INDEPENDENT SUBSPACES OF GENE EXPRESSION DATA

Hyejin Kim
Intelligent Robot Research Division
ETRI, Korea
email: marisan@etri.re.kr

Seungjin Choi
Department of Computer Science
POSTECH, Korea
email: seungjin@postech.ac.kr

ABSTRACT
Independent subspace anlaysis (ISA) is a linear model-based method which generalizes independent component analysis (ICA) by incorporating the invariant feature subspace into multidimensional ICA. In this paper we apply ISA to the problem of gene expression data analysis and show the useful behavior of the independent subspaces of gene expression data in the task of gene clustering and gene-gene interaction analysis.

KEY WORDS
DNA chip data, gene clustering, gene-gene interaction analysis, independent component analysis, independent subspace analysis.

1 Introduction

Genomic-scale gene expression data are provided by high-throughput methods such as microarray. Gene expressions measured at different time points represent biological functional behavior of genes at different expression levels, so that one can obtain new insights to regulatory networks. Current microarray technologies have difficulties in producing reliable repeated experiments, which might be a cumbersome for accurate data analysis. Genes involved in a cellular function, cooperate with each other and work within a limited time interval. Several cellular functions are involved simultaneously for cells to cope with external or internal stimuli. Thus one is not sure that all genes expressed contemporaneously, are really functionally related to each other. Moreover a single gene takes part in several cellular functions, but it takes actions under specific circumstances. Therefore, time series data measured under different cells or tissue conditions, are required, in order to identify entire gene functions.

To date, various methods have been applied to time series data of gene expression. Theses include: (1) Bayesian networks; (2) hierarchical clustering; (3) differential equations; (4) edge detection; (5) linear decomposition models such as PCA and ICA. These methods attempt to investigate functional behavior of genes by clustering or detecting gene-gene relations. Assuming that gene expression levels are continuous, Bayesian networks describe gene-gene interactions in terms of conditional probabilities [4, 1]. Hierarchical clustering methods carry out grouping of genes using a similarity measure between gene expression profiles [3]. Differential equations and edge detection methods are techniques which catch the difference between expression levels of two or more points within a local area. All these methods are concerned with the direct relationship between pairs of genes. On the other hand, linear model-based methods explicitly describe dominant functions in terms of expression modes associated with effective genes. Singular value decomposition (SVD) or PCA [6], ICA [8] are exemplary liner model-based methods.

In this paper we use a method of ISA [7] which is still a linear model-based method, but which generalizes ICA by incorporating invariant feature subspace into multidimensional ICA. In fact, ICA can be treated as a special case of ISA if the feature subspace dimension becomes 1. A major benefit of ISA, compared to ICA, is to allow some dependence between basis vectors in the same group. As will be shown in our experimental results, ISA is more useful in gene clustering and gene-gene interaction analysis, compared to ICA.

2 Independent Subspace Analysis

Linear decomposition models assume that the data matrix $\boldsymbol{X} = [X_{ij}]$ (where the element X_{ij} represents the expression level of gene i associated with the jth sample, $i = 1, \ldots, m$, $j = 1, \ldots, N$) is modelled as

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}, \tag{1}$$

where $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is a matrix consisting of latent variables (or encoding variables) and the row vectors of $\boldsymbol{A} \in \mathbb{R}^{n \times N}$ are basis vectors corresponding to *linear modes* in [8]. A pictorial illustration of generating gene expression data through a DNA chip, is shown in Fig. 1

ICA searches for a linear decomposition (1) such that statistical dependence between the columns of $\boldsymbol{S}$ is minimized. The statistical independence among latent variables is a key assumption (as well as a limitation) in ICA. Multidimensional ICA [2] generalized ICA by allowing the components in a κ-tuple to be dependent but requiring different κ-tuples to be independent. ISA [7] embeds the invariant feature subspaces in multidipmensional ICA by considering probability distributions of latent variables that are spherically symmetric, i.e., depend only on their norm. In contrast to ICA, ISA aims at finding a linear transformation $\boldsymbol{W}$ (which corresponds to the inverse system of $\boldsymbol{A}$)

