유전자 발현 데이터의 독립 특징 부공간 해석

김혜진[0] 최승진 방승양
포항공과대학교 컴퓨터공학과
{marisan[0], seungjin,sybang}@postech.ac.kr

Independent Feature Subspace Analysis for Gene Expression Data

Heijin Kim[0], Seungjin Choi, Sung Yang Bang
Dept. of CSE, Pohang University of Science & Technology

ABSTRACT

This paper addresses a new statistical method, IFSAcycle, which is an unsupervised learning method of analyzing cell cycle-related gene expression data. The IFSAcycle is based on the independent feature subspace analysis (IFSA) [3], which generalizes the independent component analysis (ICA). Experimental results show the usefulness of IFSA: (1) the ability of assigning genes to multiple coexpression pattern groups; (2) the capability of clustering key genes that determine each critical point of cell cycle.

I. INTRODUCTION

The current microarray technology enables us to estimate huge gene expressions quantitatively. Gene expression varies according to sample conditions such as time differences, chemical treatments, and the degree of states from which diseases developed. The level of gene expression reflects the biological situation and active function when the gene was expressed.

Gene chip technology data have made it possible to understand cellular function and pathways. Matrix form of those data can be factorized or decomposed into two matrices. Bayesian decomposition (BP) [1] assigns genes to multiple coexpression groups and encodes biological knowledge into the system. After applying the BP algorithm, the data are broken into two matrices, one of which takes its column as a pattern or distribution of a biological function. PCA and ICA [2] are linear model-based methods assuming that the expression of each gene is a linear function of the expression mode and linear influences of different modes. A projection to expression modes highlights particular biological functions. All the current methods demonstrate the ability to match a pattern of the feature matrix with a biological function, clustering a set of genes which play a key role in the function. However, these methods still lack the ability to reflect complex biological processes. Biological processes are horizontally or vertically related to each other [Figure 1]. It is expected that there exist interactions between genes belonging to biologically related patterns, although the connection is not stronger in between-patterns than in a pattern. To understand cell cycle regulation or metabolic pathways, we should identify the weak relations to explain vertical network in the process. For example, S.cerevisiae has three cell cycles – the chromosome cycle, the centrosome cycle and the cytoplamic cycle. Present methods can distinguish a cluster of genes in DNA synthesis and a cluster of genes of Bud emergence. However, in a chromosome cycle, a set of genes expressed in the step of replication initiation is closer to the gene set of DNA synthesis than nuclear division.

Here we employ a method of IFSA [3] in order to identify the sets of interacting physical process such as cell cycle progression or the activations of a pathway in response to a drug treatment. IFSA finds independent feature subspaces, each of which contains a few basis vectors, which are associated with some biological functions.

II. ALGORITHM AND IMPLEMENTATION

Aapo Hyvärinen proposed an IFSA algorithm [3] and showed its usefulness in adding a phase and shift invariant feature in image processing. The IFSA generalizes the sparse coding proposed by Olshausen and Field [4] who showed that the principle of maximizing nongaussianity of the underlying components was able to explain the emergence of Gabor-like filters that resemble

the receptive fields of simple cells in the mammalian primary visual cortex (V1). Taken notice of complex cell feature in V1, IFSA maximizes the independence between norms of projections on linear subspaces instead of the independence of simple linear filter outputs [Figure 2].

Figure 1 The three cell cycle in S.cerevisiae

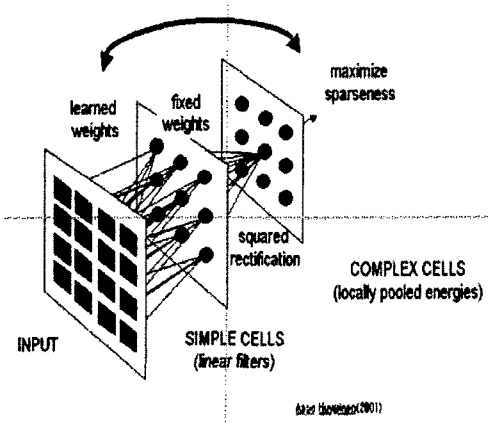


Figure 2 IFSA algorithm concept

Classical model ICA trained model in simple cell levels to maximize the sparseness with the notation of

$$s_i = < w_i, D > = \sum_{x,y} w_i(x,y) D(x,y) \qquad (1)$$

where the s_i are stochastic coefficients, different for each data $D(x,y)$ and the w_i denotes the inverse filters. The feature F with input vector is given by [Figure 3]

$$F(D) = \sum_{i=1}^{n} < w_i, D >^2 \qquad (2)$$

IFSA maximizes the independency between feature spaces, allowing s_i - n numbers in a subspace – not to be all mutually independent.

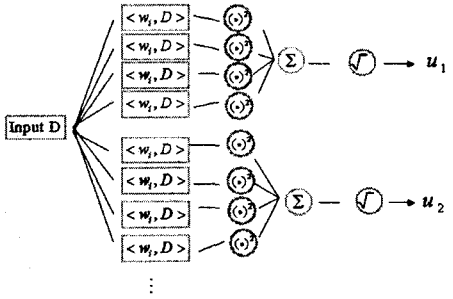


Figure 3 A graphical depiction of the feature spaces

The logarithm of the likelihood L of the data $D_k(x,y)$, given model with the probability density of $p_j(\square)$ of the n-tuple (For simplicity, equal for all subspaces) with subspace index $j \in \{1,...J\}$ can be expressed as

$$\log L(D_1,...,D_K; w_1,...,w_m) = \sum_{k=1}^{K} \sum_{j=1}^{J} \log p(\sum_{i \in S_j} s_i^{\,2}) + K \log |\det W| \qquad (3)$$

Using a stochastic gradient ascent of the log-likelihood, the learning of $\square w_i$ is represented by

$$\square w_i(x,y) \propto D(x,y) < w_i, D > g(\sum_{r \in S_{(i)}} < w_i, D >^2) \qquad (4)$$

III. EXPERIMENTAL RESULTS

We applied the IFSA to cdc28-mutant yeast cell cycle data from Spellman et al. (1998)[5] containing 800 genes and interpret our results in light of information within the SGD[6] database, the KEGG[7] and the GO database[8].

