Journal of Proteomics & Bioinformatics - Open Access

Research Article JPB/Vol.2/March 2009

Identifying Conserved and Divergent Transcriptional Modules by Cross-species Matrix Decomposition on Microarray Data

Huai Li and Ming Zhan*

Bioinformatics Unit, Research Resources Branch,
National Institute on Aging, NIH, Baltimore, MD 21224, USA

*Corresponding author: National Institute on Aging, NIH,
251 Bayview Blvd, Baltimore, MD 21224, USA, Tel: (410)-558- 8373; E-mail: zhanmi@mail.nih.gov

Received January 29, 2009; Accepted March 11, 2009; Published March 12, 2009

Citation: Huai L, Ming Z (2009) Identifying Conserved and Divergent Transcriptional Modules by Cross-species Matrix Decomposition on Microarray Data. J Proteomics Bioinform 2: 117-125. doi:10.4172/jpb.1000068

Abstract

Cross-species comparison of gene expression profiles allows deciphering fundamental and species-specific transcriptional programs of cells and offers insight into organization and evolution of the genome and genetic network. Here, we propose an algorithm for comparing microarray data from different species to unravel transcriptional modules that are conserved or divergent through evolution. The proposed algorithm is based on cross-species matrix decomposition that includes a nonlinear independent component analysis followed a generalized probabilistic sparse matrix factorization on microarray data from different species. The proposed algorithm captures transcriptional modularity that might result from highly nonlinear interactions among genes, and partitions genes into mutually non-exclusive transcriptional modules. The conserved transcriptional modules are identified by the latent variables that are associated with predominant biological prototypes shared across species. We illustrated the application of the proposed algorithm by an analysis of human and mouse embryonic stem cell (ESC) data. The analysis uncovered conserved and divergent transcriptional modules in the ESC transcriptomes, shedding light on the understanding of fundamental and species-specific regulatory mechanisms controlling ESC development.

Keywords: Comparative transcriptomics; Transcriptional modules; Generalized probabilistic sparse matrix factorization; Embryonic stem cells

Abbreviations

GPSMF: Generalized Probabilistic Sparse Matrix Factorization

ESCs: Embryonic Stem Cells

NICA: Nonlinear Independent Component Analysis

Introduction

Given the completion of genomic sequencing of various mammalian and other organisms, transcriptomes of different species can be readily compared across species through the identification of orthologous genes (Ihmels et al., 2005; Li et al., 2007a; McCarroll et al., 2004; Stuartet al., 2003). One of the most important and widespread mechanisms used by a cell in functional regulation is the coordinate modulation and interaction of genes. By organizing genes into different transcriptional modules, a living cell coordinates the activities of genes and carries out complex functions. Important sequence elements in the genome, as well as important biological processes or pathways, are often evolutionarily conserved (Ihmels et al., 2005; Li et al., 2007b; Stuart

et al., 2003; Zhan et al., 2005). The comparative transcriptomics study allows uncovering transcriptional modules conserved or divergent through evolution, and has shown to be a powerful approach in deciphering fundamental or species-specific regulatory programs of cells and for insights into organization and evolution of the genome and genetic network (Alter et al., 2003; Bergmann et al., 2004; Ihmels et al., 2005; Li et al., 2007a; McCarrollet al., 2004; Stuart et al., 2003; Vallee et al., 2006; Zhou and Gibson 2004). In most comparative transcriptomics analyses, linear correlations between genes are evaluated using conventional clustering methods such as hierarchical clustering, k-means, and SOM, and genes are partitioned into mutually exclusive modules (Atkinson et al., 2003; Ihmels et al., 2005; Li et al., 2007b; Yuh et al., 1998; Zhou and Gibson 2004). However, nonlinear interactions among genes are often observed in transcriptional networks, such as in negative feedback events or two consecutive biological events of threshold and saturation (Alter et al., 2000; Li et al., 2007a). Moreover, a single gene may participate in multiple biological processes or pathway activities, so that belong to multiple transcriptional modules. In addition, the clustering-based methods identify transcriptional modules by assuming that genes with similar expression profiles share similar functions or pathways. However, genes involved in the same biological process or pathway can have different expression patterns (Li et al., 2007a; Zhou et al., 2002). Different from clustering-based methods, matrix decomposition methods (*e.g.* singular value decomposition, independent components analysis, non-negative matrix factorization, network component analysis, and sparse matrix factorization) do not cluster genes based on the pair-wise similarity measurement in microarray data analysis (Alter et al., 2000; Carmona-Saez et al., 2006; Chiappetta et al., 2004; Dueck et al., 2005; Frigyesi et al., 2006; Kim and Tidor, 2003; Lee and Batzoglou, 2003; Liao et al., 2003; Liebermeister, 2002; Wang et al., 2006). In these methods, genes with related functions or regulatory programs can be clustered together even they have different expression profiles. A gene can be partitioned to multiple mutually non-exclusive modules if the gene participates in multiple biological processes or have multiple functions. However, the matrix decomposition methods use linear models, describing gene expression as linear combinations of latent biological sources, which is often not true for gene-gene relationships and gene expression data. To overcome the problem, we recently developed a two-stage matrix decomposition method, which is based on a nonlinear independent component analysis (NICA) on the expression data, followed by probabilistic sparse matrix factorization (PSMF), for transcriptional module discovery (Li et al., 2007b). The method combines both projection and model-based approaches and is free from both linear-models and similarity measurements, providing a more suitable solution for transcriptional module discovery from gene expression data.

