Modern Tools for the Time-Discrete Dynamics and Optimization of Gene-Environment Networks

Ozlem Defterli * **Armin Fügenschuh** **
Gerhard Wilhelm Weber ***

* *Department of Mathematics and Computer Science, Faculty of Art and Sciences, Çankaya University, Balgat 06530, Ankara, Turkey (e-mail: defterli@cankaya.edu.tr)*
** *Department for Optimization, Zuse Institute Berlin, Berlin, Germany (e-mail: fuegenschuh@zib.de)*
*** *Institute of Applied Mathematics, Middle East Technical University, 06531 Ankara, Turkey (e-mail: gweber@metu.edu.tr)*

Abstract:
In a genetic and metabolic structure, the interactions among the genes have to be identified and the influences aimed at to be predicted. In such a genetic network, expressing clearly the connections is a research problem of mathematical modeling which has significant application areas. In this study, we discuss the models whose dynamics are described by a class of time-continuous ordinary differential equations having a general form in the extended space $\dot{\mathbb{E}} = \mathbb{M}(\mathbb{E})\mathbb{E}$ where $\mathbb{E}$ is a vector of gene-expression levels and $\mathbb{M}(\mathbb{E})$ is the matrix having functional entries containing unknown parameters to be optimized. Accordingly, time-discrete versions of that model class are studied and improved by introducing 3rd-order Heun's method and 4th-order classical Runge-Kutta method. The corresponding iteration formulas are derived and corresponding matrix algebras are obtained. After that, we use nonlinear mixed-integer programming for the parameter estimation in the considered model and present the solution of a constrained and regularized given mixed-integer problem as an example. By using this solution and applying the 3rd-order Heun's method as a different discretization scheme, we generate corresponding time-series of gene-expressions and compare them with the experimental data and with the approximate results that we obtained from other numerical methods to exercise the performance of the new scheme on this example.

Keywords: Gene-networks, regulatory systems, environments, dynamical systems, discretization, mixed-integer nonlinear programming, optimization.

1. INTRODUCTION

The analysis of time-series gene expression data, which are obtained from DNA-microarray chip experiments, is a challenging problem that has significant applications in the areas of life and environmental sciences, computational biology, and engineering sciences. According to these obtained experimental data and the data coming from environmental measurements, the interactions of each gene with the others in a metabolic and genetic structure have to be identified and the influences need to be predicted.

Investigating the genetic networks is one of the interesting and promising subject of modern science. A genetic network can be defined as a weighted directed graph which consists of nodes representing genes, and of arcs with functional weights expressing the influences of each gene onto the other genes in the network. Each node can be equipped with a (level) function of the other genes' combined effects on it. These kinds of influences between genes are aimed to be predicted. There are many developed analytic and numerical tools for the construction and understanding of such networks which are studied by Ahuja et al. (1993); Chen et al. (1999); Defterli et al. (2010); DeRisi et al. (1997); Gebert et al. (2004, 2006, 2007); Taştan (2005); Uğur et al. (2009); Pickl and Weber (2001); Weber et al. (2008a,b, 2009a,b). In the papers of Uğur and Weber (2007); Uğur et al. (2009); Weber et al. (2008a,b, 2009b), genetic networks are extended to gene-environment networks. In this extension, the new nodes represent environmental items such as poison in soil, groundwater, in air or food, emissions, radiation, and also the welfare and living conditions, temperature (concerning, e.g., global warming), and so on, for a healthy lifestyle.

In this work, we introduce and analyze time-discrete target-environment regulatory systems, especially, for gene-environment and eco-finance networks. Then, we present a corresponding mixed-integer nonlinear programming (MINLP) problem. This paper widens the existing mathematical toolbox by introducing other schemes of time-discretization into the study and discussing their potential of improvement. Lastly, we gave the state-of-the-art methods and software of mixed-integer programming into

the area of gene-environment networks. By this extended toolbox, we aim at being better prepared for the modeling and prediction of our networks, and for a better service in the mentioned real-world areas.

In Section 2, we give some information about MINLP problems with their classifications. Section 3, introduces the models for our network class and present the networks' dynamics which, then, in Section 4, will become time-discretized by using the new schemes. For the model that we consider in Section 5, a MINLP problem is defined as our optimization problem. Then, it is solved numerically as an example in Section 6 where the corresponding calculations are presented with the comments on the obtained numerical results. Then, some comparative work is done for different methods with the help of the figures. The conclusion of our paper and an outlook are given in Section 7.

2. ABOUT MIXED-INTEGER NONLINEAR PROGRAMMING

Mixed-integer nonlinear programs are models of the general form

$$z = \min\ f(x), \tag{1a}$$
$$\text{subject to (s.t.) } g(x) \leq 0, \tag{1b}$$
$$x \in \mathbb{Z}^p \times \mathbb{R}^{n-p}, \tag{1c}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is an objective function, and $g : \mathbb{R}^n \to \mathbb{R}^m$ is a constraint system. We assume that f and g are continuous functions and that $X := \{x \in \mathbb{Z}^p \times \mathbb{R}^{n-p} : g(x) \leq 0\}$ is a compact set. This implies that f attains its minimum for some $x \in X$. Hence (1) is a well-defined problem. The question is how to actually solve a problem of the form (1) numerically.

Without loss of generality we can assume that f is linear. If not, we can introduce a new variable y and add the constraint $f(x) \leq y$ to the constraint system (1b). Together with the new objective function $\min y$ we then obtain a problem that is equivalent to (1), but with a linear objective function. If g is differentiable and $p = 0$, then (1) is a pure nonlinear optimization problem, and techniques from constrained nonlinear optimization can be applied. If g fulfills futher regularity assumptions, the Karush-Kuhn-Tucker (KKT) conditions provide necessary conditions for a solution to be (local) optimal; see Boyd and Vandenberghe (2004). These techniques originate from numerical analysis and yield only stationary points or local optima, if no further convexity assumption is made. Moreover, in the case of $p > 0$, they are not able to handle integrality restrictions on the variables. However, in case of a convex optimization problem, that is, if X is a convex set and f is a convex function, they are able to find a global optimum.

For a general MINLP with a non-convex set X there are several methods described in the literature in order to relax (1) to a convex and continuous problem, such that a proven global optimum can be achieved at least for the relaxed problem (Smith and Pantelides (1999); Tawarmalani and Sahinidis (2002, 2004)). The first way in this direction is to relax the integrality constraint on the variables, which gives the following relaxation:

$$z^0 = \min\ f(x), \tag{2a}$$
$$\text{s.t. } g(x) \leq 0, \tag{2b}$$
$$x \in \mathbb{R}^n. \tag{2c}$$

Clearly we have $z^0 \leq z$.

