Vol.2015-BIO-41 No.8
2015/3/20

遺伝子ネットワークのreduced S-systemモデルの効率的同定法の提案

木村 周平[1,a)]　佐藤 昌直[2]　岡田 (畠山) 眞里子[3]

概要：遺伝子ネットワーク同定とは，観察された遺伝子発現データに合うような遺伝子ネットワークのモデルのパラメータを推定する問題のことである．遺伝子ネットワークを記述するために多くの数理モデルが提案されているが，中でも S-system モデルは最も使用されてモデルの一つである．S-system モデルは生化学反応を近似したモデルであり，遺伝子ネットワークの記述に適していると考えられている．そのため S-system モデルを用いた遺伝子ネットワーク同定法が，これまでに数多く提案されている．しかしながら S-system モデルに含まれるパラメータ数は，他の頻繁に使用されているモデルのパラメータ数よりも多い．従って遺伝子ネットワークの S-system モデルを同定するためには，より多くの遺伝子発現データを与える必要があると考えられる．遺伝子ネットワーク同定に必要な遺伝子発現データの量を減らすために，本研究では S-system モデルの一部のパラメータを 0 に固定することでモデルを単純化する．本研究では単純化した S-system モデルを，reduced S-system モデルと呼ぶ．次に遺伝子ネットワークの reduced S-system モデルを効率的に同定するための新たな方法を提案する．最後に数値実験を通して，提案手法の有効性を示す．

1. はじめに

DNA マイクロアレイや RNA-seq といった生化学実験技術の進歩によって，細胞全体レベルでの遺伝子発現パターンの計測が可能になってきている．このような大量のデータから有用な情報を抽出するための方法の一つとして，遺伝子ネットワークの同定に注目が集まっている．遺伝子ネットワークとは遺伝子間の制御関係を表したネットワークであり，遺伝子，タンパク質，代謝物質などからなる現実の細胞内の生化学反応ネットワークを抽象化したものである．そのため遺伝子ネットワーク同定は，細胞システムを理解するための有効な手段の一つと考えられている．

遺伝子ネットワーク同定の目的は測定された遺伝子発現時系列データから，そのデータを説明できる数理モデルを構築することである．遺伝子ネットワークを記述するために，多くの数理モデルが提案されている．なかでも S-system モデルは生化学反応を近似したモデルであり，遺伝子ネットワークを記述するのに適していると考えられている [13]．そのため S-system モデルを用いた多くの遺伝子ネットワーク同定法が，これまでに提案されている [2], [4], [5]．S-system モデルを使用して N 遺伝子からなるネットワークの同定を行う場合，$2N(N+1)$ 個のパラメータを推定する必要がある．このパラメータ数は，線形モデル [15] や Vohradský モデル [12] といった他の頻繁に利用されるモデルの約 2 倍である．従って遺伝子ネットワークの S-system モデルを同定するためには，他のモデルに基づく遺伝子ネットワーク同定法よりも多くの遺伝子発現データを与える必要があると考えられる．ところが多くの遺伝子発現データを計測することは現在のところコスト面などから困難である．そのためパラメータ数の多さは S-system モデルの欠点となっている．

本研究では上記の S-system モデルの欠点を克服するために，一部のモデルパラメータの値を 0 に固定する．このようにして得られたモデルを，本研究では reduced S-system モデルと呼ぶ．遺伝子ネットワークの reduced S-system モデル同定には，従来の S-system モデルに基づく遺伝子ネットワーク同定法を利用することが可能であると考えられる．しかしこれらの手法は一般に，対象のネットワークに含まれる遺伝子数に比例する次元数の非線形関数最適化問題を解くことを必要とする．従って問題の高次元性のために，これらの手法を用いて大規模遺伝子ネットワークを

[1] 鳥取大学
Tottori University, 4-101, Koyama-minami, Tottori, Japan
[2] 基礎生物学研究所
National Institute for Basic Biology, 5-1, Higashiyama, Myodaiji, Okazaki, Japan
[3] 理化学研究所
RIKEN, 1-7-22, Suehiro, Tsurumi, Yokohama, Japan
[a)] kimura@ike.tottori-u.ac.jp

