2007－BIO－8 (11)
2007／3／5

遺伝子ネットワークの ニューラルネットワークモデル同定法の提案

木村 周平 †, 園田 克樹 ††, 山根 総一郎 ††, 松村 幸輝 †, 畠山 眞里子 ‡

† 鳥取大学工学部 †† JFE 技研株式会社 ‡ 理化学研究所ゲノム科学総合研究センター

連立微分方程式に基づくモデルは様々なダイナミクスを表現することが可能であり，遺伝子ネットワークを記述するために適していると考えられる．これまで連立微分方程式に基づくモデルを利用した多くの遺伝子ネットワーク同定法が提案されてきた．それらの同定法の多くが線形モデルやS-systemモデルといった何らかの固定形式の微分方程式モデルを用いている．本研究ではこのような固定形式の微分方程式モデルではなく，ニューラルネットワークによる微分方程式モデルを用いた遺伝子ネットワーク同定法を提案する．また得られたモデルを解釈するための感度解析に基づく方法も提案する．実験を通して提案手法の有効性を確認する．

Development of a Genetic Network Inference Method based on a Neural Network Model

S. KIMURA †, K. SONODA ††, S. YAMANE ††, K. MATSUMURA †, M. HATAKEYAMA ‡

† Faculty of Engineering, Tottori University

†† JFE R&D Corporation ‡ RIKEN GSC

A model based on a set of differential equations can effectively capture various dynamics. This type of model is therefore ideal for describing genetic networks. Several genetic network inference algorithms based on models of this type have been proposed. Most of these inference methods use models based on a set of differential equations of the fixed form to describe genetic networks. In this study, we propose a new method for the inference of genetic networks. To describe genetic networks, the proposed method does not use models of the fixed form, but uses neural network models. In order to interpret obtained neural network models, we also propose a method based on sensitivity analysis. The effectiveness of the proposed methods is verified through an artificial genetic network inference problem.

1 はじめに

DNAマイクロアレイ技術の進歩によって細胞全体のレベルでの遺伝子発現パターンの解析が可能になってきた．しかしこの技術を活用するためには膨大な量のデータから必要な情報を抽出することが必要である．マイクロアレイデータからの情報抽出方法の一つとして，遺伝子ネットワークの同定に注目が集まってきている．遺伝子ネットワークの同定とは，マイクロアレイの時系列データから遺伝子間の相互作用を推定する問題のことである．同定した遺伝子ネットワークモデルは仮説の生成や実験デザインに利用可能であり，また未知の遺伝子の機能を推定するためにも利用できると考えられる．

遺伝子ネットワークを記述するためのモデルとして，本研究では連立微分方程式に着目する．連立微分方程式モデルを使用する場合，遺伝子ネットワークは以下の式によって記述される．

$$\frac{dX_i}{dt} = G_i(X_1, \cdots, X_N), \ (i = 1, \cdots, N). \quad (1)$$

ただし X_i は遺伝子 i の発現量，N はネットワークに含まれる遺伝子数である．G_i は任意の関数であり，他の遺伝子から遺伝子 i への制御関係を記述している．連立微分方程式モデルを用いて遺伝子ネットワークを記述する場合，遺伝子ネットワーク同定問題は観測した遺伝子発現量の時系列データから未知の関数 G_i の良い近似を求める問題となる[2, 3, 5, 7, 12]．

連立微分方程式モデルは様々なダイナミクスを表現することが可能であり，遺伝子ネットワークを

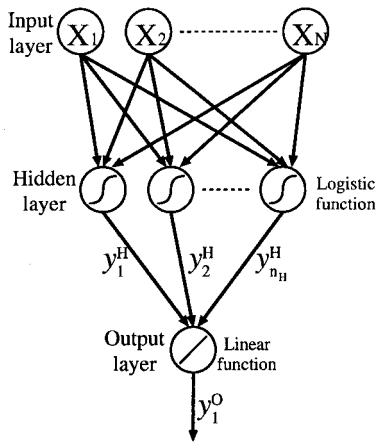

Fig. 1　ニューラルネットワークモデル.

