Nonnegative Network Component Analysis by Linear Programming for Gene Regulatory Network Reconstruction

Chunqi Chang[1], Zhi Ding[2], and Yeung Sam Hung[1]

[1] Department of Electrical and Electronic Engineering
The University of Hong Kong, Pokfulam Road, Hong Kong
{cqchang,yshung}@eee.hku.hk
[2] Department of Electrical and Computer Engineering
University of California, Davis, CA 95616, USA
zding@ucdavis.edu

Abstract. We consider a systems biology problem of reconstructing gene regulatory network from time-course gene expression microarray data, a special blind source separation problem for which conventional methods cannot be applied. Network component analysis (NCA), which makes use of the structural information of the mixing matrix, is a tailored method for this specific blind source separation problem. In this paper, a new NCA method called nonnegative NCA (nnNCA) is proposed to take into account of the non-negativity constraint on the mixing matrix that is based on a reasonable biological assumption. The nnNCA problem is formulated as a linear programming problem which can be solved effectively. Simulation results on spectroscopy data and experimental results on time-course microarray data of yeast cell cycle demonstrate the effectiveness and anti-noise robustness of the proposed nnNCA method.

1 Introduction

Gene regulatory network reconstruction is an important research problem in systems biology where structure and dynamics of cellular functions are of interest. Since gene regulatory network reveals the underlying inter-dependency and cause-and-effect relationship between various cellular functions, it has become one of the key areas of interest in systems biology.

Gaining a quantitative understanding of gene regulation is of vital importance in modern biology. In general, the problem relates to how and where a particular gene is expressed, often under combinatorial control of regulatory proteins known as transcription factors (TF). The dynamics of gene expression levels, i.e. the mRNA concentrations, in a cell can be measured simultaneously by microarray technology for all genes in the genome in a form of multi-channel time-course signal. However, the dynamics of the regulatory signal, i.e., the transcription factor activities (TFA), cannot be measured by the current technology. In addition, the control strength of a regulatory transcription factor to a gene is

another important aspect in the gene regulatory network, which is unfortunately also unknown. In order to understand the entire gene regulatory network, we need to reconstruct the transcription factor activities and the matrix of control strengths from the gene expression measurements. This is a highly challenging inverse problem, especially because microarray data are always extremely noisy.

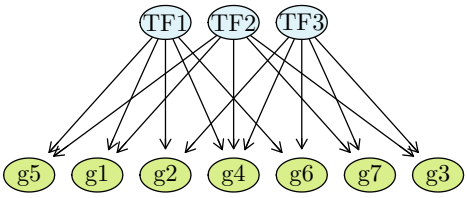

Fig. 1. Gene regulatory network

A conceptual gene regulatory network with 7 genes (g1 ... g7) and 3 transcription factors (TF1, TF2, TF3) is illustrated by Fig 1. In general, gene regulation processes are dynamic and nonlinear. It is assumed that the time scale of change of transcription factor activities (TFA) is much greater than that of gene expression. Therefore, mRNA levels at most time are in a quasi-steady state, and thus at this quasi-steady state the dynamic model becomes approximately instantaneous. In addition, the nonlinear dependence of gene expression on the TFAs is approximately log-linear [1]. Therefore, when the gene expressions are expressed as log-ratios, the model becomes $\mathbf{X} = \mathbf{AS} + \mathbf{\Gamma}$, where $\mathbf{X}, \mathbf{A}, \mathbf{S}$, and $\mathbf{\Gamma}$ are gene expression, connectivity matrix, TFAs, and noise, respectively.

Under the above instantaneous linear model, gene regulatory network reconstruction is a blind source separation problem, and independent component analysis (ICA) [2, 3] had been applied to solve the problem. However, in this special blind source separation problem, the source signals are dependent in general. Therefore, the networks inferred from ICA are not accurate and do not conform to the realistic network structure which is of known sparse structure. There are some other approaches that work on dependent sources [4, 5], but the underlying assumptions do not readily apply to gene regulatory network.

