be represented by a new, $(m+1)$st environmental item, taken into account for each gene. In the time-continuous (instantaneous) system $(\mathcal{ACE})_{\text{gene}}$, the interval value $\sum_{\ell=1}^{m} \mathrm{w}_{i\ell}(E)\check{E}_\ell \; + \mathrm{v}_i$ is interpreted as the total effect of the environment on the expression level E_i of gene i. Now, we overcome the more complex form of $(\mathcal{ACE})_{\text{gene}}$ by an idea introduced in [47–49,52] and refined in [53,61]:

$$\mathrm{W}(E)\check{E} + \mathrm{V}(E) \;= \check{M}(E)\check{E}^{\vee}.$$

Here, the *gene-environment* matrix $\check{M}(E) := \Big(\mathrm{W}(E) \,|\, \mathrm{diag}(\mathrm{V}(E))\Big)$ consists of $n \cdot (m+n)$ intervals; its second block represents $\mathrm{V}(E)$ as a diagonal matrix with intervals on the diagonal. Now, putting $\check{E}^{\vee} := (\check{E}^T, \mathrm{e}^T)^T$ with the n-vector $\mathrm{e}^T := (1, 1, \ldots, 1)$ of ones only, we get the following compact form for

$(\mathcal{ACE})_{\text{gene}}$:

$$\dot{E} \;=\; M(E)E + \check{M}(E)\check{E}^{\vee}.$$

Introducing the following $d = m + 2n$-vector

$$\mathbb{E} := \begin{pmatrix} E \\ \check{E}^{\vee} \end{pmatrix},$$

and the $(d \times d)$-matrix

$$\mathbb{M}(\mathbb{E}) = \begin{pmatrix} M(E) & \check{M}(E) \\ 0_{(m+n)\times n} & 0_{(m+n)\times(m+n)} \end{pmatrix} = \left(\begin{array}{c|cc} M(E) & \mathrm{W}(E) & \mathrm{diag}(\mathrm{V}(E)) \\ \hline 0_{m\times n} & 0_{m\times m} & 0_{m\times n} \\ 0_{n\times n} & 0_{n\times m} & 0_{n\times n} \end{array}\right),$$

we arrive at our extended system $(\mathcal{CE})$ together with an extended initial value as follows:

$$(\mathcal{CE}) \qquad \dot{\mathbb{E}} = \mathbb{M}(\mathbb{E})\mathbb{E}, \qquad\qquad \mathbb{E}^{(0)} = \mathbb{E}(t_0) = \begin{pmatrix} E^{(0)} \\ \check{E}^{\vee,0} \end{pmatrix}.$$

Now, we learn that there is an *equivalence* between this initial value problem and the corresponding initial value problem for $(\mathcal{ACE})_{\text{gene}}$ [53]. In general, $E^{(0)}$ and $\check{E}^{\vee,0}$ are chosen as the first experimental data vectors $\overline{E}^{(0)}$ and $\overline{\check{E}}^{\vee,0}$ coming from microarray experiments, followed by the environmental observations. Here, $\overline{\check{E}}^{\vee,0}$ is the initial state of the special or cumulative environmental factors having an impact on E and being expressed in a physical, chemical, financial or social dimension. If the ℓth specific environmental factor $\check{E}_\ell$ is considered to affect any gene-expression level, then, initially, the ℓth component of $\overline{\check{E}}^{(0)}$ is regarded to be 1, otherwise 0. Here, 1 (0) in $\overline{\check{E}}^{(0)}_\ell$ means that the ℓth environmental factor is "switched on" (or "off", respectively). In contrast, the cumulative environmental effect is considered to be "switched on" always. The initial state $\overline{\check{E}}^{(0)}$ (or $\overline{\check{E}}^{\vee,0}$) could also be any other vector [52].

In $(\mathcal{CE})$, equipped with the initial value $\check{E}^{\vee}(t_0) = \overline{\check{E}}^{(0)}$, the time-dependent variable $\check{E}^{\vee}(t)$ is constant: $\check{E}^{\vee} \equiv \overline{\check{E}}^{\vee,0}$. We do indeed not include any environmental dynamics, but our modeling framework allows us to do this. In fact, by turning the 0 matrices in the second and the third (block) columns of $\mathbb{M}(\mathbb{E})$ in $(\mathcal{CE})$ to matrices different from 0, we could accept variable and interacting factors of the environment. Permitting also the 0 matrices in the first column to have entries $\neq 0$, this would express that genes affect environmental items. In addition, we could allow dependence of $\mathrm{V}(E)$ and $\mathrm{W}(E)$ on the variable $\check{E}$

or even $\check{E}^{\vee}$. Later on, Section 5 would even allow to incorporate such a higher generality of $(\mathcal{CE})$.

4 The Time-Discretized Model and Stability Analysis

4.1 Time-Discretization

For a numerical analysis of the dynamics of gene-expression patterns the paper [16] introduced *Runge-Kutta methods* (*RK*) into our time-continuous modeling. Then, the works [47–49] used a different RK method called *Heun's method* in some extended model space. This method is a modification of Euler's method; it is more an illustrative, explicit and the simplest RK approach [14,16,47–49]. Here, but also in the Eulerian case and some other methods [14,16,22], we can find a representation in "multiplication-form", that allows to calculate predictions of future expression values:

$$(\mathcal{DE}) \qquad \mathbb{E}^{(k+1)} = \mathbb{M}^{(k)}\mathbb{E}^{(k)}.$$

Let the given data from DNA microarray experiments and environmental measurements be comprised by $\overline{\mathbb{E}}^{(\kappa)} := \left((\overline{E}^{(\kappa)})^T, (\check{E}^{\vee,\kappa})^T\right)^T \ (\kappa = 0, 1, \ldots, l-1)$. By $\widehat{\mathbb{E}}^{(\kappa)}$ $(\kappa = 0, 1, \ldots, l-1)$ we denote the approximations in the sense of $(\mathcal{DE})$, and we put $\widehat{\mathbb{E}}^{(0)} = \mathbb{E}^{(0)}$. Now, the kth approximation or prediction, $\widehat{\mathbb{E}}^{(k)}$, is calculated by

$$\widehat{\mathbb{E}}^{(k)} \ (:= \mathbb{E}^{(k)}) \ = \mathbb{M}^{(k-1)}(\mathbb{M}^{(k-2)} \cdots (\mathbb{M}^{(1)}(\mathbb{M}^{(0)}\mathbb{E}^{(0)}))) \ \ (k \in \mathbb{N}_0).$$

In [47–49,64,65], referring to earlier stages of modeling, we compared the first l predicted expression vectors with the l data vectors and, by this, investigated the quality of prediction, both theoretically and by numerical examples.

Via $(\mathcal{DE})$ we obtain our *gene-environment networks* by the time-discrete dynamics (while our investigation permits a time-continuous approach to the networks via $(\mathcal{CE})$, too). Indeed, the genes and environmental items are represented by the nodes (vertices) of our network; the interactions between them turn to edges weighted with effects (in the time-continuous case: with functional values). Namely, the significant entries of $\mathbb{M}^{(k)}$, say, $m_{ij}^{(k)}$, $m_{i,n+\ell}^{(k)}$ or $m_{i,n+m+i}^{(k)}$, are the effects multiplied by $\mathbb{E}_j^{(k)}$, $\mathbb{E}_\ell^{(k)}$ or 1. In this way, at the discrete time step $k \mapsto k+1$ the expression level of the ith gene becomes changed by the one of the jth gene (or ℓth environmental item or the cumulative environmental, respectively). Now, the rich arsenal of discrete mathematics and its network algorithms in both versions, statically and dynamically, becomes

applicable on subjects such as connectedness, components, clusters, cycles, shortest paths or further subnetworks.