One example from the literature for a convexification of the nonconvex, nonlinear function g is the αBB method (see Adjiman et al. (1996)). There g is replaced by convexified nonlinear functions $\hat{g}^0$ using convex terms that are added with a suitable large weight parameter α in order to ensure global convexity. In the approach that we follow, we replace g by convex, linear functions $\hat{g}^0$ which are outer approximations of the non-convex functions. In any case one has to ensure that $\hat{g}^0 \leq g$, hence $X \subseteq \hat{X} := \{x \in \mathbb{Z}^p \times \mathbb{R}^{n-p} : \hat{g}^0(x) \leq 0\}$. The corresponding global optimum $\hat{z}^0$ is at least a lower bound on z^0, since we are solving a relaxation. Denote by $\hat{x}^0$ a point with $f(\hat{x}^0) = \hat{z}^0$. If $\hat{x}^0 \in X$, then a global optimum for problem (1) is found, and we are done. Otherwise, we have to refine the relaxation. This can be done by embedding the whole procedure in a branch-and-bound process, which we outline in the sequel.

If $\hat{x}^*0 \notin X$ then two things could be the case. First, $\hat{x}^0 \notin \mathbb{Z}^p \times \mathbb{R}^{n-p}$. In this case there exists a coordinate $1 \leq j \leq p$ with $\hat{x}^0_j \notin \mathbb{Z}$. We then break the one (father) problem (2) into two child problems. One problem is

$$z^1 = \min\ f(x), \tag{3a}$$
$$\text{s.t. } \hat{g}^0(x) \leq 0, \tag{3b}$$
$$x_j \leq \lfloor \hat{x}^0_j \rfloor, \quad x \in \mathbb{R}^n, \tag{3c}$$

the other problem is

$$z^1 = \min\ f(x), \tag{4a}$$
$$\text{s.t. } \hat{g}^0(x) \leq 0, \tag{4b}$$
$$x_j \geq \lceil \hat{x}^0_j \rceil, \quad x \in \mathbb{R}^n. \tag{4c}$$

Second, $g(\hat{x}^0) > 0$. In this case we identify a coordinate $p < j \leq n$ and replace the approximation $\hat{g}^0$ by two approximations $\hat{g}^1(x)$ which is valid for $\{x \in \mathbb{Z}^p \times \mathbb{R}^{n-p} : x_j \leq \hat{x}^0\}$, and $\hat{g}^2(x)$ which is valid for $\{x \in \mathbb{Z}^p \times \mathbb{R}^{n-p} : x_j \geq \hat{x}^0\}$. Since we are branching on a continuous variable x_j, this process is also known as spatial branching. So we again obtain two child problems of the form

$$z^1 = \min\ f(x), \tag{5a}$$
$$\text{s.t. } \hat{g}^1(x) \leq 0, \tag{5b}$$
$$x_j \leq \hat{x}^0_j, \quad x \in \mathbb{R}^n, \tag{5c}$$

the other problem is

$$z^2 = \min\ f(x), \tag{6a}$$
$$\text{s.t. } \hat{g}^2(x) \leq 0, \tag{6b}$$
$$x_j \geq \hat{x}^0_j, \quad x \in \mathbb{R}^n. \tag{6c}$$

The new convexifications $\hat{g}^1, \hat{g}^2$ are only locally valid, in contrast to the globally valid convexification $\hat{g}^0$. This gives the opportunity to obtain a tighter relaxation.

By breaking up the one difficult father problem (2) into two child problems, either (3) and (4), or (5) and (6), we

can again apply the same solution algorithm as before. In case of the αBB method, this is a convex NLP solver, in our case of linear functions, this is done by an LP solver.

After solving each of the child problems, we are in general faced with the same problem as before, namely that we only obtained a solution of a relaxation that is not feasible for the original problem. Hence, we have to apply the same reasoning as above and select another variable for branching and further refinement of the relaxation. The list of problems that are created in this way is usually managed as a tree, where at the root node the first problem (2) is located, and all other problems are further nodes of this tree. If all variables are integer variables, turns out to be X is a bounded set, this tree has a finite (but potentially large) vertices only. If we have to perform spatial branching, the finiteness can only be assured if we define an approximation accuracy, such as 10^{-4}, and if the difference between the original function g and its convex approximation $\hat{g}$ is less than this finite precision, we do not branch further.

In general, we do not have to enumerate the whole tree. There are ways to bound its size and cut off unnecessary parts of it. A node can be pruned if the subproblem turn out to be infeasible. If the solution is feasible not only for the relaxation, but also for the root problem (2), then we have an upper bound $\bar{z}$ on z. We do not have to branch this node further. Moreover, the bound $\bar{z}$ can be used to prune all other open subproblems having an objective function already equal to or greater than $\bar{z}$. Due to the fact that the tree can be kept significantly smaller, and thus computation times can be reduced to a large extent, one is interested in obtaining feasible solutions to (1) early in the solution process. These solutions can also come from other sources, such as genetic algorithms, tabu search, or simulated annealing, to name just a few of the most prominent meta-heuristics in this area.

3. THE CLASS OF MODELS FOR THE DYNAMICS OF GENE-ENVIRONMENT NETWORKS

In the literature, the firstly introduced time-continuous models to represent the gene-environment networks were given by the following systems of ordinary differential equations (ODEs) of the time-autonomous form

$$\dot{\mathbb{E}} = \mathbb{F}(\mathbb{E}), \tag{7}$$

where $\mathbb{E} = (\mathbb{E}_1, \mathbb{E}_2, \ldots, \mathbb{E}_d)^T$ ($\mathbb{E} = \mathbb{E}(t), t \in I$) is the d-vector of positive concentration levels of proteins (or mRNAs, or small components) and of certain levels of the environmental factors. $\dot{\mathbb{E}}\,(= \frac{d\mathbb{E}}{dt})$ represents a continuous change in the gene-expression data, and $\mathbb{F}_i : \mathbb{R}^d \to \mathbb{R}$ are nonlinear coordinate functions of $\mathbb{F}$ (Chen et al. (1999); Hoon et al. (2003); Sakamoto and Iba (2001); Uğur et al. (2009)). The estimation of parameters associated and contained in the definition of $\mathbb{F}$ is studied by considering the experimental data vectors $\bar{\mathbb{E}}$ of these levels which are obtained from microarray experiments and from environmental measurements at the sample times. Further, $\mathbb{E}(t_0) = \mathbb{E}_0$ denotes the initial values, where $\mathbb{E}_0 = \bar{\mathbb{E}}_0$. Moreover, $E_i(t)$ stands for the gene-expression level (concentration rate) of the ith gene at time t, and $E_i(t)$ denotes anyone of the first n coordinates in the d-vector $\mathbb{E}$ of genetic and environmental states. We write $G := \{1, ..., n\}$ for the set of genes.

