社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2003－MPS－47　(7)
2003／12／12

生体内ネットワークの構造推定のための S-system モデルの構造制限付き最適化

富永大介† 高橋勝利†

生体細胞内には代謝系や遺伝子間制御などのネットワークがある．そこでは化学物質や遺伝子の発現量などがノード，化学反応のスキームや制御関係，依存関係がエッジとなる．現在それらの構造はほとんど未知であるが， DNA チップなどといった，実験技術の進歩は近年めざましく，大量のデータが短期間で得られるようになってきた．しかし観察結果からネットワーク構造などの高次知識を得るための方法は未だ乏しく，人手で推測を重ねていくのが一般的である．我々は，時系列データをもとにネットワーク構造を推定するアルゴリズムを開発した．これは各ノードの経時変化のみから遺伝的アルゴリズム (GA) で S-system モデルを推定することで行う。またネットワークにスケールフリー性を仮定することによる推定精度の向上をケーススタディで確認した．

Inference biological network structure by genetic algorithm with structure restriction using time-series data

Tominaga Daisuke† and Takahashi Katsutoshi†

Living cells consist of many kinds of networks, such as metabolic pathways, gene regulatory networks, etc. In biological networks, nodes are concentration of metabolites or expression levels of genes, and edges are chemical reaction scheme or regulatory interactions. Although structures of most of networks have not been revealed yet, technologies to observe dynamics of biological network, such as DNA microarray or mass spectrometry, are rapidly growing, and these high-throughput experimental technologies produce huge amounts of data. But there are no efficient knowledge mining method to infer structures of biological networks.

In this paper, we propose optimization algorithm to infer network structures only using time-course data of target biological networks. Our algorithm is *ab initio* method which doesn't need any other information or knowledges about target networks than time-course data. We use the S-system as a model of a network and introduce Scale-free property as a biological network. The algorithm is based on the genetic algorithm. We probed efficiency of our method on case studies.

1. はじめに

生命の最小単位は細胞であり，生命活動の仕組みを探ろうとするとき細胞の内部構造を調べるアプローチは一般的である．細胞内部には非常に多種の物質が含まれており，相互に作用しあっている．生命活動の本質を探るには，こういった複雑な相互作用のネットワークがどのように構成されているかを研究することが欠かせない．ネットワークを構成するノードは遺伝子や化学物質などである．リンクは遺伝子発現の制御関係や化学反応などであるが，これらは複雑な分子機構が担っておりその細部はほとんど知られていないため、ブラックボックスモデルによるネットワークの構造推定ができれば，生命活動をモデリングし生命活動の本質を探るためにも非常に有益である．近年技術革新で、遺伝子の発現量の時系列を観察する技術が確立されてきているが、時系列から直接ネットワークの構造を推定するよい方法はない．本研究ではそのための基本技術としてネットワークの記述法に S-system[4]，その構造推定法に遺伝的アルゴリズムを用いた手法を提案，検証する．

これまでは，前述したような生体内のネットワークは疎結合であることを考慮してネットワークの構造を絞り込んでいたが[5)6)]，構造の制約条件に妥当性がなかった．そこで今回は，生体内のネットワークはスケールフリーネットワークの特徴を持っていると言われていることから，これを制約条件として用いること

† 産業技術総合研究所生命情報科学研究センター
Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology

とした．

2. モデルと最適化法

2.1 S-system

S-system は，代謝系などの生化学反応系を表現するために考案されたもので，質量作用則によるモデルの近似式で以下の形式の連立微分方程式である．、

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

ここで X_i は代謝物質の濃度などの物理量で，ネットワークのノードと捉えられる．n はノードの総数，i, j はその添え字である．一般に X_i を増加させようとする作用（化学反応など）は複数あり得るが，S-system ではそれを式 (1) 右辺の第一項で表現する．第二項は X_i を減少させようとする作用を表す．ネットワークの構造を表すパラメータは g_{ij} と h_{ij} であり、α_i と β_i は速度係数に相当する．式 (1) に含まれる合計 $2n(n+1)$ 個の g_{ij}，h_{ij}，α_i，β_i を全て最適化することで数値積分した結果が与えられる実験データを再現するようになれば，g_{ij}，h_{ij} はなんらかのネットワーク構造を表していて，それは実際の生体内ネットワークの構造（真の解）の候補と考えられる．

