Latent Variable and nICA Modeling of Pathway Gene Module Composite

Ting Gong[1], Yitan Zhu[1], Jianhua Xuan[1, 2], Huai Li[3], Robert Clarke[4], Eric P. Hoffman[5], Yue Wang[1]
[1]Dept of ECE, Virginia Polytechnic Institute and State University, Arlington, VA, USA
[2]Dept. of EECS, the Catholic University of America, Washington, DC, USA
[3]Bioinformatics Unit, RRB, National Institutes of Health, Baltimore, MD, USA
[4]Lombardi Cancer Center, Georgetown University, Washington, DC, USA
[5]Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC, USA

Abstract – **In this paper, we report a new gene clustering approach - non-negative independent component analysis (nICA) - for microarray data analysis. Due to positive nature of molecular expressions, nICA fits better to the reality of corresponding putative biological processes. In conjunction with nICA model, VIsual Statistical Data Analyzer (VISDA) is applied to group genes into modules in the latent variable space. The experimental results show that significant enrichment of gene annotations within clusters can be obtained.**

Keywords – **non-negative ICA, latent variable model, gene clustering, module discovery, microarray data analysis**

I. Introduction

Microarray technologies provide powerful tools for genome-wide measurement of gene expressions. To discover functional modules involved in pathway signaling or gene regulation, new computational methods are required for modeling and analysis of microarray data of interest [1].

Gene clustering is widely used in the analysis of gene expression data by partitioning genes into clusters sharing similar expression patterns. The underlying assumption is that genes with similar patterns are more likely associated with common functions. Hierarchical clustering and self-organizing maps [2], have been applied to group the genes into functional modules. Recently, Independent Component Analysis (ICA) has been proposed for modeling gene clusters [3]. In contrast to traditional clustering methods, ICA-based clustering relies on a linear combination of latent biological processes and has revealed the gene clusters with significant enrichment of gene annotations or functional categories [3]. In contrast to PCA, ICA decomposes input data into components as independent as possible, showing some advantages over PCA for gene module decomposition [5].

In this paper, we report the application of non-negative ICA (nICA) for gene clustering, exploiting the non-negative nature of molecular expressions. In principle, nICA can be thought as a projection method where the expression levels are projected onto some new non-negative bases (i.e. components) with minimum statistical dependence. The nICA representation shall better reflect the biological reality. We then use VIsual Statistical Data Analyzer (VISDA) [6] to generate gene modules in the latent variable space. VISDA uses hierarchical Standard Finite Normal Mixtures (SFNM) to model clustered data where each gene belongs to each cluster with a posterior probability. The clustering procedure follows a hierarchy fasion. At each level of the hierarchy, each cluster is considered for further split, until no cluster is decomposable according to the Minimum Description Length (MDL) criterion or human justification.

This paper is organized as follows. In section II, we introduce the principle of nICA for finding gene module composites and a gradient descent algorithm of nICA. A brief description of the VISDA algorithm is also given in Section II to cluster independent components as a post-processing of nICA modeling. The application of nICA and VISDA to yeast data will be reported in Section III. Discussions and conclusions are given in Section IV.

II. Methodology

The problem of basic ICA is given according to the following linear relation:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

where $\mathbf{s} = (s_1, s_2 \ldots s_N)$ is a vector of real independent sources and $\mathbf{x} = (x_1, x_2 \ldots x_N)$ is an observation vector. The assumptions of ICA are that sources are mutually independent and of non-Gaussian distribution except for at most one source. When we apply ICA to real-world problems like gene expression analysis, the situation is different from the above ideal case because of the ambiguities of ICA: the sign and permutation of sources. To resolve these ambiguities, many researchers make further assumptions to constrain the ICA model. For example, non-negativity is a natural constraint for many real-world applications, such as blind separation of natural scenes [7]. Since we assume that the underlying biological processes are independent and their expression levels should be non-negative, nICA is believed to be a more proper model to represent a linear influence of hidden cellular variables than ICA is. By projecting the data to the latent space spanned by these non-negative independent processes, fine structure of co-regulation of genes is maintained and made prominent. VISDA clustering is then applied to catch the characteristics of those subtle differences, which may lead to identify more coherent gene groups (Fig. 1).