There is a collection of types of Eqn. (7) representing the dynamical system on the gene-expressions and having the following forms given in Chen et al. (1999); Gebert et al. (2004, 2006, 2007); Sakamoto and Iba (2001); Taştan (2005); Taştan et al. (2005); Weber et al. (2008b); Yılmaz (2004):

(i)
$$\dot{E} = ME, \tag{8}$$
where M is an $(n \times n)$-constant matrix and E is the $(n \times 1)$-vector representing the expression level of individual genes.

(ii)
$$\dot{E} = M(E)E \tag{9}$$
is a continuous differential equation and the matrix M may depend on E. This dynamical system refers to the n genes and their interaction alone so that the matrix M is an $(n \times n)$-matrix with entries as functions of polynomials, exponential, trigonometric, splines or wavelets containing some parameters to be optimized.

(iii) In Yılmaz (2004), an extended version of the model given by Eqn. (9) is derived to emphasize the nonlinear interactions with the environment. Affine linear shifts terms are added in this extended model. To keep the recursive iteration idea, that is presented in Gebert et al. (2006), by these shifts, Eqn. (9) is below reconstructed from the following system that includes an affine addition (Weber et al. (2008b); Taştan (2005); Uğur et al. (2009); Taştan et al. (2005)):
$$\dot{E} \;=\; M(E)E + C(E). \tag{10}$$
Here, $C(E)$ is an additional column vector representing environmental perturbations or contributions and provides a more accurate data fitting (cf. Yılmaz (2004) for the case of a constant C). In the extended model represented by Eqn. (10), the dimension of the vector E is increased to $n + m$ by considering the m-vector $\check{E}(t) = (\check{E}_1(t), \check{E}_2(t), \ldots, \check{E}_m(t))^T$, which represents m environmental factor affecting the gene-expression levels and their variation. To represent the weights of the effect of the jth environmental factor $\check{E}_j$ on the gene-expression data E_i, the $(n \times m)$-weight matrix $\check{M}(E)$ is introduced so that the vector $C(E)$ can be written as $C(E) = \check{M}(E)\check{E}$, where $\check{M}(E)$ is called as the gene-environment matrix and its entries c_{ij} are the weights. Therefore, the gene-environment network described by the dynamic equation in (10) becomes
$$\dot{E} \;=\; M(E)E + \check{M}(E)\check{E}. \tag{11}$$
Finally, the extended initial value problem can be written in a multiplicative form as follows:
$$\dot{\mathbb{E}} = \mathbb{M}(\mathbb{E})\mathbb{E}, \quad \mathbb{E}_0 = \mathbb{E}(t_0) = \begin{bmatrix} E_0 \\ \check{E}_0 \end{bmatrix}, \tag{12}$$
where
$$\mathbb{E} := \begin{bmatrix} E \\ \check{E} \end{bmatrix}, \mathbb{M}(\mathbb{E}) := \begin{pmatrix} M(E) & \check{M}(E) \\ 0 & 0 \end{pmatrix}, \tag{13}$$
are an $(n + m)$-vector and $(n + m) \times (n + m)$-matrix, respectively. The other versions of the extended gene-environment network in Eqn. (10) is studied in Weber et al. (2008a,b, 2009b).

The models given by the continuous dynamical equations in (i), (ii) and (iii) can be written in general as (Uğur and Weber (2007); Weber et al. (2008a,b, 2009b))

$$\dot{\mathbb{E}} = \mathbb{M}(\mathbb{E})\mathbb{E}, \tag{14}$$

with the initial value $\mathbb{E}_0 = \mathbb{E}(t_0)$, where $\mathbb{E}$ and $\mathbb{E}_0$ are $(d\times 1)$-vectors. The $(d\times d)$-matrix $\mathbb{M}(\mathbb{E})$ has entries which contain parameters to be estimated, see Aster et al. (2004); Hastie et al. (2001). The entries of $\mathbb{M}(\mathbb{E})$, which can be polynomial, trigonometric, exponential, but otherwise logarithmic, hyperbolic, spline, etc., represent the growth, cyclicity or other kinds of changes in the genetic or environmental concentration rates that we suppose by any kind of a priori information, observation or assumption (Gebert et al. (2004)).

4. DISCRETIZATION SCHEMES FOR THE TIME-DISCRETE MODELS

4.1 Formulation of the Schemes

To approximate the time-continuous models and equations listed in Section 3, discretization schemes can be used to obtain the numerical solution at a discrete set of time points. It is important to choose the appropriate method to be applied. Firstly, the Euler's method was used in the time-discretization for the gene-expression patterns; it has been seen that Euler's method is slow and inaccurate (see Dubois and Kalisz (2004) for further information). Then, Runge-Kutta methods were introduced in Ergenç and Weber (2004) and, specifically, the 2nd order Heun's method was studied in Taştan (2005); Taştan et al. (2005) known as the simplest Runge-Kutta approach. In terms of rounding error and truncation error, the choice of the method in the numerical derivations plays an important role. Comparing with the Euler's method, Runge-Kutta methods have advantages in truncation error, and in stability which is closer to the stability of the time-continuous model, and in implementation (Heath (2002)). In this paper, we study 4th order classical Runge-Kutta method in addition to the newly derived 3rd order Heun's method in Defterli et al. (2010) for the discretization of the time-continuous models to improve the rate of convergence and accuracy.