A set of genes function together which response an IFSA component and genes are physically related, belonging the same feature space. We compared the results from other algorithms to PCA and ICA results.

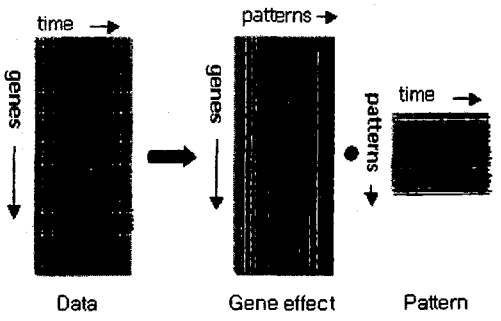
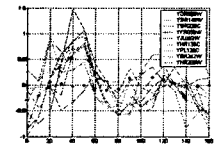
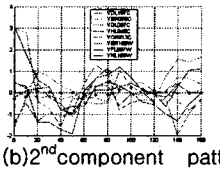


Figure 4 data matrix decomposition

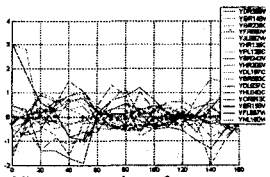
After applied PCA, ICA, or IFSA, Data matrix is decomposed into two matrices; a row vector of pattern matrix becomes a principal / independent / feature pattern [Figure 4]. A row of "Gene effect matrix" shows the contribution of a gene to composing the corresponding patterns. We normalized a gene effect and the gene belongs to a cluster which records the largest score. Unlike PCA and ICA, IFSA normalized not by a vector but by a subspace. After the above steps, we decide the final gene cluster with more than a certain threshold, which was decided in the manual. The selected gene patterns are shown below [Figure 5, Figure 6]. Gene patterns resulting from PCA and ICA cover totally different patterns but a pattern from IFSA has only similar patterns, and patterns in a feature space represent a slight time shift pattern [Figure 6] or the opposite pattern - upregulated genes vs. downregulated genes.



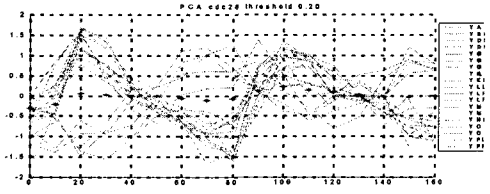
(a)1st component pattern



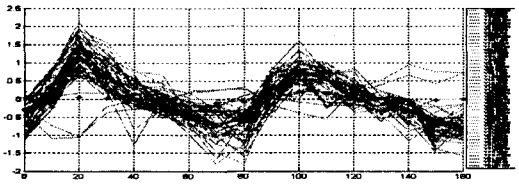
(b)2nd component pattern



(c) All patterns in a feature space



(a) Genes in a pattern of PCA



(b) Genes in a pattern of ICA

Figure 5 gene patterns in IFSA

Figure 6 gene patterns in PCA & ICA

However, our results show that the consistency of patterns in IFSA is neither better nor worse than PCA or ICA. This is because the nonlinear function to fix weights (this paper used squared norm) may not fit very well, hence the pattern has dependent information too much. We will find a good function not only to conserve ICA property but also to save the dependent information of data. Despite this fault, KEGG shows that IFSA is a useful method to depict microarray data because it shows that the set of genes of a pattern of IFSA in KEGG is closely related in a cell cycle procession or location [Figure 7].

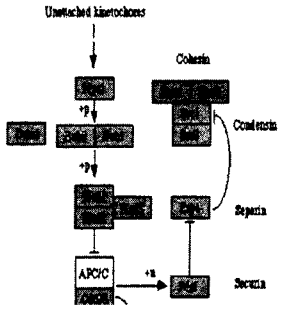


Figure 7 pathway from KEGG and genes in IFSA

IV. CONCLUSION

The IFSA algorithm maximizes the independence between feature spaces and allows the components in a space to have dependent structure. Applying IFSA to time series microarray data, we identified that a pattern of IFSA implies a physical process and a feature space make a set of functions biologically related, which is impractical in other methods.

V. REFERENCE

[1] T.D.Moloshok et al., Application of Bayesian Decomposition for analyzing microarray data, *Bioinformatics* vol**18**. no4 p566-575

[2] Wolfram Liebermeister (2002), Linear modes of gene expression determined by independent component analysis, *Bioinformatics* vol.**18** 51-60

[3] Aapo Hyvärinen et al.(1999), Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspace, *Neural Computation* August '99

[4] Oshausen, B.A. and Field, D.J.(1996).Emergence of simple-cell receptive field properties by learning a sparse code for antural images. Nature, 381:607-609

[5] Spellman, P.T., Sherlock, G. et al(1998) Comprehensive identification of cell cycle-regulated genes of the yeast *S.cerevisiae* by microarray hybridiation. *Mol.Biol. Cell*, **9**,3273-3297

[6] http://genome-www.stanford.edu/Saccharomyces/

[7] http://www.genome.ad.jp/kegg/kegg2.html