A. nICA-based decomposition and the algorithm of nICA

As it has been known, clustering by expression pattern or "co-expressed" genes under limited experimental conditions does not provide the best possible grouping of genes by biological processes [8]. ICA-based gene clustering approach, on the other hand, is built upon a latent variable model of gene module composite. The attraction of ICA clustering lies that it can account for independent hidden effects that influence gene expression. When we introduce the non-negativity into the ICA algorithm, the resulting nICA approach can incorporate prior knowledge for better

1-4244-0033-3/06/$20.00 ©2006 IEEE.

modeling hidden sources while keeping all the advantages of ICA approach.

In our nICA model, gene express is a linear combination of biological mechanisms including pathways of signaling substances, transcription factors and their binding sites in the promoter regions of genes, as well as other different kinds of regulation [3]. We use nICA to project expression data **X** to the independent mode in order to highlight these factors. We assume that x(i, j) which is the expression level of gene i under phenotype j is expressed by the sum of non-negative independent putative biological processes $s_k(i)$, k = 1, 2, …, N, weighted by the involvement strength $a_k(j)$, k = 1, 2, …, N.

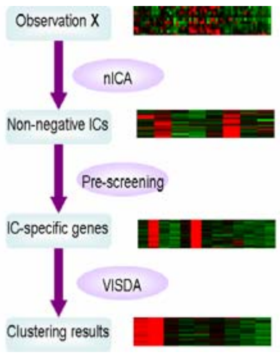


Fig. 1 Framework of nICA and VISDA for the composite module discovery

In [7], the author suggested a mathematical approach to impose non-negative constraint on sources. And if we define non-negative well-grounded sources as:

$$p(s_k < \delta) > 0 \quad \text{for } \forall \delta > 0$$
$$p(s_k < 0) = 0 \quad k = 1, 2, \ldots, N \tag{2}$$

then it has been proven [7] that we can find **y** = **Us**, where **U** is a square orthonormal rotation and permutation matrix, i.e. the elements of y_i of **y** are a permutation of sources if and only if all y_is are all non-negative. We note that **y** = **Us** can be re-written as **y** = **Wz** = **WVx** = **WVAs** with **V** a whitening matrix, **z** the pre-whitened observation vector and **W** an unknown orthonomal (rotation) matrix. Therefore we can consider nICA as a procedure with the following two steps: 1) remove the second order statistics by whitening; 2) search for a rotation matrix where all the data fit into the positive quadrant.

As described in [7], we can use the cost function ***J*** defined in the following to find the global minimum:

$$J(\mathbf{W}) = E\{\|\mathbf{z} - \mathbf{W}^T\mathbf{y}^+\|^2\} \tag{3}$$
$$\mathbf{y} = \mathbf{Wz}$$
$$y_i^+ = \max(0, y_i)$$
$$\mathbf{y}^+ = (y_1^+, y_2^+, \ldots\ y_N^+)$$

Based on the gradient descent rule, a learning algorithm to find the de-mixing matrix **W** is defined as follows [7]:

1) Pre-whitening the observed data **x**:

$$\mathbf{z} = \mathbf{Vx} \tag{4}$$

2) Using gradient descent algorithm to minimize the cost function (3):

$$\mathbf{W} = \mathbf{W} - \gamma \frac{\partial J}{\partial \mathbf{W}} \tag{5}$$

3) Projecting the unconstraint gradient descent set onto a set of orthonormal vectors.

B. Pre-screening for the clustering

After nICA, we obtain some independent components describing some distinct biological processes. In these putative biological processes, some genes showing relatively high or low expression levels are most interesting. We will use a pre-screening procedure to single out these genes.

Specifically, we can select a subsets of genes within one of the components, which includes over expressed genes (which are activated) and down expressed genes (which are repressed) according to the value of each gene in the component [3]:

a subset of genes =

$$\{genes \,|\, L_{genes} \in C\% \text{ of largest values of } y_i\}$$
$$\bigcup \{genes \,|\, L_{genes} \in C\% \text{ of smallest values of } y_i\}$$

By this pre-screening step, we actually remove some invariant genes in each component. By taking the union of the selected genes, we provide a pool of more meaningful and relevant genes to biological processes for the next step-clustering - to identify genes that belong to co-expressed modules in each component.

