32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Transcription factor activity estimation based on particle swarm optimization and fast network component analysis

Wei Chen, Chunqi Chang, *Member, IEEE*, and Y.S. Hung, *Senior Member, IEEE*

Abstract—**Transcription factors (TFs) play an important role in regulating the expression of genes. The accurate measurement of transcription factor activities (TFAs) depends on a series of experimental technologies of molecular biology and is intractable in most practical situations. Some signal processing methods for blind source separation have been applied in the prediction of TFAs from gene expression data. Most of such methods make use of statistical properties of the gene expression data only, leading to the inaccurate detection of TFAs. In contrast, network component analysis (NCA) can provide much improved result through utilizing the structural information of the gene regulatory network. However, the structure of the gene regulatory network, required by NCA, is not available in most practical cases so that NCA is not directly applicable. In this paper, we propose to use particle swarm optimization (PSO) to find the most plausible network structure iteratively from the gene expression data, with the assistance of recently developed fast algorithm for network component analysis (FastNCA). This novel approach to TFA inference can thus take advantage of NCA, even when the required network structure is unknown. The effectiveness of our novel approach has been demonstrated by applications to both simulated data and real gene expression microarray data, in the sense that TFAs can be inferred with high accuracy.**

I. Introduction

TRANSCRIPTION factors (TFs) are protein molecules that regulate the transcription of genes through binding to the promoter region of the genes [1-2]. Transcription is an important biological process which prepares for the generation of proteins (final product of gene expression). The quantitative regulation of gene transcription depends on the transcription factor activities (TFA) [3]. Therefore, obtaining accurate TFA is very important in understanding how the gene expression is regulated.

Currently, measurement of TFA is mainly performed in an *in vivo* system on the interaction among TFs and cis-regulatory elements [4], which is difficult and of high expense, intractable in most practical situations.

Since in most cases it is not applicable to measure TFA directly, we must make inference about TFA indirectly from gene expression data which can be obtained through high throughput microarray technology or more recently next generation sequencing (NGS). With high throughput microarray gene expression data, some signal processing methods for blind source separation, such as principal component analysis (PCA) [5] and independent component analysis (ICA) [6], can be adopted to deduce TFAs. However, TFAs inferred from these methods could be so inaccurate that they are not acceptable for meaningful biological interpretation. This is because the effectiveness of such methods depends on mathematical assumptions (e.g. ICA requires statistical independence of source signals) which are unlikely to be satisfied by the microarray gene expression data.

Contrary to blind source separation approaches which assume unrealistic models of transcriptional gene regulation, network component analysis (NCA) [7] is an alternative approach that does not make unrealistic assumption on the independence of TFAs. Instead, NCA makes use of the structural information of the gene regulatory network, which is available for some species such as yeast through genome-wide location analysis using ChIP-chip technology. Since NCA makes use of relevant biological information and is not dependent on unrealistic statistical assumptions, the TFAs inferred by NCA are much more accurate [7]. However, the original NCA algorithm could be computationally unstable and time consuming and may have multiple local solutions. To overcome these disadvantages, an improved algorithm called FastNCA was developed recently [8].

A critical requirement for inferring TFAs using NCA is that the TF-gene connectivity structure of the gene regulatory network must be known. However, it is quite difficult and expensive to get this required network structure by ChIP-chip experiments. In fact, such structure so far is only available for yeast and *E. Coli* [9]. In most other situations, we have to work without prior knowledge on the network structure and thus NCA cannot be applied directly. In order to deal with such difficulty, we propose to apply NCA on all possible network structures and choose the result of the most plausible one. Though this approach is in general too time consuming to be applicable, we can apply heuristic optimization techniques such as particle swarm optimization (PSO) to deduce a nearly optimal network structure without exhaustive searching. With the help of the fast NCA algorithm FastNCA, such procedure becomes computationally feasible. As a recently developed and fast-developing heuristic optimization technology, PSO has been shown to be effective in various applications. In PSO, the solution of the optimization problem is searched by a swarm of particles with inter-particle communication [10]. It has the advantage of simple operation and fast convergence [11].

Manuscript received April 1, 2010. This work is supported by Hong Kong RGC GRF grant (HKU 710709E) and University of Hong Kong Seed Funding Program for Basic Research (200808159009).