4.2 Stability Analysis

Let $\mathcal{M} := \{\mathbb{M}_0, \mathbb{M}_1, \ldots, \mathbb{M}_{z-1}\}$ be a finite set of z matrices over the intervals (as entries) be obtained from $(\mathcal{DE})$ with a sufficiently fine discretization of M, W and V and an entry-wise optimization [47–49,53] (with no confusion by the previous meaning of $\mathbb{M}^{(k)}$ as kth iterate). Let $\mathcal{M}'$ be the matrix set of all the finite matrix multiplications of elements from $\mathcal{M}$. The following definition originates in [8], but has been extended by us dimensionally and by interval-valuedness; we also include an alternative for the reader's possible preference:

Definition 4.1 *[53] The matrix set $\mathcal{M}$ (herewith, $(\mathcal{DE})$), is called* stable *if for every neighbourhood in $\mathbb{C}^d$ (or relative neighbourhood in $\mathbb{C}^n \times \{0'_{n+m}\}$), $\mathcal{U}$, of the origin 0_d (or affine origin $0'_d$, given from 0_d by shifting to 1 some of the middle m coordinates and all of the last n coordinates), there exists a (relative) neighbourhood $\mathcal{V}$ of the origin 0_d (or $0'_d$) such that for each $\mathbb{M} \in \mathcal{M}'$ it holds: $\mathbb{M}\mathcal{V} \subseteq \mathcal{U}$.*

We note that for the time-continuous system $(\mathcal{CE})$, in case of constant time shifts, i.e., $h_t \equiv h\,(t \in \mathbb{R}_0^+)$, then there is a dynamics (a continuous orbit) piecewise defined along all the intervals $[kh, (k+1)h)$. (If, in addition, the initial section $E(t)\,(t \in [0, h))$ is a constant parallelpipe, then the dynamics is piecewise constant.) Herewith, a condition of *stability* can be defined analogously as in the previous definition. For that case and provided that we concentrate on Euler discretization, having turned from the scalar- to our interval-valued model framework, if the function $\mathbb{M}$ of the right-hand side of $(\mathcal{CE})$ is Lipschitzian, we learn the following result from [62]. It is an extension of the real-valued case where it even holds for some Runge-Kutta discretizations presented [61] and, indeed, a unifying concept.

Theorem 4.1 *Let the map $x \mapsto \mathbb{M}(x)$ $(x \in \mathbb{R}^d)$ be Lipschitzian. If the Eulerian time-discrete system $\mathbb{E}^{k+1} = \mathbb{M}^k\mathbb{E}^k$ $(k \in \mathbb{N}_0)$, $\mathbb{E}^0 \in \mathbb{R}^d$, as in $(\mathcal{DE})$, some appropriate $h_{max} > 0$ being given, is stable for all values $h_k \in [0, h_{max}]$, then the time-continuous dynamics defined by the system $\dot{\mathbb{E}} = \mathbb{M}(\mathbb{E})\mathbb{E}$ (with $h > 0$ sufficiently small) is also stable.*

The parallelpipes $\mathbb{E}$ can (after some dilatation) be embedded into neighbourhoods of 0_d. Multiplying our matrices and vectors (over intervals) and observing the resulting discrete orbits can be characterized by the scalar-valued case that was introduced and investigated in, e.g., [8,22,61]. Indeed, each member in an orbit of our set-valued products is representable as the convex hull of

the corresponding common matrix products that we obtain by focusing on all of the finitely many combinations of the involved interval endpoints. By referring to these endpoint combinations, we indeed reduced the stability condition to the classical one for the scalar-valued case [53,61,62]. Herewith, we have carried over the stability theory and algorithmic methods of our and our colleagues' former investigations, e.g., the previous condition of parametric stability can be characterized analytically, spectrally and by Lyapunov functions. Our main method of analysis employs discrete orbits provided by stepwisely applying matrices on a compact neighbourhood of the origin 0_d. Choosing the initial neighbourbood and, henceforth, each element of a generated sequence, as a polytope gives the opportunity to detect parametric regions of stability and instability. If the polytope sequence is bounded, then there is stability given for that parametric constellation, otherwise instability. By this stability analysis we can make a *testing* of the goodness of data fitting of our model. A second method which we will present for such a testing will consist in the investigation on structural stability of the landscapes of gene-environment networks (cf. Subsection 6.2). We remark that our modeling and dynamical analysis can also be used for *metabolism-environment networks* [62].

5 Extracting and Optimizing Gene-Environment Networks in the Presence of Intervals

5.1 Our Hybrid Model

The hybrid approach from [24] offered a complete dynamical description of the expression levels of n genes. Then, the papers [53,61] modified it by additionally matching the n genes with m special items and the cumulative item of the environmental, and by turning to the interval-valued setting:

$$(\mathcal{HE}) \qquad \begin{aligned} \dot{E}(t) &= M_{s(t)}E(t) + \mathrm{W}_{s(t)}\check{E}(t) + \mathrm{V}_{s(t)}, \text{with} \\ Q(E(t)) &= (Q_1(E(t)), Q_2(E(t)), \ldots, Q_n(E(t))), \quad \text{where} \\ Q_i(E(t)) &:= \begin{cases} 0, \ E_i(t) < \theta_{i,1} \\ 1, \ \theta_{i,1} \leq E_i(t) < \theta_{i,2} \\ \vdots \\ d_i, \ \theta_{i,d_i} \leq E_i(t) \quad (i = 1, 2, \ldots, n). \end{cases} \end{aligned}$$

In $(\mathcal{HE})$, *thresholds of the expression levels* are given by $\theta_{i,1} < \theta_{i,2} < \ldots < \theta_{i,d_i}$. At these thresholds instantaneous changes of the parameter constellation can occur and we have to choose a local model by the special selection of the

matrices $M_{s(t)}$, $\mathrm{W}_{s(t)}$ and the vector $\mathrm{V}_{s(t)}$ (all three ones over intervals). The function $Q : \mathbb{R}^n \to \mathbb{N}_0^n$ implies the threshold constellation, and $S(Q(E))$ indicates where in the state space the system is placed at E, and which matrices and vectors M, W, V have to be chosen to specify the system such that the given data are approximated best. The mapping $S : \mathbb{N}_0^n \to \mathbb{N}_0$ has to be injective, such that a different triplet (M,W,V) is used whenever a threshold is traversed. This *piecewise linear* approach provides an approximation of the global nonlinearity of nature.

We understand ($\mathcal{HE}$) in the sense of the placement in the set of intervals (cf. Section 2) and of an extension of Q when one or more thresholds are included in the intervals $E_i(t)$ [58]. In such a case, this extension can be made by the arithmetic mean of the corresponding Q values associated with those intervals between and besides the thresholds which intersect with $E_i(t)$; this averaging is then followed by a rounding to an integer. Based on this definition of $s(t)$, we find $M_{s(t)}$, $\mathrm{W}_{s(t)}$ and $\mathrm{V}_{s(t)}$ (we could also directly use the averaging technique for these parameters [53]).

A time-*discrete* model is sometimes more preferred. Such a version ($\mathcal{HDE}$) can be found in [53]. It distinguishes between past $(k-1)$, presence (k) and future $(k+1)$, herewith, expressing a time consumption in regulatory genetic networks. State prediction for time $k+1$ needs the model at time k which memorizes the time $k-1$, where it became parametrically preadjusted. We recall that for the time-continuous system, we also interpreted the (set-valued) derivative by a (set-valued) difference quotient; this leads to a system with a forward delay (anticipation), piecewise and in a uniform manner with respect to the time.

For the *parameter estimation* of the time-continuous model and the time-discrete system we have to *estimate the thresholds* $\theta_{i,j}$ and to *calculate the matrices and vectors*, $M_{s(t)}$, $\mathrm{W}_{s(t)}$ and $\mathrm{V}_{s(t)}$, describing the system in between the thresholds [24,53,61]. The thresholds can be defined by *Akaike's Information Criterion* [27]. For closer details and concerning the parameter estimation in the time-discrete case, we refer to [2,3,21,24,39]. Since we are concentrating on the tasks in continuous optimization, we assume that we already know all the thresholds.

Now, for any subparallelpipe $\mathcal{P}^*$ given by the threshold constellation we have to extract the parametric unknowns $M_{s(t)}$, $\mathrm{W}_{s(t)}$ and $\mathrm{V}_{s(t)}$ from the measurement data. In $\mathcal{P}^*$, the hybrid system ($\mathcal{HE}$) reduces to a system of ordinary linear differential equations. Hence, we can find analytical solutions for the corresponding parts of the state space. We may assume that for the special environmental factors the times of sampling are just the genetic sampling times, and the same index sets of samplings. The environmental data $\overset{\sim}{\overline{E}}^{(\kappa)}$ ($\kappa = 0, 1, \ldots, l-1$) are considered to be binary and constant, but they

could also be variable in a more refined modeling.

We note that our hybrid model can be further extended and we can include possible delays in the interaction of the variables. Such history dependent problems have been investigated in [36] and the delays are included in the state transitions (*threshold crossing*) in the form $Q(E(t)) = (Q_1(E(t-\tau_1)), \ldots, Q_n(E(t-\tau_n)))$, where τ_i $(i = 1, \ldots, n)$ represents the delay with regard to the state i. For further details on the *time-delay hybrid model* and a stability analysis we refer to [36].