In the present study, we extend the two-stage decomposition method to cross-species studies on gene expression data for uncovering transcriptional modules conserved and divergent through evolution. A generalized probabilistic sparse matrix factorization (GPSMF) approach is particularly proposed to simultaneously decompose two independent latent component matrices from different species. A framework is then implemented for identifying evolutionarily conserved and divergent transcriptional modules from the outcomes of GPSMF and NICA analyses. In comparison with another method, our algorithm can better uncover functionally relevant transcriptional module. We applied the newly developed methodology in analyzing gene expression data of embryonic stem cells (ESCs) from human and mouse. The results demonstrated that the new algorithm can unravel conserved and divergent modules that are significantly associated to ESC development, shedding light on fundamental and species-specific mechanisms controlling ESC self-renewal and differentiation.

Materials and Methods

The NICA decomposition

Suppose we have two microarray data matrices $\mathbf{X}_1 \in \Re^{N_1 \times M}$ and $\mathbf{X}_2 \in \Re^{N_2 \times M}$ with the same sample size M, where N_1 and N_2 are the numbers of genes in the two data sets, the microarray data can be described by the noisy nonlinear mixing model

$$\begin{aligned} \mathbf{X}_1 &= f_1(\mathbf{S}_1) + \mathbf{O}_1 \\ \mathbf{X}_2 &= f_2(\mathbf{S}_2) + \mathbf{O}_2 \end{aligned} \quad (1)$$

where $\mathbf{S}_1 \in \Re^{N_1 \times M'}$ and $\mathbf{S}_2 \in \Re^{N_2 \times M'}$ denote the two latent source matrices, M' is the number of latent sources. $\mathbf{O}_1$ and $\mathbf{O}_2$ are the white Gaussian noise matrices. The nonlinear mappings $f_1(.)$ and $f_2(.)$ are modeled by a multilayer perceptron (MLP) network (Haykin, 1999) with one nonlinear hidden layer as:

$$f_1(\mathbf{S}_1) = \mathbf{C}_1 \cdot \tanh[\mathbf{S}_1 \cdot \mathbf{B}_1 + \mathbf{D}_1] + \mathbf{E}_1$$
$$f_2(\mathbf{S}_2) = \mathbf{C}_2 \cdot \tanh[\mathbf{S}_2 \cdot \mathbf{B}_2 + \mathbf{D}_2] + \mathbf{E}_2 \quad (2)$$

where $\mathbf{B}_m$ and $\mathbf{C}_m$ are the weight matrices of the hidden and output layers, and $\mathbf{D}_m$ and $\mathbf{E}_m$ are the corresponding bias matrices for m = 1,2.

Assuming that the source signals $\mathbf{S}_m$ at the input layer of the MLP network have simple Gaussian distributions, we obtain a nonlinear principal component analysis solution for $\mathbf{S}_m$ based on variational Bayesian learning for blind estimation and separation in the nonlinear mixture data model in Eq. (1). This solution models nonlinear mixtures (observed data), but provides no estimate of independent source signals. To find independent components from $\mathbf{S}_m$, we apply a standard linear ICA to $\mathbf{S}_m$ using the FastICA algorithm (Hyvarinen and Oja, 2000). The goal of the linear ICA is to decompose $\mathbf{S}_m = \overline{\mathbf{S}}_m \cdot \mathbf{A}_m$ so that columns (components) of $\overline{\mathbf{S}}_m$ are statistically as independent as possible.

The GPSMF model

The GPSMF approach we propose is a generalized extension of probabilistic sparse matrix factorization (Dueck et al., 2005) and used to decompose two data matrices simultaneously in comparative analysis of two microarray data sets. Given two matrices $\overline{\mathbf{S}}_1 \in \Re^{N_1 \times M'}$ and $\overline{\mathbf{S}}_2 \in \Re^{N_2 \times M'}$ derived from the NICA procedure, the GPSMF is to find $\mathbf{Y}_1 \in \Re^{N_1 \times L}$, $\mathbf{Y}_2 \in \Re^{N_2 \times L}$ and $\mathbf{Z} \in \Re^{L \times M'}$ such that $\overline{\mathbf{S}}_1 = \mathbf{Y}_1 \cdot \mathbf{Z}$ and $\overline{\mathbf{S}}_2 = \mathbf{Y}_2 \cdot \mathbf{Z}$. Here the columns of $\overline{\mathbf{S}}_1$ and $\overline{\mathbf{S}}_2$ represent independent latent components. N_1 and N_2 are the number of genes in the two data sets, M' is the number of latent sources. $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are factor weighting matrices. Each row of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ has at most K non-zero entries. Row vectors of $\mathbf{Z}$ contain unobserved L latent factor profiles. Specifically, let k_i^m be the number of non-zero entries ($k_i^m \le K$) of the row vector $\mathbf{y}_i^m \in \Re^{1 \times L}$ in $\mathbf{Y}_m$ and $\mathbf{l}_i^m = \left(l_{i1}, l_{i2,}, \cdots, l_{ik_i^m}\right)$ be the vector that contains column indices of non-zero entries of $\mathbf{y}_i^m$, where m=1, 2 for two data sets, we model each gene "hidden" expression profile across the independent latent component $\overline{\mathbf{s}}_i^m \in \Re^{1 \times M'}$, as a linear combination of k_i^m of the factor profiles $\mathbf{z}_l \in \Re^{1 \times M'}$, plus noise:

