The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

An Algebraic-Numeric Algorithm for the Model Selection in Network Motifs in *Escherichia coli*

Masahiko Nakatsui[1] Hiroshi Yoshida[2] Katsuhisa Horimoto[1]

[1]Computational Biology Research Center (CBRC),
National Institute of Advanced Industrial Science and Technology (AIST),
Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan
[2]Faculty of Mathematics, Organization for the Promotion of Advanced Research,
Kyushu University, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581 Japan

Abstract Recently, we have proposed a novel algorithm to select a model that is the most consistent with the time series of observed data. In the algorithm, first, a system of differential equations that express the kinetics for a biological phenomenon and a sum of exponentials that are fitted to the observed data are transformed into the corresponding system of algebraic equations, by the Laplace transformation. Then, the two systems of algebraic equations are compared by an algebraic-numeric approach. One of the merits of our algorithm estimates the model's consistency with the observed data and the determined kinetic constants. Furthermore, our algorithm allows a kinetic model with cyclic relationships between variables that cannot be handled by the usual approaches. In this paper, we examined the performance of our proposed algorithm by using three kinds of highly significant network motifs in *Escherichia coli*; feed-forward loop, single input module, dense overlapping regulons, which are found by Shen-Orr, *et al*[14].

1 Introduction

One of the most remarkable features in biological network analysis is that the network structure itself is unknown, in contrast that the network model is almost always given in the engineering field. This situation indicates that the construction of network model is the first step to clarify the molecular mechanism underlying the biological phenomena. Indeed, the aim of the experimental studies is frequently the discovery of new molecules related with the biological phenomena, and the following aim is to reveal the relationships (interactions) between the newly found molecules. The knowledge about the molecules and their relationships by experimental studies have been reported in many literatures, and they have been compiled at the web (for example, [19]).

The approach for constructing a biological network model by systematic extraction of enormous knowledge from the literatures and the following superimposition of them is recognized as one of the most promising approaches [5]. Since each relation identified by experimental studies is regarded as strong evidence for the existence of edges in the network model, biological network models have been constructed for various biological phenomena. On the other hand, it is well-known that the relationships between the molecules in a living cell change dynamically, depending on the cellular environment.

Thus, the molecular relationships in the literature represent the responses to the different conditions in the experimental studies, and in the network model generated from the biological knowledge, the consistency of the model with the data observed by experimental studies must be considered carefully. Actually, several distinctive models of the relationship between molecules for a biological phenomenon can be obtained from the large amount of information in the literature [3, 6]. In these cases, a model that is consistent with the data observed under particular conditions should be selected from the candidate models.

The consistency of a model with the observed data is investigated intensively by statistical and algebraic approaches. In statistics, the issue of the consistency of a model with the observed data is also well known , as the test for causal hypotheses by using the observed data. The origin of the test for causal hypotheses is attributed to path analysis [17]. Unfortunately, the importance of this cornerstone research has been ignored for a long time, but the natural extension of the path analysis has been established as the well-known structural equation model (SEM) [9]. Indeed, the SEM has been utilized recently in various fields, in accordance with increased computer performance. However, the SEM without any latent variables, which is the natural form for applying the SEM to the biological networks, frequently faces difficulty in the numerical calculation of the maximum likelihood for the observed data. To overcome the difficulty of this calculation, the d-sep test [15] has been developed, based on the concept of d-separation in a directed acyclic graph [12]. Notice that the graph consistency with the data in the d-sep test can consider only the directed acyclic graph (DAG), without any cyclic relationships. In algebraic approach, there exists the identifiability problem in the compartmental models for tracer kinetics [1, 7, 6]. In the compartmental models, the unknown constants are estimated from tracer data in the accessible pools. The identifiability problem addresses the issue of whether the unknown constants can be determined uniquely or non-uniquely from the tracer data. This issue has usually been solved through the transformation of differential equations into algebraic equations, by the Laplace transformation. Although a systematic algorithm for the identifiability problem was proposed [4], its application is limited to the unrealistic context of an error–free model structure and noise–free tracer data. Thus, it still seems to be difficult to solve the identifiality problem for actually observed data, in spite of the mathematical studies.

