数理モデル化と問題解決　27－10
（1999．11．26）

線形計画法による非線形システム S-system の推定

阿久津 達也* 宮野 悟* 久原 哲#

*東京大学医科学研究所ヒトゲノム解析センター

#九州大学大学院生物資源環境科学研究科

DNA マイクロアレイなどにより得られた時系列データからの遺伝子ネットワークや代謝ネットワークの推定はバイオインフォマティックスにおいて重要な課題の一つとなっている。一方、化学反応における質量保存則に基づく S-system と呼ばれる微分方程式系が以前より化学反応や代謝系などのシミュレーションに用いられてきた。本稿では、S-system に基づいて時間変化をする遺伝子や化学物質の時系列データが与えられた時に、もとの微分方程式系を推定する問題について考察する。S-system は非線形の微分方程式系ではあるが、微分値（実際には差分値）の正負だけに注目することにより、推定問題を線形計画問題に帰着させることができる。推定されるパラメータの正確さは多少犠牲になるが、線形計画法を用いるために、非線形最適化に基づく推定手法より、はるかに高速に推定を行うことができるという利点がある。

Inference of Nonlinear Biological Systems from Time Series Data Using Linear Programming

Tatsuya Akutsu* Satoru Miyano* Satoru Kuhara#

*Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirkanedai, Minato-ku, Tokyo 108-8639, Japan
{takutsu,miyano}@ims.u-tokyo.ac.jp

Graduate School of Genetic Resources Technology, Kyushu University
6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan
kuhara@grt.kyushu-u.ac.jp

Inference of genetic networks and metabolic networks from time series data obtained by biological experiments using DNA micro arrays is an important topic in bioinformatics. In this article, we propose a method for inferring S-systems from time series data, where S-systems are based on a particular kind of nonlinear differential equation and have been applied to the analysis of various biological systems. We reduce each instance of the inference problem to the linear program by means of focusing on the sign of differential (or finite difference). Although the errors of estimated parameters are relatively large, the proposed method is much faster than most other methods that are based on nonlinear optimization.

1 Introduction

Due to the recent progress of the *DNA microarray* technology [5], it has become possible (to some extent) to measure the gene expression levels of most of the genes of an organism simultaneously. Recently, many studies have been done in order to develop computational methods for reconstructing underlying *genetic networks* from time series data of gene expression patterns.

Several studies have been done using the Boolean network [9], where a gene takes one of two states (ON or OFF), and a gene regulation rule is given as a Boolean function. Liang *et*

al. [9] developed the REVEAL algorithm (reverse engineering algorithm) for inferring genetic networks from state transition tables which correspond to time series data of gene expression patterns. We also made several theoretical studies on inference of Boolean networks [1]. Since genetic networks are very complex, there are many criticisms on the Boolean network approach. Of course, many other models and inference methods have been proposed: a qualitative model [12], hybrid models [10, 14], a statistical method for inference of infer chemical networks [3], and models and inference methods based in linear differential equations [6, 7]. However, no method seems to be sufficient.

On the other hand, the S-system (*synergistic* and *saturable* system) has been developed for analyzing and modeling biological systems [8], where S-systems are based on a particular kind of nonlinear differential equation. S-systems have been successfully applied to the analysis of various biological systems [8]. Recently, Tominaga and Okamoto [13] applied GA (Genetic Algorithm) to inference of S-systems. However, their method was time consuming and was limited to inference of S-systems with a few parameters.

In this article, we propose a novel method for inferring S-systems from time series data. We reduce each instance of the inference problem to a linear program by means of focusing on the sign of differential (or finite difference). We call this method the LP-based method, where LP denotes *linear programming* in this article. Although the errors of estimated parameters are relatively large, the LP-based method is much faster than the GA-based method. It is also expected that the LP-based method is much faster than most other methods that are based on nonlinear optimization because linear programs can be solved in polynomial time.

2 Inference of S-systems

Here we briefly review the definition of the S-system [8, 13]. Let $\{X_1, \ldots, X_n\}$ be a set of genes and/or chemical substances in the underlying biological network. Let $X_i(t)$ be the value (expression level or concentration) of a gene or a chemical substance X_i at time t.

An S-system is a set of *nonlinear* differential equations of the form

$$\frac{dX_i(t)}{dt} = \alpha_i \prod_{j=1}^{n} X_j(t)^{g_{i,j}} - \beta_i \prod_{j=1}^{n} X_j(t)^{h_{i,j}}$$

where α_i and β_i are multiplicative parameters called *rate constants* and $g_{i,j}$ and $h_{i,j}$ are exponential parameters called *kinetic orders*.

