A Genetic Approach for Synthesizing Metabolic Models from Time Series

Alberto Castellini*
Dept. Comp. Sci. University of Verona, Italy
alberto.castellini@univr.it

Vincenzo Manca
Dept. Comp. Sci. University of Verona, Italy
vincenzo.manca@univr.it

Mauro Zucchelli
University of Verona, Italy
mauro.zucchelli@studenti.univr.it

Mirko Busato
University of Verona, Italy
mirko.busato@studenti.univr.it

ABSTRACT

In this paper we introduce a new approach, based on *genetic algorithms* and *multiple linear regression*, for the synthesis of flux regulation functions in metabolic models from observed time series. Genetic algorithms are used as a variable selection technique to identify the best primitive functions for flux regulation, and multiple linear regression is employed to compute primitive function coefficients. Our methodology is here successfully applied to synthesize a set of regulation functions able to regenerate an observed dynamics for the mitotic oscillator in early amphibian embryos.

Categories and Subject Descriptors

I.6.5 [**Simulation and modeling**]: Model Development—*Modeling methodologies*

Keywords

Systems Biology, P Systems, Biological Modeling, Genetic algorithms, Regression, Flux regulation.

1. INTRODUCTION

Mathematical and computational *models* of biological systems are gaining day by day more importance in life sciences. These tools aim to represent in a formal way some knowledge about biological processes usually acquired from sets of quantitative observations. The process of model building (also called *modeling* in the following) is a central methodology in systems biology and synthetic biology since it enables in-silico *simulations* and *what-if* analyses of system behaviors in various conditions, such as, environmental alterations, pathological conditions or structural changes. Several modeling frameworks have been devised for analyzing different kinds of biological systems from different perspectives. A coarse classification can be made between *continuous* and *discrete* models, *static* and *dynamical* models, *deterministic* and *stochastic* models. Ordinary Differential Equations (ODEs) have been for a long time the most employed framework for dynamical system modeling. However, an important line of research in biological modeling aims at defining new classes of discrete models which avoid ODEs limitations.

*Corresponding author

GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA.
ACM 978-1-4503-1178-6/12/07.

2. MP SYSTEM MODELING

Metabolic P systems (shortly MP systems) [3] are a modeling framework based on P systems [6] and introduced for modeling metabolic processes by means of multiset rewriting grammars. An MP model is mainly composed of a set of *substances* $X = \{x_1, x_2, \ldots, x_n\}$, a set of *reactions* $R = \{r_1, r_2, \ldots, r_m\}$ and a set of chemo-physical *parameters* $V = \{v_1, v_2, \ldots, v_k\}$. Reactions act as rewriting rules transforming at each step a certain amount of substances which is tuned by suitable *regulation functions* $\Phi = \{\varphi_1, \ldots, \varphi_m\}$. Given an initial state $X[0] \in \mathbb{R}^n$, the *dynamics* of an MP system is computed by the vector recurrent equation called *EMA[i]* (*Equational Metabolic Algorithm*), having the following form: $X[i+1] = X[i] + \mathbb{A} \times U[i]$, where $\mathbb{A}$ is the *stoichiometric matrix*, $U[i] = (\varphi_1(Z[i]), \ldots, \varphi_m(Z[i]))'$ is the *flux vector*, and $Z[i] = (X[i], V[i]) \in \mathbb{R}^{n+k}$ is the *state* vector of substances and parameters at time i.

A key problem in MP modeling concerns with the synthesis of flux regulation functions $\varphi_1, \ldots, \varphi_m$ able to regenerate observed dynamics $Z^{obs} = (Z^{obs}[i], i = 1, \ldots, t)$ of substances and parameters, where $Z^{obs}[i] = (X^{obs}[i], V^{obs}[i])$ and t is the number of observations [2, 4]. The methodology here presented enables to synthesize *polynomial* regulation functions. *Genetic algorithms* [5] are employed to identify the best monomials (regressors) for each regulation function, while *multiple linear regression* [1] is used to compute the numeric parameters to be assigned to such monomials in order to reproduce the observed dynamics.

3. THE PROPOSED METHODOLOGY

Encoding. In order to describe the subsets of regressors employed by each regulation function we use a vector K (called *chromosome* in the following), such that

$$K = (k_j \in \{0, 1\}, j = 1, \ldots, m \cdot d), \tag{1}$$

where d is the number of regressors, $k_j = 0$ if the $(j\%d)$-th regressor is not present in regulation function $\varphi_{(j/m)+1}$ ($/$ and $\%$ are, respectively, the integer division and the rest of the integer division), and $k_j = 1$ if the $(j\%d)$-th regressor is present in $\varphi_{(j/m)+1}$.

Fitness evaluation. Given an MP system M and an observed dynamics $Z^{obs} = (Z^{obs}[i], i = 1, \ldots, t)$, a chromosome K is evaluated by computing:

- the *simulation error* of K, namely, the root mean square error ($RMSE$) between Z^{obs} and Z^{sim} over the first s steps ($s \leq t$) of the dynamics:

$$e_{sim}(K) = \sqrt{\frac{\sum_{j=1}^{n+k}\sum_{i=1}^{s}(Z_j^{obs}[i] - Z_j^{sim}[i])^2}{s\cdot(n+k)}} \quad (2)$$

where $Z^{obs}[i]$ ($Z_j^{sim}[i]$) is the state of the j-th element (i.e., a substance or a parameter) at step i of the observed (simulated) dynamics. The simulated dynamics is computed using the regressors selected by chromosome K. Regressor coefficients are estimated by multiple linear regression, employing a *stoichiometric expansion* [4] of the *EMA* equation, which will be described in an extended version of this paper;

- the *total error* of K, by adding to e_{sim} the weighted contribution of term $e_{\#1}(K) = \sum_{i=1}^{m\cdot d} k_i$, which is the number of 1s in chromosome K (i.e., the number of regressors in the overall model):

$$e_{tot}(K) = e_{sim}(K) + w \cdot e_{\#1}(K) \quad (3)$$

where $w \in \mathbb{R}_0^+$.