It was noted in the pioneering work of [6] that if the sparse network structure is known and satisfies some conditions, then the network can be uniquely reconstructed if there is no noise, and a method called network component analysis (NCA) was proposed to find a connectivity matrix (conformable to the known structure) and a set of transcription factor activities that best fit the model by using alternating least squares (ALS). The original ALS approach to NCA suffers from drawbacks of instability, inefficiency, and local convergence. Tikhonov regularization has been proposed to overcome the problem of instability [7], but on the other hand it is computationally even more inefficient. Then in [8, 9] we proposed a more efficient and more effective method called FastNCA that successfully overcomes all the three drawbacks. FastNCA provides a closed-form solution to estimate the connectivity matrix and TFAs through fitting the model by a series of subspace projections.

Although NCA is by far one of the most effective approaches to gene regulatory network reconstruction, existing algorithms, however, lack accuracy and consistency. This motivates us to improve NCA and develop more accurate and robust network reconstruction methods by incorporating some prior information; thereby greatly enhance our ability to accurately reconstruct the networks. Specifically, in this paper we will assume that the entries of the connectivity matrix $\mathbf{A}$ are all nonnegative, and develop a linear programming method to solve the NCA problem with non-negativity constraints on the connectivity matrix.

2 Nonnegative Network Component Analysis

Recall the instantaneous linear gene regulation model mentioned in Section 1.

$$\mathbf{X} = \mathbf{AS} + \mathbf{\Gamma}. \tag{1}$$

Our aim is to estimate the connectivity matrix $\mathbf{A}$ and the TFAs $\mathbf{S}$ from the time-course microarray data $\mathbf{X}$.

Assume that in the network we have N genes and M transcription factors, and the length of time series is K. Then the dimension of $\mathbf{X}$ and $\mathbf{\Gamma}$ is $N \times K$, the dimension of $\mathbf{A}$ is $N \times M$, and the dimension of $\mathbf{S}$ is $M \times K$.

As proved in [6, 9], this inverse problem has a unique solution (up to scaling ambiguity of the TFAs) in the noise-less case if the following *NCA criteria* are satisfied:

(i) $\mathbf{A}$ has full column rank;
(ii) when any one column of $\mathbf{A}$ is removed together with the rows corresponding to the nonzero entries of this column, the resulting sub-matrix has full column rank;
(iii) $\mathbf{S}$ has full row rank.

The blind source separation problem or inverse problem of estimating $\mathbf{A}$ and $\mathbf{S}$ from $\mathbf{X}$ based on Eq. (1) and the three NCA criteria is called network component analysis (NCA).

Though the three NCA criteria are enough to estimate the connectivity matrix when there is no noise in the model, additional constraints, if can be used in the algorithm, will certainly yield a more robust estimate in practice when noise is inevitable.

According to [10], we have the biological knowledge that most likely a transcription factor will have the same effect (either positive or negative) on all its regulated genes. This knowledge means that the entries within the same column of $\mathbf{A}$ should have the same sign, and by moving the sign to the corresponding row of $\mathbf{S}$ we can assume that all non-zero entries of $\mathbf{A}$ are positive, i.e., all entries of $\mathbf{A}$ are nonnegative. In other words, if a transcription factor regulates the genes negatively, then we can simply multiply its transcription factor activity (TFA) by -1 and this sign-inversed TFA will regulate the genes positively.

We call the network component analysis problem under the additional nonnegativity constraint on the connectivity matrix $\mathbf{A}$ nonnegative network component analysis (nnNCA).

3 A Linear Programming Approach to nnNCA

If there is no noise in the NCA model Eq. (1), i.e., $\mathbf{X} = \mathbf{AS}$, then the range of $\mathbf{X}$ is equal to the range of $\mathbf{A}$ since $\mathbf{A}$ is of full column rank and $\mathbf{S}$ is of full row rank. Because $\mathbf{X}$ is known to us, we can get the orthonormal basis matrix of the range space of $\mathbf{X}$, denoted by $\bar{\mathbf{X}} = \text{orth}\{\mathbf{X}\}$, and we have

$$\bar{\mathbf{X}} = \text{orth}\{\mathbf{X}\} = \text{orth}\{\mathbf{A}\} \ . \tag{2}$$

With $\bar{\mathbf{X}}$ known, we can get further its orthogonal complement

$$\mathbf{C} = \bar{\mathbf{X}}^{\perp} \tag{3}$$

such that

$$\mathbf{C}^T\bar{\mathbf{X}} = \mathbf{0} \ . \tag{4}$$

From Eq. (2) and (4), we have

$$\mathbf{C}^T\mathbf{A} = \mathbf{0} \ . \tag{5}$$

Since $\mathbf{C}$ can be estimated from the known time-course microarray data $\mathbf{X}$, the connectivity matrix $\mathbf{A}$ can be obtained by solving the systems of equation consisting of Eq. (5) and the NCA criterion (ii) described in Section 2.

