Vol.2012-BIO-30 No.6
2012/8/9

概日周期の正規形微分方程式モデルによる解析

富永 大介[1,a)]

概要：原核生物から哺乳類まで、細胞の活動には概日周期変動を示すものが多く知られている。その分子機構についての知識の発展は目覚ましいが、それでも未知の機構が多いと言わざるを得ず、その分子機構を包括する精密な微分方程式 (ODE) モデルは、特に哺乳類については現在は得られない。これに正規形 ODE モデルを適用することは有効な解析手段となり得るが、モデルフィッティングの非線形性がモデル同定を困難にしている。ここでは、非線形の ODE モデルを線形に変換し、遺伝子ネットワーク特有の前提条件を置くことで、モデル同定を線形最小二乗法で解くことができることを示す。また、これにより計算される S-system の指数パラメータが時間とともに変化する様子を示す。これはネットワーク構造のダイナミックな変化のモデルとなり得る。

キーワード：S-system, numerical optimization, parameter estimation, circadian clock

Linearization of the S-system formalism for the circadian gene regulatory network

Abstract: The circadian clock system is one of the most well known system in living cells through procaryotes to mammals. Although knowlegde about the system is growing rapidly, mechanisms of large parts of the system are still unclear. Canonical form ODE models are suitable approach for such systems. The s-system is one of the most established canonical ODE model, however, nonlinearity of the parameter optimization of the s-system should be solve to its application to the circadian system.
We show linearization of the s-system that make the least square method applicable to determination of exponential parameters of the s-system, and changes of these paraemters in time, that could be show dynamical changes of the network scheme.

Keywords: S-system, numerical optimization, parameter estimation, circadian clock

1. はじめに

遺伝子の転写制御はこれまで、セントラルドグマにしたがって、遺伝子の発現したタンパク質が、転写因子として他の遺伝子の発現を制御するという機構が考えられてきた。これに加えてマイクロ RNA による制御が大きな影響を持っていることが分かってから、その機能解明が大きな研究の流れとなっているが、転写因子による機構は支配的機構の一つであることは変わらないと考えられている。

概日周期変動はこの転写因子による機構の解明が進んでいるネットワークの代表的な例である [1]。糖尿病を始めとする疾病との関連や健康維持への関心から、概日周期の維持機構はこれまで注目を集めてきた。しかし、概日周期変動に関わる遺伝子の転写因子による発現制御のしくみがよく知られてきたとはいえ、転写因子以外の制御機構がどの程度の影響を持つのかはよくわかっておらず、またそもそも転写因子による制御の分子機構が完全に明らかにされているとも言えない。

ネットワークの動的挙動を解析しようとする際、常微分方程式系 (Ordinary Differential Equation system, ODE) をもちいて、その動きを表現するモデルを得ることは、安定判別やパラメータの感度解析、定量的なシミュレーションが可能なことなどから、有効な手法の一つであると言える。酵素反応のネットワークであれば、ミカエリス・メンテン則を始めとする様々なモデルが各反応形式で定式化されており、それらを組合せることでネットワーク・ダイナ

[1] 独立行政法人産業技術総合研究所 (AIST) 生命情報工学研究センター (CBRC), 2-4-7 Aomi, Koto, Tokyo 135–0064, Japan
[a)] tominaga@cbrc.jp

ミクスを表現する ODE モデルを得ることができる。分子機構の詳細が不明なネットワークにおいてはそのアプローチは不可能であるが、詳細な反応形式に拠らない、一般形あるいは正規形で定義される微分方程式モデルを使うことで ODE モデルを得ることができる。そういった正規形微分方程式の一つに、S-system がある [2]。