In the most general model of gene-environment network that is given by Eqn. (14), we apply the 3rd order Heun's method and 4th-order classical Runge-Kutta method respectively and formulate our time-discrete model as follows:

(1) *by 3rd-order Heun's method:*

$$\begin{aligned}
\mathbb{E}^{(k+1)} &= \mathbb{E}^{(k)} + \frac{h_k}{4}(k_1 + 3k_3), \qquad (15)\\
k_1 &= \mathbb{M}(\mathbb{E}^{(k)})\mathbb{E}^{(k)},\\
k_2 &= \mathbb{M}(\mathbb{E}^{(k)} + \frac{h_k}{3}k_1)(\mathbb{E}^{(k)} + \frac{h_k}{3}k_1),\\
k_3 &= \mathbb{M}(\mathbb{E}^{(k)} + \frac{2h_k}{3}k_2)(\mathbb{E}^{(k)} + \frac{2h_k}{3}k_2).
\end{aligned}$$

Then, we get the time-discrete equation as

$$\mathbb{E}^{(k+1)} = \mathbb{M}^{(k)}\mathbb{E}^{(k)}, \tag{16}$$

where

$$\begin{aligned}
\mathbb{M}^{(k)} :=& \mathbb{I} + \frac{h_k}{4}\mathbb{M}(\mathbb{E}^{(k)}) + \mathbb{M}(\mathbb{E}^{(k)} + \frac{2h_k}{3}\mathbb{M}(\mathbb{T}^{(k)})\mathbb{T}^{(k)})\\
&\times\{\frac{3h_k}{4}\mathbb{I} + \frac{h_k^2}{2}\mathbb{M}(\mathbb{T}^{(k)}) + \frac{h_k^3}{6}\mathbb{M}(\mathbb{T}^{(k)})\mathbb{M}(\mathbb{E}^{(k)})\},
\end{aligned}$$

and $\mathbb{T}^{(k)} = \mathbb{E}^{(k)} + \frac{h_k}{3}\mathbb{M}(\mathbb{E}^{(k)})\mathbb{E}^{(k)}$.

(2) *by 4th-order classical Runge-Kutta method:*

$$\begin{aligned}
\mathbb{E}^{(k+1)} &= \mathbb{E}^{(k)} + \frac{h_k}{6}(k_1 + 2k_2 + 2k_3 + k_4), \qquad (17)\\
k_1 &= \mathbb{M}(\mathbb{E}^{(k)})\mathbb{E}^{(k)},\\
k_2 &= \mathbb{M}(\mathbb{E}^{(k)} + \frac{h_k}{2}k_1)(\mathbb{E}^{(k)} + \frac{h_k}{2}k_1),\\
k_3 &= \mathbb{M}(\mathbb{E}^{(k)} + \frac{h_k}{2}k_2)(\mathbb{E}^{(k)} + \frac{h_k}{2}k_2),\\
k_4 &= \mathbb{M}(\mathbb{E}^{(k)} + h_k k_3)(\mathbb{E}^{(k)} + h_k k_3);
\end{aligned}$$

which can be rewritten as

$$\begin{aligned}
\mathbb{E}^{(k+1)} =& \mathbb{E}^{(k)} + \frac{h_k}{6}\mathbb{M}(\mathbb{E}^{(k)})\mathbb{E}^{(k)}\\
&+ \{\frac{h_k}{3}\mathbb{M}(\mathbb{Z}^{(k)}) + \frac{h_k^2}{6}\mathbb{M}(\mathbb{Z}^{(k)})\mathbb{M}(\mathbb{E}^{(k)})\}\mathbb{E}^{(k)}\\
&+ \{\frac{h_k}{3}\mathbb{M}(\mathbb{V}^{(k)}) + \frac{h_k^2}{6}\mathbb{M}(\mathbb{V}^{(k)})\mathbb{M}(\mathbb{Z}^{(k)})\\
&+ \frac{h_k^3}{12}\mathbb{M}(\mathbb{V}^{(k)})\mathbb{M}(\mathbb{Z}^{(k)})\mathbb{M}(\mathbb{E}^{(k)})\}\mathbb{E}^{(k)}\\
&+ \frac{h_k}{6}\mathbb{M}(\mathbb{E}^{(k)} + h_k\mathbb{M}(\mathbb{V}^{(k)})\mathbb{E}^{(k)}\\
&+ \frac{h_k^2}{2}\mathbb{M}(\mathbb{V}^{(k)})\mathbb{M}(\mathbb{Z}^{(k)})\mathbb{Z}^{(k)})\\
&\times\{\mathbb{I} + h_k\mathbb{M}(\mathbb{V}^{(k)}) + \frac{h_k^2}{2}\mathbb{M}(\mathbb{V}^{(k)})\mathbb{M}(\mathbb{Z}^{(k)})\\
&+ \frac{h_k^3}{4}\mathbb{M}(\mathbb{V}^{(k)})\mathbb{M}(\mathbb{Z}^{(k)})\mathbb{M}(\mathbb{E}^{(k)})\}\mathbb{E}^{(k)}, \qquad (18)
\end{aligned}$$

where $\mathbb{Z}^{(k)} = \mathbb{E}^{(k)} + \frac{h_k}{2}\mathbb{M}(\mathbb{E}^{(k)})\mathbb{E}^{(k)}$, and $\mathbb{V}^{(k)} = \mathbb{E}^{(k)} + \frac{h_k}{2}\mathbb{M}(\mathbb{Z}^{(k)})\mathbb{Z}^{(k)}$. The time-discrete equation is obtained as

$$\mathbb{E}^{(k+1)} = \mathbb{M}^{(k)}\mathbb{E}^{(k)}, \tag{19}$$

with the matrix $\mathbb{M}^{(k)}$ defined as follows:

$$\begin{aligned}
\mathbb{M}^{(k)} :=& \mathbb{I} + \frac{h_k}{6}\{\mathbb{M}(\mathbb{E}^{(k)}) + 2\mathbb{M}(\mathbb{Z}^{(k)}) + 2\mathbb{M}(\mathbb{V}^{(k)}) + \mathbb{M}(\mathbb{T}^{(k)})\}\\
&+ \frac{h_k^2}{6}\{\mathbb{M}(\mathbb{Z}^{(k)})\mathbb{M}(\mathbb{E}^{(k)}) + \mathbb{M}(\mathbb{V}^{(k)})\mathbb{M}(\mathbb{Z}^{(k)})\\
&+ \mathbb{M}(\mathbb{T}^{(k)})\mathbb{M}(\mathbb{V}^{(k)})\} + \frac{h_k^3}{12}\{\mathbb{M}(\mathbb{V}^{(k)})\mathbb{M}(\mathbb{Z}^{(k)})\mathbb{M}(\mathbb{E}^{(k)})\\
&+ \mathbb{M}(\mathbb{T}^{(k)})\mathbb{M}(\mathbb{V}^{(k)})\mathbb{M}(\mathbb{Z}^{(k)})\}\\
&+ \frac{h_k^4}{24}\{\mathbb{M}(\mathbb{T}^{(k)})\mathbb{M}(\mathbb{V}^{(k)})\mathbb{M}(\mathbb{Z}^{(k)})\mathbb{M}(\mathbb{E}^{(k)})\},
\end{aligned}$$

where $\mathbb{T}^{(k)} = \mathbb{E}^{(k)} + h_k\mathbb{M}(\mathbb{V}^{(k)})\mathbb{V}^{(k)}$.