In general, if there is noise in the model Eq. (1), singular value decomposition (SVD) [11] will be applied to $\mathbf{X}$ to obtain a robust estimation of $\mathbf{C}$. We write $\mathbf{X}$ in the standard SVD form as follows

$$\mathbf{X} = \mathbf{U\Sigma V}^T \ . \tag{6}$$

Partition $\mathbf{U}$, the matrix of left singular vectors of $\mathbf{X}$, as

$$\mathbf{U} = [\mathbf{U}_S \quad \mathbf{U}_N] \ , \tag{7}$$

where $\mathbf{U}_S$ contains the first M columns of $\mathbf{U}$ and $\mathbf{U}_N$ contains the remaining $N - M$ columns of $\mathbf{U}$. Here we state again that N is the number of genes and M is the number of transcription factors. The matrices $\mathbf{U}_S$ and $\mathbf{U}_N$ are called the signal subspace and noise subspace of $\mathbf{X}$, respectively. Then we get a robust estimate of $\mathbf{C}$ as

$$\mathbf{C} = \mathbf{U}_N \ . \tag{8}$$

Note that in the noisy case with $\mathbf{C}$ estimated by Eq. (8), Eq. 5 does not hold in general. To estimate $\mathbf{A}$ in such case, instead of solving a system of equations as in the noiseless case, we minimize all entries of $\mathbf{C}^T\mathbf{A}$ with the constraints imposed by the NCA criteria and non-negativity of $\mathbf{A}$ by solving the following constrained optimization problem via linear programming.

Let $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_M]$ and $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_{N-M}]$, where $\mathbf{a}_i$ is the ith column of $\mathbf{A}$ and $\mathbf{c}_i$ is the ith column of $\mathbf{C}$. In addition, we denote I as the indices where the entries of $\mathbf{A}$ are zeros, and J as the indices where the entries of $\mathbf{A}$ are

nonzero (positive). Then, we can estimate the connectivity matrix $\mathbf{A}$ by solving the following linear programming problem

$$\min t \quad \text{s.t.} \quad -t < \mathbf{c}_i^T \mathbf{a}_j < t, \quad \mathbf{A}(I) = 0, \quad \mathbf{A}(J) > 0, \quad \sum_{n=1}^{N} a_{n,j} = L_j \ , \quad (9)$$

for $i = 1, \ldots, N - M$ and $j = 1, \ldots, M$, where L_j is the number of nonzero entries of $\mathbf{a}_j$, $\mathbf{A}(I) = 0$ means the entries of $\mathbf{A}$ indexed by I are zero, and $\mathbf{A}(J) > 0$ means that the entries of $\mathbf{A}$ indexed by J are positive. The last constraint $\sum_{n=1}^{N} a_{n,j} = L_j$ is imposed to avoid the trivial solution $t = 0$ with $\mathbf{A} = \mathbf{0}$, and the right-hand-side L_j is chosen to make the solution conformable to the normalization strategy adopted in [6, 9].

For real biological systems, the connectivity matrix $\mathbf{A}$ may be very sparse. Problem (9) can be simplified by considering only the non-zero (positive) entries of $\mathbf{A}$. Denote $\tilde{\mathbf{a}}_j$ as the vector of nonzero entries of the jth column of $\mathbf{A}$, and $\tilde{\mathbf{c}}_{i,j}$ as the vector of entries of the ith column of $\mathbf{C}$ corresponding to the nonzero entries of the jth column of $\mathbf{A}$. Then problem (9) can be simplified as

$$\min t \quad \text{s.t.} \quad -t < \tilde{\mathbf{c}}_{i,j}^T \tilde{\mathbf{a}}_j < t, \quad \tilde{\mathbf{a}}_j > 0, \quad \sum_{n=1}^{L_j} \tilde{a}_{n,j} = L_j \ . \quad (10)$$

Problem (10) is a linear programming problem [12, 13], and can be solved by standard algorithms implemented in many mature linear programming software packages. In this paper we use the GLPK package [14] to solve the linear programming problem.