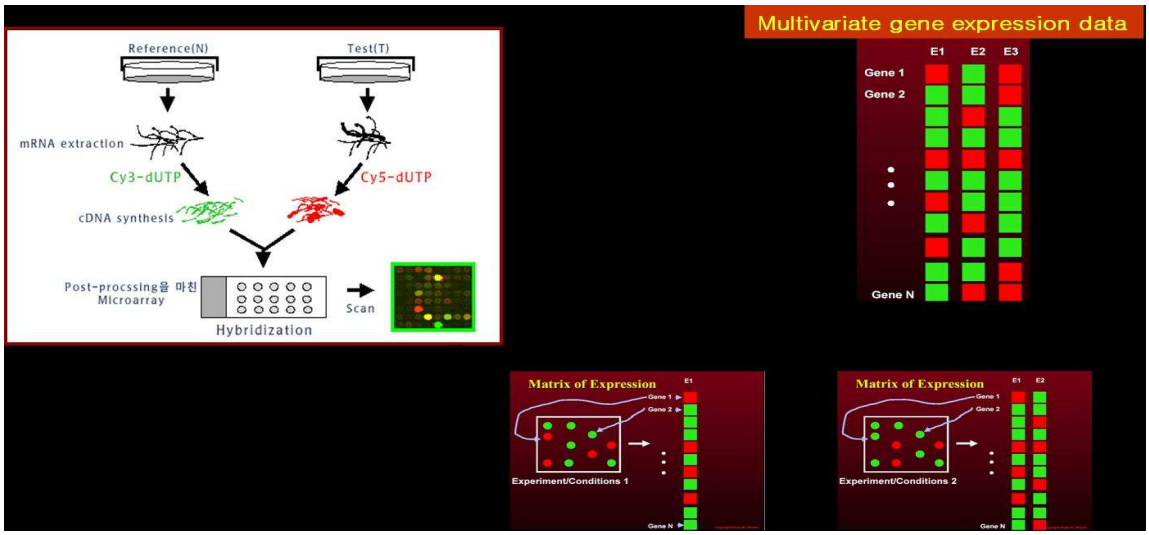


Figure 1. mRNA are extracted from a test tissue and a reference, then are dyed by Cy-5 (red) Cy-3 (green), respectively. Through a hybridization and a postprocessing, a scanned image leads to a matrix whose elements correspond the ratio of red to green light intensity. This matrix data is converted to a vector by a column stacking, thus, m different experiments result in a multivariate data $\boldsymbol{X} \in \mathbb{R}^{N \times m}$ where N is the number of genes.

such that feature subspaces become independent but components in a feature subspace is allowed to be dependent.

We assume that the data matrix $\boldsymbol{X}$ is already whitened. In other words, the row vectors of $\boldsymbol{A}$ are confined to be orthogonal each other and to be normalized to have unit norm. Non-orthogonal factor is reflected in a whitening transform. In order to avoid an abuse of notations, we use the notation $\boldsymbol{X}$ for the whitened data matrix.

Given a matrix $\boldsymbol{X} \in \mathbb{R}^{m \times N}$, its row and column vectors are denoted by $\boldsymbol{x}_i$, $i = 1, \ldots, m$ and by $\vec{\boldsymbol{x}}_j$, $j = 1, \ldots, N$. Define a linear mapping $\boldsymbol{W}$ such that $\boldsymbol{W}^T = \boldsymbol{A}^{-1}$. Then the estimate of $\boldsymbol{S}$, given $\boldsymbol{X}$ is computed by $\boldsymbol{X}\boldsymbol{W}^T$. For a orthogonal matrix $\boldsymbol{A}$, the row vectors of $\boldsymbol{W}$ coincide with the row vectors of $\boldsymbol{A}$. We consider the case where latent variables are divided into J number of κ-tuples (where κ represents the dimension of subspace). For the sake of simplicity, we assume identical dimension, κ for every feature subspace. The jth feature subspace is denoted by $\mathcal{F}_j$. The value $E_j(\boldsymbol{x})$ in $\mathcal{F}_j$ with data vector $\boldsymbol{x}$ is given by