The approximate values of the next state can be obtained from the previous one by using the above iterative formula. The DNA microarray experimental data and the environmental items obtained at the time-level t_k are represented by the vector $\bar{\mathbb{E}}^{(\kappa)}$ ($\kappa = 0, 1, \ldots, l-1$; l: the number of biological measurements) in the extended space. The approximations in the sense of (16) are denoted by $\widehat{\mathbb{E}}^{(\kappa)}$ ($\kappa = 0, 1, \ldots, l-1$), and set $\widehat{\mathbb{E}}^{(0)} = \mathbb{E}^{(0)}$.

The kth approximation or prediction, $\widehat{\mathbb{E}}^{(k)}$, is calculated as $\widehat{\mathbb{E}}^{(k)}(:= \mathbb{E}^{(k)}) = \mathbb{M}^{(k-1)}(\mathbb{M}^{(k-2)} \cdots (\mathbb{M}^{(1)}(\mathbb{M}^{(0)}\mathbb{E}^{(0)})))$, where $h_k := t_{k+1} - t_k$ and $k \in \mathbb{N}_0$. We obtain our gene-environment networks by the time-discrete dynamics using formula (16). The genes and environmental items are represented by the nodes (vertices) of our network; the interactions between them are reflected by the edges, weighted with effects. The significant entry of $\mathbb{M}^{(k)}$, say, $m_{ij}^{(k)}$, is the coefficient of proportionality (i.e., multiplied by $\mathbb{E}_j^{(k)}$). It describes that the ith gene (or environmental factor) becomes changed by the jth gene (or environmental factor or the cumulative environmental item) in the step from time level k to $k+1$.

4.2 Corresponding Matrix Algebra

We refer to the canonical form of matrix partitioning, given in Taştan (2005); Taştan et al. (2005); Uğur and Weber (2007); Weber et al. (2008b), for the time-continuous model in Eqn.(12) as

$$\mathbb{M}(\mathbb{E}) = \begin{pmatrix} M(E) & \check{M}(E) \\ 0 & 0 \end{pmatrix}, \tag{20}$$

where $M(E)$ and $\check{M}(E)$ are the matrices having dimensions $n \times n$ and $n \times m$, respectively. Herewith, the dimension of the matrix $\mathbb{M}(\mathbb{E})$ is $(n+m) \times (n+m)$. Moreover, $\mathbb{E} = (E^T, \check{E}^T)^T$ and $\mathbb{T} = (T^T, \check{T}^T)^T$ are $(n+m)$-vectors. The relations of the genes and the environmental factors, which describe the structure of the gene and gene-environment network, are represented by these matrices. The matrices $\mathbb{M}^{(k)}$ will be the basis of the networks. The product of two such canonical matrices is again canonical (Taştan (2005); Taştan et al. (2005); Uğur and Weber (2007); Weber et al. (2008a,b, 2009b)). After some notation and simplification we find that

(1) *by using 3rd-order Heun's method:*

$$\mathbb{M}^{(k)} = \mathbb{I} + \frac{h_k}{4}\begin{pmatrix} M(E^{(k)}) & \check{M}(E^{(k)}) \\ 0 & 0 \end{pmatrix} + \frac{3h_k}{4}\begin{pmatrix} A & \tilde{A} \\ 0 & 0 \end{pmatrix} + \frac{h_k^2}{2}\begin{pmatrix} B & \tilde{B} \\ 0 & 0 \end{pmatrix} + \frac{h_k^3}{6}\begin{pmatrix} C & \tilde{C} \\ 0 & 0 \end{pmatrix}, \quad \text{where}$$

$$\begin{aligned}
A &:= M(E^{(k)} + \frac{2h_k}{3}(M(T^{(k)})T^{(k)} + \check{M}(T^{(k)})\check{T}^{(k)})),\\
\widetilde{A} &:= \check{M}(E^{(k)} + \frac{2h_k}{3}(M(T^{(k)})T^{(k)} + \check{M}(T^{(k)})\check{T}^{(k)})),\\
B &:= M(E^{(k)} + \frac{2h_k}{3}(M(T^{(k)})T^{(k)} \\
&\quad + \check{M}(T^{(k)})\check{T}^{(k)}))M(T^{(k)}),\\
\widetilde{B} &:= M(E^{(k)} + \frac{2h_k}{3}(M(T^{(k)})T^{(k)} \\
&\quad + \check{M}(T^{(k)})\check{T}^{(k)}))\check{M}(T^{(k)}),\\
C &:= M(E^{(k)} + \frac{2h_k}{3}(M(T^{(k)})T^{(k)} \\
&\quad + \check{M}(T^{(k)})\check{T}^{(k)}))M(T^{(k)})M(E^{(k)}),\\
\widetilde{C} &:= M(E^{(k)} + \frac{2h_k}{3}(M(T^{(k)})T^{(k)} \\
&\quad + \check{M}(T^{(k)})\check{T}^{(k)}))M(T^{(k)})\check{M}(E^{(k)}),
\end{aligned} \tag{21}$$

and $T^{(k)} := E^{(k)} + \frac{h_k}{3}\{M(E^{(k)})E^{(k)} + \check{M}(E^{(k)})\check{E}^{(k)}\}$, $\check{T}^{(k)} := \check{E}^{(k)}$, and $\mathbb{I} = I_d$ ($(d \times d)$-unit matrix) with $d = n + m$.

(2) *by using 4th-order classical Runge-Kutta method:*

$$\mathbb{M}^{(k)} = \mathbb{I} + \frac{h_k}{6}\begin{pmatrix} A & \tilde{A} \\ 0 & 0 \end{pmatrix} + \frac{h_k^2}{6}\begin{pmatrix} B & \tilde{B} \\ 0 & 0 \end{pmatrix} + \frac{h_k^3}{12}\begin{pmatrix} C & \tilde{C} \\ 0 & 0 \end{pmatrix} + \frac{h_k^4}{24}\begin{pmatrix} D & \tilde{D} \\ 0 & 0 \end{pmatrix}, \quad \text{with}$$