S-system は主に化学反応系を想定した生体内ネットワークを対象として、質量作用則を元に考案されたモデルであり [3]、代謝系を含む生化学反応系などの応用例が多い [4], [5] が、1990 年代後半以降、DNA マイクロアレイの普及とともに、遺伝子制御ネットワークのモデリングの試みに用いられてきた [6], [7]。しかし、S-system のパラメータを決定してモデルを同定するためには、実際に S-system を数値的に解いた結果が観測された時系列データとどの程度一致するかをコスト関数として定義し、それを乱数を用いた発見的探索手法により最小化することで最適なパラメータを得る、という方法が主なアプローチであった。S-system モデルを定義するためのパラメータ数は、ネットワークを構成する要素の数を n とするとき $2n^2 + 2n$ であり、遺伝子を 3 個だけ含む非常に小規模なネットワークでも 24 個のパラメータ値を決定しなければならない。多変量自己回帰モデル (multiVariate Auto-Regression model, VAR) などでは、変数の個数と同程度のデータ点があればモデル同定が不可能ではない [8] ことなどと比較すると、S-system は非常に多くのデータを要求するモデルであると言える。要求データ量が多いと言うことはそれだけ多くの情報を含んだモデルということであり、多様な挙動を正確に表現することができるモデルであるが、遺伝子発現量の場合にはその精度を持つデータを得ることは困難である。さらに、決定すべきパラメータ数が多いことは探索空間の次元が高いと言うことでもあり、いわゆる「次元の呪い」を直接に被る最適化問題でもある。

以上に加えて S-system モデルを同定しようとする際には、実際にはパラメータ最適化が非線形性の強い問題であることが困難さをもたらす。パラメータ数が多いことに対しては、問題を分割してより容易な小規模の問題の集合に変換する方法 [9], [10], [11] が提案されているが、問題の非線形性を解決する方法は現在のところ、非常に少ない。元々の問題が線形最適化であれば、最小二乗法を始めとする多くの方法が適用でき、パラメータの値も容易に決定することができる。S-system は、定常状態 (導関数値が 0) においては対数変換によって線形方程式になり、そこから局所安定性やパラメータ感度の計算法が導出されている [12] が、このアプローチでパラメータの値を同定することはできない。しかし、転写因子による遺伝子制御ネットワークにおいては、mRNA の合成過程は他の遺伝子の産物である転写因子によって制御されるが、mRNA の分解過程は自身の量に比例した速度、つまり線形モデルで表すと前提を置く例が散見される。ここでは、この前提を置くことで S-system が線形の方程式に変換できることを示す。また、変換後も容易にはパラメータ値を決定することができないが、場合によってはそれが可能であることを示す。

2. 方法

2.1 S-system の線形方程式系への変換

S-system は非線形の常微分方程式系であり、以下の式で表される。

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

右辺第一項が従属変数 $X_i(t)$ の増加に関わる作用、右辺第二項が減少に関わる作用を表す。S-system におけるパラメータ推定問題は、式 (1) における $X_i(t)$ が観測値として与えられたときに、式 (1) を積分して得られる $X_i(t)$ の時系列が与えられた観測値と一致するようなパラメータ値 α_i、β_i、g_{ij}、h_{ij} を求める、という問題である。

S-system は生化学反応系を想定したモデルであり、従属変数は代謝物の量やタンパク質の活性などである。したがって従属変数の値は非負である。また α_i および β_i も非負の範囲だけを考えれば良い。一方指数パラメータである g_{ij} および h_{ij} は、0 を含む正負の値をとり得る。数式で表現すると以下のようになる。

$$\begin{aligned} X_i(t) &> 0 \\ \alpha_i &> 0 \\ \beta_i &> 0 \end{aligned} \tag{2}$$

なお、変数およびパラメータはすべて実数である。

ここで遺伝子制御ネットワークを想定し、減少に関わる作用は従属変数 $X_i(t)$ 自身の量の線形モデルで表されるとすると、以下のようになる。

$$\frac{dX_i(t)}{dt} = \alpha_i \prod_{j=1}^{n} X_j(t)^{g_{ij}} - \beta_i X_i(t) \tag{3}$$

左辺を微分係数を $\dot{X}_i(t)$ と書き直して、指数パラメータ g_{ij} の有無で左右の辺に分かれるように移項すると

$$\alpha_i \prod_{j=1}^{n} X_j(t)^{g_{ij}} = \dot{X}_i(t) + \beta_i X_i(t) \tag{4}$$

となり、この両辺の対数をとると

$$\log(\alpha_i) + \sum_{j=1}^{n} g_{ij} \log(X_j(t)) = \log\left(\dot{X}_i(t) + \beta_i X_i(t)\right) \tag{5}$$