$$E_j(\boldsymbol{x}) = \sum_{i \in \mathcal{F}_j} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle^2 , \tag{2}$$

where $\langle \cdot, \cdot \rangle$ represents the inner product. In fact $E_j(\boldsymbol{x})$ is a pooled energy.

With these notations, we can write the normalized log-likelihood $\mathcal{L}$ of the data given the model as

$$\begin{aligned} \mathcal{L}_{isa} &= \frac{1}{m} \sum_{t=1}^{m} \sum_{j=1}^{J} \log p \left(\sum_{i \in \mathcal{F}_j} \langle \boldsymbol{w}_i, \boldsymbol{x}_t \rangle^2 \right) \\ &\quad + \log |\det \boldsymbol{W}| , \end{aligned} \tag{3}$$

where $p\left(\sum_{i \in \mathcal{F}_j} s_i^2\right) = p_j(s_i, i \in \mathcal{F}_j)$ and $s_i = \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle$ represents the probability density inside the jth κ-tuple of s_i.

ISA finds a linear transform $\boldsymbol{W}$ which maximizes the log-likelihood (3). Learning independent feature subspaces is carried out by a stochastic gradient ascent method, whose updating rule has the form

$$\Delta \boldsymbol{w}_i \propto \boldsymbol{x} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle \varphi \left(\sum_{r \in \mathcal{F}_{j(i)}} \langle \boldsymbol{w}_r, \boldsymbol{x} \rangle^2 \right) , \tag{4}$$

where $j(i)$ is the index of the feature subspace which $\boldsymbol{w}_i$ belongs to and φ is the score function, i.e., $\varphi = \frac{p'}{p}$ and $p(\cdot)$ is the hypothesized density which is usually assumed to be heavy-tailed distributions. More details on ISA can be found in [7].

Remarks

- In contrast to ISA, ICA searches for a parameter matrix $\boldsymbol{W}$ which maximizes the normalized log-

likelihood $\mathcal{L}_{ica}$ given by

$$\mathcal{L}_{ica} = \frac{1}{m}\sum_{t=1}^{m}\sum_{i=1}^{n}\log p\left(\langle \boldsymbol{w}_i, \boldsymbol{x}_t\rangle\right) + \log|\det \boldsymbol{W}|. \quad (5)$$

- If the dimension of each feature subspace is 1, i.e., $\kappa = 1$, then ISA becomes identical to ICA.
- In ICA, each row vector, $\boldsymbol{w}_i$ is interpreted as a different linear mode which is expected to be related with a biological function. [8]. Gene expression profiles are approximated by a linear combination of linear modes with encoding variables representing contributions of linear modes. Only a single linear mode is allowed in each group since ICA does not exploit any dependence structure. On the other hand, ISA allows several linear modes that are statistically dependent in a group so that some dependence structure is exploited

3 Results

We apply ICA and ISA (for comparison) to the yeast cell cycle data [9], which contains the expression of 6178 open read frames (ORFs) during the cell replication cycle in the budding yeast *Saccaromyces cerevisiae*. This data set contains 77 tissue samples in different experimental conditions such as α factor pheromone, cdc15, cdc28, elucidation, and so on. Many cell cycle-regulated genes are involved in processes that occur only once per cell cycle. Each experiment has a different cell cycle, such as elutriation data through one cell cycle, α factor pheromone through two cycles or cdc15 through three cycles.