2.2 最 適 化 法

2.2.1 目 的 関 数

t をサンプリング時刻として，実験で観測されたノード i の量を $X_{i,exp}^t$，式 (1) を数値積分して得られたデータを $X_{i,cal}^t$ とすると，ここで行う最適化はこれらの差を最小化することである．差を相対誤差とすると，以下の式で表される．

$$f_e = \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{X_{i,exp}^t - X_{i.cal}^t}{X_{i,exp}} \qquad (2)$$

ここで T はサンプリング点の総数である．目的関数は，範囲を正規化するために以下の形式とする．

$$f = \frac{1}{1+f_e} \qquad (3)$$

最適化は，式（3）を最大化することになる．これを分散遺伝的アルゴリズム (分散 GA) で最適化する．

2.2.2 遺伝的アルゴリズム

単純 GA に加え，GA における各染色体を実数値とし，α_i，β_i，g_{ij}，h_{ij} の $2n(n+1)$ 個の各パラメータを，それぞれ一つの染色体に割り当てる．GA での一個体は，各 S-system パラメータに対応する染色体を一つずつ持つ，$2n(n+1)$ 次元のベクトルとして一つの S-system モデルを表す．各染色体にそれぞれ探索範囲を定め，初期個体集団はその範囲内で各染色体を乱数で決定して生成するが, この際にスケールフリー性を持つようにする（後述）．生成された各個体は，式 (3) で評価値を計算される．

生成された個体からは SPX[3] で子個体を生成する．エリート戦略を適用し，エリート個体以外はすべて子個体で置き換える．

GA による探索中, 初期世代からの世代交代の回数を G, エリート個体が更新されなくなってからの世代交代の回数を G_s として, $G < 2 \times G_s$ となったとき, または G が規定の最大値に達したときエリート個体を解として最適化を終了する．

2.2.3 スケールフリーネットワーク

スケールフリーネットワークは，リンクを k 本持つノードの数を $P(k)$ としたとき，

$$P(k) = Ak^{-\gamma} \qquad (4)$$

となるようなネットワークである（A は定数）[1]．酵母の代謝系では $\gamma \approx 2.2$ であることが報告されており，生体内のネットワークはスケールフリー性を持つことが示唆されている[2]．ノード数 n が既知であり, 自分から自分へのリンクはなく, リンクを持たないノードはないとすると, 一つのノードが持つ最大のリンク数は $n-1$ になるため式 (4) で $1 \leq k \leq n-1$ であり，$\sum_{k=1}^{n-1} P(k) = n$ である．これから式 (4) での A が決まり, 各 k に対する $P(k)$ が求まる．GA において初期集団として多数のネットワークモデルを生成する際, モデル中のノードがリンクを k 本持つ確率を $P(k)/n$ であるとして生成し, スケールフリー性を持たせる．

2.2.4 ネットワーク構造の決定

分散 GA の各集団で一つずつ得られるネットワークモデルに対して，各ノードのペアを結ぶリンクがいくつのモデルに存在しているかの度数を数え上げる．また, 各リンクは S-system の指数係数であり符号を持つ．各リンク g_{ij} および h_{ij} について 0, 正, 負であった回数をカウントし, 各 i について正のカウントと負のカウントの比を計算する．最終的に, 正負の比が最小でなく度数が最大のもの, また度数が最小でなく比が最大のもの, 比が計算できないもの（正, あるいは負のみのリンク）を最終的な解に含まれるリンクとし, これを探索結果のネットワークの構造とする．

3. ケーススタディ

ネットワークの構造の概略を推定するアルゴリズムとしての性能を検証するため，**表 1** に示す S-system モデルから初期値を変えて 10 セットの時系列データ

表 1　**図 1** に示すシステムを表現する S-system モデルのパラメータ.
Table 1　A set of S-system parameters which represents a network shown in figure 1.

i	α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	g_{i5}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}	h_{i5}
1	5.0	0.0	0.0	1.0	0.0	-1.0	10.0	2.0	0.0	0.0	0.0	0.0
2	10.0	2.0	0.0	0.0	0.0	0.0	10.0	0.0	2.0	0.0	0.0	0.0
3	10.0	0.0	-1.0	0.0	0.0	0.0	10.0	0.0	-1.0	2.0	0.0	0.0
4	8.0	0.0	0.0	2.0	0.0	-1.0	10.0	0.0	0.0	0.0	2.0	0.0
5	10.0	0.0	0.0	0.0	2.0	0.0	10.0	0.0	0.0	0.0	0.0	2.0