5.2 Mixed-Integer Parameter Estimation

For an estimation of parameters we have to minimize the quadratic error between the difference quotients $\dot{\overline{E}}^{(\kappa_\alpha)}$ and the right-hand side of the differential equations evaluated at the finitely many measurement intervals $\overline{E}^{(\kappa_\alpha)} \in \mathcal{P}^*$ $(\alpha = 0, 1, \ldots, l^* - 1)$ which are lying in the regarded regime $\mathcal{P}^*$:

$$(\mathcal{HLS}) \qquad \min_{(m_{ij}^*),(\mathrm{W}_{i\ell}^*),(\mathrm{V}_i^*)} \sum_{\alpha=0}^{l^*-1} \left\| M^* \overline{E}^{(\kappa_\alpha)} + \mathrm{W}^* \overline{\breve{E}}^{(\kappa_\alpha)} + \mathrm{V}^* - \dot{\overline{E}}^{(\kappa_\alpha)} \right\|_\infty^2 .$$

Parallelpipe expression vectors can affect several neighbouring subparallelpipes $\mathcal{P}^*$, such that we get corresponding problems $(\mathcal{HLS})$. Criteria for which of them to put special emphasis on consists in where the data vectors as parallelpipes are lying, and further empirical evidence given. In $(\mathcal{HLS})$, $\|\cdot\|_\infty$ stands for the *Chebychev norm* of the set inserted, i.e., it is the maximum norm with respect to the vector-valued functions defined by (independent) parametrization which we get from the interval-valued entries of M^*, W^* and V^* as well as the ones of the vectors $\overline{E}^{(\kappa_\alpha)}$, $\overline{\breve{E}}^{(\kappa_\alpha)}$ and $\dot{\overline{E}}^{(\kappa_\alpha)}$, respectively. For length measurement we use the Euclidean norm, such that our squared Chebychev norm is indeed a maximum over sums of squares, but we could also use the maximum or the sum vector norm (l_1-norm) instead of the Euclidean norm. This reconsideration turns our least-squares or Gaussian approximation problem of earlier studies (cf., e.g., [61]) to some generalized Chebychev approximation problem (see Section 2.6). The generalization comes from both the sum of squares formula where the single Chebychev norms are embedded and the fact that the left and the right sides of "$-$" are parametrically decoupled from each other. We note that each entry of M^*, W^* or V^* is defined by two or more scalar values. For the sake of simplicity, we concentrate on two ones, namely, the two interval endpoints. This means that the dimension of the problem becomes doubled, compared with the single valued case. In the following, we shall repeatedly meet this new approach and interpretation. We could indeed extend this optimization problem in the sense of our note made after introducing the system $(\mathcal{HE})$; then, we would insert the data vectors into

our uniform interval-valued framework of arithmetics and approximation.

The classical "scalar" version of ($\mathcal{HLS}$), i.e., Gaussian approximation, can be canonically treated by building the partial derivatives with respect to the unknowns and equating them to 0. Then, one has to solve the resulting *normal equations*, which are linear in the unknown parameters m_{ij}^*, $\mathrm{w}_{i\ell}^*$ and v_i^*, e.g., by Gaussian elimination method. But ($\mathcal{HLS}$) is a generalized Chebychev approximation problem; since it can equivalently be written as a semi-infinite optimization problem, we get access to the applicable methodology of SIP.

As nowadays high-throughput technologies are available gene-environment networks are huge and for practical reasons we have to rarefy them by diminishing the number of arcs [53,61]. Here, upper bounds on the outdegrees of nodes are introduced firstly; later on, these constraints are undergoing a relaxation. In this section and in Section 6, we shortly recall this process in our interval-valued generalized Chebychevian way [62]. At first, we introduce the Boolean matrices and vectors, $X = (\chi_{ij})_{i,j=1,\ldots,n}$, $\Xi = (\xi_{i\ell})_{\substack{i=1,\ldots,n \\ \ell=1,\ldots,m}}$ and $Z = (\zeta_i)_{i=1,\ldots,n}$, representing by the values 1 and 0 whether or not gene j regulates gene i, environmental item ℓ regulates gene i and the environment cumulatively regulates gene i. The *outdegrees* $\sum_{i=1}^n \chi_{ij}$, $\sum_{i=1}^n \xi_{i\ell}$ and $\sum_{i=1}^n \zeta_i$ count the numbers of genes regulated by gene j, by environmental item ℓ or by the cumulative environment, respectively. Our network rarefaction by bounding the outdegrees obeys the principles of least-squares (or maximum likelihood). We also imply any helpful *a priori* knowledge into the problem, especially, about degradation rates, and what is empirically known about the connectedness structure. Often, a lower bound $\delta_{i,\min}$ on the degradation of gene i is known or there are requests given about the feasibility of special genetic or metabolic processes [24,61]. Herewith, our parameter estimation task becomes a *mixed-integer (generalized) Chebychev approximation problem* as follows:

$$(\mathcal{MICP}) \qquad \min_{(m_{ij}^*),(\mathrm{W}_{i\ell}^*),(\mathrm{V}_i^*),(\chi_{ij}),(\xi_{i\ell}),(\zeta_i)} \sum_{\alpha=0}^{l^*-1} \left\| M^* \overline{E}^{(\kappa_\alpha)} + \mathrm{W}^* \overline{\breve{E}}^{(\kappa_\alpha)} + \mathrm{V}^* - \overline{\dot{E}}^{(\kappa_\alpha)} \right\|_\infty^2,$$

subject to

$$\begin{aligned} \textstyle\sum_{i=0}^n \chi_{ij} &\le \alpha_j \qquad (j = 1, 2, \ldots, n), \\ \textstyle\sum_{i=0}^n \xi_{i\ell} &\le \beta_\ell \qquad (\ell = 1, 2, \ldots, m), \\ \textstyle\sum_{i=1}^n \zeta_i &\le \gamma, \\ m_{ii} &\ge \delta_{i,\min} \ (i = 1, 2, \ldots, n). \end{aligned}$$

The loss of the edges emenating at a few genes which are considered to play a very important role in regulation, i.e., to have very high outdegrees, could strongly restrict the connectivity of the network. Such a loss can be the result of perturbations caused by the environment and affecting the problem

($\mathcal{MICP}$) with its rigid (exclusive) binary constraints. We therefore make them "softer" by performing a *relaxation* in the next Section 6.

6 GSIP Relaxation and Extension

6.1 The GSIP Extension

The *mixed-integer Chebychev approximation problem* ($\mathcal{MICP}$) includes rigid binary constraints. To alleviate the effects of these constraints we replace the binary variables χ_{ij}, $\xi_{i\ell}$ and ζ_i by real variables $p_{ij}, q_{i\ell}, r_i \in [0,1]$ which linearly depend on the elements of a_{ij}, $\mathrm{w}_{i\ell}$ and v_i and assume some reasonable box constraints [53,61,62]. The values $\sum_{j=1}^{n} p_{ij}(m_{ij}^*)$, $\sum_{i=1}^{m} q_{i\ell}(\mathrm{w}_{i\ell}^*)$ and $\sum_{i=1}^{m} r_i(\mathrm{v}_i^*)$ become interval-valued approximations of the number of genes regulated by gene j, environmental item ℓ and cumulative environment, respectively. Please recall that the continuous real-valued image of an interval is an interval again. Having solved the continuous optimization problem, we could return the binary variables and, hence, network rarefaction, by means of rounding or staying below some small prescribed values $\varepsilon_{ij}, \varepsilon_{i\ell}, \varepsilon_i \in [0, \frac{1}{2})$, respectively [61].