Recently, we have proposed a new method for selecting models, by estimating the consistency of a kinetic model with the time series of observed data [18]. First, the kinetics for describing a biological phenomenon is expressed by a system of differential equations, assumed that the relationships between the variables are linear. Simultaneously, the time series of the data are numerically fitted as a sum of exponentials. Next, the differential equations with the kinetic constants and the sum of exponentials fitted to the observed data are both transformed into the corresponding system of algebraic equations, by the Laplace transformation. Finally, the two systems of algebraic equations are compared by an algebraic approach. Thus, our method estimates the model's consistency with the observed data and the determined kinetic constants. Indeed, we have successfully illustrated that our method can select the actual botanic models [10], in which a kinetic model with cyclic relationships between variables that cannot be handled by the usual approaches is included, with the corresponding data generated by the differential equations for the relationships. Although we have examined the performance of our method for selecting the

model with a cyclic loop in the previous paper, it remains to be investigated in terms of the model variation, especially the typical forms in the relationships between biological molecules.

Fortunately, the gene regulatory network identified by experimental studies is composed of the limited number of network motifs[14]; each motif has simple forms of 2-layer relationship between the transcription factor and its regulating genes. Even in a complex regulatory network, therefore, the entire network can be factorized into small subnetworks by combination of network motifs. In this paper, we address the issue on the selection of the network motifs in *Escherichia coli* which are proposed by Shen-Orr, *et al*. As the same way as in the previous paper, the data are generated by the differential equations for the relationships, and the consistency of the models with the generated data is calculated by our algebraic-numeric method.

2 Methods

2.1 Overview of Model Selection Algorithm

The procedure for model selection can be summarized as follows:

(i) We fit the observed data as a sum of exponentials in 2.2.
(ii) We perform the Laplace-transformation of both the system of differential equations for the models and the sum of exponentials for the observed data in 2.3.
(iii) By using the least squares method (abbreviated as *LSM*), we calculate the consistency of the model with the observed data.

In what follows, the details of our method will be shown.

2.2 Observed Data Fitting by Genetic Algorithm (GA)

In this paper, we need Laplace-transformed observed data, because we perform the model selection over the Laplace domain. Let $Mo_i(t)$ denote the observed data corresponding to $M_i(t)$ derived theoretically. By genetic-algorithm based numerical fitting, $Mo_i(t)$ is expressed in terms of a sum of exponentials as follows:

$$\beta_b + \sum_{j=1}^{n} \beta_j \exp(-\alpha_j t), \tag{2.1}$$

where n is the number of distinct exponentials determined by $M_i(t)$, and β_b is zero in the case of the non-existence of a constant term within $M_i(t)$. $Mo_i(t)$ thus fitted is changed into the Laplace-transformed data as follows:

$$\frac{\beta_b}{s} + \sum_{j=1}^{n} \frac{\beta_j}{s+\alpha_j}, \tag{2.2}$$

where L denotes the Laplace transformation. In this problem, each set of parameter values α_i, β_i and β_b to be estimated is evaluated using the following procedure: Suppose that $Mo_i(t)$ is the calculated time-course at time t of i and that $Ms_i(t)$ represents sampling data at time t of i. The sum of the square values of the relative error between $Mo_i(t)$ and $Ms_i(t)$ gives the total relative error E_i;

$$E_i = \sum_{t=1}^{T} \left(\frac{Ms_i(t) - Mo_i(t)}{Ms_i(t)} \right)^2, \tag{2.3}$$

where T is the total number of sampling points.

The computational task is to determine a set of parameter values α_i, β_i and β_b that minimizes the objective function E_i. Instead of the use of `NMinimize` command of `Mathematica 5.2` in the previous study [18], here, we use the well-known genetic algorithm (GA). We applied RCGAs with a combination of *unimodal normal distribution crossover* (UNDX)[11] and *minimal generation gap* (MGG)[13] as a nonlinear numerical optimization method for estimating constants.

