Discrete Applied Mathematics 157 (2009) 2494–2513

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

On optimization, dynamics and uncertainty: A tutorial for gene-environment networks

G.-W. Weber [a,*], Ö. Uğur [a], P. Taylan [b], A. Tezel [c]

[a] Institute of Applied Mathematics, Middle East Technical University, 06531 Ankara, Turkey
[b] Dicle University, Department of Mathematics, 21280 Diyarbakır, Turkey
[c] Department of Mathematics, Middle East Technical University, 06531 Ankara, Turkey

ARTICLE INFO

Article history:
Received 6 January 2007
Received in revised form 15 May 2007
Accepted 11 June 2008
Available online 16 September 2008

Keywords:
Computational biology
Chebychev approximation
Generalized semi-infinite programming
Errors
Uncertainty
Modeling
Dynamical system
Intervals
Matrix
Structural stability
Splines
Conic programming
Continuous
Discrete
Hybrid

ABSTRACT

An emerging research area in computational biology and biotechnology is devoted to mathematical modeling and prediction of gene-expression patterns; to fully understand its foundations requires a mathematical study. This paper surveys and mathematically expands 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 arithmetic. 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. Finally, we analyze the topological landscape of gene-environment networks in terms of structural stability. This pioneering work is practically motivated and theoretically elaborated; it is directed towards contributing to applications concerning better health care, progress in medicine, a better education and more healthy living conditions.

1. Introduction

"*Can mathematics under the limitations of modern technology model the complexity of nature?*", "Yes", but only within the margins of our developing understanding; the margins of approximation in modeling only. Any new improvement to the model gives a chance to gain a deeper insight into nature and a hope for a continuously advanced service to the populace. In a similar manner, the complexity of the environment also includes psychological or societal phenomena, and its relationship with nature and lives of humankind are not an easy modeling task [70]. This paper is based on three 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. We aim at a greater contribution to scientific progress and services in medicine, health care, food production, industry and education.

There are two quantities coupled in modeling and prediction of gene-expression patterns: the levels (concentration, *states*) of gene-expressions and their rates of change (*dynamics*); both of them are of a "primary" importance. For the environmental effects, a "dual" role can be identified, such that we speak of some "duality" [70,79] which entirely

* Corresponding author. Tel.: +90 312 210 5652; fax: +90 312 210 2985.
E-mail addresses: gweber@metu.edu.tr (G.-W. Weber), ougur@metu.edu.tr (Ö. Uğur), ptaylan@dicle.edu.tr (P. Taylan), atezel@metu.edu.tr (A. Tezel).

0166-218X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2008.06.030

characterizes our learning problem, represented by a bilevel problem of optimization and decision [79,80]. 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 *prediction* are based. They represent linear mappings with 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 result from the matrix multiplication performed stepwise, and we analyze them by the combinatorial algorithm of Brayton and Tong [9,69]. 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.

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 [33]. In contrast, there is also the negative meaning; an organism, a living being or biosystem which is inflexible and unable to adapt to a changing environment and thus vulnerable to bacteria, viruses, radiation and other kinds of attacks. In addition, 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 or environmental levels by bounded intervals.

A *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; 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 [1,16,18,25–28,30,41,53,54,64,66,69,78–80,82,83]. 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. In [69,70,78–80], we firstly extended genetic networks to *gene-environment networks*. Now, the new nodes are environmental items such as poison in soil, groundwater, in air or food, radiation, but also welfare and living conditions, temperature (concerning, e.g., global warming), education and campaigns for a healthy lifestyle.

For a large number of genes the expression levels can easily be monitored by *DNA-microarray technology* [14]. Despite rapid advances in this area, it is nevertheless affected by uncertainties and measurement ambiguities. Therefore, we included these errors into our model [70,80]. Likewise for the environmental levels and concentrations, we face measurement and reliability problems, which we also represent in error terms. As introduced in [70,80], we will represent the various kinds of errors by *intervals*.

In general, genetic and gene-environment networks are too large to be easily investigated. Therefore, we impose bounds on 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, in addition, on errors and uncertainties located in intervals, the problem becomes one from *semi-infinite programming* (*SIP*). In addition, by allowing a 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 [70,78–80] we connected the discrete mathematics of networks with GSIP, a new and pioneering scientific approach into computational biology. GSIP is an advancing wide problem class with many motivations, results, future challenges and many practical applications even today [59,62,76]. In computational biology, sound *modeling*, *prediction* and *process optimization* are very important for a good understanding of *genetic processes*, of the *optimization of cell metabolism*, and for their application in medicine, health care, food production, industry and energy supply. Today, in a time of globalization, rapid information exchange, mobility and multicausalities in all kinds of biosystems, communities and societies, the ways in which 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 assist in advancing it.

2. Gene-expression and environmental data, modeling and dynamics

2.1. Introduction

2.1.1. Modeling by Intervals

In the early stages of modeling, time-continuous models representing the gene-environment networks were approached by systems of time-autonomous ordinary differential equations (ODEs):

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