記述するために適していると考えられている．しかし連立微分方程式モデルを用いた遺伝子ネットワーク同定法は一般に，連立微分方程式 (1) を繰り返し解くことを必要とする．そのため従来手法には，膨大な計算量を必要とするという問題があった．

本研究では遺伝子ネットワーク同定に必要な計算量の削減を目指して，新たな遺伝子ネットワーク同定法を提案する．具体的には関数 G_i を近似するために，有名な関数近似器であるニューラルネットワークを利用する．ニューラルネットワークを用いた関数 G_i の近似は，連立微分方程式 (1) を一切解かずに行うことができる．そのため提案手法は，従来手法よりも少ない計算量での遺伝子ネットワーク同定が可能である．

以下，2 章でニューラルネットワークを用いた遺伝子ネットワーク同定法を提案する．3 章では獲得したニューラルネットワークモデルから，遺伝子間相互作用に関する情報を抽出する方法を説明する．4 章では提案手法の効果を確認するために，数値実験を行う．5 章は結論である．

2　ニューラルネットワークモデルを用いた遺伝子ネットワーク同定

前述したように本論文の目的は，観測データをもとに関数 $G_i(i=1,\cdots,N)$ を良く近似するニューラルネットワークを獲得することである．なお本研究では 1 つのニューラルネットワークを用いて 1 つの関数 G_i を近似する．従って N 遺伝子からなる遺伝子ネットワーク同定問題では，N 個のニューラルネットワークが獲得されることになる．以下では関数 G_i に対応するニューラルネットワークを求める方法について具体的に説明する．

2.1　ニューラルネットワークモデル

関数 G_i を近似するために，本研究では一般的な関数近似器である 3 層構造のフィードフォワード型ニューラルネットワークを使用する (Fig. 1)．ただし入力層，中間層，出力層のニューロン数はそれぞれ N 個，n_H 個，1 個である．

本研究で使用するニューラルネットワークでは，k 番目の中間層ニューロンの出力 y_k^H および唯一の出力層ニューロンの出力 y_1^O は，それぞれ

$$y_k^H = f\left(\sum_{j=1}^{N} w_{j,k}^{IH} X_j - \theta_k^H\right),$$

$$y_1^O = g\left(\sum_{k=1}^{n_H} w_{k,1}^{HO} y_k^H - \theta_1^O\right),$$

で計算される．ただし X_j は j 番目の入力層ニューロンへの入力値，$w_{j,k}^{IH}$ は j 番目の入力層ニューロンと k 番目の中間層ニューロンを結ぶ結合の重み，θ_k^H は k 番目の中間層ニューロンの閾値パラメータ，$w_{k,1}^{HO}$ は k 番目の中間層ニューロンと出力層ニューロンを結ぶ結合の重み，θ_1^O は出力層ニューロンの閾値パラメータ，$f(x)=(1+\exp(-x))^{-1}$，$g(x)=x$ である．なお本研究では，入力 X_j は遺伝子 j の発現量，出力 y_1^O は遺伝子 i の発現変化量 (転写速度) に相当する．

2.2　ニューラルネットワークモデルの学習

本研究の目的は，観測した遺伝子発現量の時系列データから関数 G_i を良く近似するニューラルネットワークモデルを学習することである．一般に，入力ベクトル $\mathbf{x}_t(t=1,2,\cdots,T)$ に対して y_t を出力するニューラルネットワークを学習する問題は，以下の関数 E を最小にするニューラルネットワークのパラメータ $(w_{j,k}^{IH},\theta_k^H,w_{k,1}^{HO},\theta_1^O)$ を探索する問題として定式化される．

$$E=\frac{1}{2}\sum_{t=1}^{T}\left[y_1^O(\mathbf{x}_t)-y_t\right]^2. \qquad (2)$$

ただし $y_1^O(\mathbf{x}_t)$ は入力 $\mathbf{x}_t$ に対するニューラルネットワークの出力である．本研究では，入力 $\mathbf{x}_t$

を時刻 t における全遺伝子の発現量 $(X|_t = (X_1|_t, \cdots, X_N|_t)^T)$，出力 y_t を時刻 t における遺伝子 i の発現変化量 $(\frac{dX_i}{dt}\big|_t)$ とする．全遺伝子の発現量はマイクロアレイなどの計測技術によって測定可能である．また本研究では全遺伝子の発現量の時系列データが計測されていることを前提としているため，それらを滑らかに補間することで遺伝子 i の発現変化量は推定できる．従って上記のニューラルネットワーク学習問題は解くことが可能である．

