2011 International Conference on Bioscience, Biochemistry and Bioinformatics
IPCBEE vol.5 (2011) © (2011) IACSIT Press, Singapore

Migration Synchronous Genetic Algorithm for Reverse Engineering

Shinq-Jen Wu*
*Department of Electrical Engineering Da-Yeh University, Chang-Hwa, Taiwan, R.O.C
*jen@mail.dyu.edu.tw

Cheng-Tao Wu
Department of Electrical and Control Engineering National Chiao-Tung University Hsin-Chu, Taiwan, R.O.C.
dau@cn.nctu.edu.tw

***Abstract*–Much research adopts various evolution computation technologies to identify system parameters and structures of highly nonlinear S-system model. They always focus on evolution skills and neglect that the choice of performance index is the key for learning. A suitable performance index not only provides good searching direction but also reduces computation time. In this study, a migration synchronous genetic algorithm (MSGA) is proposed for achieving global optimal search. Twenty eight performances of concentration- or slope-error-based indexes for parameter identification is examined and discussed. When the chosen performance candidates are used for structure identification, only one- or two-steps pruning-operation is necessary. The pruning threshold is set to be 10^{-15} to ensure a safely pruning-action is guaranteed positively.**

Keywords-Inverse engineer; parameter estimation; evolution computation; genetic algorithm.

I. INTRODUCTION

The inverse problem of identifying the topology of a biological network from their time-course response is a cornerstone challenge in a systems biology [1]. Hill and Michaelis-Menten rate modeling is a forward approach and can provide local kinetic information of components. However, repeated modification and an undue amount of experiment data are necessary in parameter identification especially for a system with many substances or reactions involved. S-system structure [2, 3] is another preferred nonlinear dynamic model. This model can uniquely map dynamic interaction onto its parameters, and possesses good generalization characteristics.

Some researchers use gradient-based computation technologies to identify the parameters of a S-system model: Marino and Voit [4] write an algorithm to gradually increase model complexity; Chou *et al.* [5] adopt an alternating regression (AR) method; Vilela *et al.* [1] further propose an eigenvector optimization method to solve convergence issues in AR approach. Kutalik *et al.* [6] adopt Newton-flow analysis.

Recently many researchers infer a gene-regulatory network by stochastic-search intelligent technologies such as genetic programming [7-9], evolutionary algorithms [10], evolution strategies [11], differential evolution [12-15], genetic algorithms [16, 17], simulated annealing [18], radial basis function networks [19], a neural network with particle-swarm-optimization learning [20, 21].

In this paper, we propose a migration synchronous genetic algorithm to ensure global optimal search. Since a performance index determines both learning directions and computation time, we focus on performance-index selection. The chosen performance candidates from parameter identification can further integrated into a multi-objective performance to identify a system structure and parameters simultaneously.

II. METHOD

A. Migration Synchronous Genetic Algorithm

S-system is a well-known canonical nonlinear model for metabolic reaction. Base on biochemical system theory, the net influx (V_i^+) and efflux (V_i^-) of a system is approximated as two power-law functions. Each individual metabolite, protein or gene is described as

$$\dot{X}_i = V_i^+ - V_i^- = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} - \beta \prod_{j=1}^{n+m} X_j^{h_{ij}}, \text{ for i = 1,2,…,n,} \quad (1)$$

where n and m are the numbers of dependent and independent variables, respectively; α_i and β_j are rate constants, g_{ij} and h_{ij} are kinetic orders to denote the interaction from X_j to X_i where a positive value denote excitatory effect and negative for inhibitory effect.

In order to construct such a highly dimensional nonlinear system, we propose a migration synchronous genetic algorithm (MSGA) in Fig.1 to ensure global

*To whom all correspondence should be addressed.