The data matrix $\boldsymbol{X}$ contains 77 samples in its columns. Each column vector in $\boldsymbol{X}$ was shifted such that its mean value is zero. After centering the data matrix, the data sphering (whitening) was performed such that $\frac{1}{m}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix. In addition, gene profiles were filtered out if its profile variance was less than 1 so that we only considered relatively significant genes in our analysis. As a similarity measure, we used a maximal time-delayed correlation value between two vectors $\boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{y} \in \mathbb{R}^N$. The similarity score ρ_{xy} between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined by

$$\rho_{xy} = \max_l |r_{xy}(l)|, \quad (6)$$

where $r_{xy}(l)$ is the cross-correlation with time-lag l, $l = 0, 1, \ldots, L-1$ and $L \ll N$,

$$r_{xy}(l) = \frac{1}{N-l}\sum_{i=0}^{N-l-1} x_i y_{i+l}, \quad (7)$$

where x_i denotes the ith element of the vector $\boldsymbol{x}$.

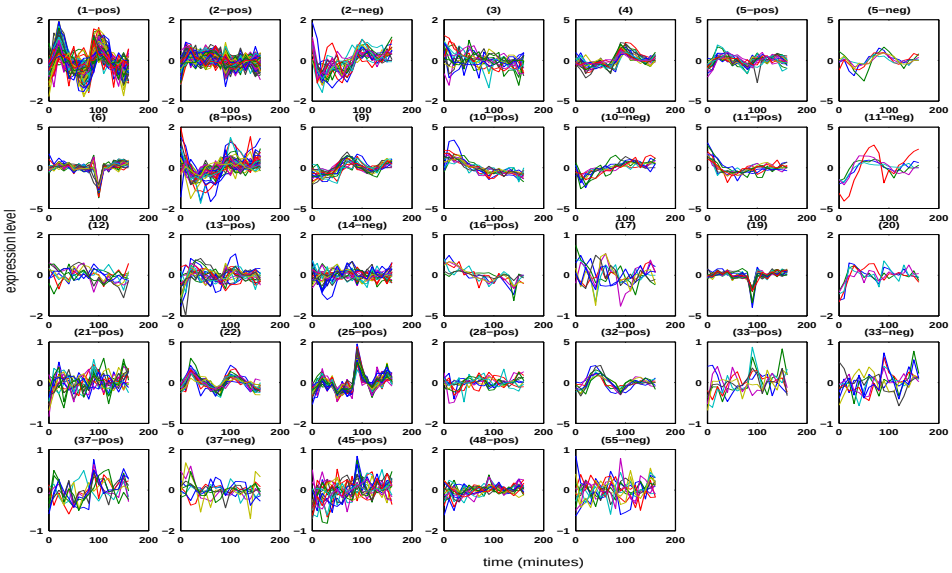
Similarity scores are validated by their associated p-value with assuming an asymptotic Normal distribution. In the task of gene clustering, if the similarity score between a linear mode $\boldsymbol{a}_i$ and a gene expression profile $\boldsymbol{x}_i$ is greater than a threshold value (here we used 0.5) with $10^{-7}\%$ significant level (p-value), then the gene is assigned to a group which the associated linear mode belongs to. In the gene-gene interaction analysis, we also used the similarity score in (6) for a possible set of pair of genes in the same group. Taking the sign of maximal cross-correlation into account, in the calculation of the similarity score, we divided three different types of interactions: (a) concurrently expressed pairs; (b) activators, or inhibitors; (c) activatee or reducer. In order to analyze the potential for determining regulatory pairs from Spellman data, we categorize all genes by phenotypes of mRNA regulated with the cell cycle and validate the results based on SGD and Gene Ontology (GO) database.