上記の関数 E をそのまま最適化しても，妥当な遺伝子ネットワークモデルを得ることは難しい．その理由は，モデルの自由度の高さと比較して，与えられるデータの量が一般に少ないためである．そこでより妥当な解を得るために，本研究では「遺伝子ネットワークは疎に結合している」という事前知識[10]を利用する．遺伝子 i が遺伝子 j から制御を受けていない場合，j 番目の入力層ニューロンから中間層ニューロンへ結合の重み $(w^{IH}_{j,1}, \cdots, w^{IH}_{j,n_H})$ は全て 0 として良い．このような場合，それらの結合強度の二乗和 $\omega_j = \sum_{k=1}^{n_H} \left(w^{IH}_{j,k}\right)^2$ も 0 となる．本研究では以上の知識を目的関数 (2) に導入し，以下の目的関数 F を得る[6]．

$$F = \frac{1}{2}\sum_{t=1}^{T}\left[y_1^O(\mathbf{x}_t) - y_t\right]^2 + c\sum_{j=1}^{N-I} W_j. \qquad (3)$$

ただし W_j は ω_j を小さい順に並び替えたもの (つまり $W_1 \le W_2 \le \cdots \le W_N$) である．また c は定数パラメータ，I は最大入り次数である．最大入り次数は遺伝子 i に影響を及ぼす遺伝子の最大数を規定するパラメータである．

2.3 学習アルゴリズム

本研究では目的関数 (3) を最適化することにより，関数 G_i を近似するニューラルネットワークモデルを得る．この目的関数は微分可能なため，最適化には誤差逆伝播法[9]が利用できる．ところが誤差逆伝播法は局所最適化手法であるため，目的関数が多峰性の場合に局所解に陥りやすいという欠点がある．

誤差逆伝播法の欠点を回避するために，最適化手法として進化的アルゴリズムがしばしば使用されている[13]．多くの進化的アルゴリズムがニューラルネットワークの学習に利用できるが，本研究では特に GLSDC[4] を使用した．GLSDC はその探索オペレータとして誤差逆伝播法を利用することができる．そのため目的関数の勾配情報を積極的に活用した効率的な探索が期待できる．

3 学習したモデルからの情報抽出

遺伝子ネットワーク同定では，どの遺伝子がどの遺伝子を制御しているかを知ることがしばしば重要となる．本研究では学習されたモデルからこのような遺伝子間の制御情報を抽出するために，感度解析を利用する[6]．

遺伝子 i が遺伝子 j から制御を受けていなければ，遺伝子 i の転写速度 $(\frac{dX_i}{dt})$ は遺伝子 j の発現量 (X_j) の影響を受けない．従ってこのとき，以下の式が常に成立する．

$$\frac{\partial}{\partial X_j}\left(\frac{dX_i}{dt}\right) = \frac{\partial G_i(X_1, \cdots, X_N)}{\partial X_j} = 0.$$

ただし G_i は式 (1) で示される，他の遺伝子から遺伝子 i への制御関係を記述した関数である．$\frac{\partial G_i}{\partial X_j}$ は一般に感度係数と呼ばれる．感度係数の絶対値の大きさを見ることで，遺伝子 j の遺伝子 i に対する影響の大きさを知ることができる．また感度係数の符号を調べることで，遺伝子 j から遺伝子 i への制御の種類 (正負) を知ることが可能である．

一般に感度係数の値は時間と共に変化する．そのため本研究では正の感度係数の時間平均 $S_i^p(j)$ と負の感度係数の時間平均 $S_i^m(j)$ を利用して，遺伝子間相互作用を推定する．しかしながら一般に，$S_i^p(j)$ と $S_i^m(j)$ の値を正確に求めることは困難である．そこで本研究では以下の近似値を利用する．

$$\begin{aligned} S_i^p(j) &= \frac{1}{t_T - t_1}\int_{t_1}^{t_T} p\left(\frac{\partial G_i}{\partial X_j}\right) dt \\ &\simeq \frac{1}{T}\sum_{k=1}^{T} p\left(\frac{\partial \hat{G}_i}{\partial X_j}\bigg|_{t_k}\right), \\ S_i^m(j) &= \frac{1}{t_T - t_1}\int_{t_1}^{t_T} m\left(\frac{\partial G_i}{\partial X_j}\right) dt \\ &\simeq \frac{1}{T}\sum_{k=1}^{T} m\left(\frac{\partial \hat{G}_i}{\partial X_j}\bigg|_{t_k}\right). \end{aligned}$$

