Prediction of Biochemical Reactions Using Genetic Programming

Masahiro Sugimoto[1,2] **Shinichi Kikuchi**[1]
msugi@sfc.keio.ac.jp kikuchi@sfc.keio.ac.jp

Masaru Tomita[1]
mt@sfc.keio.ac.jp

[1] Institute for Advanced Biosciences, Keio University, Tsuruoka 997-0035, Japan
[2] Mitsubishi Space Software Co., Ltd., 5-4-36 Tsukaguchi-honmachi, Amagasaki, Hyogo 661-0001, Japan

Keywords: biochemical reaction, generalization ability, optimization, genetic programming, biophysics, parameter estimation, mathematical model

1 Introduction

To comprehend dynamic behaviors of biological systems, many models have been proposed. These models need literatures data to represent detailed and accurate dynamics. However, those data are sufficient only in few cases. To solve this problem, many techniques have been developed, including Genetic Algorithm (GA). Those methods require defining equations before predicting biological systems. To consider the case where even equation could not be obtained, we employ the Genetic Programming (GP) that was studied as a method to predict arbitrary equation from time course data without any knowledge of the equation. However, it is difficult for conventional GP to search the equations with high accuracy because our target biochemical reactions include not only variables but also many numerical parameters. In order to improve the accuracy of GP, we extended elite strategy to focus on numerical parameters. In addition, we added a penalty term to evaluation function to save the growth of the size of tree and consuming calculation time. By applying our improvements, we were able to predict biochemical reactions whose dimensions of variables were strictly the same as those of originals. The relative square error of predicted and given time-course data were decreased from 25.4% to 0.744%. Moreover, in experiments to validate the generalization ability of the predicted equations, we successfully decreased the relative square error of the predicted and given time-course data from 25.7% to 0.836%. The results of our numerical experiments indicate that our method succeeded in predicting approximation formulas without any definition of equations with reduced square error.

2 Method and Results

To validate our proposed method, we conducted some numerical experiments. As a case study, we tried the equation about pure competitive inhibition reaction by 2 different exclusive inhibitors. We artificially prepared the numerical parameters of the above equation as below.

$$\frac{d[P]}{dt} = \frac{1.000[S]}{4.800 + [S] + 1.120[I] + 0.570[X]} \tag{1}$$

where $[S]$ and $[P]$ are the concentrations of substances.

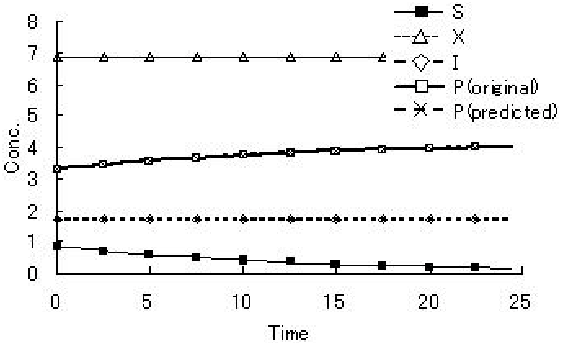

Figure 1: One of the results of given time-course data and time-course data simulated by predicted equation. Dos denote sampling points. These points were selected artificially as a case study. The square error of concentration of substance P of time-course data and time-course data simulated by predicted equation is 0.147%.

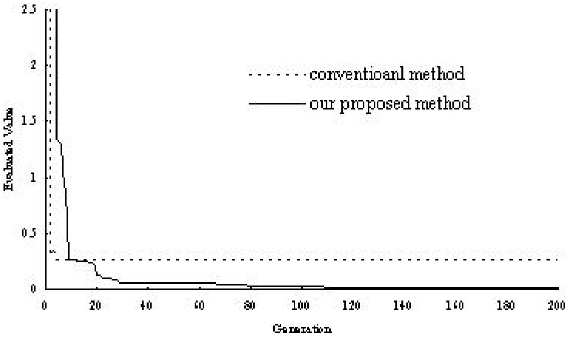

Figure 2: Transition of evaluated value of conventional method and proposed method. The dotted curve is the evaluated value calculated by conventional method. The solid one is that calculated by proposed method. The average of relative square error by conventional method is 25.4% and that by our proposed method is 0.744%.

The multiple time-course data sets with different initial concentration are generated by the equation and we predicted it using our proposed method from those time-course data sets. One of the results of predicted equations is shown as below.

$$\frac{d[P]}{dt} = \frac{[S]}{5.631 + [S] + [I] + 0.450[X]} \tag{2}$$

We prepared some time-course data and predicted equations from those time-courses. One of the results of given and simulated time-course is shown in Fig. 1. Transition of evaluated value of conventional method and proposed one is shown in Fig. 2.

3 Discussion

Our proposed method could find the equation whose dimension of variables are strictly same as originals and improve the accuracy compared to conventional GP. However, the numerical parameters of the denominator of predicted equations were a little different from the original so that the predicted equation was not able to calculate the initial dynamics satisfactorily. A simple means of avoiding such problems is to increase the number of sampling points. However, it is important to avoid over-fitting that worsens the ability to find the equation. The number of sampling points is a trade-off between those problems. In future investigations we will develop a method to find the appropriate number of sampling points automatically.

References

[1] Ando, S., Sakamoto, E., and Iba, H., Evolutionary modeling and inference of gene network, *Information Sciences*, 145(3-4):237–259, 2002.

[2] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., and Tomita, M., Dynamic modeling of genetic networks using genetic algorithm and S-system, *Bioinformatics*, 18(6):643–650, 2003.

[3] Szpiro, G.G., Forecasting chaotic time-course with genetic algorithms, *Physical Review E*, 55(3):2557–2568, 1997.