W. Chen, C.Q. Chang, and Y.S. Hung are with Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong (email: {chenwei,cqchang,yshung}@eee.hku.hk)

II. METHODOLOGY

A. NCA and FastNCA

From knowledge of systems biology, we observe that microarray gene expression data can be considered as the integration of TF-gene network and TFAs. This is the foundation of network component analysis (NCA). NCA assumes the following model:

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

where $\mathbf{X}$ is the microarray data, $\mathbf{A}$ is the matrix of TF-gene network, and $\mathbf{S}$ is the TFAs matrix. When noise, denoted as Γ, is included in microarray data, the model of NCA will be changed to:

$$\mathbf{Y} = \mathbf{X} + \Gamma = \mathbf{AS} + \Gamma \tag{2}$$

According to biological knowledge, the connectivity matrix $\mathbf{A}$ is very sparse with most elements being "0", which means that each gene is regulated by only a small number of TFs. Such a sparse structure of $\mathbf{A}$, if known, can be utilized by the NCA algorithm to deduce the TFAs ($\mathbf{S}$) from the microarray gene expression data ($\mathbf{Y}$) using alternating least squares (ALS) method [12].

With the biological significance of source signals being considered, NCA is a signal separation method suitable for transcription factor activity inference from microarray gene expression data. However, the original ALS-based NCA algorithm has some drawbacks such as instability and multiple local solutions. An improved algorithm for NCA, called FastNCA, was then developed to overcome these drawbacks [8]. By using matrix factorization instead of ALS iteration, FastNCA is much faster and more robust.

B. Objective Function

In this paper, FastNCA is applied to the microarray data for estimating the connectivity matrix $\mathbf{A}$ and the TFAs matrix $\mathbf{S}$, assuming a sparse connectivity structure Z, a binary matrix with 1 representing a connection and 0 for no connection. To find the most plausible unknown connectivity structure, we minimize the objective function defined as below:

Given Z, suppose $\mathbf{A}$ and $\mathbf{S}$ are determined using FastNCA subject to $\mathbf{A}$ conformal with the structure Z, so that $\mathbf{A}$ and $\mathbf{S}$ can be regarded as functions of Z through the FastNCA procedure, i.e. $\mathbf{A} = \mathbf{A}(Z)$ and $\mathbf{S} = \mathbf{S}(Z)$. Hence,

$$f(Z) = \frac{\sum_{i=1}^{N} \frac{\left\| \mathbf{Y}_i - (\mathbf{A}(Z)\mathbf{S}(Z))_i \right\|_F}{\left\| \mathbf{Y}_i \right\|_F}}{N} \tag{3}$$

where $\mathbf{Y}_i$ is the i^{th} row of $\mathbf{Y}$, $(\mathbf{A}(Z)\mathbf{S}(Z))_i$ is the i^{th} row of $\mathbf{A}(Z)\mathbf{S}(Z)$, $\| \ \|_F$ is Frobenius norm, and N is the number of rows of $\mathbf{Y}$ (number of genes).

The objective function $f(Z)$ describes the level of deviation between the deduced data ($\mathbf{A}(Z)\mathbf{S}(Z)$) and the real data ($\mathbf{Y}$). Therefore, a smaller $f(Z)$ implies a better estimate of the connectivity matrix ($\mathbf{A}$) and TFAs matrix ($\mathbf{S}$).

C. Particle Swarm Optimization (PSO)

PSO will be applied to find an optimal connectivity structure Z that minimizes the objective function $f(Z)$ as stated in (3). This is described as follows:

1. Initialize Z and its velocity v for each of the M particles, and set both the optimal connectivity structure of each specific particle, $pbest_k$, and the optimal connectivity structure of the whole swarm, $gbest_k$, as Z.
2. In the kth iteration, update Z and v for each particle as

$$v_{k+1} = v_k + \varphi_1(pbest_k - Z_k) + \varphi_2(gbest_k - Z_k) \tag{4}$$

$$Z_{k+1} = Z_k + v_{k+1} \tag{5}$$

where φ_1 and φ_2 are cognitive confidence coefficients determining the particle's tracking tendency on the local and global optimum, respectively, chosen as random numbers with uniform distribution over the interval [0, 2]. Such settings for φ_1 and φ_2 guarantee convergence [11]. The structure matrix is then transformed to a binary matrix by applying a threshold that keeps only a certain pre-defined percentage, denoted as s, of all possible connections.
3. For each particle, apply FastNCA based on the connectivity structure Z_{k+1} to estimate $\mathbf{A}$ and $\mathbf{S}$, update $pbest_{k+1} = Z_{k+1}$ if $f(Z_{k+1}) < f(pbest_k)$, and update $gbest_{k+1} = pbest_{k+1}$ if $f(pbest_{k+1}) < f(gbest_k)$.
4. Iterate Procedure 2 and Procedure 3 until convergence.