解析することは困難である．

Reduced S-system モデルのパラメータ推定問題における高次元性を解決するために，本研究ではモデルの特徴を利用した新たなパラメータ推定法を提案する [7]．具体的には N 遺伝子からなるネットワークの reduced S-system モデル同定問題を，N 個の 2 次元関数最適化問題として定式化する．定式化した 2 次元関数最適化問題を解くためには，進化的アルゴリズム REX^{star}/JGG [8] を利用する．最後に数値実験を通して，提案手法の有用性を確認する．

2. Reduced S-system モデル

S-system モデル [13] は以下の連立微分方程式で表されるモデルである．

$$\frac{dX_n}{dt} = \alpha_n \prod_{m=1}^{N} X_m^{g_{n,m}} - \beta_n \prod_{m=1}^{N} X_m^{h_{n,m}}, \quad (n = 1, 2, \cdots, N). \tag{1}$$

ただし X_n は n 番目の状態変数，N はネットワークの要素数，α_n (> 0), β_n (> 0), $g_{n,m}$, $h_{n,m}$ $(n, m = 1, 2, \cdots, N)$ は定数パラメータである．遺伝子ネットワーク同定では，X_n は遺伝子 n の発現量，N は解析しようとしているネットワークに含まれる遺伝子の総数に相当する．S-system モデルを用いた遺伝子ネットワーク同定では，測定した遺伝子発現データに合うように $2N(N+1)$ 個の定数パラメータ α_n, β_n, $g_{n,m}$, $h_{n,m}$ $(n, m = 1, 2, \cdots, N)$ の値を調整することが目的となる．

遺伝子ネットワーク同定ではしばしば，遺伝子間の制御の有無や，その制御の種類 (活性と抑制) を知ることが重要となる．S-system モデルではこれらの関係性を，パラメータ $g_{n,m}$ と $h_{n,m}$ のどちらを用いても表現することが可能である．そのため上記の関係性を知るという目的に対しては，S-system モデルは冗長であると言うことが出来る．S-system モデルの冗長性を無くすために，本研究ではパラメータ $h_{n,m}$ $(n \neq m)$ の値を 0 に固定する．このようにして得られた新たなモデルを，本研究では reduced S-system モデルと呼ぶ．従って reduced S-system モデルは以下で表される．

$$\frac{dX_n}{dt} = \alpha_n \prod_{m=1}^{N} X_m^{g_{n,m}} - \beta_n X_n^{h_{n,n}}, \quad (n = 1, 2, \cdots, N). \tag{2}$$

これにより，遺伝子ネットワーク同定において推定しなければならないパラメータ数は，$2N(N+1)$ 個から $N(N+3)$ 個に減る．

なお既に幾つかの研究において実際に S-system モデルを簡略化したモデルが用いられており，そのようなモデルを用いても妥当な遺伝子ネットワークの同定が可能であることが実験的に示されている [1], [10]．しかしながらこれらの手法はパラメータ推定に，S-system モデルのパラメータ推定法を利用している．これに対して本研究では，reduced S-system モデルの特徴を利用した効率的なパラメータ推定法を新たに提案する．

3. モデルパラメータ推定

提案手法は N 遺伝子からなるネットワークの reduced S-system モデル同定問題を，N 個の部分問題に分割する．n 番目の部分問題を解くことで，パラメータ α_n, β_n, $\mathbf{g}_n = (g_{n,1}, g_{n,2}, \cdots, g_{n,N})$, $h_{n,n}$ が推定される．以下では n 番目の部分問題について説明する．

3.1 問題定義

n 番目の部分問題では以下の連立方程式を解くことで，パラメータ α_n, β_n, $\mathbf{g}_n$, $h_{n,n}$ の推定を行う．

$$\begin{aligned}
\tfrac{dX_n}{dt}\big|_{t_1} &= \alpha_n \textstyle\prod_{m=1}^{N} \left(X_m|_{t_1}\right)^{g_{n,m}} - \beta_n \left(X_n|_{t_1}\right)^{h_{n,n}}, \\
\tfrac{dX_n}{dt}\big|_{t_2} &= \alpha_n \textstyle\prod_{m=1}^{N} \left(X_m|_{t_2}\right)^{g_{n,m}} - \beta_n \left(X_n|_{t_2}\right)^{h_{n,n}}, \\
&\vdots \\
\tfrac{dX_n}{dt}\big|_{t_K} &= \alpha_n \textstyle\prod_{m=1}^{N} \left(X_m|_{t_K}\right)^{g_{n,m}} - \beta_n \left(X_n|_{t_K}\right)^{h_{n,n}}.
\end{aligned} \tag{3}$$