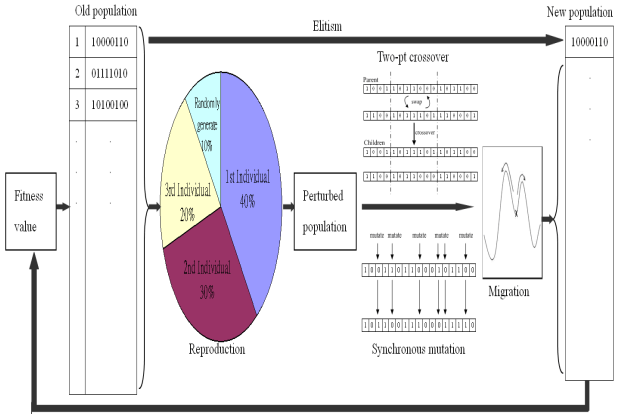

Figure 1. Flowchart of the proposed Migration Synchronous Genetic Algorithm (MSGA).

optimal search. In this algorithm synchronous mutation is used to increase population diversity and migration operation is for widening a search space.

Different from a conventional GA which randomly chooses only one individual for mutation, each individual in MSGA is chosen and given a random mutation probability. This synchronous mutation brings a population excessive diversity but may result in failure in convergence. So adopt an elitism strategy to decrease this effect. Elitism keeps the best-so-far individual to survive for each generation and ensures the best characteristic to pass down. Further, to widen a search space a migration operator is proposed for generating a new diverse population. The degree of population diversity η is defined as

$$\eta = \sum_{i=1}^{NP-1} \sum_{j=1}^{Dim\_I} \frac{temp_{ij}}{Dim\_I * (NP-1)} < \varepsilon_1,$$

$$temp_{ij} = \begin{cases} 0, \ if \left| \dfrac{x_{ij} - x_{bj}}{x_{bj}} \right| < \varepsilon_2, \\ 1, \ otherwise, \end{cases} \quad (2)$$

where $\varepsilon_2 \in [0,1]$ is the tolerance of real-valued gene diversity; x_{ij} and x_{bj} are, respectively, the j-th chromosomes in the i-th individual and the best individual; NP is the number of individuals and Dim_I is the dimension of individuals. $\varepsilon_1 \in [0,1]$ is tolerance threshold of population diversity for migration. If the degree η is small than ε_1, migration is taken and a new chromosome is generated as follows.

$$x_{ij} = \begin{cases} x_{bj} + r_2 \times (x_{j,\min} - x_{bj}), \ if \dfrac{x_{bj} - x_{j,\min}}{x_{j,\max} - x_{j,\min}} > r_1 \\ x_{bj} + r_2 \times (x_{j,\max} - x_{bj}), \ otherwise \end{cases}, \quad (3)$$

where $x_{j,\max}$ and $x_{j,\min}$ are, respectively, the upper and lower bound of the j-th chromosome; r_1, $r_2 \in [0, 1]$ are two random numbers.

TABLE I. PERFORMANCE INDEXES; $J_1^l,...,J_6^l$ ARE DEFINED ON CONCENTRATION ERRORS AND $J_7^l,...,J_{14}^l$ ARE DEFINED ON SLOPE ERRORS; $l =$ M (MAXIMUM) OR S (SUMMATION).

An entire performance		The ith-individual performance
Summation operation $J_l^S = \sum_{i=1}^{NP} J_{li}$	Maximum operation $J_l^M = \max\{J_{l1},...,J_{l,Np}\}$	
J_1^S	J_1^M	$J_{1i} = \left(\dfrac{x^i - x^i_{\exp}}{\max(x^i_{\exp})} \right)^2$
J_2^S	J_2^M	$J_{2i} = t_s \left(\dfrac{x^i - x^i_{\exp}}{\max(x^i_{\exp})} \right)^2$
J_3^S	J_3^M	$J_{3i} = t_a \left(\dfrac{x^i - x^i_{\exp}}{\max(x^i_{\exp})} \right)^2$
J_4^S	J_4^M	$J_{4i} = \sqrt{\left(\dfrac{x^i - x^i_{\exp}}{\max(x^i_{\exp})} \right)^2}$
J_5^S	J_5^M	$J_{5i} = t_s \sqrt{\left(\dfrac{x^i - x^i_{\exp}}{\max(x^i_{\exp})} \right)^2}$
J_6^S	J_6^M	$J_{6i} = t_a \sqrt{\left(\dfrac{x^i - x^i_{\exp}}{\max(x^i_{\exp})} \right)^2}$
J_7^S	J_7^M	$J_{7i} = \left(\dot{x}^i - \dot{x}^i_{\exp} \right)^2$
J_8^S	J_8^M	$J_{8i} = t_s \left(\dot{x}^i - \dot{x}^i_{\exp} \right)^2$
J_9^S	J_9^M	$J_{9i} = t_a \left(\dot{x}^i - \dot{x}^i_{\exp} \right)^2$
J_{10}^S	J_{10}^M	$J_{10i} = \left(\dfrac{\dot{x}^i - \dot{x}^i_{\exp}}{\max\left(\dot{x}^i_{\exp} \right)} \right)^2$
J_{11}^S	J_{11}^M	$J_{11i} = t_s \left(\dfrac{\dot{x}^i - \dot{x}^i_{\exp}}{\max\left(\left\| \dot{x}^i_{\exp} \right\| \right)} \right)^2$
J_{12}^S	J_{12}^M	$J_{12i} = t_a \left(\dfrac{\dot{x}^i - \dot{x}^i_{\exp}}{\max\left(\left\| \dot{x}^i_{\exp} \right\| \right)} \right)^2$
J_{13}^S	J_{13}^M	$J_{13i} = \sqrt{\left(\dot{x}^i - \dot{x}^i_{\exp} \right)^2}$