$$\overline{\mathbf{s}}_i^m = \sum_{k=1}^{k_i^m} y_{il_{ik}}^m \mathbf{z}_{l_{ik}} + \mathbf{n}_i^m \qquad m = 1,2 \quad (3)$$

Supposing the noise is Gaussian with variance σ_i^{m2} for $\overline{\mathbf{s}}_i^m$, then the likelihood of $\overline{\mathbf{s}}_i^m$ can be written as:

$$P(\overline{\mathbf{s}}_i^m \mid \mathbf{y}_i^m, \mathbf{Z}, \mathbf{l}_i^m, k_i^m, \sigma_i^{m2}) = \mathrm{N}(\sum_{k=1}^{k_i^m} y_{il_{ik}}^m \mathbf{z}_{l_{ik}}, \sigma_i^{m2}\mathbf{I})$$
$$m = 1,2 \quad (4)$$

Assume that $\mathbf{z}_l$ is normally distributed, $\mathbf{l}_i^m$ is uniformly distributed, and k_i^m is multinomially distributed. Multiplying these priors by Eq. (4) forms the joint distribution $P(\overline{\mathbf{S}}_m, \mathbf{Y}_m, \mathbf{Z}, \mathbf{L}_m, \mathbf{K}_m \mid \boldsymbol{\Sigma}_m)$. From the joint distribution, we first estimate elements in $\mathbf{Y}_1$ and $\mathbf{Z}$ by utilizing a factorized variational inference method (Dueck et al., 2005; Jordan et al., 1999) from $\overline{\mathbf{S}}_1$ that contains the independent latent components in the primary ("reference") organism. Then, we estimate elements in $\mathbf{Y}_2$ by the same method from $\overline{\mathbf{S}}_2$ that contains the independent latent components in the second ("target") organism and $\mathbf{Z}$.

When applying the GPSMF, the choices of the parameters L and K affect the structure of decomposition. L is the predefined number of possible latent variables that determines the number of modules identified by our algorithm. L should be much smaller than N_1 and N_2 (i.e. total gene numbers in the two data sets), since the expression of most genes is thought to be influenced by a small set of genes that act in combination as key regulators or network hubs to maintain the overall expression pattern of a transcriptional module. K is the maximum number of "effective" latent variables and should be less or equal to L. In the case of K = 1, each row in the data matrix is associated with only a single factor, and the sparse matrix factorization is a clustering of the data rows. When $K = L$, the factorization is simply a low rank approximation. Since our assumption is that the expression of each gene is determined by only a small set of possible key genes, we heuristically set K = 3 in our study.

Identification of conserved and divergent modules

Given the factor weighting matrices $\mathbf{Y}_1$, $\mathbf{Y}_2$, and the factor profile matrix $\mathbf{Z}$, we propose an approach for identifying conserved and divergent transcriptional modules. Let us define

setA$_l$:= (orthologous geneID1[i_1], satisfy $y_{i_1 l}^1 \neq 0$, for $i_1 = 1, \cdots, N_1$)

setB$_l$:= (orthologous geneID2[i_2], satisfy $y_{i_2 l}^2 \neq 0$, for $i_2 = 1, \cdots, N_2$)

where $y_{i_1 l}^1$ and $y_{i_2 l}^2$ are the elements of $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively, $l = 1, \cdots, L$. We then determine 1) conserved tran-

scriptional modules as the common orthologous genes in both setA$_l$ and setB$_l$ (i.e., set $A_l \cap$ set B_l); 2) divergent modules in the primary organism as the orthologous genes in setA$_l$ but not in setB$_l$ (i.e., set $A_l -$ set B_l); and 3) divergent modules in the second organism as the orthologous genes in setB$_l$ but not in setA$_l$ (i.e., set $B_l -$ set A_l).

Results and Discussion

Cross-species matrix decomposition of microarray data

Figure 1 shows a general schema of the proposed algorithm. The algorithm is based on two-stage matrix decomposition on two microarray datasets from different species to identify conserved or divergent transcriptional modules. We first apply the NICA transformation to capture the nonlinear structure in the data and represent the data with independent latent components. We then apply GPSMF to simultaneously decompose the two independent latent component matrices of different species. We finally identify conserved and divergent transcriptional modules from the outcomes of the matrix decomposition.