The environment can affect the connectedness between the genes or destroy some of the connecting paths but also cycles among the genes ("knockout"; [20]), and an external stimulus can activate a higher regulation among the genes. For reasons like these [53,61] implied all the possible convex combinations of the environmental effects into the inequalities about the bounded outdegrees. The *set of combined environmental effects* is defined as the convex hull of all the vectors $\mathrm{w}_{i\ell}^* e_{m(i-1)+\ell}$ and $\mathrm{v}_i^* e_{mn+i}$, i.e.,

$$
\begin{aligned}
Y(\mathrm{V}^*, \mathrm{W}^*) &:= \operatorname{conv}\Big(\Big\{ \mathrm{w}_{i\ell}^* e_{m(i-1)+\ell} \,\Big|\, i = 1,2,\ldots,n;\ \ell = 1,2,\ldots,m \Big\} \\
&\qquad\qquad \cup \Big\{ \mathrm{v}_i^* \sigma_{i,m+1} e_{mn+i} \,\Big|\, i = 1,2,\ldots,n \Big\} \Big) \\
&= \left\{ \sum_{\substack{i=1,\ldots,n,\\ \ell=1,\ldots,m}} \sigma_{i\ell} \mathrm{w}_{i\ell}^* e_{m(i-1)+\ell} + \sum_{i=1,\ldots,n} \sigma_{i,m+1} \mathrm{v}_i^* e_{mn+i} \,\Bigg| \right. \\
&\qquad \left. \sigma_{i\tau} \geq 0 \ (i = 1,2,\ldots,n;\ \tau = 1,2,\ldots,m+1),\ \sum_{\substack{i=1,\ldots,n\\ \tau=1,\ldots,m+1}} \sigma_{i\tau} = 1 \right\},
\end{aligned}
$$

with e_η denoting the ηth $((m+1)n)$-dimensional unit vector $(0,\ldots,1,\ldots,0)^T$. Formally, we can write $Y(\mathrm{V}^*, \mathrm{W}^*)$ as a parallelpipe:

$$
Y(\mathrm{V}^*, \mathrm{W}^*) = \bigtimes_{\substack{i=1,\ldots,n\\ \ell=1,\ldots,m}} [0, \mathrm{w}_{i\ell}^*] \times \bigtimes_{i=1,\ldots,n} [0, \mathrm{v}_i^*];
$$

however, we underline that the elements y of the Cartesian factors (formal intervals) are just our parametric intervals. The wealth of how the environment is implied bases on and applies any given *a priori* knowledge about the genes that helps scientists, practitioners and decision makers when determining and elaborating the rarefied network. We recall that all intervals y can be encoded by a tuple of scalar values. Now, we get our *relaxed (generalized) Chebychev approximation problem* in the following form:

$$(\mathcal{RCP}) \qquad \min_{(m_{ij}^*),(\mathrm{W}_{i\ell}^*),(\mathrm{V}_i^*)} \sum_{\alpha=0}^{l^*-1} \left\| M^* \overline{E}^{(\kappa_\alpha)} + \mathrm{W}^* \overline{\check{E}}^{(\kappa_\alpha)} + \mathrm{V}^* - \overline{\dot{E}}^{(\kappa_\alpha)} \right\|_\infty^2,$$

subject to

$$\begin{aligned}
\textstyle\sum_{i=1}^n p_{ij}(m_{ij}^*, y) &\leq \alpha_j(y) && (y \in Y(\mathrm{V}^*, \mathrm{W}^*)),\\
\textstyle\sum_{i=1}^m q_{i\ell}(\mathrm{w}_{i\ell}^*, y) &\leq \beta_\ell(y) && (y \in Y(\mathrm{V}^*, \mathrm{W}^*)),\\
\textstyle\sum_{i=1}^m r_i(\mathrm{v}_i^*, y) &\leq \gamma(y) && (y \in Y(\mathrm{V}^*, \mathrm{W}^*)),\\
\delta_{i,\min} &\leq m_{ii} && (i = 1, 2, \ldots, n),\\
\underline{m}_{ij}^* &\leq m_{ij}^* \leq \overline{m}_{ij}^* && (i, j = 1, 2, \ldots, n),\\
\underline{\mathrm{w}}_{i\ell}^* &\leq \mathrm{w}_{i\ell}^* \leq \overline{\mathrm{w}}_{i\ell}^* && (i = 1, 2, \ldots, n;\ \ell = 1, 2, \ldots, m),\\
\underline{\mathrm{v}}_i^* &\leq \mathrm{v}_i^* \leq \overline{\mathrm{v}}_i^* && (i = 1, 2, \ldots, n).
\end{aligned}$$

We note that, firstly, we could compare $\underline{m}_{ii}^*$ and $\delta_{i,\min}$ and, then, take the largest of the two values as a single lower bound instead (provided that $\delta_{i,\min} < \overline{m}_{ii}^*$). As given in the objective function by generalized Chebychev approximation, this uniform interpretation of the "$\leq$" conditions amounts to the SIP character of $(\mathcal{RCP})$. By the additional coupling of our inequality constraint set $Y(\mathrm{V}^*, \mathrm{W}^*)$ with the states $(\mathrm{V}^*, \mathrm{W}^*)$, $(\mathcal{RCP})$ even becomes a GSIP problem. In the objective function, the terms with the κth Chebychev norm $\|\cdot\|_\infty$ are nonsmooth max-type functions ($\kappa = 0, 1, \ldots, l^*-1$). By the following standard technique, $(\mathcal{RCP})$ becomes smoothly modeled. For each of them, we introduce a new coordinate τ_κ, in addition to the unknowns of $(\mathcal{RCP})$, considered as a new coordinate and as a uniform bound for the squared Euclidean norms of the elements inside the Chebychev norms (see Subsection 2.4). Herewith, we minimize the sum of the bounds. As new inequalities we just introduce these bounding conditions; we write them so that the Euclidean norms of all the elements inside the Chebychev norms have uniformly to stay below ("$\leq$") the corresponding bounds. We note that we could also use *one* single new coordinate τ for an overall uniform bound. Indeed, we can choose between both alternatives according to our preferences. In case we replace the squares in the objective function by absolute values and make a further linearity assumption on the constraints, $(\mathcal{RCP})$ comes close to a GSIP kind of conic quadratic

programming [31,50,54,57] (cf. Subsection 6.2).

6.2 On GSIP and Structural Stability for Gene-Environment Networks

6.2.1 Introduction

GSIP optimization, revisited for our gene-environment network problem ($\mathcal{RCP}$) in Section 6, reveals the following general program form [44,46,57]:

$$\mathcal{P}_{\mathcal{GSI}}(f,h,g,u,v) \qquad \left\{ \begin{array}{l} \text{minimize } f(x) \text{ on } M_{\mathcal{GSI}}[h,g], \text{ where} \\ M_{\mathcal{GSI}}[h,g] := \Big\{ x \in \mathbb{R}^d \big|\, h_i(x) = 0 \ (i \in I), \\ \qquad\qquad g^j(x,y) \geq 0 \ (y \in Y^j(x),\ j \in J) \Big\} \end{array} \right\}, \quad (\mathcal{A}_1)$$

with finite cardinalities $|I|, |J| < \infty$, and with the sets $Y^j = Y^j(x)$ being defined as feasible sets in the sense of *finitely constrained* ($\mathcal{F}$) programming. Hence, also the sets of inequality constraints possess finitely many elements only. Moreover, for each $x \in \mathbb{R}^d$ it holds

$$\left. \begin{array}{rl} Y^j(x) &= M_{\mathcal{F}}[u^j(x,\cdot), v^j(x,\cdot)] \\ &:= \Big\{ y \in \mathbb{R}^q \big|\, u_k(x,y) = 0 \ \ (k \in K^j),\ v_\ell(x,y) \geq 0 \ (\ell \in L^j) \Big\} \end{array} \right\}, \quad (\mathcal{A}_2)$$

where $|K^j|, |L^j| < \infty$. Moreover, the model ($\mathcal{A}_1$)-($\mathcal{A}_2$) allows equality constraints on both the upper (x-) level and lower (y-) level representing, e.g., further metabolic restrictions, reactions or balance equations [53,61]. Let us suppose that the outdegree constraints in ($\mathcal{RCP}$) are of class C^2, too. The upper and lower bounds guarantee that the feasible set $M_{\mathcal{GSI}}[h,g]$ is compact in the projective sense of the original $2(n^2+mn+n)$ unknowns (with intervals encoded by tuples of endpoints), but not in the "height" dimensions of the new coordinates τ_κ. This noncompactness can be overcome in the way explained in [54,57]. By their form, the sets $Y^j(x)$ are compact indeed. What is more, we can even state that they fulfill the *Linear Independence Constraint Qualification* (*LICQ*), an appropriate choice of the overall box constraints that are given. The works [46,53,57,61] provide more detailed discussions and possible generalizations of GSIP.

6.2.2 Stability Theory

Perturbations of our gene-environment networks $(f,h,g,u,v) \mapsto (\tilde{f},\tilde{h},\tilde{g},\tilde{u},\tilde{v})$ are generated or caused, e.g., by *outliers of parallelpipes*, *"perturbed" problems and networks* and certain kinds of *errors, imprecision and uncertainty* [53,61].