ただし $X_m|_{t_k}$ は時刻 t_k における遺伝子 m の発現量，$\frac{dX_n}{dt}\big|_{t_k}$ は時刻 t_k における遺伝子 n の発現量の変化量 (転写速度)，K は測定時点数である．なお $X_m|_{t_k}$ は RNA-seq などの生化学実験技術によって測定可能であり，$\frac{dX_n}{dt}\big|_{t_k}$ は測定した遺伝子発現時系列データを滑らかに補間することで推定可能である．同様のアイデアに基づき，幾つかの遺伝子ネットワーク同定法が既に提案されている [14], [15].

3.2 連立方程式の効率的解法

本研究では連立方程式 (3) を解くことで，モデルパラメータ α_n, β_n, $\mathbf{g}_n = (g_{n,1}, g_{n,2}, \cdots, g_{n,N})$, $h_{n,n}$ を推定する．しかし連立方程式 (3) は非線形であることから，これを解くことはそれほど容易ではない．そこで本研究では連立方程式 (3) を解くことの困難さを解決するために，新たな方法を提案する．

3.2.1 コンセプト

本研究では連立方程式 (3) を効率的に解くために，これを以下のように変形する．

まず，連立方程式 (3) の k 番目の式を変形することで以下を得る．

$$\left.\frac{dX_n}{dt}\right|_{t_k} + \beta_n \left(X_n|_{t_k}\right)^{h_{n,n}} = \alpha_n \prod_{m=1}^{N} \left(X_m|_{t_k}\right)^{g_{n,m}}. \tag{4}$$

次に上式の両辺の対数を取ることで

$$\log Y_k = \log \alpha_n + \sum_{m=1}^{N} g_{n,m} \log \left(X_m|_{t_k}\right), \tag{5}$$

を得る．ただし

$$Y_k = \left.\frac{dX_n}{dt}\right|_{t_k} + \beta_n \left(X_n|_{t_k}\right)^{h_{n,n}},$$

である．式 (5) はパラメータ β_n, $h_{n,n}$ に関しては非線形であるものの，パラメータ $\log\alpha_n$, $\mathbf{g}_n = (g_{n,1}, g_{n,2}, \cdots, g_{n,N})$ に関しては線形である．従ってもしパラメータ β_n, $h_{n,n}$ の値が分かれば，その他のパラメータ α_n, $\mathbf{g}_n$ の値は線形連立方程式を解くことで簡単に得られる．提案手法では次項で説明するように，上記の性質を利用することで連立方程式 (3) を効率的に解く．

3.2.2 目的関数

前述したように n 番目の部分問題においてパラメータ β_n, $h_{n,n}$ の値が分かれば，その他のパラメータ α_n, $\mathbf{g}_n$ の値は簡単に推定できる．そこで本研究では，パラメータ β_n, $h_{n,n}$ の値のみを直接的に推定することで連立方程式 (3) を解く．従って本研究では連立方程式 (3) を解く問題を，以下の 2 次元関数の最小化問題として定式化する．

$$S_n(\beta_n, h_{n,n}) = \sum_{k=1}^{K}\left[\left.\frac{dX_n}{dt}\right|_{t_k} - \alpha_n^* \prod_{m=1}^{N}\left(X_m|_{t_k}\right)^{g^*_{n,m}} + \beta_n\left(X_n|_{t_k}\right)^{h_{n,n}}\right]^2 + \max\left\{0, d_n(\beta_n, h_{n,n})\right\}, \quad (6)$$

ただし

$$d_n(\beta_n, h_{n,n}) = \max\left\{\beta_n\left(X_n|_{t_1}\right)^{h_{n,n}}, \cdots, \beta_n\left(X_n|_{t_K}\right)^{h_{n,n}}\right\} - c_d \times \max\left\{\left|\left.\frac{dX_n}{dt}\right|_{t_1}\right|, \left|\left.\frac{dX_n}{dt}\right|_{t_2}\right|, \cdots, \left|\left.\frac{dX_n}{dt}\right|_{t_K}\right|\right\},$$

