Time series microarray data analysis using Independent subspace analysis

Heijin Kim[0], Seungjin Choi, Sung Yang Bang
Dept. of CSE, Pohang University of Science & Technology {marisan, seungjin,sybang}@postech.ac.kr

ABSTRACT

ISA is an unsupervised learning method for gene expression data analysis, based on independent subspace analysis (ISA) which aims at finding independent feature subspace of multivariate data in dependent sub-components. In a feature subspace, patterns of each component are shown the properties: slight time invariance or the reverse sign invariance. As shown by Liebermeister, each component correspond to one of biological functions. Our results make a set of function(a feature space) assumed that biologically related each other.

I. INTRODUCTION

With the advent of microarray technology, quantitatively huge gene expressions can be estimated. Gene expression according to sample conditions such as time differences, chemical treatments, and the degree of states of which diseases developed. The level of gene expression reflects the biological situation and active function when the gene was expressed.

Microarray and gene chip technology data have made it available 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 BP algorithm, the data are broken into two matrices, one of which has taken its column as a pattern or distribution of a biological function. PCA and Independent Component Analysis (ICA)[2] derived a linear model 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 of the currents methods demonstrate the ability to match a pattern of the feature matrix with a biological function, clustering a set of genes identified to play a key role in the function. However, these methods still lack an ability to reflect the real complex of biological process. Biological processes are horizontally or vertically related each other [Figure 1]. There should exist interactions between genes belonging to patterns biologically related, although the connection is not stronger in between-patterns than in a pattern. To understand cell cycle regulation or metabolic pathway, we should identify the weak relations to explain vertical network in the process. For example, S.cerevisiae has three cell cycles – chromosome cycle, centrosome cycle and cytoplamic cycle. Current 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 present an algorithm Independent subspace analysis (ISA) to identify sets of interacting physical process such as cell cycle progression or the activations of a pathway in response to a drug treatment. ISA consists of feature subspaces, each of which contains a few basis vectors and ach vector is associated with a biological function.

II. ALGORITHM AND IMPLEMENTATION

Aapo Hyvärinen devises ISA algorithm [3] for the purpose of emergence of phase and shift invariant feature in image processing. He refers to Olshausen et al.[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 mammalian primary visual cortex(V1). Taken notice of complex cell feature in V1, ISA 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

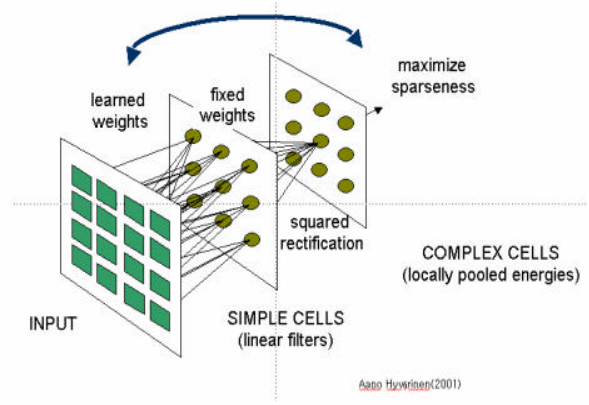


Figure 2 ISA algorithm concept

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]

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

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

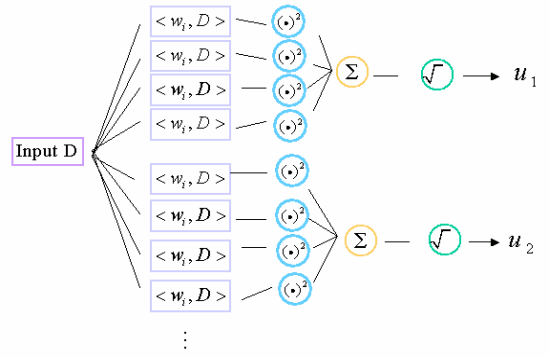


Figure 3 A graphical depiction of the feature spaces

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

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

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

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

III. EXPERIMENTAL RESULTS

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

A set of genes in response to a ISA component function together and genes belonging the same feature space are physically related, compared with other algorithms, PCA and ICA. PCA results.