The procedure is illustrated in the following example.

I. Initialize 5 particles and their related matrices. As an example, the first particle is initialized as:

$$Z_0 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad v_0 = \begin{pmatrix} -0.9556 & -0.7866 & 0.2235 \\ 0.2451 & -0.6893 & -0.1158 \\ -0.5006 & 0.3324 & 0.9652 \end{pmatrix}, \text{ and}$$

$pbest_0 = Z_0$. Set $gbest_0 = Z_0$.

II. 1st iteration based on (4) and (5):

$$Z_1 = \begin{pmatrix} 1.0701 & -0.0398 & 0.3401 \\ 0.3886 & 0.6873 & 0.1643 \\ -0.0699 & 0.5360 & 2.5187 \end{pmatrix};$$

Assuming the sparse ratio s to be 0.35, Z_1 is further transformed to the binary connectivity structure.

$$Z_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

III. Apply FastNCA based on Z_1 to get the estimate of $\mathbf{A}$ and $\mathbf{S}$, and then update $pbest_1$ and $gbest_1$ based on the value of the objective function.

IV. Iterate Procedure II and Procedure III until convergence.

Note that the prerequisite of calculating $f(\mathbf{A}(Z),\mathbf{S}(Z))$ is to know the particular values of $\mathbf{A}(Z)$ and $\mathbf{S}(Z)$, which are deduced by a NCA algorithm. Because the number of particles and iterations is large, the fast execution and precise prediction of the algorithm is important. The low algorithm complexity, high accuracy of deduction and the fast speed of FastNCA make it possible to adopt heuristic optimization (i.e.PSO) to find out a suitable TF-gene network structure.

D. Framework of the algorithm

The whole execution procedure is illustrated in Fig.1.

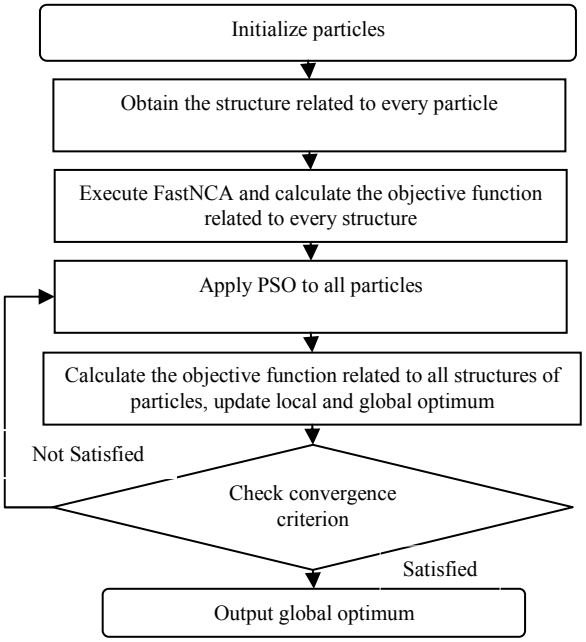

Fig. 1. Execution flow graph of the whole algorithm

III. Experiments and Results

A. Data Description

In our experiments, we use both simulated microarray data and real microarray data. The sizes of the two types of microarray data are both 100×25. Hence, the microarray data have 100 genes and 25 sample points. The numbers of TFs related to the two microarray data are both 16.