The inference problem is, given time series data $X_i(t)$ that are assumed to be generated from an S-system $\mathcal{S}$, to estimate to parameters α_i, β_i, $g_{i,j}$ and $h_{i,j}$ of $\mathcal{S}$.

Since S-systems are nonlinear, we can not apply linear regression [7] to inference of S-systems. As mentioned in the introduction, Tominaga and Okamoto [13] applied GA (Genetic Algorithm) to inference of S-systems with a few parameters. However, their method was time consuming. Therefore, we developed a new method for the inference of S-systems based on LP.

The method is quite simple. Assume that $\frac{dX_i(t)}{dt} > 0$ at time t. By taking 'log' of each side of $\alpha_i \prod X_j(t)^{g_{i,j}} > \beta_i \prod X_j(t)^{h_{i,j}}$, we have

$$\log \alpha_i + \sum_{j=1}^{n} g_{i,j} \log X_j(t) > \log \beta_i + \sum_{j=1}^{n} h_{i,j} \log X_j(t) .$$

Since $X_j(t)$'s are known data, this inequality is linear if we treat $\log \alpha_i$'s and $\log \beta_i$'s as parameters. In the case of $\frac{dX_i(t)}{dt} < 0$, we can obtain a similar inequality. Therefore, solving these linear inequalities by LP, we can determine parameters.

However, parameters are not determined uniquely even if a lot of data are given, because the inequality can be re-written as $(\log \alpha_i - \log \beta_i) + \sum (g_{i,j} - h_{i,j}) \log X_j(t) > 0$. Therefore, only relative ratios of $\log \alpha_i - \log \beta_i$ and $g_{i,j} - h_{i,j}$'s are determined (for each i). But, this information is useful for qualitative understanding of S-systems. Since $\prod X_j(t)^{g_{i,j}}$ contributes to the net production of X_i, $\prod X_j(t)^{h_{i,j}}$ contributes to the net degradation of X_i and it is not usual that X_j contributes to both the net production and the net degradation, either $g_{i,j} = 0$ or $h_{i,j} = 0$ holds for each (i, j) in most cases. Thus, the fact that $|g_{i,j} - h_{i,j}|$ is large means that X_i is influenced by X_j.

It should be noted that the LP-based method is not robust for noises since parameter values are not determined even if one linear inequality is not satisfied. However, such noisy cases may be handled by using *robust linear programming* [4].

3 Computational Experiments

We made computational experiments on the LP-based method using a SUN ULTRA-2 Workstation with 1 CPU (296MHz). In order to solve LP, we used SOPT [11]. In these experiments, we used $\frac{\Delta X(t)}{\Delta t}$ in place of $\frac{dX(t)}{dt}$.

First we examined the following simple cases of $n = 2$, where case (A) was examined in Ref. [13] too.

	i	α_i	$g_{i,1}$	$g_{i,2}$	β_i	$h_{i,1}$	$h_{i,2}$
(A)	1	3.0	0.0	-2.5	3.0	0.125	0.0
	2	3.0	2.5	0.0	3.0	0.0	0.125
(B)	1	3.0	0.0	-2.5	3.0	1.25	0.0
	2	3.0	2.5	0.0	3.0	0.0	1.25

As input data, time series data beginning from randomly generated initial values in $[0.5, 2.0]$ were used. The Euler method was used to generate the time series data, where $\Delta t = 0.02$ was used. Since the LP-based method can only compute relative values of $g_{i,j} - h_{i,j}$'s, we compare the ratios $r_1 = \frac{g_{1,1} - h_{1,1}}{g_{1,2} - h_{1,2}}$ and $r_2 = \frac{g_{2,2} - h_{2,2}}{g_{2,1} - h_{2,1}}$. The following table shows the result, where average values and standard deviations over 20 trials are shown. m denotes the total number of time points in the data, where 50 point data are generated from each set of initial values.

		Correct	$m = 1 \times 50$	$m = 5 \times 50$	$m = 10 \times 50$
(A)	(r_1, σ)	(0.05, -)	(0.129, 0.032)	(0.081, 0.009)	(0.077, 0.011)
	(r_2, σ)	(-0.05,-)	(-0.261,0.232)	(-0.086,0.023)	(-0.085,0.011)
(B)	(r_1, σ)	(0.5, -)	(0.653, 0.099)	(0.598, 0.054)	(0.574, 0.040)
	(r_2, σ)	(-0.5,-)	(-0.648,0.108)	(-0.568,0.032)	(-0.538,0.029)

In each case, parameters were inferred within 1 second, which is much faster than the GA-based algorithm [13]. On the other hand, the errors (in case (A)) are larger. But, it is not a serious problem because we do not aim at determining precise values. We only want to know whether each $|g_{i,j} - h_{i,j}|$ is relatively large or small. Note that the errors are small for $m = 50$ in case (B), whereas the errors are not small even for $m = 500$ in case (A). This observation suggests that good values are not inferred if parameters in the different levels are included.