The strong Whitney topology C^2_S [31] serves as a "measure" of perturbations so that asymptotic aspects are taken into account.

The character –"genetic (and environmental) fingerprint"– of $(\mathcal{RCP})$ is given by all the lower level sets of its objective function, which are subsets of the feasible set. If under arbitrarily slight perturbations and some correspondence between the levels the perturbed and the unperturbed lower level sets are homeomorphic to each other, we call $(\mathcal{RCP})$ *structurally stable* [31,33,54,57]. Now, we can carry over and state the *Characterization Theorem on Structural Stability for Gene-Environment Networks* from [53,61] for $(\mathcal{RCP})$. In order not to overload the exposition, we may avoid giving the full definitions and details but refer to [34,55–57]. Our main theorem basically states that structural stability can just be *characterized* by two well-known regularity conditions and a more technical one:

Theorem 6.1 *(Characterization Theorem on Structural Stability for Gene-Environment Networks)*

The optimization problem $\mathcal{P}_{\mathcal{GSI}}(f,h,g,u,v)$ *on gene-environment networks is* structurally stable, if and only if *the following triplet of conditions,* $\mathcal{C}_1$ — $\mathcal{C}_3$*, is satisfied:*

- $\mathcal{C}_1$. *The* Extended Mangasarian-Fromovitz Constraint Qualification *(EMFCQ) holds for the set* $M_{\mathcal{GSI}}[h,g]$ *defined in* $\mathcal{P}_{\mathcal{GSI}}(f,h,g,u,v)$.
- $\mathcal{C}_2$. *All the* $\mathcal{G}$*-*$\mathcal{O}$ *Kuhn-Tucker points* $\overline{x}$ *of* $\mathcal{P}_{\mathcal{GSI}}(f,h,g,u,v)$ *are* ($\mathcal{G}$*-*$\mathcal{O}$) strongly stable.
- $\mathcal{C}_3$. *For each two different* $\mathcal{G}$*-*$\mathcal{O}$ *Kuhn-Tucker points* $\overline{x}^1 \neq \overline{x}^2$ *of* $\mathcal{P}_{\mathcal{GSI}}(f,h,g,u,v)$ *the corresponding critical values are different (*separate*), too:* $f(\overline{x}^1) \neq f(\overline{x}^2)$.

Our Characterization Theorem helps for a well understanding of the topological "landscape" of gene-environment networks, for their perturbational behaviour and for the development of numerical procedures. For instance, we can consider "mountain paths" (saddle points) between any two candidate networks being given by local minimizers of $(\mathcal{RCP})$. All the points around candidate solutions can be regarded as potential networks which may be obtained after perturbations, e.g., inward shifts from a genetic or environmental boundary to an interior position [34,55–57], or, generally, be the result of a "forward operator". They may be outcomes of underlying constellations in the experimental design which may have to be reconstructed, which is an inverse problem [6].

Let us give a short explanation about the regularity conditions $\mathcal{C}_{1,2,3}$: EMFCQ basically guarantees that the feasible set $M_{\mathcal{GSI}}[h,g]$ is a topological manifold with generalized boundary. If this set is compact, then EMFCQ can be characterized by its stability under slight perturbations of the defining functions

(results for noncompact case are prepared, too) [31,54,57]. The condition of strong stability on our critical points guarantees their local uniqueness and continuous dependence on any slight perturbation of the defining functions [33,34,54,57]. Finally, the more technical condition of separated critical values makes the unperturbed and any slighly perturbed situation comparable. It prevents from some different topological situations [31,54].

In terms of testing the goodness of data fitting, the lower level sets can be interpreted as confidence regions around the parameters estimated. The size of these regions is basically governed by the steepness of the function around the solution. In cases where a local or global minimizer is very steep, we can associate this with stability, whereas flatness is more likely related with instability. During the resolution of $(\mathcal{RCP})$, we have to understand possible pathologies in terms of the violation of one or more of the conditions $\mathcal{C}_{1,2,3}$.

Future research may investigate dynamics within of our networks such as *"tectonics"* generating "clashes", "folds", "reefs", "volcanoes" and "areas lifted or dropped". Here, a well interpretation and prediction of the biological, economical and social factors are necessary and intended, and a suitable numerical methodology has to be prepared by all of this. But there are also dynamical phenomena along different networks such as *"cascades"* of gene-environment networks. By the time-discrete dynamics, the networks generate expression level vectors which can reversely be interpreted as simulated data on which further models could base.

7 Overview of Classification and Model Selection Methods for Gene Network Data

There is a growing interest in the application of machine learning techniques together with optimization to real-world applications such as biological problems [42], engineering problems etc.. In this review, we will introduce recent developments in one of the most efficient methods, Support Vector Machines (SVM). In [42], an efficient and novel model selection algorithm embedded in a classical SVM to predict pro-peptide cleavage sites in *filamentus fungi* [42]. Prediction results of the confidence level by an SVM are compared with the results achieved by the pro-peptide prediction tool ProP1.0 [15]. *ProP1.0* is a bioinformatics and computational biology tool which predicts pro-peptide cleavage sites on a furin specific based network and a general PC network separately by using a *neural network*. ProP1.0 consists of 227 proteins of all eukaryotes including those of humans and animals. The data set is presented to the neural network by sparsely encoded moving windows. The output of a neural network is assessed by a threshold of 0.5 to determine the potential pro-peptide cleavage site.

The study in [42] concentrates more on fungal proteins due to the industrial importance of these organisms in heterologous protein production, including those of humans. The data set is collected from largely non-homologous fungal proteins consisting of 72 sequences. Our prediction tool, *confidence level* SVM is fed with both binary input vectors and the substitution matrix PAM250 separately and results are reported for both. The sequences are given to the learning machine by encoded sliding windows through each sequence. Each protein is tested with different training sets. Rather than splitting the data set into groups, we have used a different strategy that enables us to use the whole data for both training and testing. This is explained in detail in the next section. The construction of the data set from non-homologous sequences is justified by using ClustalW to construct a phylogenetic tree which is based on multiple sequence alignment.

7.1 Materials and Methods

The data set is collected from the NCBI databank based on fungal proteins which are publicly available [1] . 72 fungal sequences are selected among non-homologous protein families. This is one of the reasons for the small number of sequences contained. To reduce further redundancy in the data set and prevent the training and testing from being homologous, we made a phylogenetic tree analysis based on multiple sequence alignment by ClustalW. There, in a phylogenetic tree many individual main branches are resulted (data not shown) indicating that the selected proteins are not homologous. In our learning process by SVM, we chose symmetric windows around possible cleavage sites, where the window length varies between 11 to 21 and the results indicates that the optimum window length lies between 13 and 19. The best accuracy results are found with window length chosen as 15. These parameters can vary according to the type of the data set and the kind of problem.

To see the discriminative motifs existing in the sequences, we used *MEME* software [2] . This yielded the motif **KR**. To check this result, *Multiple Sequence Alignment (MLA)*, with the package *ClustalW* is applied to the data set which confirms this observation. The motif **KR** gives us a clue for the preparation of the input sequences for the SVM. With MLA, most of the cleavage site patterns are in the form of either **K**, **R** or **KR**. Therefore, it is sensible to train the SVM restricted to inputs with **K** or **R** residues.

[1] http://www3.iam.metu.edu.tr/iam/images/1/1a/Datasetsureyya.pdf
[2] http://meme.sdsc.edu/meme/meme-output-example.html

There are different ways to represent *text based* data when introducing the data to a learning algorithm. In bioinformatics, these data can be amino acid (a.a.) sequences, DNA sequences, etc.. The most popular method of encoding amino acid sequences into numerical values is given by binary vectors [7]. However, this ignores the *context* information. There has been a lot of research on encoding amino acids to give each individual amino acid a numerical value regarding the biochemical and physiochemical properties [37]. One of the most powerful substitution matrices is PAM250 matrix due to its property of preserving mutations of the sequences. In this study, two types of encoding are considered, namely, a binary encoding matrix and the *PAM250* substitution matrix. Please note that, encoding a.a. by substitution matrices is needed for the input vectors for the SVM. Thus, the windows of a.a. sequences are presented to the SVM with the numerical values corresponding to the input vectors.