$$\begin{aligned}
A &:= M(E^{(k)}) + 2M(Z^{(k)}) + 2M(V^{(k)}) + M(T^{(k)}),\\
\widetilde{A} &:= \check{M}(E^{(k)}) + 2\check{M}(Z^{(k)}) + 2\check{M}(V^{(k)}) + \check{M}(T^{(k)}),\\
B &:= M(Z^{(k)})M(E^{(k)}) + M(V^{(k)})M(Z^{(k)}) \\
&\quad + M(T^{(k)})M(V^{(k)}),\\
\widetilde{B} &:= M(Z^{(k)})\check{M}(E^{(k)}) + M(V^{(k)})\check{M}(Z^{(k)}) \\
&\quad + M(T^{(k)})\check{M}(V^{(k)}),\\
C &:= M(V^{(k)})M(Z^{(k)})M(E^{(k)}) + M(T^{(k)})M(V^{(k)})M(Z^{(k)}),\\
\widetilde{C} &:= M(V^{(k)})M(Z^{(k)})\check{M}(E^{(k)}) + M(T^{(k)})M(V^{(k)})\check{M}(Z^{(k)}),\\
D &:= M(T^{(k)})M(V^{(k)})M(Z^{(k)})M(E^{(k)}),\\
\widetilde{D} &:= M(T^{(k)})M(V^{(k)})M(Z^{(k)})\check{M}(E^{(k)}),
\end{aligned} \tag{22}$$

where $Z^{(k)} := E^{(k)} + \frac{h_k}{2}\{M(E^{(k)})E^{(k)} + \check{M}(E^{(k)})\check{E}^{(k)}\}$, $V^{(k)} := E^{(k)} + \frac{h_k}{2}\{M(Z^{(k)})Z^{(k)} + \check{M}(Z^{(k)})\check{Z}^{(k)}\}$, $T^{(k)} := E^{(k)} + h_k\{M(V^{(k)})V^{(k)} + \check{M}(V^{(k)})\check{V}^{(k)}\}$, $\check{Z}^{(k)} = \check{V}^{(k)} = \check{T}^{(k)} := \check{E}^{(k)}$, $\mathbb{I} = I_d$ ($(d \times d)$-unit matrix) with $d = n + m$.

Therefore, $\mathbb{M}^{(k)}$ has its final canonical block form:

$$\begin{pmatrix} \widetilde{M(E^{(k)})} & \widetilde{\check{M}(E^{(k)})} \\ 0 & I_m \end{pmatrix}. \tag{23}$$

5. THE STUDIED MODEL

The model that we discuss as an example here is represented by the differential equation

$$\dot{E} = M(E)E, \tag{24}$$

that is described in Section 3 and where M is an $(n \times n)$-constant matrix. Our aim is to compute a gene network (represented by the matrix M) based on gene expression data. In order to do this, we solve a MINLP problem that is derived in the following way.

The objective function of our MINLP problem is the following (Defterli et al. (2010)):

$$\min_{M=(m_{ij})} \sum_{k=1}^{l} \left\| M\bar{E}^{(k)} - \dot{\bar{E}}^{(k)} \right\|_2^2, \tag{25}$$

that means, we want to find a matrix M such that the distances between the forecasted and the actual observed values are as small as possible with respect to the $\|\cdot\|_2$-norm. Here, l is the number of biological measurements and the $\dot{\bar{\mathbb{E}}}^{(k)}$ are the difference quotients based on the kth experimental data $\bar{\mathbb{E}}^{(k)}$ with step lengths h_k between neighbouring sampling times (Gebert et al. (2004, 2007); Uğur and Weber (2007)).

Because of a high degree of freedom in the problem, it is needed to restrict the solution space according to the underlying biological motivation (Gebert et al. (2004, 2006, 2007)). Otherwise, a very big amount of expression data is necessary to solve the minimization problem in (25). The values m_{ij} are nonnegative since no gene consumes another one, and $m_{ij} = 0$ means that the two genes i and j do not interact at all. A constant vector $\lambda \in \mathbb{R}^n$ represents the lower bound for the amount of decrease of the transcript concentration (Gebert et al. (2004, 2006, 2007)) between two time steps. Therefore, for $i, j \in G$ (where $G = \{1, 2, \ldots, n\}$ is the set of genes but environmental factors could be included here, too) we have

$$m_{ij} \geq \begin{cases} -\lambda(i), & i = j, \\ 0, & i \neq j. \end{cases} \tag{26}$$

To obtain a relatively sparse network, it is needed to limit the maximum outdegree and indegree of each node. In order not to lose the decomposition property of the minimization problem by limiting the maximum outdegree, we bound the indegree of each gene i by a given parameter $deg_{max,i} \in \mathbb{Z}_+$. So, in order to bound the indegree of each node, we introduce binary variables $y_{ij} \in \{0, 1\}$ in the subsequent way:

$$y_{ij} = \begin{cases} 0, & \text{if } \ m_{ij} = 0, \\ 1, & \text{if } \ m_{ij} \neq 0. \end{cases} \tag{27}$$

We formulate (27) as the following nonlinear constraints for our model:

$$(1 - y_{ij}) \cdot m_{ij} = 0, \quad \forall\, i, j \in G. \tag{28}$$

Now, the number of nonzero entries per row of the matrix $M = (m_{ij})_{1 \leq i,j \leq n}$ can be limited by the degree number, $deg_{max,i}$, which is content of the following constraints:

$$\sum_{j \in G} y_{ij} \leq deg_{max,i}, \quad \forall\, i \in G. \tag{29}$$

After considering all these constraints, we aim to solve the MINLP problem

$$\min (25), \text{ subject to } \{(26), (28), (29)\}, \tag{30}$$

to proven global optimality.

6. NUMERICAL RESULTS

Here, we numerically solve the problem in (30) within the model described by (24). We have four different genes and their expression levels at four different times according to the Table 1 (from Gebert et al. (2004)). We use an equally-

Table 1. Expression scores of the genes A, B, C and D at four time points

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

spaced time discretization as $h_k = 1 \ \forall k = 1, 2, ..., l-1$. In Defterli et al. (2010), we apply the 3rd-order Heun's method to approximate the $\dot{\bar{E}}_t$ according to the above given data and obtain $\dot{\bar{E}}_1^T = [0 \quad -50 \quad 50 \quad -255]$, $\dot{\bar{E}}_2^T = [0 \quad -20 \quad 20 \quad 255]$, $\dot{\bar{E}}_3^T = [0 \quad -20 \quad 20 \quad -255]$.