The NICA method that we adopt is based on a variational Bayesian learning (Jutten and Karhunen, 2004; Lappalainen and Honkela, 2000). The method uses a multilayer perceptron (MLP) network as a nonlinear mapping to model nonlinear mixtures of data. The MLP network can model any nonlinear mapping from sources to observed data with certain accuracy, given enough nodes in the hidden layer (Haykin, 1999). The MLP network also provides a flexible nonlinear mapping because its model complexity scales linearly with the dimension of the latent source space (Lappalainen and Honkela, 2000).

The GPSMF approach we propose is used to simultaneously decompose two matrices derived from the NICA procedure on microarray data of different species. The GPSMF models the expression profiles in each species as a linear weighted combination of profiles from a small number of prototypes that represent the influence of different biological or experimental factors shared by the two species. The GPSMF modeling is based on the following assumptions: a) the expression of each orthologous gene responding to experimental conditions is equivalent in both species; b) the expression profile of a gene is determined by a linear combination of hidden biological sources or variables, represented by the latent components; and c) the two

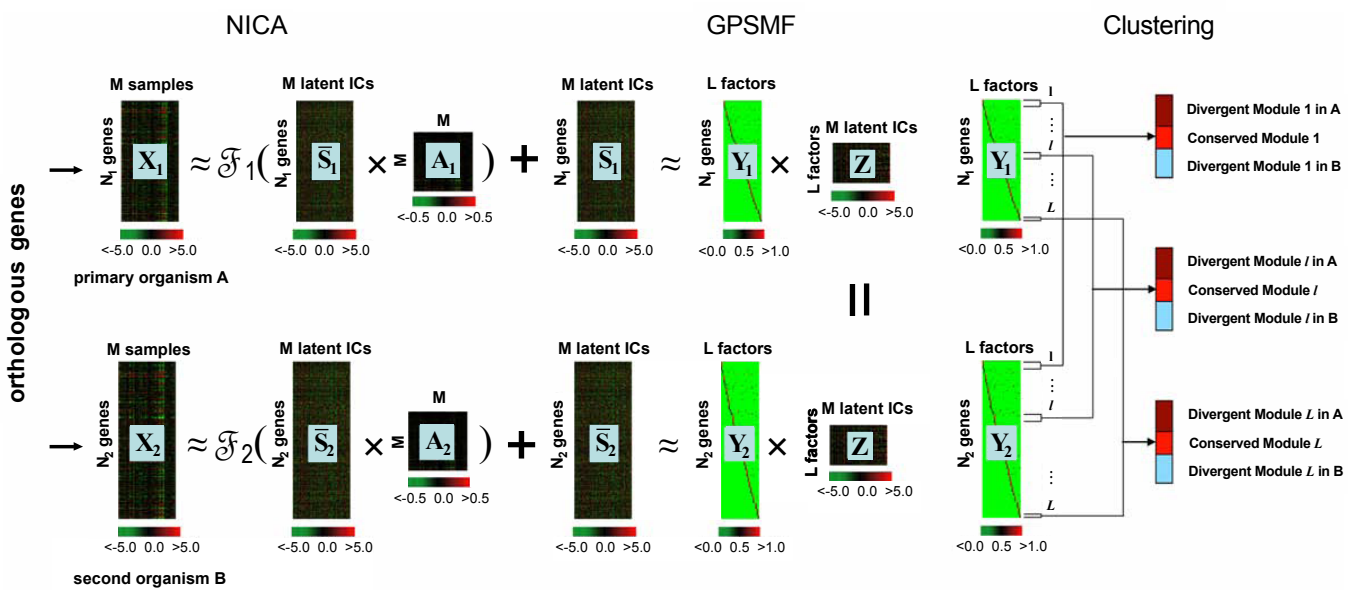

Figure 1: A general framework of the proposed algorithm

From orthologous gene expression profile data sets $\mathbf{X}_1$ (N_1 genes and M samples) in organism A and $\mathbf{X}_2$ (N_2 genes and M samples) in organism B, the NICA extracts nonlinear independent components (columns in $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$). At the GPSMF stage, $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$ are jointly approximated by the product of sparse matrix $\mathbf{Y}_1$ and low-rank $\mathbf{Z}$ and the product of sparse matrix $\mathbf{Y}_2$ and $\mathbf{Z}$, respectively. The values of all matrices are color coded by using a color heatmap, from dark green (minimum) to dark red (maximum). In the clustering process, conserved and divergent gene modules are identified by finding the common and different orthologous genes corresponding to nonzero indices of each column of $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

species examined share, to a certain degree, a small set of biological prototypes that have predominant influence on the expression patterns of most of genes. The GPSMF procedure is appropriate for modeling gene expression data across species, since while many genes are involved in gene regulation, a small set of transcriptional regulators or network hub genes have a predominant impact on the overall expression patterns of most of genes. The biological prototypes with the predominant impact and their activities are either conserved across species or divergent, which provides a basis for the identification of conserved and divergent transcriptional modules in our algorithm. The conserved transcriptional modules are likely related to fundamental biological processes, pathways or molecular mechanisms. The divergent modules are suggestive of species-specific transcriptional programs.