C. VISDA clustering

In this step, we will cluster genes into modules associated with their values in the independent components. VISDA employs the hierarchical SFNM model for hierarchical clustering. The hierarchical SFNM model uses the following probability density function to describe the relationship between successive levels in the hierarchy,

$$\mathrm{f}\left(\mathbf{r}\middle|\boldsymbol{\pi},\boldsymbol{\theta}\right) = \sum_{k=1}^{K_0}\pi_k \sum_{j=1}^{L_k}\pi_{j|k}\,\mathrm{g}\left(\mathbf{r}\middle|\boldsymbol{\theta}_{j|k}\right)$$
$$\sum_{k=1}^{K_0}\pi_k = 1 \quad and \quad \sum_{j=1}^{L_k}\pi_{j|k} = 1 \tag{6}$$

where **r** denotes the genes to be grouped, the upper level has K_0 clusters, the *k*th cluster in the upper level has L_K sub-clusters in the lower level, π_k is the mixing proportion of the *k*th up level cluster, $\pi_{j|k}$ is the mixing proportion of the *j*th sub-cluster in the *k*th upper level cluster, g(•) is Gaussian distribution function, $\boldsymbol{\theta}_{j|k}$ are the parameters associated with the sub-cluster. The fitting process of this model is executed by the Expectation Maximization (EM) algorithm [6], which achieves a local maximum of the likelihood function.

For each cluster at a level of the hierarchy, VISDA uses two different projection methods, Principle Component Analysis (PCA) and Principle Component Analysis – Projection Pursuit (PCA – PPM) [6], to visualize the sub-clusters within the clusters. The user chooses one of the projections that he/she thinks better revealing the data structure. On the chosen projection, user initializes models with different number of clusters by clicking on the computer screen at the centers of the clusters. These 2-D models will be refined by EM algorithm and compete according to MDL criterion or human justification. The winning model in 2-D space will be transferred back to original data space to initialize the data model in that space. Then EM algorithm in original data space will refine the model and obtain the partition of data at that level. When no more new clusters can be found in the model validation step, the algorithm ends and a final partition is obtained.

III. Results

A. Data treatment

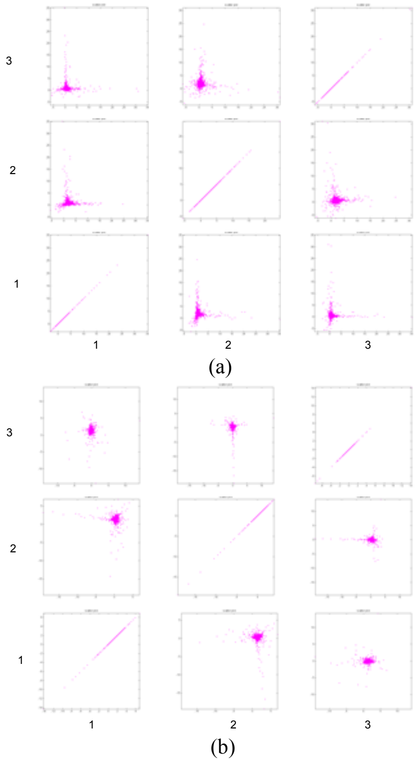
We applied nICA to *Saccharomyces cerevisiae* gene expression dataset [9]. The dataset contains 6152 genes with Open Reading Frames (ORFs) and 173 samples that include the different experimental conditions [9]: temperature shocks, amino acid starvation, and progression into stationary phase etc. As in [3], we also used KNNimpute to fill in missing values. And due to the triviality of clustering environmental stress response (ESR) genes defined by [9], we eliminated them in our analysis. The final dataset contains 5284 genes and 173 samples. To evaluate the experimental results, we measured the biological significance of each cluster using Gene Ontology (GO) annotation database. We mainly measured each cluster of nICA in terms of the biological process and molecular functional categories using *p*-value [3].