3.1 Gene Clustering

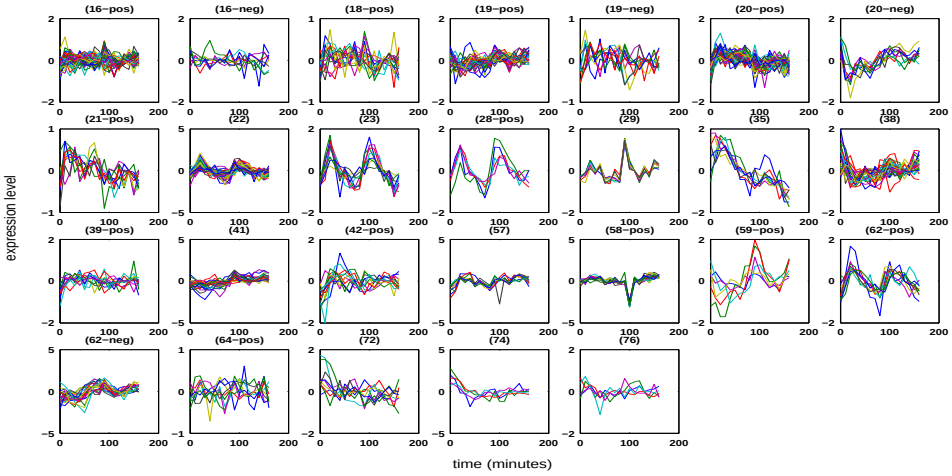
In applying ISA to the yeast cell cycle data, we consider 11 invariant feature subspaces ($J = 11$), each of which is 7-dimensional space ($\kappa = 7$), which produce 11 different groups with 7 linear modes for each group. For ICA, 76 independent components were extracted. Fig. 2 displays gene profiles which best match the linear modes computed by ICA and ISA. When gene profiles which match the ith-mode with positive correlation or negative correlation are almost evenly separated, we further divide them into i(pos)th- and i(neg)th-patterns.

Genes PRS* and RPL* are known to be involved with protein synthesis. Those genes were found in the 6th-, 19th-, and 25(pos)th-modes in the case of ICA and they were associated with the 29th- and 58(pos)th-modes in the case of ISA. In the context of functional groups of genes, gene profiles associated with the 25th-mode in ICA are comparable to those associated with the 29th-mode in ISA. On the other hand, gene profiles associated with the 6th- and 19th-modes computed by ICA, are comparable to those associated with the 58(pos)th-mode in ISA. As shown in Fig.2, genes associated with the 6th- and 9th-modes were assigned into different clusters by ICA, although they are known to exhibit the same biological function. Meanwhile, those patterns appeared in the 58(pos)th-mode in ISA, which means they were preserved in the same group in the case of ISA. Another interesting example is involved with histone groups which are related with the 5th-mode in ICA and the 57th-mode in ISA. This result is summarized in Tables 1 and 2.

We compared linear modes computed by ISA with those found by ICA in [8]. We calculated the p-value by incorporating with Gene Ontology (GO) annotation with assuming hypergeometric distribution where the parameters are: (1) the total population of genes; (2) the number of items with the desired category in the population; (3) the number of sample genes; (4) the number of particular genes in the category of GO among sampled genes. By random selection, averaged p-values after eliminating maximum and minimum values, are 6.23232×10^{-5} for ICA



(a) Gene profile patterns in ICA.



(b) Gene profile patterns in ISA.

Figure 2. Gene expression profile patterns for the case of ICA (above) and ISA (below). A gene profile is assigned to a mode that produces the best similarity score (with considering positive or negative correlation). When gene profiles that best match the ith-mode with positive or negative correlations are almost evenly separated, then they further divided to i(pos)th- or i(neg)th-mode. Gene profile patterns for the case of ICA and ISA are not much different. However, those in the case of ISA tend to be ordered in a topographical fashion, which is desirable in the study of gene-gene interaction analysis.

and 1.63403×10^{-5} for our method based on ISA. This states that linear modes computed by ISA are more useful to find informative and significant genes, compared to ICA.