There are many similarity matrices developed according to different similarity approaches and gap penalties given between two amino acids. Dayhoff et al. [12] created a table where they aligned the proteins in several families of proteins and constructed phylogenetic trees for each family [12]. The resulting similarity table presents relative frequencies with which amino acids replace each other in a short evolutionary period since each phylogenetic tree is checked for the substitutions found on each branch. The traditional Dayhoff PAM250 matrix assumes the occurrence of 250 point mutations per 100 amino acids or 300 nucleotides in the gene [38].

PAM matrices are theoretically more advantageous than the others. They arise from Dayhoff's method [12] which is based on observed evolutionary mutations. Hence, they preserve information given by the processes that generate the mutations. Statistically, PAM matrices and other log-odds matrices are the most accurate description of the changes in the amino acid composition after a given number of mutations. Details about the formulation of log odds matrices and PAM matrices can be found in [4,12].

Since there are 20 amino acids, there are entries in a 20×20 PAM250 matrix. Each amino acid is represented by a 20 dimensional vector corresponding to the entries in a column of the PAM250 matrix. If there is a sequence of n amino acids, then we will have an $n \times 20$ dimensional real-valued vector as input.

In [42], *sliding window* approach is used while scanning input sequences. The sliding window approach is a method to construct the training and test set with a previously chosen window size. Training windows are chosen from the

neighbourhood of the potential cleavage sites in such a way that the cleavage sites are at the centre of the window. For example, if we have a window size of 11, then the considered cleavage site is between the 5th and the 6th position of the window. In this way, each sequence contributes one positive window. For the negative class, three windows are chosen from each sequence by selecting positions which have residues **K** or **R** at their centre. Here, windows are chosen as symmetric in all cases. A test sequence is constructed by sliding the window through the whole sequence. In our case, all the sequences have at least one **K** or **R** which are the motifs that we learned from ClustalW through multiple sequence alignment. Sliding windows through the whole sequence generate many test windows, i.e., test inputs. Furthermore, the cleavage window(s) in the test sequence are going to be labeled as a positive class from the output of SVM and the others as a negative class. It is clear that restricting the windows by including to those windows that have **K** or **R** at their center will decrease the number of test examples and, hence, makes it easier to select the positive one(s) (cleavage window(s)) when compared to the high number of windows for a particular test sequence. In other words, if we call the set of all sliding windows S and choose a special subset $A \subseteq S$ which depends on motifs known in advance from a bioinformatics tool, then searching a cleavage window(s) among A will be easier than searching from the bigger set S for a particular test sequence. If the set A is empty, i.e., $A = \emptyset$, the set S which contains all possible windows of the particular test sequence can be used as test examples. In our special data set on fungal proteins, the subset A of S is nonempty, i.e., $A \neq \emptyset$. Moreover, the cardinality of A is always greater than 3, i.e., $|A| \geq 4$.

Our data set comprises 72 proteins and, hence, 72 amino acid sequences, each giving rise to be one positive window and three negative windows. So, 72 sequences are used for both training and testing using the *leave one out* principle that leaves each sequence in turn as testing while using the remaining 71 for training. In this way, we have trained using 71 sequences and have tested 1 sequence 72 times. The accuracy is calculated as the precentage of the total number of correct predictions over the 72 sequences.

The definition of the kernel and the SVM algorithm both involve an additional parameter vector (C_+, C_-, σ), the parameters C_+ and C_- for the SVM and the kernel width σ for the Gaussian kernels. The usual way to set these parameters is using cross-validation [27]. This assesses the quality of different parameter settings by dividing the training data into m groups. It then leaves out one group in turn to train the classifier with a range of possible values for the parameters and uses the group left out as a test set. The average accuracy for each parameter setting over all m test groups is then used to select the parameter settings. We employed this approach where we took $m = 71$, i.e., we performed a subround of "leave one out" error estimation on each training set in order to select the parameters to use training for the set of 71 sequences

before testing on 72nd left out sequence. Note that this is the only *leave one out* at the level of sequences, since each sequence corresponds to 4 windows, one of which is positive.

Our second method of model selection is a novel approach for problems in which each test involves multiple inputs, but with the additional information that only one is positive: in our case, there are many windows, but only one is a cleavage site. Rather than to pre-select the parameters, we train the SVM on all the training data (other than the single test sequence) with all the parameter settings. For each SVM we compute the real-valued outputs, for all the windows arising from the sequence. We define the confidence of the classifier as the difference between the maximal output and the second largest. Now, we select the parameter settings for which the confidence is largest and identify the window with maximal output as the cleavage site. It should be stated that not every test sequence has to have a cleavage site. It corresponds to having all test window outputs being negative. In such cases, our algorithm outputs that these sequences have not cleavage site. Similar analysis can be easily done for DNA sequence data and gene networks to recognize particular pattern by windows as in [42]. We refer [42] for numercial experiments and results.

7.3 SVM Model Selection Based on Observed Margin

Support vector machines (SVMs) carry out binary classification by maximizing the margin of a hyperplane between the two classes of examples and then classifying test points according to the half-spaces in which they reside (irrespective of the distances that may exist between the test examples and the hyperplanes). In cross validation, the principle idea is to find the *one* SVM model and its optimal parameters that help to achieve the smallest training error amongst all of the models that can be constructed. In contrast, in [43] *all* of the models found in the model selection phase are collected and predictions are obtained for test points by finding the SVM models whose hyperplanes achieve the maximum distance from the test points. In this setting, the complex and time consuming paradigm of model selection via cross validation are avoided. Experimental results demonstrate the plausibility of the method proposed and show a significant decrease in computational time as well as a competitive generalization error.

For all kinds of data mining tools, parameter selection is one of the critical questions; it determines the right model for data analysis and prediction. In this chapter, we mainly develop a fast algorithm for model selection which uses the benefit of all hypothesis space by means of functions or models [43]. The new model selection approach in [43] called *maximum margin* to the binary

classification problems by using support vector machines (SVMs) which, as we recall, is one of the most efficient methods in machine learning.

7.3.1 Methods

In this section, three different norms will be discussed for model selection at the testing phase. Given a set of functions in the variable $\mathbf{x}$, $f_1, f_2, \ldots f_\ell$ being the outputs by the SVM, with $\ell = |C| \cdot |\sigma| = \ell_1 \cdot \ell_2$, then being the number of models that can be constructed from the set of parameter values $C \in \{C_1, C_2, \ldots, C_{\ell_1}\}$ and $\sigma \in \{\sigma_1, \sigma_2, \ldots, \sigma_{\ell_2}\}$, where C is the error constant and σ is the Gaussian kernel width. We can use some or a combination of models derived by these parameters in order to make predictions. The first approach which we propose uses the L_∞-norm for choosing which function to use. This is equivalent to evaluating the distance of a test point according to the function that achieves the largest (functional) margin.

We assume here, without loss of generality, that the functions f computes the functional margins and not the geometrical margins (hence, the reason that the example values we have presented are not bounded by 1 and -1). Finally, we would predict the class of $\mathbf{x}^0$ by looking for the maximal (positive) and the minimal (negative) value of all functions. The L_∞ prediction function $F_\infty(x)$ evaluated at our given example $\mathbf{x}^0$ can be defined in the following way:

$$F_\infty(x) := \text{sgn}\left(\max\{f_i(x)\}_{i=1}^{\ell} + \min\{f_i(x)\}_{i=1}^{\ell}\right),$$

where sgn denotes the sign function, i.e., positivity or negativity of a function

The second approach which we introduce is for the L_1-norm where the decision depends on the sign of the Riemann sum of all outputs evaluated for a test point [43]. This results in the following L_1-norm prediction function F_1 given a test example $\mathbf{x}$, e.g., $\mathbf{x}^0$:

$$F_1(x) = \text{sgn}\left(\sum_{i=1}^{\ell} f_i(x)\right).$$

The final approach [43] corresponds to the L_2-norm and is similar to the one with the L_1-norm discussed above, but with a down-weighting if the absolute values are less than 1 and an up-weighting if they are above 1. This means that we are giving a greater confidence to functions that predict functional values greater than 1 or less than -1 but less confidence to those that are closer to the threshold of 0. Another way of thinking about this approach is that it is equivalent to a weighted combination of functional margins with the absolute values of themselves.

Therefore, given a test example $\mathbf{x}$, e.g., $\mathbf{x}^0$, we have the following L_2-norm prediction function $F_2(\mathbf{x})$ defined by

$$F_2(\mathbf{x}) := \operatorname{sgn}\left(\sum_{i=1}^{\ell} f_i(\mathbf{x})|f_i(\mathbf{x})|\right).$$

We refer to [43] for experimental setup and results. In this section, we propose different model selection techniques for binary classification problems which can be directly applicable to real world situations such as detection of diseased cells, cancerous cells, or finding important patterns in experimental data in biology or medicine.

