

# Evolutionary Inference of a Biological Network as Differential Equations by Genetic Programming

Erina Sakamoto<sup>1</sup>

erina@miv.t.u-tokyo.ac.jp

Hitoshi Iba<sup>2</sup>

iba@miv.t.u-tokyo.ac.jp

<sup>1</sup> School of Engineering, Department of Information and Communication Engineering, University of Tokyo, Hongo 7-3-1 Bunkyo-ku, Tokyo 113-8656, Japan

<sup>2</sup> Graduate School of Frontier Sciences, Department of Frontier Informatics, University of Tokyo, Hongo 7-3-1 Bunkyo-ku, Tokyo 113-8656, Japan

**Keywords:** genome pathway, interaction analysis, genetic programming, and inverse problem.

## 1 Introduction

Inferring a biological network from a set of observed time series is becoming more important as technologies such as DNA microarrays have been developed rapidly in recent years. Many models have been proposed to describe the biological network. Among them is a system of differential equations, which is flexible to represent the complex relationships among their components. In most of the previous studies [1], the form of the equations was fixed because it was difficult to determine the suitable form giving the similar time series to the target. We have used the arbitrary form in the right hand side of the equation (1) as the model of the network [2], in which Genetic Programming (GP) has been successfully applied to inferring the biological network.

$$dX_i/dt = f_i(X_1, X_2, \dots, X_n) \quad (1)$$

## 2 Method

We use GP to identify a biological network in the form of the system of differential equations. Though GP is capable of finding a desirable structure effectively, it can't always be effective in finding the proper coefficients because GP uses the combination of randomly selected ones. We use the least mean square method (LMS) to tackle this defect of the ordinary GP. For this purpose, coefficients are not included in the terminal set that is used to compose a GP individual tree. The coefficients of each term of the GP tree are calculated by using the LMS method and a table of them composes a GP individual along with a tree.

The fitness of each individual is defined as the sum of the squared error and the penalty for the number of terms in the equation as shown in (2):

$$fitness = \sum_{i=0}^{T-1} (y(t_0 + i\Delta t) - x(t_0 + i\Delta t))^2 + a \times n, \quad (2)$$

where  $x(t_0 + i\Delta t)$  is the given target time series.  $y(t_0 + i\Delta t)$  is the time series acquired by calculating the system of differential equation represented by a GP individual.  $n$  is the number of the terms in the equation and  $a$  is the coefficient for the penalty. The individual which is of less terms and closer time series is more likely to be selected and inherited to the next generation.

We used several sets of time series as the training data for GP to enhance the accuracy of the GP search. Each data set was generated from the same target by using different initial values.

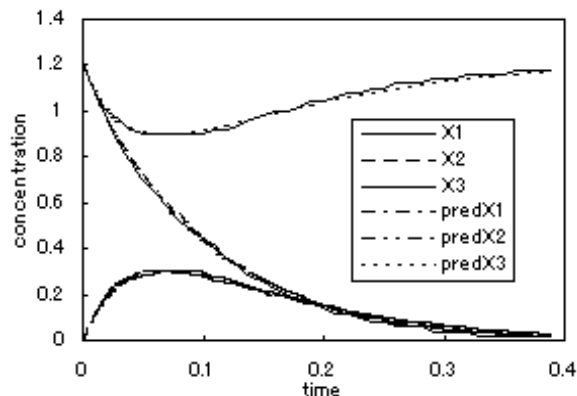


Figure 1: Acquired time series for the e-cell.

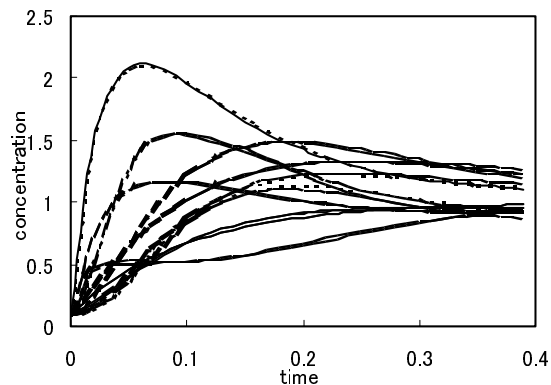


Figure 2: Acquired time series for the S-system.

### 3 Experimental Results

We have conducted the experiment on the data of a metabolic network that consists of three substances. This target network is a part of the biological phospholipid pathway. The target data was derived from the e-cell simulation model. This network can be approximated as (3).

Three sets of time series with a different initial value were used for the training of GP. Experimental parameters are shown in Table 1.

By applying our method, we have acquired equations shown in (4) and the average MSE (Mean Square Error) of 10 runs is  $2.545 \times 10^{-3}$ . Its time series is shown in Figure 1 along with that of the target.

$$\begin{cases} dX_1/dt = -k_1X_1X_3 \\ dX_2/dt = k_1X_1X_3 - k_2X_2 \\ dX_3/dt = -k_1X_1X_3 + k_2X_2 \end{cases} \quad (3) \quad \begin{cases} dX_1/dt = -10.3176X_1X_3 \\ dX_2/dt = 9.7149X_1X_3 - 17.5084X_2 \\ dX_3/dt = -9.7018X_1X_3 + 17.4766X_2 \end{cases} \quad (4)$$

We have also conducted a comparative experiment without the LMS method to confirm its effectiveness (in this case, coefficients are added to the terminal set). The average MSE of 10 runs is  $5.328 \times 10^{-3}$ , whereas that of the experiment with the LMS method is  $2.545 \times 10^{-3}$ . Besides, the correct form of equations was not always acquired without the LMS method. For example, in no runs, the correct form of equations for  $X_3$  was acquired without the LMS method. From these results, we can conclude that the LMS method worked effectively to acquire the better individual.

We also tested on the network which consists of 10 nodes and had been generated from the S-system. The acquired and the given target time series are shown in Figure 2. As can be seen, the acquired time series is quite close to the target one.

### 4 Discussion and Conclusion

As shown in the experimental results, we can confirm that our method can derive the differential equation which is closer to the target and the exact relation among the components. We will work on the extension of our method and apply our method to the network composed of more components.

### References

- [1] Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S., and Eguchi, Y., Development of a system for the inference of large scale genetic networks, *Proc. Pacific Symp. Biocomputing '01*, World Scientific, 446–458, 2001.
- [2] Sakamoto, E. and Iba, H., Inferring a system of differential equations for a gene regulatory network by using genetic programming, *Proc. Congress on Evolutionary Computation '01*, 720–726, 2001.