Transcriptional modules in embryonic stem cells

The gene expression in embryonic stem cells (ESCs) is carefully regulated so that the cells either maintain the pluripotent state by self-renewal or undergo differentiation. An understanding of gene regulatory mechanisms is essential for realizing the great potential of ESCs in regenerative medicine. The Oc4/Sox2/Nanog-directed network is a central regulatory circuitry controlling ESC self-renewal and differentiation (Boyer et al., 2005; Loh et al., 2006; Sun et al., 2006; Zhan, 2008; Zhan et al., 2005). However, whether there are fundamental or species-specific mechanisms underlying the activity of this critical network in ESCs has not been adequately explored. As described above, our method is particularly designed for identifying conserved and divergent transcriptional modules in which a small set of prototypes (*e.g.* transcriptional factors or network hub genes) control the overall expression pattern. The method is thus suitable for analyzing the Oct4/Sox2/Nanog-directed regulatory network for insight into fundamental and species-specific mechanisms in regulating ESC development.

For the analysis, we selected 1681 orthologous genes bound by Oct4, Sox2 and Nanog in human and mouse genomes, and examined their expression profiles determined from multiple cell lines of undifferentiated ESCs and their earliest differentiated counterparts, embryoid bodies (EBs). Totally 18 samples were examined for human and mouse cells, respectively. The human microarray data were obtained from our previous studies on ESCs and other publications (Li et al., 2006; Li et al., 2007a; Liu et al., 2006; Sato et al., 2003). The mouse microarray data were obtained from the GEO database (accession numbers: GSE3231, GSE2972 and GSE3749) (http://www.ncbi.nlm.nih.gov/geo). The expression data determined by BeadArray were normalized using the quantile method, and the data by Affymetrix were normalized using the RMA method. The expression data were then converted into log2 ratios of expression values over the average expression value for each gene. The missing data in the data sets were filled by KNN imputing. Human and mouse orthologous genes were obtained from the Affymetrix human-mouse ortholog links.

In the analysis, we set the number of independent latent components equal to the number of experimental conditions

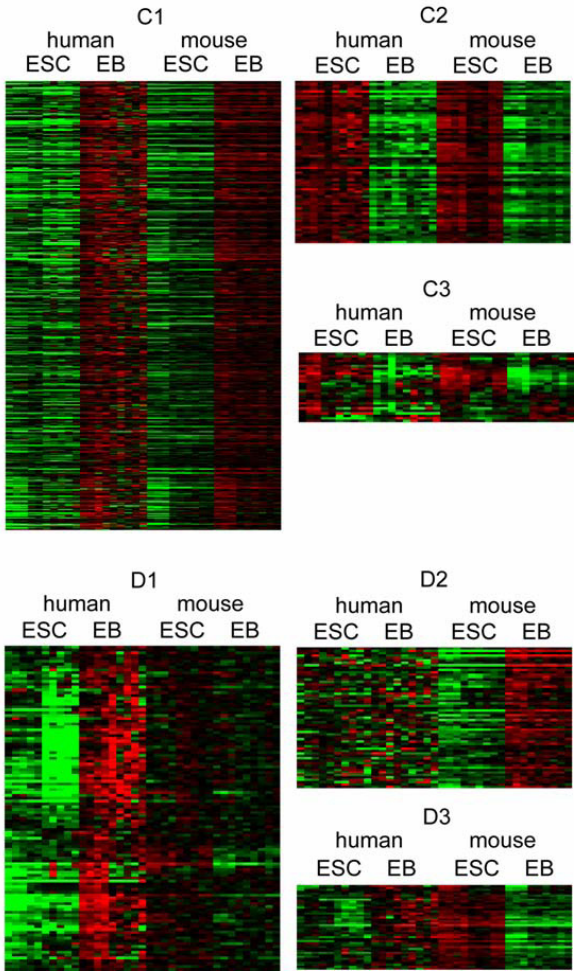

Figure 2: Heatmap of conserved and divergent transcriptional modules

Gene expression profiles of conserved (C1, C2, C3) and divergent (D1, D2, D3) transcriptional modules identified from Oct4/Sox2/Nang-directed network genes in human and mouse ESCs. Each module is presented by a heatmap of the expression profile (red: gene up-regulated in comparison with the mean, green: gene down-regulated, black: no change on the expression level).

for simplicity. We set the number of hidden neurons in the MLP network as twice as the number of independent latent components for an accurate nonlinear mapping. We also set *K* to 3 and *L* to 3 through 10 in the computation. For each identified transcriptional module, we identified biological processes or pathways that were significantly over-represented, using the Fisher's exact test followed by the false discovery rate adjustment. Figure 2 shows the identified conserved and divergent transcriptional modules with heatmap presentations of the expression profiles. The gene list of each transcriptional module is provided in Supplementary File 1.