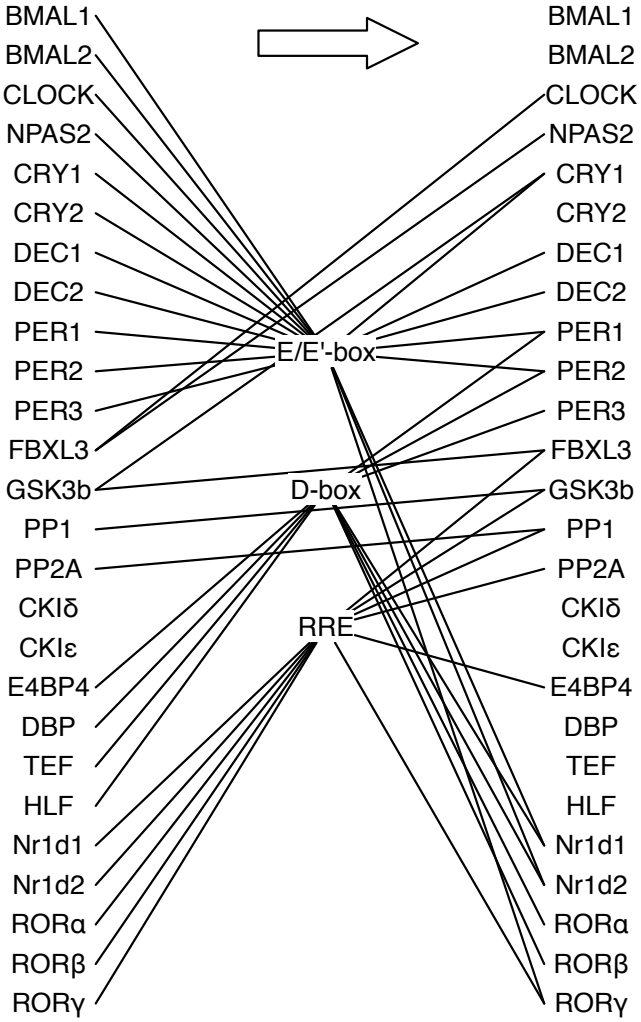
となる。これは、g_{ij} と α_i だけが未知変数であると考えると、連立一次方程式系である。この未知変数の値は、以下の条件がすべて満たされたときに、最小二乗法により求めることができる。

条件 1: $\dot{X}_i(t)$ の値が得られる

Vol.2012-BIO-30 No.6
2012/8/9

図 1 遺伝子および制御因子の制御関係。

Fig. 1 Regulatory relationships between genes and regulation elements.



条件 2: $X_i(t)$ および $\dot{X}_i(t)$ のデータ点数が n+1 よりも大きく、係数行列のランクが n+1 以上である

条件 3: $\dot{X}_i(t) + \beta_i X_i(t) > 0$ である

条件 4: β_i の値が既知である

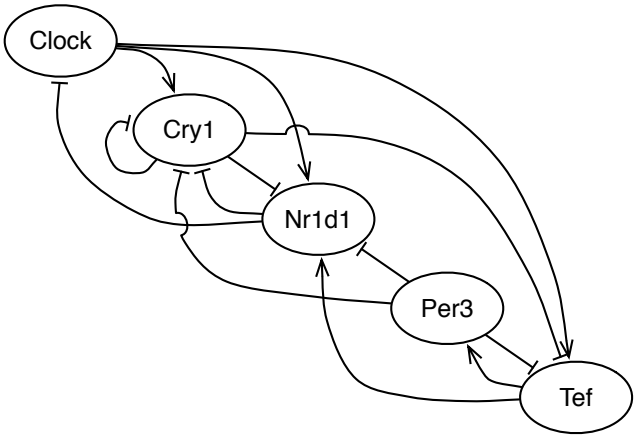
条件 1 は多くの場合観測データとしては成立しないが、$X_i(t)$ が与えられれば数値微分により解決できる。条件 2 は観測条件に依存し、遺伝子発現時系列の観測例ではほとんどの場合、これを満たさないが、モデリングの対象とする遺伝子を絞り込み、かつフィッティングや補完を行うことで解決できる。条件 3 は発現量がある程度大きな遺伝子であれば成立している。しかし条件 4 は一般には成立しない。そこで、別途に値を推測する必要がある。

2.2 パラメータ β_i の推定

遺伝子制御ネットワークを想定した S-system モデルにおいて、増加作用を表す項においてただ一つの従属変数 $X_j(t)$ からのみ影響を受ける従属変数 $X_i(t)$ を考えると、それは以下のように表される。

図 2 想定したネットワークの様式。この構造に対してモデルのあてはめを行う。

Fig. 2 Network scheme that an S-system model fits to.