J_{14}^{S}	J_{14}^{M}	$J_{14i} = \sqrt{\left(\frac{\dot{x}^i - \dot{x}_{\exp}^i}{\max\left(\left\| \dot{x}_{\exp}^i \right\| \right)} \right)^2}$

B. Performance index

A performance index defined as the summation of various errors is a criterion to show how closing the estimated data to a real profile. A performance index determines both search directions and computation time. Therefore, the choice of a performance index is a key point for computation optimization. We use twenty eight performance indexes in Table 1 for S-system's parameter identification. $J_1^l \sim J_9^l$ (l = M or S) are error-related indexes and $J_{10}^l \sim J_{14}^l$ are slope-related indexes. Some with normalization are to ensure comparable competition in different-scales species. x^i, i=1,…,n, is the i-th estimated concentration, $x_{\exp}^i$ is the i-th measured concentration and $\max(x_{\exp}^i)$ is the maximum of measured concentrations; $\dot{x}^i$ is i-th the estimated slope, $\dot{x}_{\exp}^i$ is the i-th measured slope and $\max(\dot{x}_{\exp}^i)$ is the maximum of measured slopes; t_s and t_a are time-weighting factors increasing and decreasing with time, respectively. The performance of these indexes is examined. Those well-performance indexes can chosen as candidates and further integrated into a performance index of multi-objective optimization for biological-system identification; that is, structure- and parameter-identification simultaneously.

III. Results And Discussion

We now apply the MSGA algorithm with different performance criteria to identify the parameters of S-systems for two dry-lab experiments. All computation is performed on an Intel core duo 3.16GHz computer using Microsoft Windows XP. The parameter ranges are [0, 30] for rate constants and [-4, 4] for kinetic orders. 250000 maximum iterations are performed to minimize those error criteria.

A. Dry-lab experiment 1

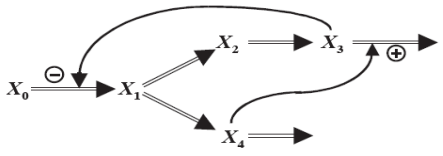

Figure 2. A generic-branch pathway with two regulatory signals: four dependent variables and one constant source x_0.

A generic branch pathway in Fig. 1 is used by Voit and Almeida [17] to demonstrate decoupling dynamic behavior. There are four dependent constituents x_1, x_2, x_3 and x_4, and one constant source x_0. The corresponding S-system is described as

TABLE II. Performance comparison for the generic-branch-pathway system in Fig. 2 (MSAG with 250000 iterations).