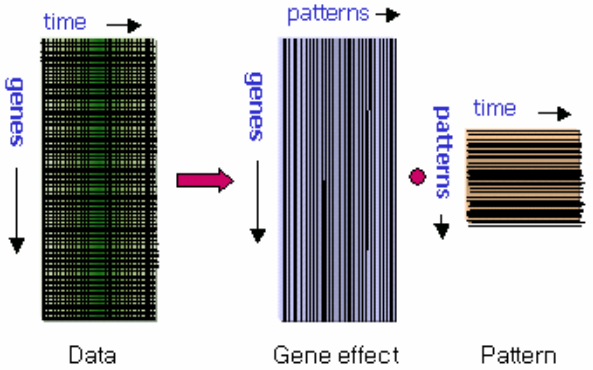
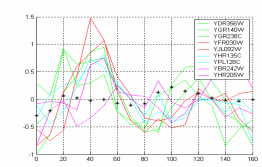
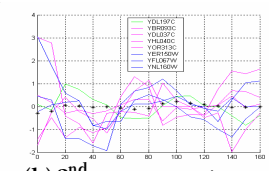


Figure 4 data matrix decomposition

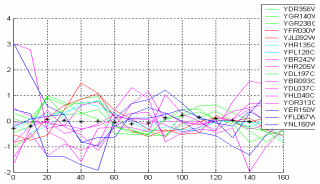
Applied PCA, ICA, or ISA, Data matrix is decomposed into two matrices, a row vector of pattern matrix becomes a principal / independent / feature pattern[Figure **4**]. A row of Gene effect matrix shows the contribution of a gene to make the corresponding patterns. We normalized a gene effect and the gene belongs to a cluster recording the biggest score. Differing from PCA and ICA, ISA normalized not a vector but a subspace. After previous steps, we decide final gene cluster with more than a certain threshold, which was decided in 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 ISA has only similar patterns and patterns in a feature space represent slight time shift pattern [Figure 6] or the opposite pattern - upregulated genes vs. downregulated genes.



(a) 1st component pattern

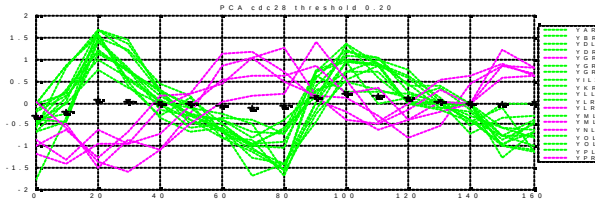


(b) 2nd component pattern

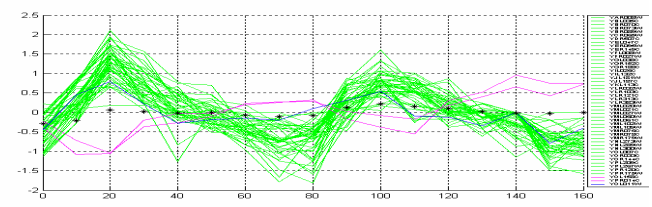


(c) All patterns in a feature space

Figure 5 gene patterns in ISA



(a) Genes in a pattern of PCA



(b) Genes in a pattern of ICA

Figure 6 gene patterns in PCA & ICA

However, Our result shows that the consistency of patterns in ISA is not better than or even worse than PCA or ICA. That is because the nonlinear function to fix weights (this paper used squared norm) may be not fitted very well hence the pattern has dependent information too much. We'll find a good function not only to conserve ICA property but also not to lose the dependent information of data. Despite this fault, KEGG shows that ISA is a useful method to depict microarray data because shows that the set of genes of a pattern of ISA in KEGG is very closely related in a cell cycle procession or location [Figure 7].

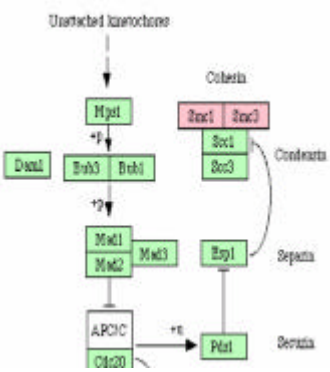


Figure 7 pathway from KEGG and genes in ISA

IV. CONCLUSION

ISA algorithm maximize the independence between feature spaces and allow the components in a space to have dependent structure. Applying ISA to time series microarray data, we identified that a pattern of ISA imply 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