$$\dot{X}_i(t) = \alpha_i X_j(t)^{g_{ij}} - \beta_i X_i(t) \tag{6}$$

これは容易に g_{ij} について解くことができ、前節における条件 2 が成立していれば、以下のように表される。

$$g_{ij} = \frac{\log(\beta_i + \dot{X}_i(t)) - \log(\alpha_i)}{\log(X_j(t))} \tag{7}$$

ここで g_{ij} は定義上定数である。したがって S-system でのモデリングが適切であるような系においては、データ $\dot{X}_i(t)$ および $X_j(t)$ が与えられたとき、α_i と β_i が適切な値であれば g_{ij} は定数になるべきである。現実にはデータの誤差やモデルの表現能力などの影響で、g_{ij} が完全に一定値になるような α_i と β_i はまず求められないが、g_{ij} の分散を最小化するような値を求めることは、様々な発見的探索手法で可能である。

遺伝子制御ネットワークを想定した S-system モデルにおいては、従属変数の減少を表す項は mRNA の分解を表す項であり、その分子機構に関する情報が少ない状況においては、すべての mRNA でその項は共通であると見なせるかどうかを検討すべきである。そこで、複数の β_i を g_{ij} の分散最小化によって求め、その値が近いものであれば、モデリング対象としている系のすべての mRNA についてその値を共通と見なすことができる、と考える。その値を用いることで、前節における指数パラメータの最小二乗法による推定が可能になる。

2.3 指数パラメータの時間依存性

一般的は、データ数が多ければ得られる数値の信頼性も高くなると考えられるが、データの振る舞いによっては、バラつきの少ない範囲に限定した方がいい場合も有り得る。その範囲の限定を時間軸上でのウィンドウの設定を捉えると、ウィンドウ内で指数パラメータを求めることができ、ウィンドウをスライドさせていくことで、指数パラメータの時間変化を表すことができる。そこで、ウィンドウを設

定せずデータ全体を使った場合に求められるパラメータ値が信頼性が高くないと考えられる場合には、ウィンドウイングを行う。

3. 結果

3.1 観測データ

表 1 全データを対象として求めた S-system の指数パラメータ。表中の 0 は最適化対象ではなく、0 に固定されていることを表す。

Table 1 Exponential coefficients values which are determined based on whole data. 0s in the table mean those parameters are fixed at zero and are not optimized.

	CLOCK	CRY1	Nr1d1	PER3	TEF
α_i	1.801	1.056	0.4351	0.717534	1.03424
g_{1i}	0	7.564e-2	1.111e-6	0	5.85ae-6
g_{2i}	0	-5.342e-5	-2	0	-2.199e-5
g_{3i}	9.369e-3	-2.806e-5	0	0	0
g_{4i}	0	-7.068e-5	-7.381e-3	0	-1.977e-5
g_{5i}	0	0	2.674e-6	5.633e-5	0

米国立衛生研究所 (National Institute of Health) 傘下のバイオテクノロジー情報センター (National Center for Biotechnology Information) において公開されている遺伝子発現データの公開データベースである GEO (Gene Expression Omnibus)[13] に GDS404 として登録されている、マウスの DNA マイクロアレイデータを用いた [14]。マウスは健常であり、12 時間周期で明暗を繰り返す環境に置かれた後に常暗に置かれ、4 時間おきに大静脈血管が採取された。この組織における遺伝子発現量が測定された。測定には使われたのは、米 Affymetrix 社の MG-U74A チップである。チップ (DNA マイクロアレイ) の各プローブには、同社によるアノテーションが GO term を用いて付けられている。

この GDS404 データセットには、常暗に置かれてから 4 時間後を最初として、48 時間までの 13 点のサンプルが含まれているが、最初のサンプリング時刻におけるサンプルが重複している (他の点は $n = 1$ だがこの点だけ $n = 2$ ということである) 。そのためサンプル点としてはデータセットの先頭に格納されているサンプルは解析対象から除いた。

このチップでは複数のプローブで観察するようになっている遺伝子が複数あり、概日周期に関わるとされる遺伝子もそれに含まれている。そういった遺伝子については、データセット中でもっとも先頭に近いプローブ一つを選んで解析対象とした。