8 Conclusion

In this review, we surveyed recent approaches in mathematical modeling, optimization and dynamical representation of gene-expression patterns and environmental information. Gene-environment networks provide a general framework for the analysis of complex systems in computational biology. In particular, various kinds of data uncertainties in DNA microarray experiments and environmental observations can be included. We arrived at approximation problems of a generalized Chebychevian kind and investigated them by generalized semi-infinite optimization. For a deep understanding of the topological landscape of gene-environment networks determined by that optimization, we could state a characterization result on structural stability. Complementary to our optimization theory, we gave a stability theory on dynamical systems which support, e.g., the prediction of genetic and environmental levels and the testing of the goodness of data fitting. With all these explanations we demonstrated the importance of optimization and dynamics in a modern interdisciplinary approach which has discrete, continuous and hybrid features as well. In our analysis we saw how GSIP can help realize the close interaction between genetic and environmental information.

Machine learning methods can complement the dynamic approach by a classification of microarray data. Model selection and kernel learning methods for binary classification can be used to classify microarray data, e.g., for cancer genes or for discrimination of other kind of diseases.

The authors tried to give a more theoretical but helpful contribution to a better understanding of nature and for improvements in health care, medicine and living conditions.

References

[1] Ahuja, R.K., Magnanti, T.L., and Orlin J.B., *Network Flow: Theory, Algorithms and Applications*, Prentice Hall, N.J., 1993.

[2] Akçay, D., *Inference of Switching Networks by Using a Piecewise Linear Formulation*, Institute of Applied Mathematics, METU, MSc thesis, 2005.

[3] Akhmet M.U., Gebert, J., Öktem, H., Pickl, S.W., and Weber, G.-W., An improved algorithm for analytical modeling and anticipation of gene expression patterns, Journal of Computational Technologies 10, 4 (2005) 3-20.

[4] Altschul, S.F., Amino acid substitution matrices from an information theoretic perspective, Journal of Molecular Biology, 219, 1991, 555-665.

[5] Amann, H., *Gewöhnliche Differentialgleichungen*, Walter de Gruyter, Berlin, New York, 1983.

[6] Aster, A., Borchers, B., and Thurber, C., *Parameter Estimation and Inverse Problems*, Academic Press, 2004.

[7] Atalay, V. and Cetin-Atalay, R., Implicit motif distribution based hybrid computational kernel for sequence classification, Bioinformatics, 21 (8), 2005, 1429-1436.

[8] Brayton, R.K., and Tong, C.H., Stability of dynamical systems: A constructive approach, IEEE Transactions on Circuits and Systems 26, 4 (1979) 224-234.

[9] Bröcker, Th., and Lander, L., *Differentiable Germs and Catastrophes*, London Math. Soc. Lect. Note Series 17, Cambridge University Press, 1975.

[10] Carbayo, M.S., Bornman, W., and Cardo, C.C., DNA Microchips: technical and practical considerations, Current Organic Chemistry 4, 9 (2000) 945-971.

[11] Chen, T., He, H.L., and Church, G.M., Modeling gene expression with differential equations, in: *Proc. Pacific Symposium on Biocomputing* (1999) 29-40.

[12] Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C., A model of evolutionary change in proteins, Atlas of Protein Sequence and Structure, Dayhoff, M.O. eds., National Biomedical Research Foundation, Washington, 5 (3), 1978, 345-352.

[13] DeRisi, J., Iyer, V., and Brown, P., Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680-686.

[14] Dubois, D.M., and Kalisz, E., Precision and stability of Euler, Runga-Kutta and incursive algorithm for the harmonic oscillator, International Journal of Computing Anticipatory Systems 14 (2004) 21-36.

[15] Duckert, P., Brunak, S. and Blom, N., Prediction of proprotein convertase cleavage sites, Protein Engineering, Design and Selection, 17 (1), 2004, 107-112.

[16] Ergenç, T., and Weber, G.-W., Modeling and prediction of gene-expression patterns reconsidered with Runge-Kutta discretization, special issue at the occasion of seventith birthday of Prof. Dr. Karl Roesner, TU Darmstadt, Journal of Computational Technologies 9, 6 (2004) 40-48.

[17] Feil, R., *Environmental and Nutritional Effects on the Epigenetic Regulation of Genes*, Mutation Research, 2006.

[18] Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T.D., Wu, Y.Z., Plass, C., and Esteller, M., Epigenetic differences arise during the lifetime of monozygotic twins, PNAS 102 (2005) 10604-10609.

[19] Gebert, J., Lätsch, M., Pickl, S.W., Weber, G.-W., and Wünschiers R., Genetic networks and anticipation of gene expression patterns, in: *Computing Anticipatory Systems: CASYS(92)03 – Sixth International Conference*, AIP Conference Proceedings 718 (2004) 474-485.

[20] Gebert, J., Lätsch, M., Quek, E.M.P., and Weber, G.-W., Analyzing and optimizing genetic network structure via path-finding, Journal of Computational Technologies 9, 3 (2004) 3-12.

[21] Gebert, J., Öktem, H., Pickl, S.W., Radde, N., Weber, G.-W., and Yılmaz, F.B., Inference of gene expression patterns by using a hybrid system formulation – an algorithmic approach to local state transition matrices, in: *Anticipative and Predictive Models in Systems Science I*, Lasker, G.E., and Dubois, D.M. (eds.), IIAS (International Institute for Advanced Studies) in Windsor, Ontario (2004) 63-66.

[22] Gebert, J., Lätsch, M., Pickl, S.W., Weber, G.-W., and Wünschiers, R., An algorithm to analyze stability of gene-expression pattern, in: special issue *Discrete Mathematics and Data Mining II* of Discrete Applied Mathematics 154, 7, Anthony, M., Boros, E., Hammer, P.L., and Kogan, A. (guest eds.) (2006) 1140-1156.

[23] Gebert, J., and Radde, N., A network approach for modeling procaryotic biochemical networks with differential equations, in: *Computing Anticipatory Systems*, CASYS'05, Seventh International Conference on Computing Anticipatory Systems, Liege, Belgium, August, 2005 (2006) 526-533.

[24] Gebert, J., Radde, N., and Weber, G.-W., Modelling gene regulatory networks with piecewise linear differential equations, in the special issue (feature cluster) *Challenges of Continuous Optimization in Theory and Applications* of European Journal of Operational Research 181, 3 (2007) 1148-1165.

[25] Green, M. L., Singh, A.V., Zhang, Y., Nemeth, K.A., Sulik, K. K. and Knudsen, T.B., Reprogramming of Genetic Networks During Initiation of the Fetal Alcohol Syndrome, Developmental Dynamics, 236, 613-631, 2007.

[26] Guckenheimer, J., and Holmes, P., *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer, 1997.

[27] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning – Data Mining, Inference and Prediction*, Springer Series in Statistics, 2001.

[28] Hettich, R., and Zencke, P., *Numerische Metoden der Approximation und semi-infiniten Optimierung*, Teubner, Stuttgart, 1982.

[29] Hoon, M.D., Imoto, S., Kobayashi, K., Ogasawara, N., and Miyano, S., Inferring gene regulatory networks from time-ordered gene expression data of Bacillus subtilis using differential equations, in: *Proc. Pacific Symposium on Biocomputing* (2003) 17-28.

[30] Huang, S., Gene expression profiling, genetic networks and cellular states: an integrating concept for tumorigenesis and drug discovery, J. Mol. Med. 77 (1999) 469-480.

[31] Jongen, H.Th., Jonker, P., and Twilt, F., *Nonlinear Optimization in Finite Dimensions – Morse Theory, Chebyshev Approximation, Transversality, Flows, Parametric Aspects*, Nonconvex Optimization and Its Applications 47, Kluwer Academic Publishers, Boston, 2000.

[32] Jongen, H.Th., and Weber, G.-W., On parametric nonlinear programming, Annals of Operations Research 27 (1990) 253-284.

[33] Jongen, H.Th., and Weber, G.-W., Nonlinear optimization: Characterization of structural stability, Journal of Global Optimization 1 (1991) 47-64.

[34] Jongen, H.Th., Rückmann, J.-J., and Stein, O., Generalized semi-infinite optimization: a first order optimality condition and examples, Mathematical Programming 83 (1998) 145-158.