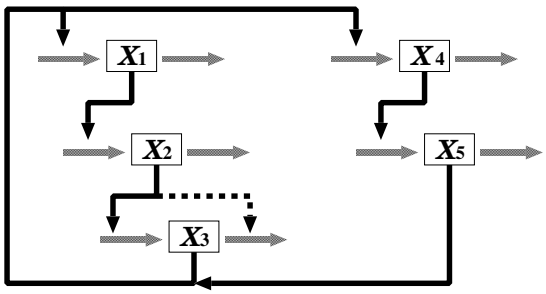

図 1　典型的な遺伝子発現系のモデル．実線: 各ノードを増やす作用に対するリンク, 破線: 各ノードを減少する作用に対するリンク.
Fig. 1　A model of a typical gene expression scheme

を作成し，それを用いて図 1 に示す構造を見つけられるかどうかをテストした．ここでは，1) 交叉に SPX[3] を用いること以外は単純 GA を用いる方法，2) それに加えてモデル中のリンクの本数を各要素につき 2 本以内に制限する方法, 3) さらに一つのモデルがスケールフリーネットワークとなるように制限する方法の 3 種類を比較した．ネットワークモデルのスケールフリー性は，式 (4) において $\gamma = 2.2, A = 3.51$ とした．また生化学反応系を想定しここでは α_i と g_{ij} のみの最適化を行った。式 (1) において最適化対象となるパラメータの総数は 30 で，探索範囲は α_i は $[0, 15]$，g_{ij} は $[-3, 3]$ である．最適化は，目的関数から計算されるサンプリングポイント 1 点あたりの平均相対二乗誤差が 0.01 以下になったとき，および規定の回数（3000 回）の世代交代を行ったときに終了する．1 から 3 の各方法について, それぞれ 100 個の個体からなる 64 集団の分散 GA の結果から 2.2.4 節に示す計算を行い，ネットワーク構造を決めた．その際，$i = j$ となる g_{ij} はその際に対象から外した．その結果を**表 2** に示す．また各方法で得られたネットワーク構造を**図 2** に示す．

False positive，false negative ともに，スケールフリー性を仮定して最適化することにより改善することができる．ネットワークモデルから計算される時系列データと, 与えられたデータとの誤差は, 単にランダムに探索するよりも構造制約を導入することで改善される．単にリンクの本数を制限すると, 制限なしよりも評価値は高いがオリジナルの構造を探せなくなる．その点ではランダムに探索した方が性能が良く, さらにスケールフリー性を持たせることが有効である．

表 2　3 つの方法の比較．algorithm 1: ランダム, 2: リンク本数の制限, 3: スケールフリー．a) 収束率（最適化が最大世代数に到達する前に終了した割合），b) 平均二乗誤差の平均, c) 平均世代数，d) False positive，e) False negative．d+e は false の合計.
Table 2　Comparison of three algorithms.　algorithm 1: simple GA, 2: limited numbers of links, 3: scale-free. a) Conversion ratio (terminated GA runs before maximum generation), b) Average of average squared error, c) Average number of generations, d) False positive, e) False negative.　d+e means total sum of false (miss estimated links.)

algorithm	a	b	c	d	e	d + e
1	0.713	5.80	1539	3	5	8
2	0.438	0.0123	2159	7	5	12
3	0.5	0.000890	2073	3	1	4

4. ま と め

表 2 では，推定しようとするネットワーク構造にスケールフリー性を仮定するかどうかで，推定能力が大きく左右されることを示した。一般には GA の特徴は局所解に捕らわれない大域探索にあるが，ここでの GA は局所探索を行っている．これは，**図??**にあるように目的関数には未定義の領域が広く，そういった領域では GA の探索は非常に効率が悪いからである．したがってランダムに発生した初期推定ネットワークのうち，実行可能でもっとも目的関数値が高いものの周辺を探索する．これは, 初期集団に含まれるエリート個体と同じ構造を持つものが, 探索が終了するまでずっとエリート個体であることが多いことからも分かる．したがって初期集団の生成の方法が最重要であるが, その際にリンクの本数のみを制限した場合は, 十分に解を絞れないために探索性能が落ち, スケールフリー性を持たせた場合は十分に解を絞ることができると言える．また表 2 において false negative を減らすことができるのも重要なことである．実験観測データには誤差が少なからず含まれるため, 完全な正解を探