Error criterion, $E = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x^i - x_{\exp}^i}{\max(x_{\exp}^i)} \right)^2$			
Sum	Optimal fitness	Max	Optimal fitness
J_1^S	4.5835935E-10	J_1^M	5.0488280E-08
J_2^S	1.1989171E-09	J_2^M	5.7832483E-08
J_3^S	4.5960822E-11	J_3^M	1.0699997E-10
J_4^S	3.4262501E-07	J_4^M	3.0536607E-06
J_5^S	1.1229400E-04	J_5^M	7.7593655E-05
J_6^S	1.8479043E-07	J_6^M	2.2431690E-05
J_7^S	4.7749491E-07	J_7^M	5.5962543E-08
J_8^S	1.0961241E-08	J_8^M	3.9076341E-05
J_9^S	2.9665960E-10	J_9^M	7.8697814E-08
J_{10}^S	4.8324012E-11	J_{10}^M	1.7753159E-09
J_{11}^S	1.2970259E-08	J_{11}^M	4.7993926E-07
J_{12}^S	1.2807886E-10	J_{12}^M	3.5800163E-07
J_{13}^S	1.8329152E-06	J_{13}^M	5.1474803E-06
J_{14}^S	3.9287059E-05	J_{14}^M	6.0951341E-06

TABLE III. Performance comparison for the cascade pathway system in Fig. 3 (MSAG with 250000 iterations).

Error criterion, $E = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x^i - x_{\exp}^i}{\max(x_{\exp}^i)} \right)^2$			
Sum	Optimal fitness	Max	Optimal fitness
J_1^S	5.7719659E-09	J_1^M	7.1645410E-09
J_2^S	1.1456407E-05	J_2^M	7.7100757E-05
J_3^S	5.7623443E-12	J_3^M	5.1791243E-10
J_4^S	1.0183205E-05	J_4^M	1.5735404E-05
J_5^S	1.4208229E-04	J_5^M	7.1913330E-05
J_6^S	1.7395375E-09	J_6^M	6.5643269E-08
J_7^S	5.1764368E-11	J_7^M	1.9206372E-08
J_8^S	1.1109550E-05	J_8^M	5.7691971E-06
J_9^S	2.3851452E-09	J_9^M	9.9108335E-09
J_{10}^S	1.4667368E-07	J_{10}^M	1.5248148E-06
J_{11}^S	6.7262942E-05	J_{11}^M	5.6203930E-05
J_{12}^S	1.4635411E-06	J_{12}^M	3.1027507E-06
J_{13}^S	5.1961041E-09	J_{13}^M	4.9981500E-07
J_{14}^S	1.5947509E-05	J_{14}^M	3.9171596E-05

$$\begin{aligned}
\dot{x}_1 &= \alpha_1 x_3^{g_{13}} x_0 - \beta_1 x_1^{h_{11}}, \\
\dot{x}_2 &= \alpha_2 x_1^{g_{21}} - \beta_2 x_2^{h_{22}}, \\
\dot{x}_3 &= \alpha_3 x_2^{g_{32}} - \beta_3 x_3^{h_{33}} x_4^{h_{34}}, \\
\dot{x}_4 &= \alpha_4 x_1^{g_{41}} - \beta_4 x_4^{h_{44}}.
\end{aligned} \tag{4}$$

Eight sets of experiment data, $x^i_{\exp}, i = 1,\ldots,n$, and the corresponding slope information $\dot{x}^i_{\exp}$ are from the M-M dynamic system in Voit and Almeida [17]. The simulation time of an experiment is from t=0 sec to t=8 sec and sample time is set to be 0.02. An error criterion, $E = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x^i - x^i_{\exp}}{\max(x^i_{\exp})}\right)^2$ is used to examine the performance of those indexes. Table II is the results of these performance indexes. The results show some information. First, summation operations perform much better than the corresponding maximum operations except for J_5^l, J_7^l and J_{14}^l. Second, J_1^l, J_2^l and J_3^l are, respectively, better than J_4^l, J_5^l and J_6^l; J_7^l is better than J_{13}^l and J_{10}^l is better than J_{14}^l. That is, even if errors are always less than one and squaring-error operation will scale down errors to generate a small-quantity criterion for updating, this operation really improves performance. Third, those indexes with normalization have better performance (J_{10}^l, J_{11}^l and J_{12}^l are, respectively, better than J_7^l, J_8^l and J_9^l). Fourth, we find that among the summation-operation set the performance of those weighting-related indexes are always $t_a > 1 > t_s$ (notation ">" denotes better). But if slope normalization is taken ($J_{10}^l \sim J_{12}^l$) then it becomes $1 > t_a > t_s$. Based on these, we suggest to choose J_1^S, J_2^S, J_3^S, J_9^S, J_{10}^S, J_{12}^S, J_3^M and J_{10}^M as learning-criterion candidates.