The constraints in the mixed-integer problem in (30), are given biologically as

$$\lambda(i) = 2, \ i = 1, ..., 4, \quad deg_{max,i} = 2. \tag{31}$$

Then, the problem is formulated and solved with the necessary software (see Defterli et al. (2010) and the references therein) in order to calculate the following matrix M:

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.26 & -0.46 & 0 & 0 \\ 0.19 & 0 & -0.46 & 0 \\ 1 & 0 & 0 & -2 \end{pmatrix}, \tag{32}$$

where the objective function value of matrix M for (25) is 92.31. Next, the 3rd-order Heun's time discretization formula for our model in (24) is derived as follows

$$E_{k+1} = (I + h_k M + \frac{h_k^2}{2} M^2 + \frac{h_k^3}{6} M^3) E_k, \tag{33}$$

Lastly, by using the obtained matrix M and the iteration formula in Eqn. (33), Defterli et al. (2010) get the approximate values of gene expressions in the below table:

Table 2. Approximation and extrapolation of gene expressions

time / genes	A	B	C	D
1	255	250	0	255
2	255	211.00	38.99	85
3	255	186.49	63.51	141.67
4	255	171.08	78.92	122.78
5	255	161.39	88.60	129.07
6	255	155.30	94.69	126.98
7	255	151.48	98.52	127.67
8	255	149.07	100.92	127.44
⋮	⋮	⋮	⋮	⋮
23	255	145.00	104.99	127.50
⋮	⋮	⋮	⋮	⋮
100	255	145.00	104.99	127.50

According to the generated time series in Table 2, we can say that the structural behavior of the obtained results is almost the same (constant first column, decreasing second column and increasing third column) with the given data in Table 1. For the values presented in the last column of Table 2., instead of an alternating behaviour, we obtain a damped oscillatory behaviour by using the 3rd-order Heun's discretization scheme. The results for the last column converges to the mean value of 0 and 255. The mean value that is reached can have two possible explanations:
(i) The gene-expression data shown in Table 1 are chosen as artificial data. The amplitudial maxima for the alternating gene expression in the last column in Table 2 were chosen at very close time point, which does most likely not represent a real biological behaviour. For the present paper, we did not change this in order to be able to compare our results with those of Gebert et al. (2004) where Table 1 is given. For our future work, we intend to use experimentally obtained gene-expression data.
(ii) Nevertheless, fading oscillating gene expression can be observed in biological systems. One well-known example is the damped oscillation of circadian rhythm after the trigger (day-light) has been removed.

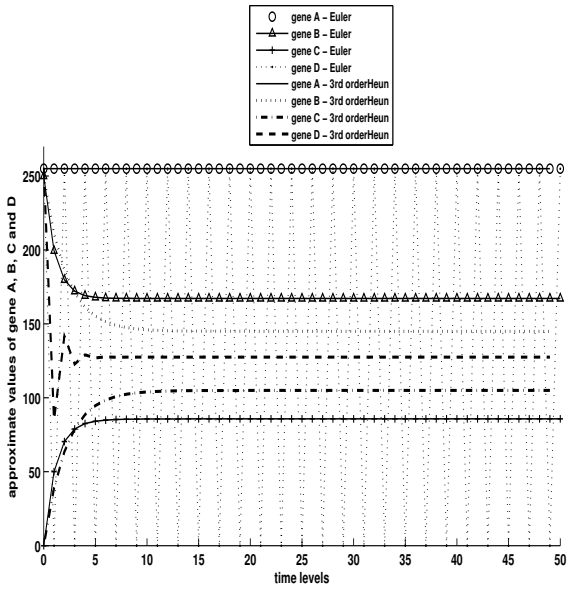

Fig. 1. Approximate results of gene-expressions of all genes by using Euler's and 3rd order Heun's methods.

We presented here Fig. 1, in order to compare the output coming from both Euler and 3rd order Heun methods using the calculated matrix M in (32). It is seen that the results of 3rd order Heun method are convergent and we reach the stable values after a few time steps. As a further step in this work, we calculated the approximate results of gene-expressions by using three different discretization methods and the same data for $\dot{\bar{E}}_t$, then compare the obtained results among them. Therefore, we apply the same procedure described above for Euler's method, 2rd order Heun's method and 3rd order Heun's method for the following fixed data of $\dot{\bar{E}}_t$:

$$\begin{aligned} \dot{\bar{E}}_1^T &= [0 \quad -50 \quad 50 \quad -255], \\ \dot{\bar{E}}_2^T &= [0 \quad -20 \quad 20 \quad 255], \\ \dot{\bar{E}}_3^T &= [0 \quad -10 \quad 10 \quad -255], \end{aligned} \tag{34}$$

obtained from the data in Table 1 and for the correspondingly calculated matrix M (with objective function value 2.564) in below :

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -0.20 & 0.38 & 0 \\ 0.19 & 0 & -0.58 & 0 \\ 1 & 0 & 0 & -2 \end{pmatrix}. \tag{35}$$

We present in the following graphs, the generated time series results for the gene-expression values that we get from these three different discretization schemes for the fixed data in (34). The newly derived 4th-order classical Runge-Kutta method will be compared in our future work.

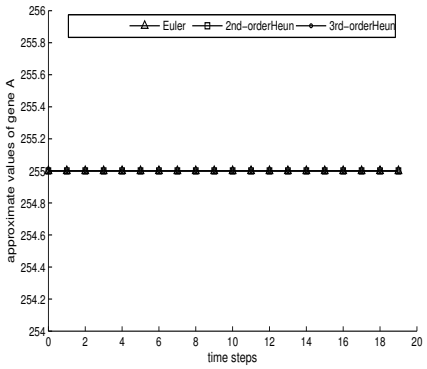

Fig. 2. Results of Gene A using different methods for fixed data

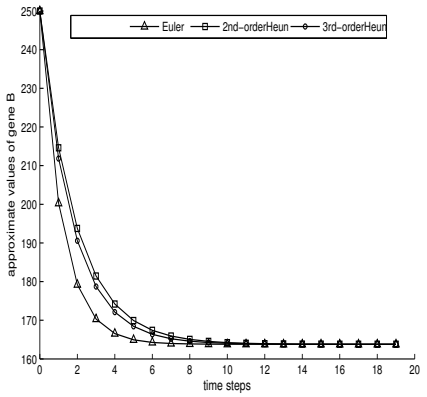

Fig. 3. Results of Gene B using different methods for fixed data

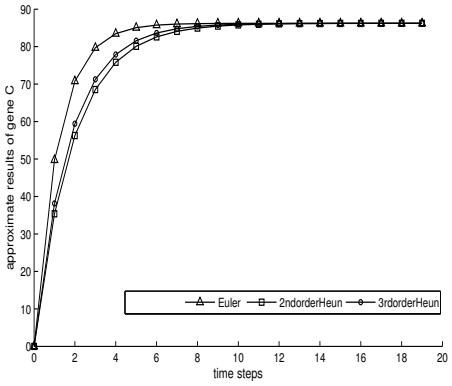

Fig. 4. Results of Gene C using different methods for fixed data

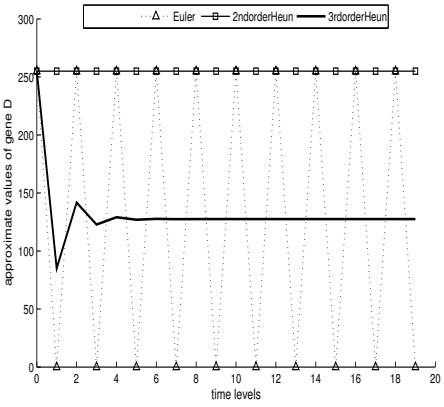

Fig. 5. Results of Gene D using different methods for fixed data

7. CONCLUSION

This research introduces and contributes to mathematical modeling, prediction and optimization of networks and systems whose motivations come from the real world. Within this work, we gave a contribution to an improved modeling of gene-environment networks, and to the numerical solution of their dynamics. By this, we supported a better future prediction of how such networks can develop in time (Hastie et al. (2001)), with important consequences in living conditions of the people.