3.2 Gene-Gene Interaction Analysis

Gene-gene interactions, given a linear mode, were investigated using cross-correlations with setting the threshold as 0.6 and p-value as $10^{-7}\%$ in the case of ICA and ISA. As an example of useful behavior of ISA in gene-gene interaction analysis, we chose a group associated with *chromatin assembly/disassembly and DNA binding* process. Genes involving with such process, take a small portion in yeast genome. In the SGD database, the number of those genes is only 24 out of 7270 genes in total, approximately 0.33%. Those genes were associated with the 5(pos)th-mode in ICA and the 57th-mode in ISA. Tables 1 and 2 show the list of genes which are associated with those modes for the case of ICA and ISA: (a) in the case of ICA, 8 out of 13 with p-value $1.78 \times 10^{-15}\%$, were found in a set of genes associated with the 5(pos)th-mode; (b) in the case of ISA, 8 out of 9 with p-value $1.26 \times 10^{-17}\%$, were found in a set of genes associated with the 57th-mode.

These methods did a successful grouping, showing that HHT2, HTA1 HTB2, HTA2 and HHT2 were linked each other. This relation is confirmed by GO [5]. We also constructed genetic networks for these groupings, using the Osprey program that is available in the web [1]. These results are shown in Fig. 3 where an edge represents an interaction between genes and the nodes with the same color imply that they belong to one biological category. In the context of biological meaning, we confirmed edges and nodes by GO. We put genes in the 'chromatin assembly/dissembly and DNA binding' process in a single map. The edges were shown only if its biological evidences exist. For genes in Table 1 (by ICA), 5 genes preserved their connections, meanwhile outlier nodes, PDS1, PDE2 and WSC2 were incorrectly connected (see Fig. 3 (a)). On the other hand, genes listed in Table 2 (by ISA) did not show any outlier nodes (see Fig. 3 (b)).

4 Discussion

We have introduced a new approach to the identification of information from time series DNA microarray data. To overcome previous limitations, we introduced a novel method based on ISA. As mentioned earlier, phase and shift invariant characteristics of ISA made it possible to take time-delay and asymmetry between gene profiles into account in the task of clustering using a linear model. We demonstrated that ISA produced satisfactory clustering which exhibited more meaningful biological relations, compared to ICA. A genetic network, in general, was constructed by gene-gene interactions, which we referred to as a bottom-up approach. In contrast, we tackled this problem using a top-down method, analyzing gene-gene interaction, given a linear mode. We did this analysis, based on ISA-based clustering and cross-correlations and confirmed that the genetic network using our method well matched the results which were already known in SGD database.

[1] *http://biodata.mshri.on.ca/osprey/servlet/Index*

5 Acknowledgments

This work was supported by Systems Bio-Dynamics Research Center under KOSEF NCRC Program and POSTECH Research Fund.

References

[1] Y. Barash and N. Friedman, "Context-specific Bayesian clustering for gene expression data," *Journal of Computational Biology*, vol. 9, no. 2, pp. 169–191, 2002.

[2] J. F. Cardoso, "Multidimensional independent component analysis," in *Proc. ICASSP*, Seattle, WA, 1998.

[3] M. B. Eisen, P. T. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 14 863–14 868, 1998.

[4] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, pp. 601–620, 2000.

[5] Gene Ontology Consortium, "Creating the gene ontology resource: Design and implementation," *Genome Research*, vol. 11, pp. 1425–1433, 2001.

[6] N. Holter, M. Mitra, A. Maritan, M. Cieplan, J. Banavar, and N. Fedoroff, "Fundamental patterns underlying gene expression profiles: Simplicity from complexity," *Proc. Natl. Acad. Sci. USA*, vol. 97, pp. 8409–8414, 2000.

[7] A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invarianct features by decomposition of natural images into independent feature subspaces," *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, 2000.

[8] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, no. 1, pp. 51–60, 2002.

