Mathematical Modeling and Approximation of Gene Expression Patterns

F.B. Yılmaz[1], H. Öktem[1], and G.-W. Weber[1]

Institute of Applied Mathematics, Middle East Technical University, 06531, Ankara, Turkey

Abstract. This study concerns modeling, approximation and inference of gene regulatory dynamics on the basis of gene expression patterns. The dynamical behavior of gene expressions is represented by a system of ordinary differential equations. We introduce a gene-interaction matrix with some nonlinear entries, in particular, quadratic polynomials of the expression levels to keep the system solvable. The model parameters are determined by using optimization. Then, we provide the time-discrete approximation of our time-continuous model. Finally, from the considered models we derive gene regulatory networks, discuss their qualitative features and provide a basis for analyzing networks with nonlinear connections.

Keywords:
Gene Expression, Gene Regulation, Mathematical Modeling, Gene Network, Inference, Optimization, Dynamical Systems.

1 Introduction

Organisms contain a genetic material which has two main characteristics related to our study: *storage* and *expression* of information. The genetic material has some expressive parts, which are called *genes*. Each gene *stores* interrelated information necessary for *synthesis* of some particular *protein*. *Expression* of this information, i.e., *protein synthesis*, is a foundation for most of the phenotypic processes. Proteins function as *enzymes* and they *stabilize* the *transition states* of the reactions by increasing the rate of these reactions. In addition to this role, together with the genes, proteins *regulate* the synthesis of themselves. This interrelated structure, i.e., genes encode for proteins which regulate expression of genes, constructs a *gene regulatory network* where nodes stand for the genes and the regulatory influences are represented by directed edges between these nodes.

DNA arrays technology, described in [3], makes it possible to measure the expression levels on a genome-wide scale and, consequently, to infer the underlying network structures from experimental measurements, which is called *gene expression profiling*. The expression levels of genes show the relation between genotype and phenotype; this helps understanding biological processes. In our study, given a finite set of expression levels, we propose *a continuous*

model to describe the *dynamics* of the gene regulatory network, provide a *time-discrete equation* and compare linear and nonlinear approaches.

2 A Time Continuous Model

Given a finite number of gene expression levels for n genes, say $\bar{E}_0, \bar{E}_1, ..., \bar{E}_{l-1}$, where each $\bar{E}_m \in \mathbb{R}^n$ is a column vector representing the gene expression profile at time $\bar{t}_m$, these times satisfying $\bar{t}_m < \bar{t}_{m+1}$ $(m \in \{0, 1, ..., l-2\})$, then the question is how to *infer* the underlying gene regulatory network at the system level. Although the current sampling times of the experimental data constitutes an important challenge in *identification*, continuous nature of the underlying dynamics should be described by a continuous model [1,8].

Chen et al. [4] propose to model the gene expression by the differential equation $\dot{E} = ME$, where $E(t)$ and $\dot{E}(t)$ represents the concentrations and concentration changes of *mRNAs* and *proteins* at time t, respectively. Here, M is a matrix with constant entries representing the interactions between regulatory factors in the cell. In other words, $f_{j,i}(x) = a_{j,i}x$ $(a_{j,i} \in \mathbb{R})$ is used as the regulatory function for the influence level of the i^{th} regulatory factor to j^{th} regulatory factor, where $x = E_i$ represents the expression level of the i^{th} regulatory factor. *Chen et al.* [4] use *Minimum Weight Solutions to Linear Equations (MWSLE)* to determine the regulatory influences. Unfortunately, the algorithm is NP-complete and does not guarantee the solution.

Another continuous model proposed by *De Hoon et al.* [5], is based on the described model. They first approximate the differential equation by a difference equation and use *maximum likelihood estimation* to estimate M, then determine the number and places of the nonzero parameters in the matrix by the so-called *Akaikes Information Criterion [9].* In a more flexible approach, *Sakamoto and Iba* [10] choose the model $\dot{E}_j = f_j(E_1, E_2, E_3, ..., E_n)$ $(j = 1, 2, ..., n)$ with n being the number of genes and f being a function in $E_1, E_2, ..., E_n$. *Sakamoto and Iba* found the functions f_j by genetic programming combined with the least mean square method.

In [1,8], a new approach is introduced based on these models [4,5,10]: $\dot{E} = M(E)E$. Here, in the right-hand side of the equation, we see that the interaction matrix M depends on the current state, E. The study does not use particular functions, e.g., linear or piecewise linear, for inference, instead they suggest using any model class function and *restrict* the solution space to identify a *unique* regulatory network. We remind that the characteristics related to the model function reflect the dynamics in the interactions and incorporate any preinformation about the expected expression levels in mid- and long-term. In a process of *statistical learning* [9], where *training* and *testing* are iteratively coupled, the model class assumptions about the functions $f_{j,i}$ can step by step become improved.

Based on this approach, we assume that sampling times of gene expressions are sufficiently small to prevent aliasing in the sampling of the con-