We identified three conserved and three divergent transcriptional modules that showed distinctive expression patterns (Figure 2 - C1, C2, C3, D1, D2, D3; Supplementary File 1). The conserved module C1 showed repressed expression in undifferentiated ESCs of both human and mouse, as illustrated by the heatmap (Figure 2 - C1). The module, composed of 401 genes, was enriched by genes involved in development (27.3% of the total genes in the module, p-value 2.52×10^{-20}), morphogenesis (14.04%, p-value 5.92×10^{-12}), and cell differentiation (13.5%, p-value 2.24×10^{-11}). Embryonic development (p-value 4.93×10^{-7}), mesoderm development (p-value 2.70×10^{-3}), cell proliferation (p-value 2.61×10^{-5}), pattern specification (p-value 3.70×10^{-5}), embryonic pattern specification (p-value 1.50×10^{-3}), and apoptosis (p-value 1.60×10^{-3}) were particularly enriched in this module. Also included in this module were members of the Wnt pathway (p-value 6.87×10^{-5}; including Jun, GSK3b, Dkk1, Fzd1, Fzd2, Fzd8, Sfrp1, and Tbl1x), BMP pathway (p-value 2.70×10^{-3}; including Twsg1, Tob1, Gpc3, Bmp2, Prss11), TGFb pathway (Bmp2, Bmp5, Bmpr2, Smad3, Id2, and Pitx2), JAK-STAT pathway (Bcl2l1, Cntfr, Pias4, Stat2, and Stat3), and PI3K pathway (Eif2ak3, Pik4ca, Pip5k1c, and Pik3r1). All these enriched biological processes and signaling pathways are critical for ESC development (Li et al., 2007b; Sun et al., 2006). In addition, the module contained 30 transcription factors, including Hand1, GATA6 and ZIC1, which are know to be repressed in ESCs of both human and mouse (Li et al., 2007a; Sun et al., 2006). The conserved module C2, on the other hand, showed elevated expression in undifferentiated ESCs of both human and mouse (Figure 2 - C2). This transcriptional module, containing 67 orthologous genes, was enriched by genes participating in cell cycle (14.7% of the module genes, p-value 2.11×10^{-4}) and regulation of biological process (38.2%, p-value 2.71×10^{-5}). The conserved transcriptional module C3, however, showed a mixed expression pattern in both human and mouse ESCs (Figure 2 - C3). The module contained 28 genes, mainly participating in metabolism (67.9%, p-value 1.30×10^{-3}). The module also included members of the Wnt pathway (Myc, Senp2, Ppp2r1a). The divergent transcriptional modules, on the other hand, shared little similarities between human and mouse ESCs on the predominant regulatory programs. The divergent module D1 showed transcriptional modularity in human but not in mouse ESCs, as illustrated by the heatmap (Figure 2 - D1). The module, composed of 106 orthologous genes, showed repressed expression in undifferentiated ESCs in human but little transcriptional changes during ESC differentiation in mouse. The genes of the module were mainly involved in development (26.4% of the total genes in this module; p-value 1.29×10^{-5}) and morphogenesis (16.9%; p-value 1.48×10^{-5}). Also enriched in the module were pattern specification (p-value 5.55×10^{-4}), cell differentiation (p-value 4.80×10^{-3}) and cell fate commitment (p-value 1.40×10^{-2}). The module included six transcription factors: Lef1, Hoxb5, Hoxb6, Hoxb9, Hoxc5 and Rax, all of which were related to development. The module also included members of the TGFb pathway (Bmp4, Thbs3, and Inhba). The divergent module D2, on the other hand, showed transcriptional modularity in mouse but not in human ESCs (Figure 2 - D2). The module consisted of 59 orthologous genes, which were repressed in undifferentiated ESCs in mouse but showed no consistent trend of expressional changes in human ESCs. The module was enriched by genes participating in development (25.4%, p-value 3.60×10^{-3}) and morphogenesis (15.6%, p-value 8.60×10^{-3}), including embryonic development (p-value 1.50×10^{-2}) and pattern specification (p-value 2.60×10^{-3}). Also enriched were the BMP signaling pathway (p-value 3.40×10^{-3}) and TGFb signaling pathway (p-value 6.00×10^{-3}; including ligand Bmp7, receptor Acvr1b, and signal transducers Smad1 and Smad 5). The divergent module D3, strikingly, showed transcriptional modularity of opposite transcriptional trends between human and mouse ESCs (Figure 2 - D3). The 43 orthologous genes of this module were over-expressed in mouse ESCs while under-expressed in human ESCs. The module was mainly enriched by genes involved in translation (14.0%, p-value 5.24×10^{-5}), particularly translational initiation (9.3%, p-value 2.52×10^{-4}) and regulation of translational initiation (7.0%, p-value 1.00×10^{-3}).

The conserved and divergent transcriptional modules underlie the fundamental and species-specific gene regulatory mechanisms in ESCs. The results of this study suggest

that the Oct4-Sox2-Nanog-directed regulatory network is not only responsible for primary "stemness" properties of ESCs, but also maintains species-specific programs in regulating pluripotency. The results are consistent with the fact that significant differences exist on the potential targets of Oct4, Sox2 and Nanog between human and mouse, despite of a certain overlap (Loh et al., 2006).