2.3 Laplace-transformation of Model Formula

Suppose that the model formulae are described over the time domain as follows:

$$\frac{\mathrm{d}M_i(t)}{\mathrm{d}t} = F_i(\vec{M}, \vec{k}), \tag{2.4}$$

where $\vec{M} = \{M_1, M_2, \ldots, M_n\}$ and $\vec{k} = \{k_1, k_2, \ldots, k_m\}$. Function $F_i(\vec{M}, \vec{k})$ can be determined according to the graph describing the model, and $\vec{k}$ denotes the kinetic constants between the chemicals. We transform this system of differential equations into a system of algebraic equations over the Laplace domain, and solve the equations in $L[M_i(t)](s)$ $(i = 1, 2, \ldots, n)$.

2.4 Calculation of Consistency Measure and Model Selection

To evaluate the consistency of the model with the observed data, we define *consistency measure*. If the model is completely consistent with the observed data and the data lack noise and inaccuracies, then $L[M_i(t)](s) = L[Mo_i(t)](s)$ $(i = 1, 2, \ldots, n)$ holds. This fact has led us to the following definitions of consistency measure:

Let *comp* denote the set of polynomials obtained by matching the coefficients of $L[M(t)](s)$ and $L[Mo(t)](s)$ over the Laplace domain, in which every element is zero in the case of $L[M_i(t)](s) = L[Mo_i(t)](s)$ $(i = 1, 2, \ldots, n)$; that is, when Formula $L[M_i(t)](s) = L[Mo_i(t)](s)$ is an identity in s.

The consistency measure (in short, *CM*) of the model is defined as the smallest sum-square value of the elements in *comp* under the following constraint:

$$k_1 \geq 0, k_2 \geq 0, \ldots, k_m \geq 0. \tag{2.5}$$

In order to obtain the smallest value, we have utilized the least squares method using the following equations:

$$\frac{\partial}{\partial k_1} g(\vec{k}) = \frac{\partial}{\partial k_2} g(\vec{k}) = \cdots = \frac{\partial}{\partial k_m} g(\vec{k}) = 0, \tag{2.6}$$

where $g(\vec{k})$ is the sum-square value of the elements in *comp*.

Then, we survey all of the possible candidates of the minimum by calculating *all* of the real positive roots of the system of algebraic equation (2.6). Several method and

tools exist to calculate all real roots of algebraic equations adjoined by a zero-dimensional ideal.

The consistency measure can be calculated by the following recursive procedure [18]:

Let $MinimumValue(q(\vec{l}))$ denote the *minimum value* of function q with variables: $\vec{l} = \{l_1, l_2, \ldots, l_m\}$ by the following procedure:

1. If the cardinality of $\vec{l}$, namely m, is zero, then the *minimum value* is infinity.
2. Otherwise, let v_0 denote the minimum value of q under Constraint (2.5) via homotopy method. Furthermore, let v_i $(i = 1, 2, \ldots, m)$ denote the value calculated by $MinimumValue(q(\vec{l_i}))$, where $\vec{l_i}$ is the vector: $\{l_1, l_2, \ldots, l_{i-1}, 0, l_{i+1}, \ldots, l_m\}$.
3. The *minimum value* is the smallest value among $v_0, v_1, \ldots, v_m$.

Using the consistency measures, *CM*, we performed model selection. We, first, calculated the consistency measures of the candidate models with the observed data. Then, we listed the smallest consistency measures and the corresponding values of kinetic constants of each candidate model for the two consistent measures. Last, we select simply one candidate model showing the smallest values by the consistent measures.

2.5 Case Study

Shen-Orr *et al.* found three highly significant motifs in the transcriptional regulation network of *Escherichia coli*.[14] We modified these three kinds of network motif to four nodes (M_1, M_2, M_3, and M_4). Fig. 1 shows the three network motif analyzed in this paper. One is a motif of a chain graph with feed-forward loop, the other one is a motif of single input module, and last one is a motif of dense overlapping regulons.