$\max\{\cdot\}$ は集合内の最大値を返す関数，c_d は定数パラメータ，α_n^* と $\mathbf{g}_n^* = (g^*_{n,1}, g^*_{n,2}, \cdots, g^*_{n,N})$ は β_n と $h_{n,n}$ の与えられたもとでの α_n と $\mathbf{g}_n$ の最適値である．α_n^* と $\mathbf{g}_n^*$ を求める方法については次項で説明する．

提案手法は最小二乗法に基づいており，目的関数 (6) の第 1 項は連立方程式 (3) の右辺と左辺の二乗誤差の総和になっている．他方，目的関数 (6) の第 2 項は，β_n の値が大きくなり過ぎないようにするためのペナルティ項である．予備実験に基づき c_d の値は 10 に設定した．

3.2.3 α_n^* と $\mathbf{g}_n^*$ の推定

3.2.2 項で述べたように，目的関数 (6) の値を計算するためには α_n^*, $\mathbf{g}_n^* = (g^*_{n,1}, g^*_{n,2}, \cdots, g^*_{n,N})$ を与える必要がある．本研究ではこれらの値は β_n と $h_{n,n}$ の値が与えられたもとでの，変形後の連立方程式 (5) の解として求められる．β_n と $h_{n,n}$ の値が与えられた場合，連立方程式は (5) は未知パラメータに関して線形となるため，α_n^* と $\mathbf{g}_n^*$ は簡単に求めることが可能である．本研究では連立方程式 (5) を解く問題を，以下の制約付き関数最小化問題として定式化する．

$$\underset{\log\alpha_n, \mathbf{g}_n, \xi_k^+, \xi_k^-}{\text{minimize}} \; C\sum_{k=1}^{K}\gamma_k\left(\xi_k^+ + \xi_k^-\right) + \sum_{m=1}^{N}|g_{n,m}|, \quad (7)$$

subject to

$$L_k - \log\alpha_n - \sum_{m=1}^{N} g_{n,m}\log\left(X_m|_{t_k}\right) \le \xi_k^+,$$
$$L_k - \log\alpha_n - \sum_{m=1}^{N} g_{n,m}\log\left(X_m|_{t_k}\right) \ge -\xi_k^-,$$
$$\xi_k^+ \ge 0, \quad \xi_k^- \ge 0 \quad (k = 1, 2, \cdots, K).$$

ただし

$$L_k = \log\left(Z_k\right),$$
$$Z_k = \begin{cases} Y_k, & (\text{if } Y_k \ge \delta), \\ \delta, & (\text{otherwise}), \end{cases}$$

ξ_k^+ と ξ_k^- は補助変数，γ_k, δ, C は定数パラメータである．この問題では，パラメータ β_n と $h_{n,n}$ は定数として扱われることに注意されたい．

ξ_k^+ と ξ_k^- は連立方程式 (5) の k 番目の式の右辺と左辺の差を表す．従って問題 (7) の第 1 項は，連立方程式 (5) の右辺と左辺の絶対値誤差の重み付きの総和である．他方，問題 (7) の第 2 項は，殆どの $g_{n,m}$ の値を 0 に強制するためのペナルティ項である．パラメータ $g_{n,m}$ は遺伝子 m から遺伝子 n への制御を表すパラメータである．遺伝子 m から遺伝子 n への制御が無い場合は，$g_{n,m}$ の値は 0 となる．本研究では，「遺伝子ネットワークは疎結合である．」という事前知識 [9] に基づいて，上記のペナルティ項を導入した．