B. *Dry-lab experiment 2*

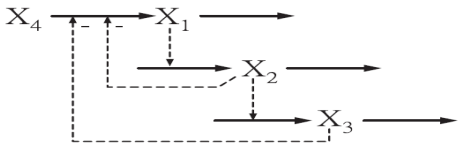

Figure 3. A Cascade pathway with three steps and two feedback signals; three dependent variables x_1, x_2 and x_3, and one source constant x_4.

We now consider another biological system [13] in Fig. 3. The cascade pathway has three dependent constituents x_1, x_2 and x_3, and one constant source x_4. The following is S-system representation:

$$\begin{aligned}
\dot{x}_1 &= \alpha_1 x_2^{g_{12}} x_3^{g_{13}} x_4 - \beta_1 x_1^{h_{11}}, \\
\dot{x}_2 &= \alpha_2 x_1^{g_{21}} - \beta_2 x_2^{h_{22}}, \\
\dot{x}_3 &= \alpha_3 x_2^{g_{32}} - \beta_3 x_3^{h_{33}}.
\end{aligned} \tag{5}$$

We sample the M-M dynamic system [13] to get eight sets of experiment data and estimate slope data. Sample time is 0.02 and sample period is from t=0 sec to t=8 sec. Table 3 is the results after 250000 iterations under different performance indexes.

Table III has similar results as Table II except the third phenomenon. This is due to that there exists a dramatically high slope in x_2 profile. So slope normalization is not suitable. Based on Table 3, we know J_1^S, J_3^S, J_6^S, J_7^S, J_9^S, J_{13}^S, J_1^M, J_3^M and J_9^M are all suitable for being

TABLE IV. STEP 0 SHOWS THE TRUE PARAMETERS IN THE S-SYSTEM OF A CASCADE-PATHWAY NETWORK. STEPS 1 TO 3 SHOW THE INFERRED STRUCTURE AND ESTIMATED PARAMETERS.

Step	Variable	αi	βi	gi1	gi2	gi3	gi4	hi1	hi2	hi3	hi4
	x1	10	5		-0.1	-0.05	1	0.5			
0	x2	2	1.44	0.5					0.5		
	x3	3	7.2		0.5					0.5	
	x1	1.3184830E+01	8.4097688E+00	2.7902452E-17	-6.3869379E-02	-3.6599590E-02	6.5849933E-01	3.2571761E-01	1.0463913E-16	-3.0548374E-17	-5.5032187E-18
1	x2	1.9722535E+00	1.4123528E+00	4.9636622E-01	6.1479326E-18	9.4662519E-19	-2.8361325E-18	1.3525760E-18	5.0099319E-01	-2.2086675E-17	-1.6228509E-17
	x3	3.7347711E+00	7.3040329E+00	7.4718549E-03	3.7992538E-01	3.2379347E-18	4.6108057E-18	6.5479067E-18	3.5107504E-17	3.8270780E-01	-4.4658564E-17
	x1	9.9950744E+00	4.9952221E+00		-1.0006935E-01	-5.0029445E-02	1.0006459E+00	5.0033529E-01			
2	x2	2.0000010E+00	1.4399719E+00	4.9998783E-01					5.0001106E-01		
	x3	2.9853968E+00	7.1886307E+00	-3.2348485E-17	5.0175137E-01					5.0196660E-01	
	x1	9.9999762E+00	4.9999742E+00		-1.0000043E-01	-5.0000252E-02	1.0000041E+00	4.9999861E-01			
3	x2	2.0000040E+00	1.4400034E+00	5.0002488E-01					5.0000203E-01		
	x3	2.9998517E+00	7.1999582E+00		4.9999891E-01					5.0002460E-01	

learning indexes. Those well-performance indexes can be as error-performance and slope-performance candidates to develop a multi-objective-optimization technology. Table 4 shows the results for structure identification. Only two-steps pruning-operation is necessary and truncation threshold is set to be 10^{-15}.