3.2 解析対象とする遺伝子

ここではマウスの概日周期を解析対象とした。ここで参考とした概日周期変動に関わる遺伝子制御ネットワーク [1] には、26 の遺伝子と 3 つの転写制御モチーフが含まれており、各遺伝子は転写制御モチーフを通じて他の遺伝子の発現を制御している (**図 1**) 。この図において、左右の列はそれぞれ概日周期変動に関わると見られる遺伝子であり、同一の列である。中央に 3 個の制御因子が並んでいる。各制御因子をつかって制御する遺伝子について、左の列から中央に向かって対応する遺伝子と制御因子を線で結んでいる。各制御因子が転写を制御する遺伝子について、中央から右の列に向かって同様に対応する遺伝子と制御因子を線で結んでいる。これによると、各遺伝子は一つあるいは複数のモチーフによる制御を受けている。遺伝子発現を制御しているモチーフの組合せは 5 種類 (E/E'box、D-box、RRE それぞれ単独、および E/E'-box と D-box の組合せ、および E/E'-box と RRE の組合せ) であり、したがって同じ制御を受けている遺伝子が 5 群に同定される。

また、12 点からなる各遺伝子のデータについて、24 時間周期の成分が有意に大きいかどうかをフーリエ変換と情報量基準を使って判定 [15] し、24 時間周期の成分が有意には大きくないとされた遺伝子については解析対象から除いた。またデータに欠損値を含むものも除いた。そして各群に含まれる複数の遺伝子の発現時系列に三角関数 $A\sin(\frac{2\pi}{24}(t+B))+C$ を当てはめ、もっとも当てはめ誤差の小さい遺伝子を各群から一つ選んで、CLOCK、CRY1、Nr1d1、PER3、TEF の 5 個の遺伝子からなるネットワークを想定 (**図 2**) し、解析対象とした (**図 3**) 。このネットワークは図 1 に示した図のサブセットとなる。なおモデルを当てはめる時系列データは、当てはめた三角関数の値とし、時点数を 45 点とした。

3.3 パラメータ β_i の推定

解析対象とした 5 遺伝子のうち、他の一つの遺伝子からのみ制御を受けているとされるのは CLOCK と PER3 の二つであり、これらを対象として式 (7) における g_{ij} の分散を最小化する α_i および β_i を、数値最適化により求めた。α_i および β_i の初期値を $[0, 10]$ の範囲の一様乱数で生成し、式 (7) における g_{ij} の分散をコスト関数として、ネルダーとミードによるシンプレックス法で最小化することを、1000 回繰り返した (**図 4**)。その結果、CLOCK については α_i の平均値と標準偏差がそれぞれ 6.10 と 3.03、β_i では 1.01 と 0.000112 となった。また PER3 については α_i では 6.17 と 3.05、β_i では 0.888 と 0.000111 となった。計算にはいずれも GNU R ver. 2.15.1 を用いた。α_i と β_i の絶対値はおおよそ同程度の桁の値でありながら、標準偏差は β_i の方が 4 桁も小さく、非常に高い精度で求められていることが分かる。また CLOCK と PER3 でそれぞれ 1.01 と 0.888 であり、似通った値となっている。ここでは g_{ij} の同定を可能とするために、これらは同一の値で 1.0

図 3 遺伝子発現時系列の観測値とフィッティングデータ。
Fig. 3 Observed gene expression time series and fitted curves.

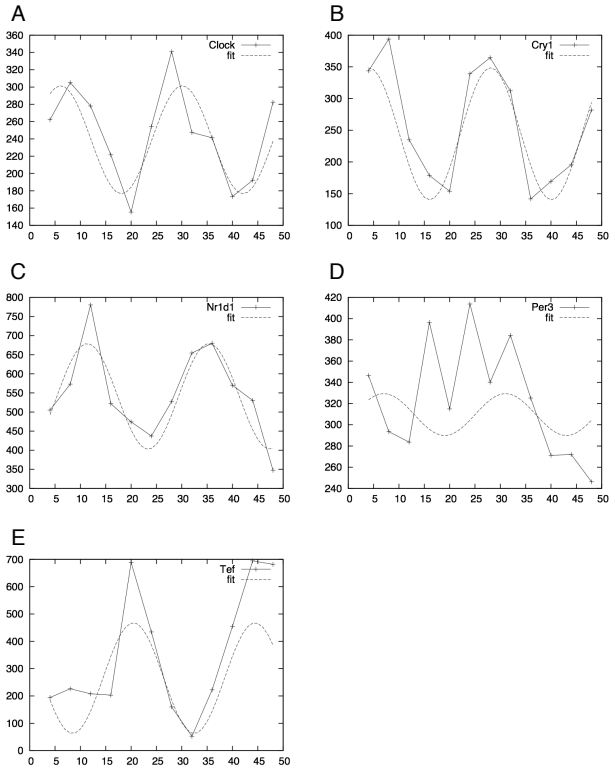
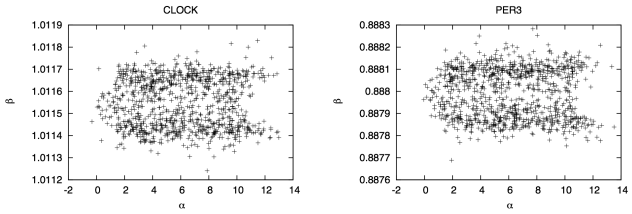