3.2.1 項で述べたように，本研究では連立方程式 (3) の解を得るために，これを変形して式 (5) を得た．ただしこの式変形は $\left.\frac{dX_n}{dt}\right|_{t_k} + \beta_n\left(X_n|_{t_k}\right)^{h_{n,n}} > 0$ $(k = 1, 2, \cdots, K)$ が成り立つ場合にのみ可能である．しかしながら一般に測定された遺伝子発現データにはノイズが含まれるため，仮に β_n と $h_{n,n}$ に最適な値を与えたとしても，上記の条件は必ずしも常には成立しない．そこで本研究では閾値 δ を導入し，その値を 1.0×10^{-6} とした．ところで $\left.\frac{dX_n}{dt}\right|_{t_k} + \beta_n\left(X_n|_{t_k}\right)^{h_{n,n}}$ の値が 0 に近づくと，$\log\left[\left.\frac{dX_n}{dt}\right|_{t_k} + \beta_n\left(X_n|_{t_k}\right)^{h_{n,n}}\right]$ の値は $-\infty$ に近づく．従って $\left.\frac{dX_n}{dt}\right|_{t_k} + \beta_n\left(X_n|_{t_k}\right)^{h_{n,n}}$ の値が 0 に近い場合，式 (3) から式 (5) への変形の過程で，観測データに含まれるノイズが増幅される可能性がある．従って $\left.\frac{dX_n}{dt}\right|_{t_k} + \beta_n\left(X_n|_{t_k}\right)^{h_{n,n}}$ が 0 に近い場合は，得られた方程式 (5) はあまり信頼できない．この考えをパラメータ推定に導入するために，本研究では定数パラメータ γ_k を以下のように設定した．

$$\gamma_k = \frac{Z_k}{\sum_{j=1}^{K} Z_j}.$$

問題 (7) は線形計画問題に変換できる．そのため本研究では，この問題を解くためにシンプレクス法を用いた [11].

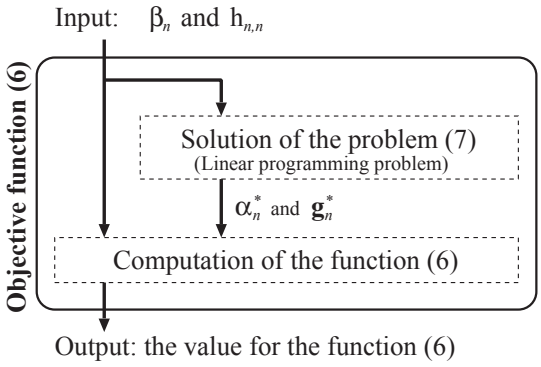

図 1 目的関数 (6) の計算.

3.2.4 アルゴリズム

前にも述べたように本研究では N 遺伝子からなるネットワークの同定問題を，N 個の部分問題に分割する．n 番目の部分問題では 2 次元の目的関数 (6) を最小化することで．パラメータ α_n, β_n, $\mathbf{g}_n = (g_{n,1}, g_{n,2}, \cdots, g_{n,N})$, $h_{n,n}$ を得る．ただし目的関数 (6) の値を計算するためには，制約付き関数最適化問題 (7) を解く必要があるという点に注意されたい．従って提案手法においては，関数 (6) の最適化の過程で問題 (7) が繰り返し解かれる (図 1)．なお問題 (7) は線形計画問題に変換できるため，シンプレクス法を用いて簡単に解くことができる．他方，目的関数 (6) の最適化のためには，どのような非線形最適化手法も利用することができる．しかしながら予備実験の結果，目的関数 (6) は多峰性であることが分かったため，本研究では最適化に進化的アルゴリズム REX^{star}/JGG [8] を利用した．

4. 数値実験

提案手法によって遺伝子ネットワークが正しく推定できるかどうかを確認するために実験を行った．

4.1 実験設定

本実験では reduced S-system モデルで記述された 30 遺伝子からなるネットワークをターゲットネットワークとした ($N = 30$)．提案手法の性能はターゲットネットワークの構造に依存する可能性があるので，本研究ではモデルパラメータを変えることで異なる構造を持つ複数のターゲットネットワークを作成した．実験は試行ごとに異なるネットワークを対象とした．

提案手法の性能は与えられた遺伝子発現データの量に依存するため，本研究では異なる量の遺伝子発現データを与えた実験を行った．時系列データはターゲットネットワークのモデルパラメータのもとで連立微分方程式 (2) を解くことで得られる．各遺伝子の発現量の初期値は $[0.0, 2.0]$ の範囲でランダムに決定した．測定ノイズとして，各データには 10%のガウスノイズを加えた．各データセットは 11 時点からなる全遺伝子の発現量となる．本研究ではデータセット数を，10 から 200 の範囲で変えて実験を行った．与えられた遺伝子発現時系列データを局所線形回帰 [3] を用いて平滑化し，各遺伝子の発現量の変化量 (転写速度) の推定を行った．