4 Results

4.1 Simulation results

To test the proposed linear programming based nnNCA algorithm, we use the simulation data described in [6]. In this simulation data, the conceptual gene regulatory network shown in Fig 1 is simulated by the mixing of spectroscopy signals. Three transcription factors are simulated by three kinds of hemoglobins, and the expression of seven genes are simulated by the spectroscopy of the mixed hemoglobins with different mixing ratios that are conformable to the structure of the network in Fig 1. The length of each measured spectroscopy signal is 321. It has been shown in [6, 8] that conventional blind source separation methods, based on either higher-order statistics or second-order statistics, cannot recover the true pure spectroscopy signal of the three hemoglobins, while the NCA methods, either by ALS algorithm or FastNCA, can extract the source signals perfectly.

To demonstrate the effectiveness of the nnNCA method, we apply it to the spectroscopy mixtures, and the estimated source signals are shown in the middle column of Fig 2. It is found that the results of nnNCA in this case is almost exactly the same as that of FastNCA [9], with negligible difference of the order of

numerical calculation error. Then independent white Gaussian noises are added to the mixing mixture to the level of SNR=5dB (signal to noise ratio). Both FastNCA and nnNCA are applied to the noisy data, and a great number of Monte Carlo runs are performed. Though nnNCA is not guaranteed to always perform better than FastNCA, the overall performance of nnNCA is superior to FastNCA. The estimated source signals by nnNCA and FastNCA from a typical Monte Carlo run are compared in Fig 2. The results demonstrate that the inclusion of the constraints on the positivity of the entries of the connectivity matrix **A** makes the NCA method more robust to measurement noise.

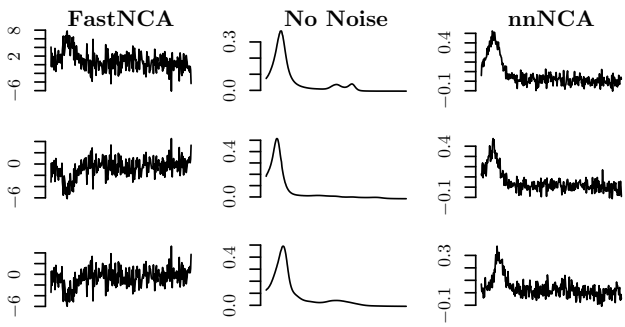

Fig. 2. Simulation results, SNR=5 dB

4.2 Experimental results

To test the improved robustness to noise of the proposed nnNCA over conventional NCA methods for the analysis of real biological networks, we apply it to analyze the time-course microarray data of yeast cell cycle in [15], and compare the results with that of FastNCA. In this study, there are three experiments with different synchronization methods represented by alpha, cdc15, and elutriation. The time-course microarray data contains 6178 genes and 56 time points. In this analysis, we are interested in the 11 transcription factors that are known to regulate the expression of genes that are involved in the cell cycle process. In order to apply NCA to recover the TFAs of these 11 transcription factors, we need to work on a sub-network that contains the 11 TFs. For this purpose we construct a network that contains only these 11 TFs and those genes that are regulated by only these 11 TFs, based on the network topology information inferred from the ChIP-chip experiment in [16]. Both nnNCA and FastNCA are then applied to this network and the estimated TFAs are displayed in Fig 3 shoulder to shoulder for the ease of comparison, where the curves in black are for FastNCA and the curves in blue are for nnNCA.

The experiment "alpha" contains 2 cell cycles, "cdc15" contains 3 cycles, and "elutriation" contains 1 cycle. The names of the TFs are shown on the right-

hand side of the figure. Since these 11 TFs regulate the cell cycle process, it is expected that the TFA of them should be cyclic, and we know that the gene expression signal of many of them are not cyclic at all [6, 9]. We observe that the results of these two method are similar in general, and there is no case that the estimated TFA by nnNCA is less cyclic than that by FastNCA, while in some cases, such as `FKH1` for all 3 experiments and `MCM1` for "alpha", the results of nnNCA are significantly more cyclic than that by FastNCA. These demonstrate the superior robustness of nnNCA for the analysis of real biological networks.