図 4 g_{ij} の分散を小さくするような α_i と β_i の分布。
Fig. 4 Distribution of α_i and β_i which minimize the variance of g_{ij}.



であると見なすこととした。

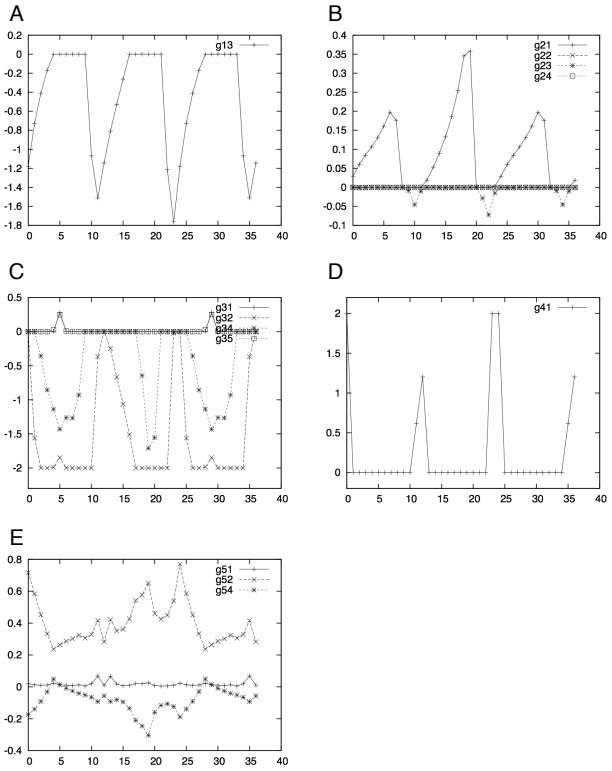
3.4 指数パラメータの推定

β_i の値を 1.0 であると決めると、式 (5) から最小二乗法によって g_{ij} の値を求めることができる。各遺伝子 45 点のデータを使って、Mathematica 7.0.1.0 の FindFit 関数で g_{ij} の値を求めた (**表 1**) 。この際、g_{ij} の値には制約条件を設け、図 2 でネガティブな制御とされている関係を表すパラメータは $[-2, 0]$、ポジティブな制御のパラメータは $[0, 2]$ の範囲内の値となるようにに限定した。

制約条件に関わらず、求められた指数パラメータはいずれも非常に小さな値となっており、これは遺伝子間の影響の大きさが非常に小さいと解釈され、ネットワークモデルとして意味を成さない。

そこで、短い時間範囲のデータをとりだすウィンドウを設定し、ウィンドウを観測開始時刻から終点へ向けてずらしていくことで、各ウィンドウにおいて求められる指数パラメータがどうなるかを同じ制約条件のもとで調べた (**図 5**) 。ウィンドウサイズを 8 としたところ、指数パラメータはいずれも周期的な挙動を示した。これは概日周期に合わせてネットワークモデル自体が変化していることを表している。

図 5 各遺伝子についての指数パラメータの時間変化。
Fig. 5 Changes in time of exponential parameters for each gene.