[35] Kaati, G., Bygren, L., and Edvinsson, S., Cardiovascular and diabetes mortality determined by nutrition during parents and grandparents slow growth period, European Journal of Human Genetics 10 (2002) 682-688.

[36] Kahraman, M., Öktem, H., Weber, G.-W., Akhmet, M., Using piecewise linear systems with delay to grab the functional dynamics in biological systems, preprint at Institute of Applied Mathematics, METU, to appear in the proceedings of International Symposium on Health Informatics and Bioinformatics, Turkey08, May 18-20, 2008, Istanbul, Turkey.

[37] Kawashima, S., Ogata, H. and Kanehisa, M., AAindex: amino acid index database, Nucleic Acids Res., 27, 1999, 368-369.

[38] Nicholas, H. and Ropelewski, A., Sequence Analysis: Which scoring method should I use?,
http://www.psc.edu/research/biomed/homologous/scoring_primer.html.

[39] Öktem, H., A survey on piecewise-linear models of regulatory dynamical systems, Nonlinear Analysis 63 (2005) 336-349.

[40] Özcan, S., Yıldırım, V., Kaya, L., Becher, D., Hecker, M., and Özcengiz, G., Phanerochaete chrysoporium proteome and a large scale study of heavy metal response, in: *HIBIT – Proceedings of International Symposium on Health Informatics and Bioinformatics, Turkey'05*, Antalya, Turkey, November (2005) 108-114.

[41] Özöğür, S., Sağdıçoğlu Celep, A.G., Karasözen, B., Yıldırım, N., and Weber, G.-W., Dynamical modelling of enzymatic reactions, simulation and parameter estimation with genetic algorithms, in: *HIBIT – Proceedings of International Symposium on Health Informatics and Bioinformatics, Turkey'05*, Antalya, Turkey (November 2005) 78-84.

[42] Özöğür- Akyüz, S., Shawe-Taylor, J., Weber, G.-W. and Ögel, Z., Pattern Analysis for the Prediction of Eukoryatic Pro-peptide Cleavage Sites, *Article in Press in Discrete Applied Mathematics*, 2008, doi:10.1016/j.dam.2008.06.043 (SCI).

[43] Özöğür- Akyüz, Hussain, Z. and S., Shawe-Taylor, J., Model Selection via Test Margin, *to appear in the Special Issue on Data Mining of journal of Annals of Informations Systems, Springer Book Series*, 2009.

[44] Rückmann, J.J., and Gómez, J.A., On generalized semi-infinite programming, invited paper, TOP 14, 1 (June, 2006).

[45] Sakamoto, E., and Iba, H., Inferring a system of differential equations for a gene regulatory network by using genetic programming, in: *Proc. Congress on Evolutionary Computation* (2001) 720-726.

[46] Stein, O., *Bi-level Strategies in Semi-infinite Programming*, Kluwer Academic Publishers, Boston, 2003.

[47] Taştan, M., *Analysis and Prediction of Gene Expression Patterns by Dynamical Systems, and by a Combinatorial Algorithm*, Institute of Applied Mathematics, METU, MSc Thesis, 2005.

[48] Taştan, M., Ergenç, T., Pickl, S.W., and Weber, G.-W., Stability analysis of gene expression patterns by dynamical systems and a combinatorial algorithm, in: *HIBIT – Proceedings of International Symposium on Health Informatics and Bioinformatics, Turkey '05*, Antalya, Turkey (November 2005) 67-75.

[49] Taştan, M., Pickl, S.W., and Weber, G.-W., Mathematical modeling and stability analysis of gene-expression patterns in an extended space and with Runge-Kutta discretization, in: *proceedings of Operations Research 2005*, Bremen, (September 2005), Springer, 443-450.

[50] Taylan, P., Weber, G.-W., and Beck, A., New approaches to regression by Generalized Additive Models and continuous optimization for modern applications in finance, science and techology, to appear in the special issue of Optimization at the occasion of the 5th Ballarat Workshop on Global and Non-Smooth Optimization: Theory, Methods and Applications, November 28-30, 2006.

[51] Tezel, A., Weber, G.-W., Karasözen, B., and Ergenç, T., On semi-infinite optimization of anticipatory systems and their modern applications, presentation given at *8th SIAM Conference on Optimization*, Stockholm, Sweden, May 15-19, 2005.

[52] Uğur, Ö., Pickl, S.W., Weber, G.-W., and Wünschiers, R., An algorithmic approach to analyze genetic networks and biological energy production: an introduction and contribution where OR meets biology, Optimization 58, 1 (January 2009) 1-22.

[53] Uğur, Ö., and Weber, G.-W., Optimization and dynamics of gene-environment networks with intervals, to appear in the special issue of Journal of Industrial Management and Optimization at the occasion of the 5th Ballarat Workshop on Global and Non-Smooth Optimization: Theory, Methods and Applications, November 28-30, 2006.

[54] Weber, G.-W., *Charakterisierung struktureller Stabilität in der nichtlinearen Optimierung*, Aachener Beiträge zur Mathematik 5, Bock, H.H., Jongen, H.Th., and Plesken, W. (eds.), Augustinus publishing house (now: Mainz publishing house) Aachen, 1992.

[55] Weber, G.-W., Generalized semi-infinite optimization: On iteration procedures and topological aspects, in: *Similarity Methods. International Workshop*, Kröplin, B., Rudolph, S., and Brückner, S. (eds.), Institute for Statics and Dynamics of Aerospace Structures, Stuttgart (1998) 281-309.

[56] Weber, G.-W., Generalized semi-infinite optimization: On some foundations, Journal of Computational Technologies 4, 3 (1999) 41-61.

[57] Weber, G.-W., *Generalized Semi-Infinite Optimization and Related Topics*, Heldermann Publishing House, Research and Exposition in Mathematics 29, Lemgo, Hofmannn, K.H., and Wille, R. (eds.), 2003.

[58] Weber, G.-W., Kropat, E., Akteke-Öztürk, B., Görgülü, Z.-K., A survey on OR and mathematical methods applied on gene-environment networks, to appear in the special issue of Central European Journal of Operations Research (CEJOR) at the occasion of EURO XXII 2007 (Prague, Czech Republic, July 8-11, 2007) in 2009 (issue 3 or 4).

[59] Weber, G.-W., Uğur Ö., Taylan, P., Tezel, A., On optimization, dynamics and uncertainty: a tutorial for gene-environment networks, to appear in the special issue *Networks in Computational Biology* of Discrete Applied Mathematics, DOI number: doi:10.1016/j.dam.2008.06.030.

[60] Weber, G.-W., and Tezel, A., On generalized semi-infinite optimization of genetic networks, TOP 15, 1 (2007).

[61] Weber, G.-W., Tezel, A., Taylan, P., Soyler, A., and Çetin, M., On dynamics and optimization of gene-environment networks, to appear in the special issue of Optimization in honour of the 60th birthday of Prof. Dr. H.Th. Jongen.

[62] Weber, G.-W., Taylan, P., Alparslan-Gök, Z., Özögür, S., and Akteke-Öztürk, B., Optimization of gene-environment networks in the presence of errors and uncertainty with Chebychev approximation, preprint no. 64, Institute of Applied Mathematics, METU, 2006, submitted to the special issue of Discrete Applied Mathematics "GO V" in honour of the 70th birthday of Prof. Dr. P.L. Hammer and Prof. Dr. J. Krarup.

[63] Yagil, G., and Yagil, E., On the relation between effector concentration and the rate of induced enzyme synthesis, Biophysical Journal 11 (1971) 11-27.

[64] Yılmaz, F.B., *A Mathematical Modeling and Approximation of Gene Expression Patterns by Linear and Quadratic Regulatory Relations and Analysis of Gene Networks*, Institute of Applied Mathematics, METU, MSc Thesis, 2004.

[65] Yılmaz, F.B., Öktem, H., and Weber, G.-W., Mathematical modeling and approximation of gene expression patterns and gene networks, in: *Operations Research Proceedings*, Fleuren, F., den Hertog, D., and Kort, P. (eds.) (2005) 280-287.

[66] Zupan, B., Bratkoa, I., Demšar, J., Juvana, P., Curka, T., Borštnik, U, Beckc, J.R., Halterd, J., Kuspae, A. and Shaulskyf, G., GenePath: a system for inference of genetic networks and proposal of genetic experiments, Artificial Intelligence in Medicine 29 (2003) 107-130.