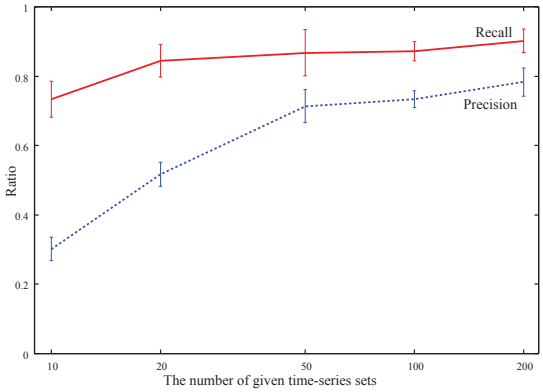

図 2 与える遺伝子発現データの量を変えた場合の提案手法の性能．実線は再現率 (recall)，点線は適合率 (precision) を表す．

提案手法の性能を統計的に評価するため，実験はそれぞれの遺伝子発現データセット数の問題で 10 試行行った．パラメータ $h_{n,n}$ の探索範囲は $[-5, 5]$ とした．他方，パラメータ β_n は正の値を取るため，本研究ではこの値を対数空間で探索した．$\log \beta_n$ の探索範囲は $[-20, 10]$ とした．目的関数 (6) の最適化に使用した進化的アルゴリズム REX^{star}/JGG [8] のパラメータには推奨値を用いた: 集団サイズ $n_p = 40$, 生成子個体数 $n_c = 6$, ステップサイズ $t = 2.5$．各試行は REX^{star}/JGG の世代数が 500 に達するまで行った．予備実験に基づき，問題 (7) に含まれる定数パラメータ C の値は 30 とした．

4.2 実験結果

ノイズの含まれた時系列データから，ターゲットネットワークのモデルパラメータを正確に推定することは困難である．そこで本実験では，推定された遺伝子ネットワークとターゲットネットワークの構造を比較することで，提案手法の性能を評価した．本実験では以下のようにして推定されたモデルパラメータからネットワークの構造を抽出した [6]: もし $g_{n,m} \geq Th_n$ and/or $h_{n,m} \leq -Th_n$ の場合，遺伝子 n は遺伝子 m から正の制御を受けていると判断する; 同様に $g_{n,m} \leq -Th_n$ and/or $h_{n,m} \geq Th_n$ の場合，遺伝子 n は遺伝子 m から負の制御を受けていると判断する; さもなければ遺伝子 n は遺伝子 m から制御を受けていないと判断する．なお Th_n は閾値パラメータであり，本研究では

$$Th_n = 0.05 \times \max\{|g_{n,1}|, \cdots, |g_{n,N}|, |h_{n,1}|, \cdots, |h_{n,N}|\},$$

とした．

与えた遺伝子発現データセットの数をそれぞれ 10, 20, 50, 100, 200 とした場合の，提案手法の再現率 (recall) と適合率 (precision) を図 2 に示す．再現率と適合率はそれ

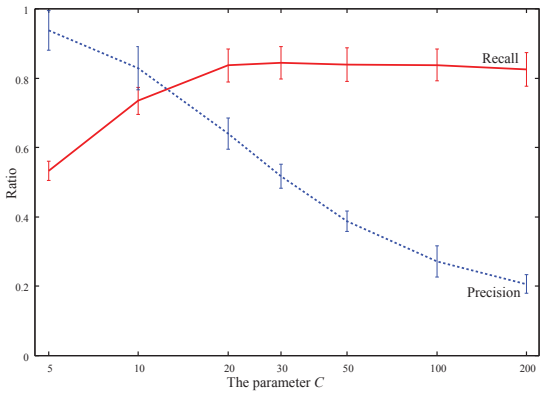

図 3 定数パラメータ C を変えた場合の提案手法の性能．実験は 20 セットの時系列データを与えた問題で行った．

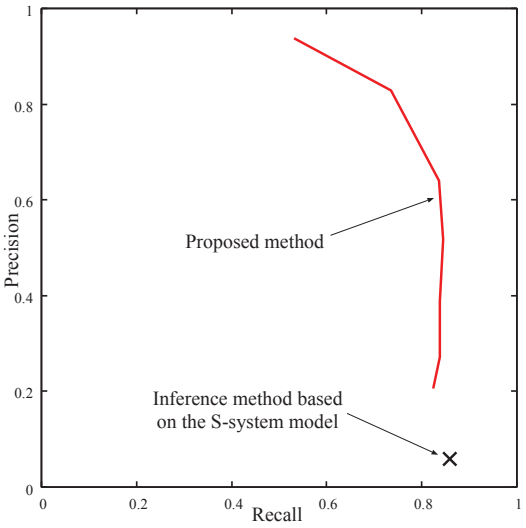

図 4 提案手法 (実線) と，S-system モデルに基づく遺伝子ネットワーク同定法 [6](× 印) の，20 セットの遺伝子発現データの与えられた問題における性能比較．