B. Results on Simulated Data

Based on our test of the optimization performance of PSO, 300 particles and 200 iterations are adopted in our experiments. The simulated microarray data do not include noise. The sparse ratio of the simulated TF-gene network is $s = 0.0875$. After convergence of the PSO, we get the estimate of both the connectivity matrix $\mathbf{A}$ and the TFA matrix $\mathbf{S}$, denoted as $\hat{\mathbf{A}}$ and $\hat{\mathbf{S}}$, respectively. Because the order of the rows of $\hat{\mathbf{S}}$ and that of $\mathbf{S}$ may be different, we need to determine which row of $\hat{\mathbf{S}}$ is the estimate of a particular original TFA. It can be done by finding the row of $\hat{\mathbf{S}}$ that is mostly correlated to this specific TFA. Mathematically, the estimate of the k^{th} original TFA s_k can be determined as

$$g_k = \arg\max_{\mathbf{s}\in\{\hat{\mathbf{s}}_1,\hat{\mathbf{s}}_2,\cdots,\hat{\mathbf{s}}_M\}} \left|corr(\mathrm{s_k},\mathbf{s})\right| \tag{6}$$

where $\hat{\mathbf{s}}_j$ is the j^{th} estimated TFA, M is the number of TFAs, and $corr(\mathrm{s_k},\hat{\mathbf{s}}_j)$ is the correlation coefficient between $\mathrm{s_k}$ and $\hat{\mathbf{s}}_j$. Replace $corr(\mathrm{s_k},\hat{\mathbf{s}}_j)$ with rho. The result is shown in Table I.

TABLE I
CORRELATION RESULTS FOR SIMULATED DATA

$\mathrm{s_k}$	g_k	*rho*
s_1	$\hat{\mathbf{s}}_{12}$	0.8889
s_2	$\hat{\mathbf{s}}_{14}$	0.9613
s_3	$\hat{\mathbf{s}}_{16}$	0.9945
s_4	$\hat{\mathbf{s}}_4$	0.8149
s_5	$\hat{\mathbf{s}}_8$	0.8679
s_6	$\hat{\mathbf{s}}_2$	0.9896
s_7	$\hat{\mathbf{s}}_{12}$	0.7498
s_8	$\hat{\mathbf{s}}_6$	0.8609
s_9	$\hat{\mathbf{s}}_8$	0.9803
s_{10}	$\hat{\mathbf{s}}_{15}$	0.9990
s_{11}	$\hat{\mathbf{s}}_{10}$	0.9917
s_{12}	$\hat{\mathbf{s}}_5$	0.9456
s_{13}	$\hat{\mathbf{s}}_1$	0.9768
s_{14}	$\hat{\mathbf{s}}_9$	0.9889
s_{15}	$\hat{\mathbf{s}}_3$	0.9733
s_{16}	$\hat{\mathbf{s}}_{13}$	0.9955

From Table I, we see that there exist repetitions for two estimated TFAs – $\hat{\mathbf{s}}_{12}$ and $\hat{\mathbf{s}}_8$, each of which should be assigned to the estimate of only one of the original TFA. After eliminating repetitions, we get estimates for 14 TFAs, with high correlation with their original TFAs (absolute correlation coefficient above 0.8). The detection ratio is

87.5% (14/16).

C. Results on Real Microarray Data

The real microarray data is from a microarray experiment of *E. Coli* [13]. The biological background is that *E. Coli* carbon source transmits from glucose to acetate. The sparse ratio of TF-gene connectivity matrix is 0.0388. The noise ratio is 0.3031.

We run PSO 4 times. The result is shown in Table II. We observe that there always exist TFAs which are not detected in each experiment. However, only TFA15 is not detected in all experiments. Through integrating the results of 4 experiments, we detect other TFAs. So the detection ratio of original TFAs is 93.75% (15/16). The detected TFAs (circle) are compared with their corresponding original TFAs (star) in Fig 2.

TABLE II
PSO PREDICTION FROM E. COLI MICROARRAY DATA

Experiment No.	1	2	3	4
Number of TFAs not detected	4	5	3	5
List of not detected TFAs	TFA1, TFA4 TFA15, TFA16	TFA3, TFA4, TFA8, TFA9, TFA15	TFA2, TFA8, TFA15	TFA7, TFA9, TFA12, TFA14, TFA15