In our future studies, we will work on the further improvements of the algorithms and different kinds of rarefications and combined methods together with the comparative studies. Moreover, we will combine our new numerical methods with the concepts of uncertainty and robustness to make modeling and prediction both more accurate and more stable.

ACKNOWLEDGEMENTS

This study is a part of Ph.D. thesis of Ozlem Defterli at the Department of Mathematics in Middle East Technical University (METU), Ankara, Turkey. The authors would like to thank to Professor Röbbe Wünschiers for his valuable comments on the biological interpretation of the calculated approximate results.
This work is partially supported by the Scientific and Technical Research Council of Turkey.

REFERENCES

C.S. Adjiman, I.P. Androulakis, C.D. Maranas, and C.A. Floudas. A global optimization method, alphabb, for process design. *Computers and Chemical Engineering*, 20:419 – 424, 1996.

R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network Flow: Theory, Algorithms and Applications*. Prentice Hall, New Jersey, 1993.

R.C. Aster, B. Borchers, and C.H. Thurber. *Parameter Estimation and Inverse Problems*. Academic Press, New York, 2004.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

T. Chen, H.L. He, and G.M. Church. Modeling gene expression with differential equations. *Proc. Pacific Symposium on Biocomputing*, 4:29–40, 1999.

O. Defterli, A. Fügenschuh, and G.-W. Weber. New discretization and optimization techniques with results in the dynamics of gene-environment networks. *Proc. PCO 2010-3rd Global Conference on Power Control and Optimization*, ISBN: 978-983-444483-1-8, 2010.

J. DeRisi, V. Iyer, and P. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

D.M. Dubois and E. Kalisz. Precision and stability of euler, runge-kutta and incursive algorithm for the harmonic oscillator. *International Journal of Computing Anticipatory Systems*, 14:21–36, 2004.

T. Ergenç and G.-W. Weber. Modeling and prediction of gene-expression patterns reconsidered with runge-kutta discretization. *Journal of Computational Technologies*, 9(6):40–48, 2004.

J. Gebert, M. Lätsch, S.W. Pickl, G.-W. Weber, and R. Wünschiers. Genetic networks and anticipation of gene expression patterns. *AIP Conference Proceedings*, 718:474–485, 2004.

J. Gebert, M. Lätsch, S.W. Pickl, G.-W. Weber, and R.Wünschiers. An algorithm to analyze stability of gene-expression pattern. *Discrete Applied Mathematics*, 154(7):1140–1156, 2006.

J. Gebert, N. Radde, and G.-W. Weber. Modelling gene regulatory networks with piecewise linear differential equations. *European Journal of Operational Research*, 181(3):1148–1165, 2007.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference and Prediction, Springer Series in Statistics*. Springer Verlag, New York, 2001.

M. Heath. *Scientific Computing: An Introductory Survey*. McGraw-Hill, New York, 2002.

M.D. Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations. *Proc. Pacific Symposium on Biocomputing*, 8:17–28, 2003.

S.W. Pickl and G.-W. Weber. Optimization of a time-discrete nonlinear dynamical system from a problem of ecology - an analytical and numerical approach. *Journal of Computational Technologies*, 6(1):43–52, 2001.

E. Sakamoto and H. Iba. Inferring a system of differential equations for a gene regulatory network by using genetic programming. *Proc. Congress on Evolutionary Computation'01*, pages 720–726, 2001.

E.M.B. Smith and C.C. Pantelides. A Symbolic Reformulation/Spatial Branch-and-Bound Algorithm for the Global Optimization of nonconvex MINLPs. *Computers & Chemical Engineering*, 23:457–478, 1999.

M. Taştan. *Analysis and prediction of gene expression patterns by dynamical systems, and by a combinatorial algorithm*. MSc Thesis, Institute of Applied Mathematics, METU, Ankara, Turkey, 2005.

M. Taştan, T. Ergenç, S.W. Pickl, and G.-W. Weber. Stability analysis of gene expression patterns by dynamical systems and a combinatorial algorithm. *Proc. International Symposium on Health Informatics and Bioinformatics (HIBIT'05)*, pages 67–75, 2005.

M. Tawarmalani and N.V. Sahinidis. *Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming: Theory, Algorithms, Software, and Applications*. Kluwer Academic Publishers, Boston MA, 2002.

M. Tawarmalani and N.V. Sahinidis. Global optimization of mixed-integer nonlinear programs: A theoretical and computational study. *Mathematical Programming*, 99(3):563–591, 2004.

Ö. Uğur and G.-W. Weber. Optimization and dynamics of gene-environment networks with intervals. *Journal of Industrial Management and Optimization(IJOR)*, 3(2): 357–379, 2007.

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

G.-W. Weber, P. Taylan, Z. Alparslan-Gök, S. Özögür, and B. Akteke-Öztürk. Optimization of gene-environment networks in the presence of errors and uncertainty with chebychev approximation. *TOP*, 16(2):284–318, 2008a.

G.-W. Weber, A. Tezel, P. Taylan, A. Soyler, and M. Çetin. Mathematical contributions to dynamics and optimization of gene-environment networks. *Optimization*, 57(2): 353–377, 2008b.

G.-W. Weber, S.Z. Alparslan-Gök, and B. Söyler. A new mathematical approach in environmental and life sciences: gene-environment networks and their dynamics. *Environmental Modeling and Assessment*, 14(2): 267–288, 2009a.

G.-W. Weber, Ö. Uğur, P. Taylan, and A. Tezel. On optimization, dynamics and uncertainty: a tutorial for gene-environment networks. *Discrete Applied Mathematics*, 157(10):2494–2513, 2009b.

F.B. Yılmaz. *A mathematical modeling and approximation of gene expression patterns by linear and quadratic regulatory relations and analysis of gene networks*. MSc Thesis, Institute of Applied Mathematics, METU, Ankara, Turkey, 2004.