Comparison with generalized singular value decomposition (GSVD)

A GSVD-based matrix decomposition method was previously proposed for cross-species analysis of gene expression data (Alter et al., 2003). Different from our method, the GSVD method is based on a linear model and one-step matrix decomposition, conducted by singular matrix decomposition. We compared our method with the GSVD method through analysis on the same set of ESC expression data to further evaluate our method. We firstly identified three conserved modules and three divergent modules using the GSVD method, as we did by using our method, from the human and mouse ESC data. We then conducted functional analyses on each of the identified modules, using DAVID tools (Huang da et al., 2007). Table 1 shows the statistical enrichment of functional categories in the transcriptional modules identified by both methods. The functional categories included Gene Ontology (GO) terms, protein-protein interactions, protein functional domains, bio-pathways, and literatures. The enrichment level was calculated by transforming the enrichment p value after FDR correction to a negative log value and averaged over all functional categories for corrected $p < 0.05$. If no functional categories were found for corrected $p < 0.05$, the smallest value of corrected p was taken for calculating the enrichment level. As illustrated, our method outperformed the GSVD method on the functional enrichment in both the conserved and divergent modules. This implies that our method can better identify functionally coherent transcriptional modules that are either conserved or divergent through evolution.

Future Perspectives

The method presented in this study demonstrates success in identifying evolutionarily conserved and divergent transcriptional modules. Nevertheless there remain limitations of the reported method. For example, the current approach can only apply to microarray data of two species. Moreover, challenges still remain in designing *in vitro* or *in vivo* experiments to validate the results predicted by ours or other approaches. In the further research, we will extend our approach to compare more than two species. We will also integrate gene expression data with transcription

Our method		GSVD-based method	
Gene module	Enrichment level	Gene module	Enrichment level
Conserved modules			
C1 (401 genes)	3.63	C1 (250 genes)	3.08
C2 (68 genes)	3.32	C2 (74 genes)	2.47
C3 (28 genes)	1.87	C3 (39 genes)	1.66
Averaged over conserved modules	2.94	Averaged over conserved modules	2.40
Divergent modules			
D1 (106 genes)	1.44	D1 (101 genes)	1.25
D2 (59 genes)	1.07	D2 (57 genes)	0.84
D3 (43 genes)	2.30	D3 (44 genes)	1.68
Averaged over conserved modules	1.60	Averaged over conserved modules	1.26

Table 1: Comparison with the GSVD method based on DAVID functional analysis

Three conserved modules and three divergent modules identified by each method were evaluated. The functional enrichment level in each gene module is calculated by transforming the enrichment p value after FDR correction to a negative log value and averaged over all functional categories for corrected $p < 0.05$. If no functional categories are found for corrected $p < 0.05$, the smallest value of corrected p is taken for calculating the enrichment level.

factor binding information into this method to identify TF-mediated regulatory modules conserved or divergent across species. This would allow experimental validation of genes in a regulatory module by either ChIP methods or RNAi–mediated depletion of the specific transcription factors.

Conclusion

In this study, we present an algorithm for cross-species analysis of microarray data to address the challenge of discovering transcriptional modules conserved and divergent through evolution. The proposed algorithm tackles two microarray data sets from different species as inputs, imposing two stage matrix decomposition on the microarray data, firstly by NICA and then by GPSMF. The new algorithm captures transcriptional modularity that might result from highly nonlinear interactions among genes, and partitions genes into mutually non-exclusive transcriptional modules. The conserved transcriptional modules are identified by the latent variables that are associated with predominant biological prototypes shared across species. The identified transcriptional modules are highly associated with biological functions, in comparison with those identified by another method. As demonstrated by the analysis on human and mouse ESC data, the newly developed methodology can identify evolutionarily conserved and divergent transcriptional modules and facilitate the comparative transcriptomics studies.

Acknowledgements

This study was supported by the Intramural Research Program, National Institute on Aging, NIH.

References

1. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 97: 10101-10106. » CrossRef » Pubmed » Google Scholar

2. Alter O, Brown PO, Botstein D (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci USA 100: 3351-3356. » CrossRef » Pubmed » Google Scholar

3. Atkinson MR, Savageau MA, Myers JT, Ninfa AJ (2003) Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli. Cell 113: 597-607. » CrossRef » Pubmed » Google Scholar

4. Bergmann S, Ihmels J, Barkai N (2004) Similarities and differences in genome-wide expression data of six organisms. PLoS Biol 2: E9. » CrossRef » Pubmed » Google Scholar

5. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947-956. » CrossRef » Pubmed » Google Scholar

6. Carmona-Saez P, Pascual-Marqui RD, Tirado F, CarazoJM, Pascual-Montano A (2006) Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. BMC Bioinformatics 7: 78. » CrossRef » Pubmed » Google Scholar

7. Chiappetta P, Roubaud MC, Torresani B (2004) Blind source separation and the analysis of microarray data. J Comput Biol 11: 1090-1109. » CrossRef » Pubmed » Google Scholar

8. Dueck D, Morris QD, Frey BJ (2005) Multi-way clustering of microarray data using probabilistic sparse matrix factorization. Bioinformatics 1: i144-i151. » CrossRef » Pubmed » Google Scholar