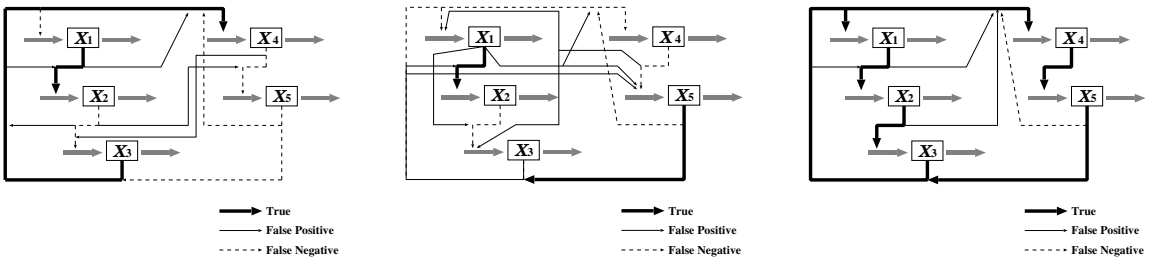

図 2 最適化により得られたネットワーク構造．a) 方法 1，b) 方法 2，c) 方法 3 による．
Fig. 2 Obtained network structures by a) algorithm 1, b) algorithm 2, c) algorithm 3.

索するよりも「あるものを見逃さない」特性が必要である。

今後の課題としてスケールアップがある．今回のケーススタディでは 3000 世代に約 90 分かかっており，世代交代に MGG を導入するなどして，速度の向上を図る必要がある．また有効な初期推定の生成がキーポイントであるため，時系列データのプロファイルから，あらかじめ制御関係を大まかに推定しておくことも考えられる．

参 考 文 献

1) Albert, R. and Barabasi, A.-L.: Power-law distribution of the World Wide Web, *Science*, Vol. 287, No. 2115a (2001).
2) Podani, J., Oltvai, Z., Jeong, H., B.Tombor, Barabasi, A.-L. and Szathmary, E.: Comparable system-level organization of Archaea and Eukaryotes, *Nature Genetics*, Vol. 29, pp. 54–56 (2001).
3) Tsutsui, S., Yamashita, M. and Higuchi, T.: Multi-parent Recombination with Simplex Crossover in Real Coded Genetic Algorithms, *Proc. of the Genetic and Evolutionary Computation Conference*, Vol. 1, pp. 657–664 (1999).
4) Savageau, M. A.: *Biochemical System Analysis. A Study of Function and Design in Molecular Biology*, Addiso-Wesley, Reading, M.A., USA (1976).
5) Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K. and Tomita, M.: Dynamic modeling of genetic networks using genetic algorithm and S-system, *Bioinformatcis*, Vol. in press (2003).
6) Tominaga, D., Koga, N. and Okamoto, M.: Efficient Numerical Optimization Algorithm Based on Genetic Algorithm for Inverse Problem, *Proc. of Genetic and Evolutionary Computation Conference (GECCO 2000)*, pp. 251–258 (2000).

(平成？年？月？日受付)
(平成？年？月？日採録)

富永　大介（正会員）

昭和 45 年生．平成 9 年九州工業大学大学院情報工学研究科情報科学専攻博士後期課程修了．同年独立行政法人産業技術総合研究所生命情報科学研究センター入所（現職）．遺伝子ネットワークの推定，プロテオーム解析の研究に従事．博士（情報工学）．日本バイオインフォマティクス学会，CBI 学会，人工知能学会会員．

高橋　勝彦（正会員）

昭和 41 年生．平成 5 年京都大学大学院理学研究科化学専攻博士後期課程修了．同年東ソー（株）入社．薬物設計技術の研究に従事．平成 7 年弘前大学理学部教務職員．分子グラフィクス，蛋白質構造の統計解析の研究に従事．平成 8 年金沢工業大学工学部助手．DNA 二次元電気泳動像自動解析システムの研究に従事．平成 9 年金沢工業大学工学部講師．蛋白質二次元電気泳動像自動解析システムの研究に従事．平成 10 年北陸先端科学技術大学院大学知識科学研究科助手．電子顕微鏡による蛋白質三次元構造決定法の研究に従事．平成 13 年独立行政法人産業技術総合研究所生命情報科学研究センター入所（現職）．細胞情報科学における蛋白質プロファイリングの研究に従事．