B. Experiment and results evaluation

To use nICA model, we took the inverse-logarithm of the data before further analysis. Hence, in our nICA model, the microarray expression corresponds to a linear additive model of interactions among biological processes. Since our goal is to find some most relevant components, dimension reduction using PCA was first applied to the data with 90% of energy maintained. Then nICA was applied to the dimension reduced dataset. As a result, we obtained the eighteen non-negative independent components. For comparison, we also did the experiment using an ICA algorithm with the same parameters.

Fig. 2 shows the first 3 independent components from nICA and ICA respectively. It is clear that, comparing to ICA, nICA is effective in separating sources as independent non-negative "biological processes" in which process-specific genes are highly biased onto two orthogonal axes respectively showed in each sub-panel in Fig. 2.

In Table 1 we only listed seven most significant clusters resulted from our nICA and VISDA approach. We measured the biological significance of each cluster using GO annotation database. The *p*-value of each cluster was calculated according to its overlap with the functional annotations in GO (see [3] for the details). Among those functional categories detected significantly by both nICA and ICA clusters, there are five out seven clusters that nICA produced significant lower *p*-values than ICA did. From these experiments, it seems to us that nICA followed by VISDA can extract more coherent groups of genes in terms of their functional categories.



(a)

(b)

Fig. 2 Comparison of the scatter plots of the first three independent components from nICA/ICA. (a) Results from nICA; (b) Results from ICA. Each sub-panel shows the two subsequent components plotted against each other. In (a), process-specific genes are highly biased on two non-negative axes, whereas the results of ICA in (b) are not.

To further evaluate our nICA-based clustering method, we used the z-score introduced in [8] to conduct a comparative study. As described in [8], the z-core is based on the mutual information between clustering results and the gene annotation. The higher scores indicate clustering results more significantly. We compared the clustering results of nICA and ICA under the same parameters and the z scores are shown in Fig. 3. As we can see from Fig. 3, nICA algorithm

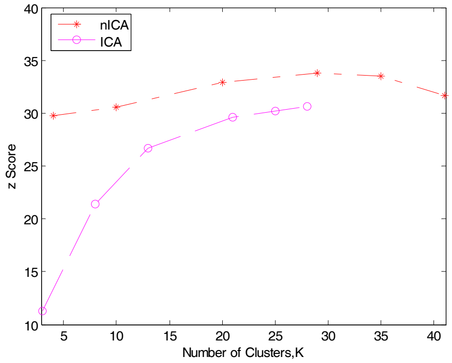


Fig. 3 Z score for the nICA (asterisk) and ICA (circle). At each level of the hierarchy of VISDA, we recorded all the intermediate clusters. At last, we got 41 clusters for nICA and 28 clusters for ICA. We inputted all these clusters to compute the z scores and drew the curves here. So based on the figures, it is reasonable to draw the conclusion that clustering methods by nICA has found a finer structure than ICA has.