ぞれ

$$(\text{再現率}) = \frac{TP}{TP+FN},\quad (\text{適合率}) = \frac{TP}{TP+FP},$$

で定義される．ただし TP, FN, FP はそれぞれ，推定されたネットワークの真陽性，偽陰性，偽陽性の相互作用の数である．再現率は偽陰性相互作用の数が少ないほど，適合率は偽陽性相互作用の数が少ないほど大きな値を取る．図 2 より，与える遺伝子発現データの量の増加に伴って，提案手法の再現率と適合率が改善していることが分かる．ところで本実験では，問題 (7) に含まれる定数パラメータ C の値を 30 としている．しかしながら図 3 に示されるように，提案手法の性能は C の値に大きく依存する．従って実データを用いて遺伝子ネットワークの解析を行う場合，我々はパラメータ C の値を注意深く決定する必要がある．なお，この問題において一つの遺伝子ネットワークを同定するために必要とした計算時間は，35.0 ± 1.6 分 (Core i5-4670 3.4GHz) であった．

本研究では S-system モデルのパラメータの一部を 0 に固定することで簡略化した，reduced S-system モデルを用いて遺伝子ネットワーク同定を行っている．遺伝子ネットワーク同定において reduced S-system モデルを用いることの効果を確認するために，S-system モデルを用いた遺伝子ネットワーク同定法 [6] と提案手法との性能を比較した (図 4)．実験は 20 セットの遺伝子発現時系列データを与えた問題で行った．図 4 において提案手法の結果は，定数パラメータ C の値を 5 から 200 の範囲で変えることで得た．なお文献 [6] の方法は，提案手法と同様のアイデアに基づいて遺伝子ネットワークの S-system モデルを高速に同定する方法である．図 4 より S-system モデルを用いた方法は，再現率に関して提案手法と同等の性能を持つのに対して，適合率に関しては大きく劣ることが分かる．このことは提案手法によって推定される偽陽性相互作用の数が，S-system モデルに基づく方法によって推定されるものよりも非常に少ないことを意味している．これは遺伝子ネットワーク同定において，推定すべきパラメータの数を減らしたことによる効果であると考えられる．

これまでの実験で示したように，与える遺伝子発現時系列データの量を充分に多くしても，提案手法によって推定される誤った遺伝子間相互作用の数を 0 にすることは出来なかった．しかしながら与えるデータにノイズが含まれていなければ，提案手法は非常に高い精度でモデルパラメータを推定することが可能である．ノイズの含まれない 20 セットの遺伝子発現時系列データを与えた場合，提案手法によって得られた解の目的関数 (6) の評価値は平均で $1.237 \times 10^{-11} \pm 5.529 \times 10^{-12}$ であった ($C = 3000$ とした場合)．なお同量のノイズ有りのデータの場合，評価値は平均で $2.809 \times 10^{0} \pm 2.176 \times 10^{0}$ であった．推定されたパラメータ値とその真の値との差は，平均で $2.081 \times 10^{-7} \pm 4.684 \times 10^{-7}$ であった．

5. おわりに

S-system モデルは，生化学反応ネットワークを記述するのに適したモデルであると考えられている．しかしながらこのモデルのパラメータ数は，他の頻繁に利用されるモデルに比べて多いという問題がある．従って S-system モデルを用いて遺伝子ネットワーク同定を行う場合，妥当な結果を得るにはより多くの遺伝子発現データを与える必要があると考えられる．遺伝子ネットワーク同定に必要なデータ量を減らすために本研究ではまず，S-system モデルを簡略化した reduced S-system モデルを提案した．N 遺伝子からなるネットワークの同定を行う場合，S-system モデルでは $2N(N+1)$ 個のパラメータを推定する必要がある．それに対して reduced S-system モデルにおいて推定すべきパラメータの数は $N(N+3)$ 個である．本研究では次に，reduced S-system モデルに基づく新たな遺伝子ネットワーク同定法を提案した．提案手法は reduced S-system