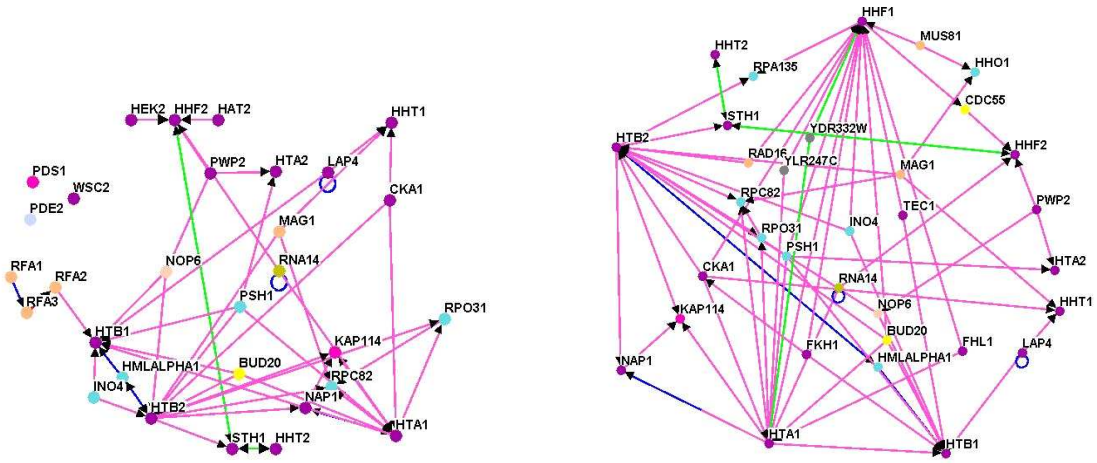
[9] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273–3297, Dec. 1998.

Table 1. List of genes associated with the 5(pos)th-mode in ICA, is shown. Interactive genes are selected, according to maximal cross-correlation between gene profiles. PSA1, PDS1, RFA3, and WSC2 do not match chromatin structure, although they are group together with HHF2, HTA1, HHT1, HTB1, HTB2, HTA2, HHT2.

SGD	Peak	Interactive gene	Process	Function
PSA1	G1	HHF2	mannose metabolism	mannose-1-phosphate guanyltransferase
PDS1	G1	RFA3	cell cycle	mannose-1-phosphate guanyltransferase
RFA3	G1	PDS1	DNA replication	anaphase inhibitor (putative)
HHF2	S	HHT2	chromatin structure	replication factor A, 13 kD subunit
HTA1	S	HTB1	chromatin structure	histone H4
HHT1	S	HTB1	chromatin structure	histone H2A
HTB1	S	HTA1	chromatin structure	histone H3
HTB2	S	HTA2	chromatin structure	histone H2B
HTA2	S	HTA1	chromatin structure	histone H2A
HHT2	S	HHF2	chromatin structure	histone H3
WSC2	S	HTB1	cell wall biogenesis	alpha-1,4-glucan-glucosidase

Table 2. List of genes associated with the 57th-mode in ISA, is shown. All of these genes correspond to chromatin structure.

SGD	Peak	Interactive gene	Process	Function
HHO1	S	HTA1	chromatin structure	histone H1
HTA2	S	HTA1	chromatin structure	histone H2A
HTB2	S	HTA2	chromatin structure	histone H2B
HHT2	S	HHF2	chromatin structure	histone H3
HTT1	S	HTB1	chromatin structure	histone H3
HHB1	S	HTA1	chromatin structure	histone H2B
HHF1	S	HHF2	chromatin structure	histone H4
HTA1	S	HTB1	chromatin structure	histone H2A
HHF2	S	HHT2	chromatin structure	histone H4



(a) (b)

Figure 3. Mutiple Interactions: The same colored circles represent that their characteristics is identical (based on GO and MIPS database). Each edge shows interaction between two genes. Most genes work for chromatin assembly/disassembly and DNA binding. Those exist only 24 among 7270 in total of yeast genes in SGD database. In (a), they were found 8 out of 13 with significance level 1.78^{-15}% ; in (b), 8 out of 9 with 1.26^{-17}%.