TABLE I
THE SEVEN MOST SIGNIFICANT CLUSTERS OF NON-NEGATIVE ICA

Cluster index	Gene Ontology term	Cluster frequency	Genome frequency of use	P-value
6	Ty element transposition	(24 / 73, 32.8%)	(95 / 7291, 1.3%)	3.7E-27
	DNA transposition	(24 / 73, 32.8%)	(108 / 7291, 1.4%)	7.3E-26
	DNA recombination	(24 / 73, 32.8%)	(192 / 7291, 2.6%)	4.2E-20
	RNA-directed DNA polymerase activity	(14 / 73, 19.1%)	(52 / 7291, 0.7%)	2.2E-16
	DNA-directed DNA polymerase activity	(15 / 73, 20.5%)	(67 / 7291, 0.9%)	2.5E-16
16	glycolysis	(9 / 109, 8.2%)	(21 / 7291, 0.2%)	4.5E-11
	Glucose metabolism	(12 / 109, 11.0%)	(65 / 7291, 0.8%)	3.6E-10
17	proteolysis	(40 / 250, 16%)	(164 / 7291, 2.2%)	4.3E-22
	ubiquitin-dependent protein catabolism	(24 / 250, 13.6%)	(128 / 7291, 1.7%)	5.5E-20
	endopeptidase activity	(25 / 250, 10%)	(62 / 7291, 0.8%)	4.5E-19
22	amino acid and derivative/metabolism	(21 / 43, 48.8%)	(199 / 7291, 2.7%)	8.5E-22
	amino acid metabolism	(20 / 43, 46.5%)	(183 / 7291, 2.5%)	5.4E-21
	oxidoreductase	(10 / 43, 23.2%)	(247 / 7291, 3.3%)	1.4E-06
23	amino acid biosynthesis	(19 / 85, 22.3%)	(102 / 7291, 1.3%)	1.0E-17
	catalytic activity	(43 / 85, 50.5%)	(1937 / 7291, 26.5%)	2.1E-06
24	cellular response to nitrogen starvation	(4 / 47, 8.5%)	(5 / 7291, 0.0%)	3.9E-08
	cellular response to nitrogen levels	(4 / 47, 8.5%)	(5 / 7291, 0.0%)	3.9E-08
	asparagine	(4 / 47, 8.5%)	(5 / 7291, 0.0%)	3.9E-08
33	generation of precursor metabolites and energy	(32 / 68, 47.0%)	(231 / 7291, 3.1%)	8.8E-30
	oxidative phosphorylation	(19 / 68, 27.9%)	(46 / 7291, 0.6%)	4.0E-26
	hydrogen ion transporter activity	(20 / 68, 29.4%)	(55 / 7291, 0.7%)	2.1E-26

The selected clusters are listed along with the functional categories with the smallest p-value. Numbers in parentheses in the third column show the number and percentage of genes within the cluster that are presented in one of the functional category. For instance, (24/73, 32.8%) means the cluster has 73 genes, among which 24 (32.8%) genes are annotated with "Ty element transposition". And the numbers in the fourth column are presented in the similar way which corresponds to the total number within the whole genome set that are annotated with one of the special categories in GO system.

consistently performed better than ICA with an average increase of z-score of 5.

IV. CONCLUSION AND DISCUSSION

This paper presents a new gene clustering approach, namely nICA-based approach for composite module discovery. By projecting the gene expression data onto nICA space, co-regulation structure of modules are revealed and highlighted. Using a pre-screening and VISDA clustering procedure, we can identify biological process enriched clusters with coherent functional annotations. The experimental results on a yeast data set have demonstrated its advantages over conventional ICA-based approach.

Although nICA-based approach exhibits some promise for gene clustering, there is future work to be conducted. For example, we notice that in the de-mixing matrix **W**, there are some negative values that need to be properly explained, i.e., how these composite modules are involved in the corresponding biological process. Another possible direction is that we may perform gene clustering to find groups of genes under distinctive regulators or combinations of genes of these regulators.

ACKNOWLEDGMENT

This work was partially supported by the National Institutes of Health under Grants (CA109872, NS29525, EB00830, and CA096483) and the Department of Defense under Grant (BC030280).

REFERENCES

[1] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Boststein, and D. Koller, "Module networks: identifying regulatory modules and their condition-specific regulations from gene expression data," *Nature Genetics*, vol. 34, pp. 166-176, 2003.

[2] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, and E. Dmitrovsky, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 2907-2912, 1999.

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

[4] O. Troyanskaya, M. Cantor, O. Alter, G. Sherlock, P. Brown, D. Botstein, R. Tibshirani, T. Hastie, and R. Altman, "Missing values estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, pp. 520-525, 2001.

[5] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, pp. 51-60, 2002.

[6] Y. Wang, L. Luo, M. T. Freedman, and S.-Y. Kung, "Probabilistic principal component subspaces: a hierarchical finite mixture model for data visualization," *IEEE Trans on Neural Networks*, vol. 11, pp. 625-636, 2000.

[7] E. Oja and M. Plumbley, "Blind separation of positive sources by globally convergent gradient search," *Neural Computation*, vol. 16, pp. 1811-1825, 2004.

[8] F. D. Gibbons and F. P. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Research*, vol. 12, pp. 1574-1581, 2002.

[9] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic expression programs in the response of yeast cells to environmental changes," *Molecular Biology of the Cell*, vol. 11, pp. 4241-4257, 2000.