Next we examined whether or not qualitative relations are correctly inferred, by applying the LP-based method to the case of $n = 10$ and $K = 2$ and the case of $n = 10$ and $K = 4$. Note that only the case of $n = 2$ was examined in Ref. [13]. In these cases, we did not try to infer precise values of parameters, but we tried to infer whether or not X_i is influenced by X_j examining the value of $|g_{i,j} - h_{i,j}|$. We say that the set of input nodes $\{X_{i_1}, \cdots, X_{i_K}\}$ to X_i is

correctly inferred if the LP-based method outputs the same set for X_i, where we say that X_j is an input node to X_i if $h_{i,j} \neq 0$ and $g_{i,j} \neq 0$ hold in the original S-system. We count the number of nodes for which the sets of input nodes are correctly inferred. The result is shown in the table below. In the table, the average ratios (%) of correctly inferred nodes over 10 randomly generated S-systems are shown, where the following values are used: $\Delta t = 0.01$, $\alpha_i = \beta_i = 3.0$, $0.5 < |g_{i,j}| < 3.0$, $0.5 < |h_{i,j}| < 3.0$. Even in the case of $m = 100 \times 20$, each inference can be done within 30 sec. (CPU time).

	$m = 25 \times 20$	$m = 50 \times 20$	$m = 100 \times 20$
$K = 2$	30%	86%	100%
$K = 4$	26%	69%	87%

From this table, it is seen that the sets of input nodes are correctly inferred for most nodes if m is large enough.

Finally, we examined the case of $n = 100$, $K = 4$, and $m = 1000 \times 20$. In this case, the LP-based method inferred the sets of input nodes correctly for 96 nodes using less than 5 hours (with 1 CPU), where $\Delta t = 0.005$. This result demonstrates the power of the LP-based method because we are tackling a very hard problem, inference of nonlinear systems with more than $100 \times 100 \times 2$ parameters.

References

[1] T. Akutsu, S. Miyano and S. Kuhara, Identification of genetic networks from a small number of gene expression patterns under the boolean network model, *Proc. Pacific Symp. on Biocomputing* **4**, 17–28, 1999.

[2] T. Akutsu, S. Miyano and S. Kuhara, Algorithms for inferring qualitative models of biological networks, to appear in Pacific Symp. on Biocomputing 2000.

[3] A. Arkin, P. Shen and J. Ross, A test case of correlation metric construction of a reaction pathway from measurements, *Science* **277**, 1275–1279, 1997.

[4] K.P. Bennett and O.L. Mangasarian, Robust linear programming discrimination of two linear separable sets, *Optimization Method and Software* **1**, 23–34, 1992.

[5] J.L. DeRisi, V.R. Lyer and P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science* **278**,680–686, 1997.

[6] T. Chen, H.L. He and G.M. Church, Modeling gene expression with differential equations, *Proc. Pacific Symp. Biocomputing* **4**, 29–40, 1999.

[7] P. D'haeseleer, X. Wen, S. Fuhrman and R. Somogyi, Linear modeling of mRNA expression levels during CNS development and injury, *Proc. Pacific Symp. Biocomputing***4**, 41–52, 1999.

[8] D.H. Irvine and M.A. Savageau, Efficient solution of nonlinear ordinary differential equations expressed in S-system canonical form, *SIAM J. Numer. Anal.* **27**, 704–735, 1990.

[9] S. Liang, S. Fuhrman and R. Somogyi, REVEAL, a general reverse engineering algorithm for inference of genetic network architectures, *Proc. Pacific Symp. on Biocomputing* **3**, 18–29, 1998.

[10] H.H. McAdams and L. Shapiro, Circuit simulation of genetic networks, *Science* **269**, 650–656, 1995.

[11] *Smart Optimizer User's Guide*, SAITECH Inc. (http://www.saitech-inc.com/math.htm), 1998.

[12] D. Thieffry and R. Thomas, Qualitative analysis of gene networks, *Pacific Symp. on Biocomputing* **3**, 77–88, 1998.

[13] D. Tominaga and M. Okamoto, Design of canonical model describing complex nonlinear dynamics, *Proc. IFAC Int. Conf.*, CAB7, 85–90, 1998.

[14] C-H. Yuh, H. Bolouri and E.H. Davidson, Genomic Cis-regulatory logic: experimental and computational analysis of a sea urchin gene, *Science* **279**, 1896–1902, 1998.