IV. CONCLUSION

The inverse problem of identifying a dynamic biological system from time-series data is a central theme in systems biology. Many optimization methods with various performance indexes have been applied for parameter identification of biochemical systems. They always focus on evolution skills and neglect that the choice of performance index. Performance index is the key for learning optimization. A performance index determines both learning directions and computation time. In this study, twenty eight performances of concentration- or slope-error-based indexes for parameter identification is examined and discussed. The proposed MSGA algorithm and chosen performance candidates from parameter identification can further integrated into a multi-objective performance to identify a system structure and parameters simultaneously.

Acknowledgement: This research is supported by the National Science Council of the R.O.C. under Grant NSC-99-2221-E-212-021-

REFERENCES

[1] Vilela M, Chou IC, Vinga S, Vasconcelos ATR, Voit EO, Almeida JS, "Parameter optimization in S-system models," BMC Syst Biol, 2008, 2:35

[2] Savageau MA, "Biochemical Systems Analysis: a Study of Function and Design in Molecular Biology," Addison-Wesley, Reading, Massachusetts 1976

[3] Voit EO, "Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists," Cambridge University Press, Cambridge, UK 2000

[4] Marino S, Voit EO, "An automated procedure for the extraction of metabolic network information from time series data," Bioinform Comput Biol, 2006, 4:665

[5] Chou IC, Martens H, Voit EO, "Parameter estimation in biochemical systems models with alternating regression," Theor Biol Med Model, 2006, 3:25

[6] Kutalik Z, Tucker W, Moulton V, "S-system parameter estimation for noisy metabolic profiles using newton-flow analysis," IET Syst Biol, 2007, 1:174-80

[7] Sakamoto E, Iba H, "Inferring a System of Differential Equations for a Gene Regulatory Network by using Genetic Programming," CEC: Proc Congress Evolutionary Comput, 2001, 1:720-726

[8] Ando S, Sakamoto E, Iba H, "Evolutionary Modeling Inference of Gene Network," Information Sciences, 2002, 145:237-259

[9] Cho DY, Cho KH, Zhang BT, "Identification of biochemical networks by S-tree based genetic programming," Bioinformatics, 2006, 22:1631–1640

[10] Kimura S, Ide K, Kashihara A, Kano M, Mariko H, Masui R, Nakagawa N, Yokoyama S, Kuramitsu S, Konagaya A, "Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm," Bioinformatics, 2005, 21:1154–1163

[11] Moles CG, Mendes P, Banga JR, "Parameter estimation in biochemical pathways: A comparison of global optimization methods," Genome Res, 2003, 13: 2467-2474

[12] Noman N, Iba H, "Inference of gene regulatory networks using S-system and differential evolution," GECCO: Proc conf Genetic Evolutionary Comput, 2005, 1: 439-446

[13] Tsai KY, Wang FS, "Evolutionary optimization with data collocation for reverse engineering of biological networks," Bioinformatics, 2005, 21:1180-1188

[14] Wu SJ, Wu CT, Lee TT, "Computation Intelligent for Eukaryotic Cell-Cycle Gene Network," Conf Proc IEEE Eng Med Biol Soc, 2006, 1:2017-20

[15] Liu PK, Wang FS, "Hybrid differential evolution with geometric mean mutation in parameter estimation of bioreaction systems with large parameter search space," Comput Chem Eng, 2009, 33: 1851-1860

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

[17] Voit EO, Almeida J, "Decoupling dynamical systems for pathway identification from metabolic profiles," Bioinformatics, 2004, 20:1670–1681

[18] Gonzalez OR, Küper C, Jung K, Naval PC, Mendoza E, Jr, Mendoza E, "Parameter estimation using simulated annealing for S-system models of biochemical networks," Bioinformatics, 2007, 23:480–486

[19] Matsubara Y, Kikuchi S, Sugimoto M, Tomita M, "Parameter estimation for stiff equations of biosystems using radial basis function networks," BMC Bioinformatics, 2006, 7:230

[20] Murata H, Koshino M, Mitamura M, Kimura H, "Inference of S-system models of genetic networks using product unit neural networks," SMC: IEEE Conf Syst Man Cybernetics, 2008, 1390-1395

[21] Xu R, Wunsch II DC, Frank RL, "Inference of Genetic Regulatory Networks with Recurrent Neural Network Models Using Particle Swarm Optimization," IEEE Trans Comput Biol Bioinform, 2007, 4: 1545-5963