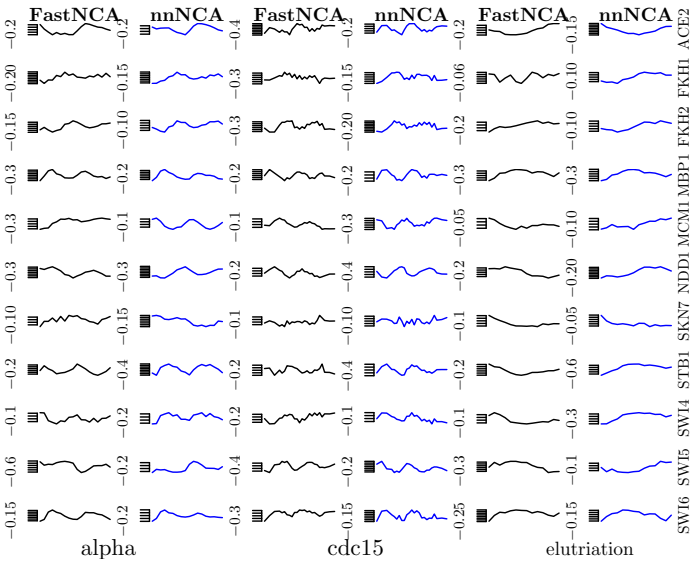

Fig. 3. Analysis of the yeast cell cycle data

5 Discussion and Conclusion

Gene regulatory network reconstruction is an inverse problem similar to blind source separation, but conventional blind source separation methods cannot be applied because the source signals are dependent in general. Network component analysis (NCA) is a suitable source separation method for this specific problem. In this paper a new NCA method, nnNCA, is developed that incorporates a reasonable biological knowledge. It is demonstrated by both simulation and experimental results that nnNCA is more robust. The linear programming based algorithm is also very fast, slightly slower than but comparable to FastNCA, and much faster than the original ALS based NCA. The developed method may also find its applications in some other similar signal processing problems.

Acknowledgements

This work is supported by the University of Hong Kong Seed Funding for Basic Research.

References

1. Savageau, M.A.: Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology. Addison-Wesley, Reading, MA (1976)
2. Liebermeister, W.: Linear modes of gene expression determined by independent component analysis. Bioinformatics **18**(1) (2002) 51–60
3. Lee, S.I., Batzoglou, S.: Application of independent component analysis to microarrays. Genome Biology **4**(11) (2003) R76
4. Abrard, F., Deville, Y.: Blind separation of dependent sources using the "time-frequency ratio of mixtures" approach. Proceedings of Seventh International Symposium on Signal Processing and Its Applications (July 2003) 81–84
5. Chang, C.Q., Ren, J., Fung, P., Hung, Y., Shen, J., Chan, F.: Novel sparse component analysis approach to free radical EPR spectra decomposition. Journal of Magnetic Resonance **175**(2) (2005) 242–255
6. Liao, J.C., Boscolo, R., Yang, Y.L., Tran, L.M., Sabatti, C., Roychowdhury, V.P.: Network component analysis: Reconstruction of regulatory signals in biological systems. Proceedings of the National Academy of Sciences of the United States of America **100**(26) (2003) 15522–15527
7. Tran, L.M., Brynildsen, M.P., Kao, K.C., Suen, J.K., Liao, J.C.: gnca: A framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. Metabolic Engineering **7**(2) (2005) 128–141
8. Chang, C.Q., Ding, Z., Hung, Y.S., Fung, P.C.W.: Fast network component analysis for gene regulation networks. In: Proc. 2007 IEEE International Workshop on Machine Learning for Signal Processing, Thesaloniki, Greece (Aug 2007)
9. Chang, C.Q., Ding, Z., Hung, Y.S., Fung, P.C.W.: Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data. Bioinformatics **24**(11) (2008) 1349 – 1358
10. Alon, U.: An Introduction to Systems Biology: Design Principles of Biological Circuits. Chapman & Hall/CRC (2007)
11. Golub, G.H., van Loan, C.F.: Matrix Computation. 3rd edn. The Johns Hopkins University Press (1996)
12. Dantzig, G.: Linear Programming and Extensions. Princeton Univ Pr (1963)
13. Luenberger, D.: Introduction to Linear and Nonlinear Programming. Addison-Wesley Pub. Co. (1973)
14. GNU: GLPK (GNU Linear Programming Kit) [web page and software]. http://www.gnu.org/software/glpk/glpk.html. (November 2008)
15. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., D., B., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell **9**(12) (1998) 3273–3297
16. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A.: Transcriptional regulatory networks in saccharomyces cerevisiae. Science **298**(5594) (2002) 799–804