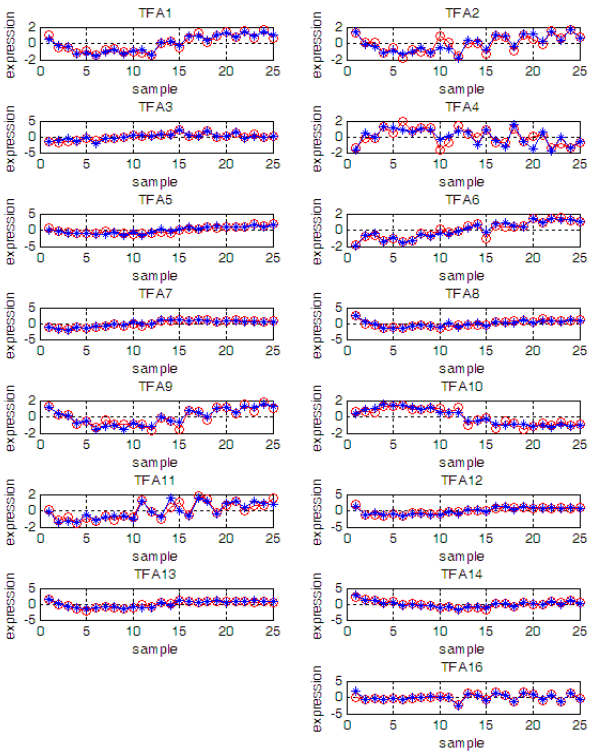

Fig. 2. Comparison of detected TFAs with real TFAs

From Fig. 2, we can see that the predicted normalized TFAs almost overlap with the real normalized TFAs.

IV. Conclusion

In this paper, we consider the problem of estimating transcription factor activities (TFAs) from microarray gene expression data by integrating FastNCA (a recently developed fast network component analysis algorithm) and particle swarm optimization (PSO) to search for the optimal TF-gene network and estimate TFAs simultaneously. Experiments on both simulated data and real microarray data demonstrate that our method can estimate the unknown TFAs accurately. Our future work is to improve the method in order to make it work robustly for more complicated biological networks.

References

[1] D. S. Latchman, "Transcription factors: an overview," *Int. J. Biochem. Cell Biol.,* vol. 29, pp. 1305-1312, Dec. 1997.

[2] M. Karin, "Too many transcription factors positive and negative interactions," *New Biol.,* vol. 2, pp. 126-131, Feb. 1990.

[3] P. Jorgensen and M. Tyers, "The fork'ed path to mitosis," *Genome Biol.,* vol. 1, pp. 1022.1-1022.4, Sep. 2000.

[4] J. Locker, *Transcription Factors* (chapter 2). San Diego: Academic Press, 2001.

[5] O. Alter, P.O. Brown and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc. Natl Acad. Sci,* vol. 97, pp. 10101-10106, August. 2000.

[6] S.I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biology,* vol. 4, pp. R76-R96, Oct. 2003.

[7] J. Liao, R. Boscolo, Y.L. Yang, L.M. Tran, C. Sabatti and V.P. Roychowdhury, "Network component analysis: Reconstruction of regulatory signals in biological systems," *Proc. Natl Acad. Sci,* vol. 100, pp. 15522-15527, Dec. 2003.

[8] C.Q. Chang, Z. Ding, Y.S. Hung and P.C.W. Fung, "Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data," *Bioinformatics,* vol. 24, pp. 1349-1358, April. 2008.

[9] M. Zheng, L.O. Barrera, B. Ren and Y.N. Wu, "Chip-chip: data, model, and analysis," *Biometrics,* vol. 63, pp. 787-796, Sep. 2007.

[10] J. Kennedy and R. Eberhart, "Particle swarm optimization," In *Proceedings of 2nd IEEE International Conference on Neural Networks*, pp. 1942-1948, Hawaii, 1995.

[11] M. Clerc and J. Kennedy, "The particle swarm – explosion, stability, and convergence in a multidimensional complex space," *Evolutionary Computation, IEEE Transactions on,* vol. 6, pp. 58-73, Feb. 2002.

[12] L. M. Tran, M.P. Brynildsen, K.C. Kao, J.K. Suen and J.C. Liao, "gNCA: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation," *Metabolic Engineering,* vol. 7, pp. 128-141, March. 2005.

[13] K.C. Kao, Y.L. Yang, R. Boscolo, C. Sabatti, V. Roychowdhury and J.C. Liao, "Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis," *Proc. Natl Acad. Sci,* vol. 101, pp. 641-646, Dec. 2003.