3 Results

3.1 Formulation

According to the models in Fig. 1, the kinetics can be expressed by two systems of differential equations as follows:

Model (a)

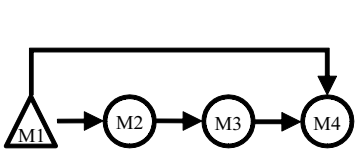
$$\begin{cases} \mathrm{d}/\mathrm{d}t\ M_1(t) &= -k_{12}\ M_1(t) - k_{14}\ M_1(t), \\ \mathrm{d}/\mathrm{d}t\ M_2(t) &= k_{12}\ M_1(t) - k_{23}\ M_2(t), \\ \mathrm{d}/\mathrm{d}t\ M_3(t) &= k_{23}\ M_2(t) - k_{34}\ M_3(t), \\ \mathrm{d}/\mathrm{d}t\ M_4(t) &= k_{14}\ M_1(t) + k_{34}\ M_3(t). \end{cases} \tag{3.1}$$

Model (b)

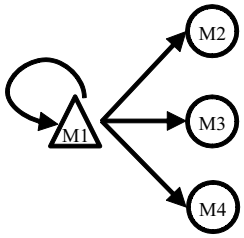
$$\begin{cases} \mathrm{d}/\mathrm{d}t\ M_1(t) &= k_{11}\ M_1(t) - k_{12}\ M_1(t) - k_{13}\ M_1(t) - k_{14}\ M_2(t), \\ \mathrm{d}/\mathrm{d}t\ M_2(t) &= k_{12}\ M_1(t), \\ \mathrm{d}/\mathrm{d}t\ M_3(t) &= k_{13}\ M_1(t), \\ \mathrm{d}/\mathrm{d}t\ M_4(t) &= k_{14}\ M_3(t). \end{cases} \tag{3.2}$$

Model (c)

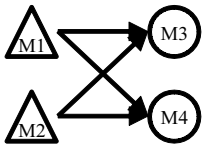
$$\begin{cases} \mathrm{d}/\mathrm{d}t\ M_1(t) &= -k_{13}\ M_1(t) - k_{14}\ M_1(t), \\ \mathrm{d}/\mathrm{d}t\ M_2(t) &= -k_{23}\ M_2(t) - k_{24}\ M_2(t), \\ \mathrm{d}/\mathrm{d}t\ M_3(t) &= k_{13}\ M_1(t) + k_{23}\ M_2(t), \\ \mathrm{d}/\mathrm{d}t\ M_4(t) &= k_{14}\ M_1(t) + k_{24}\ M_2(t). \end{cases} \tag{3.3}$$



(a) Feed-forward loop



(b) Single input module (SIM)



(c) Dense overlapping regulons (DOR)

Figure 1: Three kinds of network motif which were proposed by Shen-Orr *et al*[14]. The nodes shown as M1 to M4 are transcription factors. (a) Feed-forward loop: a transcription factor M1 regulates M2, and both jointly regulate M4. (b) Single input module: A single transcription factor M1 regulates a set of regulons shown as M2 to M4. (c) Dense overlapping regulons: a set of regulons M3 and M4 were each regulated by combination of a set of regulator M1 and M2.

where $M_1(t)$, $M_2(t)$, $M_3(t)$ and $M_4(t)$ represents the expression level of transcription factor M_1, M_2, M_3 and M_4 at time t, respectively. Then the above differential equations are transformed into the corresponding systems of algebraic equations by the Laplace transformation.