ただし

$$p(x) = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise}, \end{cases}$$

$$m(x) = \begin{cases} x, & \text{if } x < 0, \\ 0, & \text{otherwise}, \end{cases}$$

である．また T は時系列データのサンプル数，$\left.\frac{\partial \hat{G}_i}{\partial X_j}\right|_t$ は遺伝子 i に対応するニューラルネットワークモデルから計算される，時刻 t における感度係数である．

遺伝子 j からの遺伝子 i に対する影響力が大きければ，$|S_i^p(j)| + |S_i^m(j)|$ は大きな値となるはずである．また $|S_i^p(j)|$ と $|S_i^m(j)|$ の値を比較することで，遺伝子 j から遺伝子 i への制御が正なのか負なのかを知ることができると考えられる．そこで本研究では以下のルールに従って遺伝子間相互作用を推定する．

- $|S_i^p(j)| + |S_i^m(j)| > Thresh(i)$ かつ

$$\frac{|S_i^p(j)|}{|S_i^p(j)| + |S_i^m(j)|} > \alpha,$$

であれば，遺伝子 i は遺伝子 j から正の制御受けているとする．

- $|S_i^p(j)| + |S_i^m(j)| > Thresh(i)$ かつ

$$\frac{|S_i^m(j)|}{|S_i^p(j)| + |S_i^m(j)|} > \alpha,$$

であれば，遺伝子 i は遺伝子 j から負の制御受けているとする．

ただし

$$Thresh(i) = \beta \max_j \left(|S_i^p(j)| + |S_i^m(j)|\right),$$

である．α と β は定数パラメータであり，本研究では $\alpha = 0.3$，$\beta = 0.05$ とした．α が 0.5 以下であるため，本手法は遺伝子 j から遺伝子 i への正の制御と負の制御を同時に推定することがある．

4 数値実験

提案手法が正しく遺伝子ネットワークを同定することができるかを確認するために実験を行った．

4.1 実験設定

本研究では，解析対象ネットワークとして 30 遺伝子からなる S-system モデル [11] を使用した．S-system モデルとは以下の連立微分方程式で記述されたモデルである．

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{N} X_j^{g_{i,j}} - \beta_i \prod_{j=1}^{N} X_j^{h_{i,j}}, \quad (i = 1, 2, \cdots, N). \quad (4)$$

Table 1 対象ネットワークのモデルパラメータ．

α_i	1.0
β_i	1.0
$g_{i,j}$	$g_{1,14} = -0.1$, $g_{5,1} = 1.0$, $g_{6,1} = 1.0$, $g_{7,2} = 0.5$, $g_{7,3} = 0.4$, $g_{8,4} = 0.2$, $g_{8,17} = -0.2$, $g_{9,5} = 1.0$, $g_{9,6} = -0.1$, $g_{10,7} = 0.3$, $g_{11,4} = 0.4$, $g_{11,7} = -0.2$, $g_{11,22} = 0.4$, $g_{12,23} = 0.1$, $g_{13,8} = 0.6$, $g_{14,9} = 1.0$, $g_{15,10} = 0.2$, $g_{16,11} = 0.5$, $g_{16,12} = -0.2$, $g_{17,13} = 0.5$, $g_{19,14} = 0.1$, $g_{20,15} = 0.7$, $g_{20,26} = 0.3$, $g_{21,16} = 0.6$, $g_{22,16} = 0.5$, $g_{23,17} = 0.2$, $g_{24,15} = -0.2$, $g_{24,18} = -0.1$, $g_{24,19} = 0.3$, $g_{25,20} = 0.4$, $g_{26,21} = -0.2$, $g_{26,28} = 0.1$, $g_{27,24} = 0.6$, $g_{27,25} = 0.3$, $g_{27,30} = -0.2$, $g_{28,25} = 0.5$, $g_{29,26} = 0.4$, $g_{30,27} = 0.6$, other $g_{i,j} = 0.0$
$h_{i,j}$	1.0 if $i = j$, 0.0 otherwise.