9. Frigyesi A, Veerla S, Lindgren D, Hoglund M (2006) Independent component analysis reveals new and biologically significant structures in microarray data. BMC Bioinformatics 7: 290. » CrossRef » Pubmed » Google Scholar

10. Haykin S (1999) Neural networks: a comprehensive foundation. 2nd edition (Upper Saddle River, New Jersey: Prentice Hall). » CrossRef » Google Scholar

11. Huang da W, Sherman BT, Tan Q, Kir J, Liu D, et al. (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res 35: W169-175. » CrossRef » Pubmed » Google Scholar

12. Hyvarinen A, Oja E (2000) Independent component analysis: algorithms and applications. Neural Netw 13: 411-430. » CrossRef » Pubmed » Google Scholar

13. Ihmels J, Bergmann S, Berman J, Barkai N (2005) Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program. PLoS Genet 1: e39. » CrossRef » Pubmed » Google Scholar

14. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. In Jordan, M.I. (ed), Learning in Graphical

Models. MIT Press Cambridge pp105-161.

15. Jutten C, Karhunen J (2004) Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures. Int J Neural Syst 14: 267-292. » CrossRef » Pubmed » Google Scholar

16. Kim PM, Tidor B (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res 13: 1706-1718. » CrossRef » Pubmed » Google Scholar

17. Lappalainen H, Honkela A (2000) Bayesian nonlinear independent component analysis by multi-layer perceptrons. In Girolami, M. (ed), Advances in Independent Component Analysis. Springer-Verlag Berlin pp93-121.

18. Lee SI, Batzoglou S (2003) Application of independent component analysis to microarrays. Genome Biol 4: R76. » CrossRef » Pubmed » Google Scholar

19. Li H, Liu Y, Shin S, Sun Y, Loring JF, et al. (2006) Transcriptome coexpression map of human embryonic stem cells. BMC Genomics 7: 103. » CrossRef » Pubmed » Google Scholar

20. Li H, Sun Y, Zhan M (2007a) Analysis of gene coexpression by B-spline based CoD estimation. J Bioinform Sys Biol 2007: 1-10. » CrossRef » Pubmed » Google Scholar

21. Li H, Sun Y, Zhan M (2007b) The discovery of transcriptional modules by a two-stage matrix decomposition approach. Bioinformatics 23: 473-479. » CrossRef » Pubmed » Google Scholar

22. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, et al. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci USA 100: 15522-15527. » CrossRef » Pubmed » Google Scholar

23. Liebermeister W (2002) Linear modes of gene expression determined by independent component analysis. Bioinformatics 18: 51-60. » CrossRef » Pubmed » Google Scholar

24. Liu Y, Shin S, Zeng X, Zhan M, Gonzalez R, et al. (2006) Genome wide profiling of human embryonic stem cells (hESCs), their derivatives and embryonal carcinoma cells to develop base profiles of U.S. Federal government approved hESC lines. BMC Dev Biol 6: 20. » CrossRef » Pubmed » Google Scholar

25. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, et al. (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet 38: 431-440. » CrossRef » Pubmed » Google Scholar

26. McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, et al. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nat Genet 36: 197-204. » CrossRef » Pubmed » Google Scholar

27. Sato N, Sanjuan IM, Heke M, Uchida M, Naef F, et al. (2003) Molecular signature of human embryonic stem cells and its comparison with the mouse. Dev Biol 260: 404-413. » CrossRef » Pubmed » Google Scholar

28. Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302: 249-255. » CrossRef » Pubmed » Google Scholar

29. Sun Y, Li H, Yang H, Rao MS, Zhan M (2006) Mechanisms controlling embryonic stem cell self-renewal and differentiation. Crit Rev Eukaryot Gene Expr 16: 211-231. » CrossRef » Pubmed » Google Scholar

30. Vallee M, Robert C, Methot S, Palin MF, Sirard MA (2006) Cross-species hybridizations on a multi-species cDNA microarray to identify evolutionarily conserved genes expressed in oocytes. BMC Genomics 7: 113. » CrossRef » Pubmed » Google Scholar

31. Wang G, Kossenkov AV, Ochs MF (2006) LS-NMF: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates. BMC Bioinformatics 7: 175. » CrossRef » Pubmed » Google Scholar

32. Yuh CH, Bolouri H, Davidson EH (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279: 1896-1902. » CrossRef » Pubmed » Google Scholar

33. Zhan M (2008) Genomic studies to explore self-renewal and differentiation properties of embryonic stem cells. Front Biosci 13: 276-283. » CrossRef » Pubmed » Google Scholar

34. Zhan M, Miura T, Xu X, Rao MS (2005) Conservation and variation of gene regulation in embryonic stem cells assessed by comparative genomics. Cell Biochem Biophys 43: 379-405. » CrossRef » Pubmed » Google Scholar

35. Zhou X, Kao MC, Wong WH (2002) Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci USA 99: 12783-12788. » CrossRef » Pubmed » Google Scholar

36. Zhou XJ, Gibson G (2004) Cross-species comparison of genome-wide expression patterns. Genome Biol 5: 232. » CrossRef » Pubmed » Google Scholar