3.2 Data Generation for Simulation

In order to evaluate our proposed algorithm for the model selection, we prepared three sets of artificial simulated time-series data which were considered to be experimental observations. The initial conditions for each molecules and the kinetic constants are set as follows: $M_1(0) = 10, M_2(0) = 7, M_3(0) = 3$, $M_4(0) = 1$, $k_{12} = 135/928$, $k_{23} = 1/29$, $k_{34} = 1/8$, $k_{14} = 13/928$ for feed-forward loop, $k_{11} = 1/11$, $k_{12} = 1/17$, $k_{13} = 1/21$, $k_{14} = 1/23$ for single input module, $k_{13} = 1/23$, $k_{14} = 1/25$, $k_{23} = 1/21$, $k_{24} = 1/27$ for dense overlapping regulons, respectively. By using kinetic constants, we sampled the data for examining the models. Since the digits of the constants are different in the above sets of equations, we sampled the data at 100 points when t is in the range from 0 to 10, at 100 points when t is from 10 to 30, and at 70 points when t is from 30 to 100. Furthermore, 5% of fluctuation is added for each data as the noise of data. Results of fitting by using RCGAs, three sets of generated data are fitted well to three different models.

3.3 Model Selection by Algebraic-Numeric Approach

To examine the performance of our method for three kinds of network motif, we selected one motif among the two motifs with the data generated from one model. In actual use of the present method, first, the data are observed by the experiments, and then

Table 1: Consistency measure with kinetic constants. The given values of kinetic constants are $k_{12} = 135/928(\sim 0.145)$, $k_{23} = 1/29(\sim 0.0345)$, $k_{34} = 1/8(\sim 0.125)$, $k_{14} = 13/928(\sim 0.0140)$ for feed-forward loop, $k_{11} = 1/11(\sim 0.0909)$, $k_{12} = 1/17(\sim 0.0588)$, $k_{13} = 1/21(\sim 0.0476)$, $k_{14} = 1/23(\sim 0.0435)$ for single input module, $k_{13} = 1/23(\sim 0.0434)$, $k_{14} = 1/25(\sim 0.0400)$, $k_{23} = 1/21(\sim 0.0476)$, $k_{24} = 1/27(\sim 0.0370)$ for dense overlapping regulons. The symbol '0^*' indicates the exact value of zero.

data-generating model	examined model	smallest ssq	k_{11}	k_{12}	k_{13}	k_{14}	k_{23}	k_{34}
(a)	(a)	0.000907	-	0.149	-	0.00925	0.0345	0.125
(a)	(b)	0.537	0.0^*	0.00666	0.0313	0.0844	-	-
(a)	(c)	0.531	-	-	0.0472	0.0755	0.0^*	0.00285
(b)	(a)	1.52	-	0.0700	-	0.0^*	0.0371	0.0^*
(b)	(b)	0.00000177	0.0934	0.0593	0.0486	0.0446	-	-
(b)	(c)	0.0157	-	-	0.0295	0.0355	0.00557	0.0^*
(c)	(a)	0.000689	-	0.0172	-	0.0672	0.111	0.00471
(c)	(b)	0.357	0.0572	0.0^*	0.0728	0.0676	-	-
(c)	(c)	0.000184	-	-	0.0722	0.0108	0.0^*	0.846

a model is selected among some candidates of models. Thus, we examine the performance of the present method by solving which models one set of data is consistent with.

Table 1 shows the consistency of the models with the three motifs by consistency measure, together with the estimated values of kinetic constants. As seen in the table, in the all cases, the consistency estimation was succeeded. The kinetic constants in network motifs are well estimated when the data are generated from network motif (a) and (b). Unfortunately, our method does not operate well about estimation of the values of kinetic constants, when the data are generated from model (c).

In summary, our method can identify the network motif from observed time-course data sets. Furthermore, our method also can estimate the value of kinetic constants well excluding dense overlapping regulons.

4 Discussion

We examined the performance of our method for selecting the model with three kinds of network motifs which are proposed as highly significant motifs in the transcriptional regulation network of *Escherichia coli*. We have perfectly succeeded in selecting the correct network motif by using consistency measure. This result shows that, by factorizing large-scale network to simple network motif, we could apply our proposed algorithm to analyze organizationally complex system.