ただし X_i は遺伝子 i の発現量であり，α_i，β_i，$g_{i,j}$，$h_{i,j}$ は定数パラメータである．本研究で使用したモデルのパラメータを Table 1 に，そのネットワーク構造を Fig. 2 に示す [7]．

全ての遺伝子の発現量の初期値をランダムに決定し，連立微分方程式 (4) を解くことにより，全遺伝子の発現量の時系列データを得ることができる．本研究では，このようにして得た 15 セットの時系列データだけを利用して，対象ネットワーク構造の同定を試みた．なお本研究では，各データセットは 11 時点の観測からなるものとした．従ってデータは，遺伝子毎に $15 \times 11 = 165$ 時点の観測を含んでいる．

2.2 節で説明したように，提案手法を用いて遺伝子ネットワークを同定するためには，観測データから全遺伝子の発現変化量を推定する必要がある．本研究で与えられる時系列データにはノイズが含まれないため，遺伝子の発現変化量の推定にはスプライン補間 [8] を利用した．通常，マイクロアレイなどによって測定された遺伝子発現データには観測ノイズが含まれるが，そのような場合には局所線形回帰 [1] などの平滑化手法を利用して遺伝子の発現変化量を推定すれば良い．

2 章で述べたように，N 遺伝子の遺伝子ネットワーク同定を行う場合，提案手法では N 個のニューラルネットワークモデルを学習する必要がある．1 つのニューラルネットワークモデルは N 個の入力層

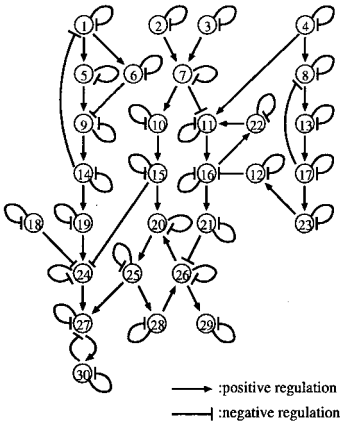

Fig. 2　対象ネットワークの構造.

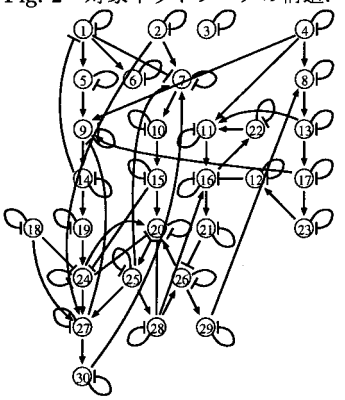

Fig. 3　提案手法によって同定されたネットワークの例.

ニューロン，n_H 個の中間層ニューロン，1 個の出力層ニューロンから構成される．本研究では $n_H = 5$ としたため，1 つのニューラルネットワークの学習は，$N \times n_H + n_H \times 1 + 1 = 161$ 次元の関数最適化問題となる．2.3 節で述べたように，それぞれのニューラルネットワークの学習には GLSDC [4] を使用した．モデルパラメータの探索範囲は $[-10, 10]$ とし，GLSDC のパラメータには推奨パラメータを利用した; 集団サイズ $n_p = 483$，生成子個体数 $n_c = 10$，交叉繰り返し数 $N_0 = 322$．その他のパラメータは，最大入り次数 $I = 5$，ペナルティ係数 $c = 0.1$ とした．遺伝子ネットワークの同定は，学習器に与える乱数の種を変えながら 10 試行行った．

4.2　結果

獲得されたニューラルネットワークモデルから，3 章で説明した方法を用いてネットワーク構造に関する情報を抽出した．Fig. 3 に，抽出したネットワークの例を示す．図に示されるように，対象ネットワークに含まれる相互作用の多くは正しく推定することができた．しかしながら同定したネットワークには多くの偽陽性の相互作用が含まれていた．偽陽性の相互作用の本数と偽陰性の相互作用の本数は，それぞれ平均して 13.0±3.9 本と 5.5±2.2 本であった．対象ネットワークには 68 本の相互作用が含まれているため，提案手法の敏感度 (sensitivity) と特異度 (specificity) はそれぞれ 0.919 ± 0.030 と 0.993 ± 0.002 となる．なお敏感度と特異度は

$$\text{敏感度} = \frac{TP}{TP + FN},$$

$$\text{特異度} = \frac{TN}{FP + TN},$$

で定義される指標である．ただし TP, FN, TN, FP はそれぞれ，推定されたネットワークに含まれる真陽性，偽陰性，真陰性，偽陽性の相互作用の本数である．敏感度は偽陰性の相互作用の数が，特異度は偽陽性の相互作用の数が少ないほど大きな値となる．