Errors $e_{sim}(K)$ and $e_{tot}(K)$ are fitness functions (to be minimized) used in different stages of our evolutionary process.

Chromosome selection. A population of N random chromosomes is generated at the beginning of the evolutionary process (see Figure 1). Afterwards, chromosomes are iteratively selected by *rank selection with elitist replacement* and recombined in order to improve their fitness.

Genetic operators. The chromosome population is evolved by means of two genetic operators (see Figure 1), namely, *crossover* and *mutation*. *Single point crossover* is performed with a probability p_{cross}. The output of this operation is a new couple of chromosomes that are subsequently mutated and inserted in the new population of chromosomes. Mutation is a very important operator in our approach since it makes possible to generate regulation functions incorporating some biological *a-priori* knowledge.

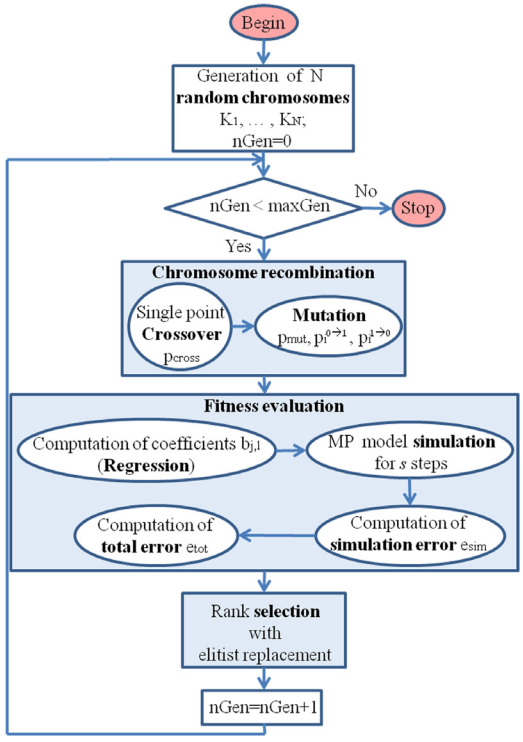

Figure 1: Overview of the evolutionary process.

4. EXPERIMENTS AND RESULTS

We have tested our evolutionary methodology on the case study of the *mitotic oscillator in early amphibian embryos*. This biological system has five substances, namely, C (*cyclin*), X (*active cyclin protease*), Xp (*inactive cyclin protease*), M (*active cdc2 kinase*) and Mp (*inactive cdc2 kinase*), and seven reactions that determine univocally the stoichiometry of the system.

Table 1 shows the regulation functions achieved in this test and Figure 2 displays a comparison between the dynamics generated using these functions (crosses) and the target dynamics (circles), that agree almost completely.

$$\begin{aligned}
\varphi_1(Z) &= 0.09C - 0.56X + 0.11X/(C+0.001) + \\
&\quad -0.00063/(C+0.001) \\
\varphi_2(Z) &= 0.10C^2 - 0.00067C/(M+0.001) \\
\varphi_3(Z) &= -0.22C + 0.67M - 0.93X^2 + \\
&\quad +0.0010C/(M+0.001) \\
\varphi_4(Z) &= 0.00047X/(M+0.001) \\
\varphi_5(Z) &= -0.5C^2 + 0.52M^2 + 0.13M/(C+0.001) + \\
&\quad -0.0011/(C+0.001) \\
\varphi_6(Z) &= -0.82M + 0.16C^2 \\
\varphi_7(Z) &= 0.75X - 1.92M^2 - 0.057M/(C+0.001)
\end{aligned}$$

Table 1: Regulation functions synthesized in our best test.

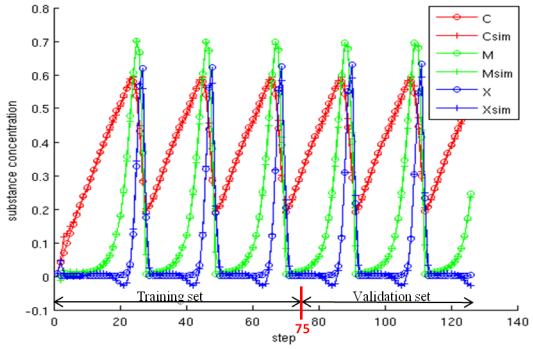

Figure 2: Comparison between the target dynamics (circles) and the dynamics generated using regulation functions of Table 1 (crosses).

5. REFERENCES

[1] A. D. Aczel and J. Sounderpandian. *Complete business statistics*. McGraw-Hill/Irwin, 2006.

[2] A. Castellini, G. Franco, and R. Pagliarini. Data analysis pipeline from laboratory to MP models. *Natural Computing*, 10(1):55–76, 2011.

[3] V. Manca. Fundamentals of metabolic P systems. In *Handbook of Membrane Computing*, chapter 19. Oxford University Press, 2009.

[4] V. Manca and L. Marchetti. Log-gain stoichiometric stepwise regression for MP systems. *Int. J. Found. Comput. Sci.*, 22(1):97–106, 2011.

[5] M. Mitchell. *An introduction to genetic algorithms*. MIT Press, 1998.

[6] G. Păun. *Membrane Computing. An Introduction*. Springer, Berlin, 2002.