モデルの性質を利用することで，N 遺伝子からなるネットワークの同定問題を，N 個の 2 次元関数最適化問題として定式化する．数値実験により，提案手法はより少ない量の遺伝子発現データから妥当な遺伝子ネットワークを同定することが可能であることを示した．一般に充分な量の遺伝子発現データを測定することは困難であることから，この性質は望ましいと言えるだろう．

謝辞 本研究は JSPS 科研費 26330275 の助成を受けて行われた．

参考文献

[1] Chemmangattuvalappil, N., Task, K. and Banerjee, I.: An Integer Optimization Algorithm for Robust Identification of Non-linear Gene Regulatory Networks, *BMC Systems Biology*, Vol. 6: 119 (2012).
[2] Cho, D.-Y., Cho, K.-H. and Zhang, B.-T.: Identification of Biochemical Networks by S-tree Based Genetic Programming, *Bioinformatics*, Vol. 22, No. 13, pp. 1631–1640 (2006).
[3] Cleveland, W.S.: Robust Locally Weight Regression and Smoothing Scatterplots, *J. of American Statistical Association*, Vol. 79, No. 368, pp. 829–836 (1979).
[4] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K. and Tomita, M.: Dynamic Modeling of Genetic Networks Using Genetic Algorithm and S-system, *Bioinformatics*, Vol. 19, No. 5, pp. 643–650 (2003).
[5] Kimura, S., Ide, K., Kashihara, A., Kano, M., Hatakeyama, M., Masui, R., Nakagawa, N., Yokoyama, S., Kuramitsu, S. and Konagaya, A.: Inference of S-system Models of Genetic Networks Using a Cooperative Coevolutionary Algorithm, *Bioinformatics*, Vol. 21, No. 7, pp. 1154–1163 (2005).
[6] Kimura, S., Araki, D., Matsumura, K. and Okada-Hatakeyama, M.: Inference of S-system Models of Genetic Networks by Solving One-dimensional Function Optimization Problems, *Mathematical Biosciences*, Vol. 235, No. 2, pp. 161–170 (2012).
[7] Kimura, S., Masanao, S. and Okada-Hatakeyama, M.: An Effective Method for the Inference of Reduced S-system Models of Genetic Networks, *IPSJ Transactions on Bioinformatics*, in press.
[8] 小林: 実数値 GA のフロンティア, 人工知能学会論文誌, Vol. 24, No. 1, pp. 147-162 (2009).
[9] Thieffry, D., Huerta, A.M., Pérez-Rueda, E. and Collado-Vides, J.: From Specific Gene Regulation to Genomic Networks: a Global Analysis of Transcriptional Regulation in *Escherichia Coli*, *BioEssays*, Vol. 20, No. 5, pp. 433–440 (1998).
[10] Thomas, R., Mehrotra, S., Papoutsakis, E.T. and Hatzimanikatis, V.: A Model-based Optimization Framework for the Inference on Gene Regulatory Networks from DNA Array Data, *Bioinformatics*, Vol. 20, No. 17, pp. 3221–3235 (2004).
[11] Todd, M.J.: The Many Facets of Linear Programming, *Mathematical Programming*, Vol. 91, No. 3, pp. 417–436 (2002).
[12] Vohradský, J.: Neural Network Model of Gene Expression, *FASEB J.*, Vol. 15, No. 3, pp. 846–854 (2001).
[13] Voit, E.O.: *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*, Cambridge University Press, Cambridge, UK (2000).
[14] Voit, E.O. and Almeida, J.: Decoupling Dynamical Systems for Pathway Identification from Metabolic Profiles, *Bioinformatics*, Vol, 20, No. 11, pp. 1670–1681 (2004).
[15] Yeung, M.K.S., Tegnér, J. and Collins, J.J.: Reverse Engineering Gene Networks Using Singular Value Decomposition and Robust Regression, *Proc. of National Academy of Sciences of USA*, Vol. 99, No. 9, pp. 6163–6168 (2002).