本実験と同じ問題に対して S-system モデルに基づく遺伝子ネットワーク同定法 [3] を適用した場合の敏感度と特異度は，それぞれ 1.000 と 0.870 であった．提案手法は敏感度に関しては S-system を用いた従来手法に劣るものの，特異度に関しては良い性能であった．しかしながら S-system モデルに基づく手法の場合，ほとんどの偽陽性の相互作用に対応するパラメータの絶対値は，真陽性の相互作用に対応するパラメータの絶対値よりも非常に小さい．従って偽陽性の相互作用の多くは無視できると考えられる．以上の点を考慮すると，提案手法は S-system を用いた従来手法よりも同定性能に関して必ずしも良いとは言えない．しかし提案手法の計算時間は S-system を用いた従来手法よりも非常に短かった．提案手法が本研究で使用した問題を解くために必要とした計算時間は約 47.1×30 分 (Pentium IV 2.8GHz) であったのに対して，S-system モデルに基づく従来手法は同じ問題を解くために 73.8×30 時間 (Pentium III 1GHz) が必要である．特に PC クラスタのような大型計算機が

利用できない環境においては，提案手法のように計算時間が短いという特長は有用であると考えられる．

5 おわりに

本研究ではニューラルネットワークモデルを用いた新たな遺伝子ネットワーク同定法を提案した．人工的な遺伝子ネットワーク同定問題への適用を通して，提案手法が敏感度や特異度といった同定性能に関しては従来手法よりも必ずしも勝っているとは言えないものの，計算時間に関しては従来手法よりも非常に短いということを示した．

謝辞

本研究を行うにあたり，有用な助言を頂いたJFEソルデック株式会社 寺沢英樹氏に感謝する．

参考文献

1) W.S. Cleveland, "Robust Locally Weight Regression and Smoothing Scatterplots," *J. of American Statistical Association*, vol. 79, pp. 829-836, 1979.

2) S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi and M. Tomita, "Dynamic Modeling of Genetic Networks using Genetic Algorithm and S-system," *Bioinformatics*, vol. 19, pp. 643-650, 2003.

3) S. Kimura, M. Hatakeyama and A. Konagaya, "Inference of S-system Models of Genetic Networks from Noisy Time-series Data," *Chem-Bio Informatics J.*, vol. 4, pp. 1-14, 2004.

4) 木村, 高橋, 小林, 小長谷, "並列化に適した遺伝的ローカルサーチによる非線形関数最適化," 計測自動制御学会論文集, Vol. 40, pp. 448-457, 2004.

5) S. Kimura, K. Ide, A. Kashihara, M. Kano, M. Hatakeyama, R. Masui, N. Nakagawa, S. Yokoyama, S. Kuramitsu and A. Konagaya, "Inference of S-system Models of Genetic Networks using a Cooperative Coevolutionary Algorithm," *Bioinformatics*, vol. 21, pp. 1154-1163, 2005.

6) S. Kimura, K. Sonoda, S. Yamane, K. Matsumura and M. Hatakeyama, "Function Approximation Approach to the Inference of Neural Network Models of Genetic Networks," *IPSJ Trans. on Bioinformatics*, in press.

7) Y. Maki, D. Tominaga, M. Okamoto, S. Watanabe and Y. Eguchi, "Development of a System for the Inference of Large Scale Genetic Networks," *Proc. of PSB 6*, pp. 446-458, 2001.

8) W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in C, 2nd edn*, Cambridge University Press, 1995.

9) D.E. Rumelhart, J.L. McClelland and PDP Research Group, *Parallel Distributed Processing*, MIT Press, 1986.

10) D. Thieffry, A.M. Huerta, E. Pérez-Rueda and J. Collado-Vides, "From Specific Gene Regulation to Genomic Networks: a Global Analysis of Transcriptional Regulation in *Escherichia Coli*," *BioEssays*, vol. 20, pp. 433-440, 1998.

11) E.O. Voit, *Computational Analysis of Biochemical Systems*, Cambridge University Press, 2000.

12) E.O. Voit and J. Almeida, "Decoupling Dynamical Systems for Pathway Identification from Metabolic Profiles," *Bioinformatics*, vol. 20, pp. 1670-1681, 2004.

13) X. Yao, "Evolving Artificial Neural Networks," *Proc. of the IEEE*, Vol. 87, pp. 1423-1447, 1999.