Moreover, we have partly succeeded in estimating the kinetic constants in the network motifs. Note that the present performance is examined by one set of data generated from the given values of kinetic constants. At any rate, we should further test the performance of our method for the generated data by different kinetic constants as well as for actually observed data. Furthermore, we should test the performance of our method for various structures of motifs.

References

[1] Audoly, S., D'Angiò, L., Saccomani, M. P. and Cobelli, C.: Global identifiability of linear compartmental models — A computer algebra algorithm, *IEEE Trans. Biomed. Eng.*, Vol. 45 (1998), 36–47.

[2] Bernshteĭn, D.N.: The number of roots of a system of equations, *Functional Anal. Appl.*, Vol. 9(3) (1975), 183–185.

[3] Bisits, A. M., Smith, R., Mesiano, S., Yeo, G., Kwek, K., MacIntyre, D. and Chan, E. C.: Inflammatory aetiology of human myometrial activation tested using directed graphs, *PLoS Comput. Biol.*, Vol. 1 (2005), 132–136.

[4] Buchberger, B.: An Algorithmic Criterion for the Solvability of a System of Algebraic Equations, in Buchberger, B. and Winkler, F. eds., *Gröbner Bases and Applications*, London Mathematical Society Lecture Notes Series 251, Cambridge University Press, 1998, 535–545.

[5] Calvano, S. E., Xiao, W., Richards, D. R., Felciano, R. M., Baker, H. V., Cho, R. J., Chen, R. O., Brownstein, B. H., Cobb, J. P., Tschoeke, S. K., Miller-Graziano, C., Moldawer, L. L., Mindrinos, M. N., Davis, R. W., Tompkins, R. G., Lowry, S. F. and Inflammation and Host Response to Injury Large Scale Collab. Res. Program, : A network-based analysis of systemic inflammation in humans, *Nature*, Vol. 437 (2005), 1032–1037.

[6] Cobelli, C., Foster, D. and Toffolo, G.: *Tracer Kinetics in Biomedical Research: From Data to Model*, Kluwer Academic/Plenum Publishers, 2000.

[7] Cobelli, C. and Toffolo, G.: *Theoretical aspects and practical strategies for the identification of unidentifiable compartmental systems*, Pergamon Press, Oxford, 1987, chapter 8, 85–91.

[8] Hanzon, B. and Jibetean, D.: Global minimization of a multivariate polynomial using matrix methods, *Journal of Global Optimization*, Vol. 27 (2003), 1–23.

[9] Joreskog, K. G.: A general method for analysis of covariance structures, *Biometrika*, Vol. 57 (1970), 239–251.

[10] Meziane, D. and Shipley, B.: Direct and Indirect Relationships Between Specific Leaf Area, Leaf Nitrogen and Leaf Gas Exchange. Effects of Irradiance and Nutrient Supply, *Annals of Botany*, Vol. 88 (2001), 915–927.

[11] Ono, I. and Kobayashi, S.: A real-coded genetic algorithm for function optimization using unimodal distribution crossover, *Proc 7th ICGA*, (1997) 249–253.

[12] Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, 1988.

[13] Satoh, H., Ono, I. and Kobayashi, S.: A new generation alternation model of genetic algorithm and its assessment, *J. of Japanese Society for Artificial Intelligence*, 15(2) (1997) 743-744.

[14] Shen-Orr, S. S., Milo, M., Mangan, S and Alon, U.: Network motifs in the transcriptional regulation network of *Escherichia coli*, *Nature Genetics*, Vol. 31 (2002), 64–68.

[15] Shipley, B.: A new inferential test for path models based on directed acyclic graphs, *Structural Equation Modeling*, Vol. 7 (2000), 206–218.

[16] Verschelde, J. and Haegemans, A.: Homotopies for solving polynomial systems within a bounded Domain, *Theor. Comp. Sci.*, Vol. 133(3) (1994), 165–185, (See also http://www.math.uic.edu/ jan/)