4. 考察

S-system は一般化質量作用則 (Generailzed Mass Action model, GMA) モデルを簡略化したものであるが、パラメータの数はネットワークの要素数を n をとすると $2n^2 + 2$ である。代謝系やシグナルパスウェイ、遺伝子ネットワークなどを想定するとパラメータ数がすなわち探索空間の次元数であり、直接にモデルを時系列データにフィッティングすることは非線形最適化問題であることから、そのパラメータ値の決定は決して容易とは言えない。

本論文ではこの問題に対し、遺伝子制御ネットワーク特有の事情を前提とすることで、S-system の最適化を線形システムのフィッティング問題に変換し、最小二乗法の適用を可能とした。パラメータ数が多いこと、および問題の非線形性による困難性はこれにより排除される。

ウィンドウ内で求められた指数パラメータは、たびたび制約条件の境界値となっている。特にネットワークの形

(既知の制御様式) から指数パラメータの符号を固定して最小二乗法の制約条件としたが、図 5 の A および C では、その制約により挙動を制限されたかのように見られる。これは制約条件を緩めれば値が変わるものと思われ、特に既知の制御様式でポジティブ、あるいはネガティブな制御であると知られていることでも、時間変化を追ってモデルを解析すると、正負が反転している可能性があることを示唆している。しかし S-system はいわゆるべき乗則の形をとったモデルであるため、指数パラメータの値が大きくなると、微分係数の値を計算するときにオーバーフローが起きやすくなると言う、数値計算上の欠点がある。そのため、本論文の方法による指数パラメータの決定とは別に、微分方程式として数値的に解けるような指数パラメータの範囲を、制約条件として求めるような基準が必要である。

実際に、本手法により求められたパラメータ値には、そのままでは S-system パラメータとして成立していない、つまり微分方程式が数値的に解けないような値が含まれている。これまでは S-system のパラメータはすべて時間に関しては不変であるとして解析が行われてきたが、これが可変である可能性を本論文では示した。そのための数値解法の実装をはじめ、動力学的特性の解析法を開発していく必要がある。

参考文献

[1] Ukai, H. and Ueda, HR.: Systems Biology of Mammalian Circadian Clocks, *Annual Review of Physiology*, **72**:579-603 (2010).

[2] Voit, EO.: *Canonical Nonlinear Modeling: S-System Approach to Understanding Complexity*, Van Nostrand Reinhold, NY, USA (1991).

[3] Savageau MA.: *Biochemical systems analysis: a study of function and design in molecular biology*, Addison-Wesley, Reading, MA, USA (1976).

[4] Gonzalez, OR., *et. al.*: Parameter estimation using Simulated Annealing for S-system models of biochemical networks, *Bioinformatics*, **23**:480-486 (2007).

[5] Shikata N., Maki Y., Nakatsui M., Mori M., Noguchi Y., Yoshida S., Takahashi M., Kondo N., Okamoto M.: Determining important regulatory relations of amino acids from dynamic network analysis of plasma amino acids, *Amino Acids*, **38**:179-187 (2010).

[6] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K. and Tomita, M.: Dynamic modeling of genetic networks using genetic algorithm and S-system, *Bioinformatics*, **19**:643-650 (2003).

[7] Nakatsui M., Ueda T., Maki Y., Ono I., Okamoto M (2008) Method for inferring and extracting reliable genetic interactions from time-series profile of gene expression, *Mathematical Biosciences*, **215**:105-114 (2008). doi: 10.1016/j.mbs.2008.06.007

[8] 北川源四郎: 多変量時系列モデル, in 時系列解析の方法 (尾崎, 北川編), pp. 107-117, 朝倉書店, 東京 (1998).

[9] Maki Y., *et. al.*: Inference of Genetic Network Using the Expression Profile Time Course Data of Mouse P19 Cells, *Genome Informatics*,**13**:446-458 (2002). (2001).

[10] Kumura, S., *el. al.*: Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm, *Bioinformatics*, **21**:1154-1163 (2005).

[11] Ho, SY., *et. al.*: Al intelligent Two-Stage Evolutionary Algorithm for Dynamic Pathway Identification from Gene Expression Profiles, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **4**(4):648-660 (2007).

[12] 岡本正宏: 非線形数理モデルのたて方と解析手順 in バイオプロセスシステム工学 (清水編), pp. 351-360, アイピーシー, 東京 (1994).

[13] Barrett, T., *et. al.*: NCBI GEO: archive for functional genomics data sets-10 years on, *Nucleic Acids Research*, **39 (suppl 1)**:D1005-D1010 (2010).

[14] Rudic, RD. *et. al.*: Bioinformatic Analysis of Circadian Gene Oscillation in Mouse Aorta, *Circulation*, **112**:2716-2724 (2005).

[15] Tominaga, D.: Periodicity detection method for small-sample time series datasets, *Bioinformatics and Biology Insights*, **4**:127-136 (2010).