Bioinformatics, 32, 2016, i137–i146
doi: 10.1093/bioinformatics/btw278
ISMB 2016

OXFORD

A cross-species bi-clustering approach to identifying conserved co-regulated genes

Jiangwen Sun[1,†], Zongliang Jiang[2,†], Xiuchun Tian[2] and Jinbo Bi[1,*]

[1]Department of Computer Science and Engineering and [2]Center for Regenerative Biology and Department of Animal Science, University of Connecticut, Storrs, CT 06269, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as Joint First Authors.

Abstract

Motivation: A growing number of studies have explored the process of pre-implantation embryonic development of multiple mammalian species. However, the conservation and variation among different species in their developmental programming are poorly defined due to the lack of effective computational methods for detecting co-regularized genes that are conserved across species. The most sophisticated method to date for identifying conserved co-regulated genes is a two-step approach. This approach first identifies gene clusters for each species by a cluster analysis of gene expression data, and subsequently computes the overlaps of clusters identified from different species to reveal common subgroups. This approach is ineffective to deal with the noise in the expression data introduced by the complicated procedures in quantifying gene expression. Furthermore, due to the sequential nature of the approach, the gene clusters identified in the first step may have little overlap among different species in the second step, thus difficult to detect conserved co-regulated genes.

Results: We propose a cross-species bi-clustering approach which first denoises the gene expression data of each species into a data matrix. The rows of the data matrices of different species represent the same set of genes that are characterized by their expression *patterns* over the developmental stages of each species as columns. A novel bi-clustering method is then developed to cluster genes into subgroups by a joint sparse rank-one factorization of all the data matrices. This method decomposes a data matrix into a product of a column vector and a row vector where the column vector is a consistent indicator across the matrices (species) to identify the same gene cluster and the row vector specifies for each species the developmental stages that the clustered genes co-regulate. Efficient optimization algorithm has been developed with convergence analysis. This approach was first validated on synthetic data and compared to the two-step method and several recent joint clustering methods. We then applied this approach to two real world datasets of gene expression during the pre-implantation embryonic development of the human and mouse. Co-regulated genes consistent between the human and mouse were identified, offering insights into conserved functions, as well as similarities and differences in genome activation timing between the human and mouse embryos.

Availability and Implementation: The R package containing the implementation of the proposed method in C++ is available at: https://github.com/JavonSun/mvbc.git and also at the R platform https://www.r-project.org/.

Contact: jinbo@engr.uconn.edu

1 Introduction

The process of mammalian pre-implantation embryonic development is characterized by the degradation of maternal RNA stored in the oocytes and the gradual activation of the embryonic genome. Rapid advances in the whole-genome RNA sequencing techniques has led to a growing number of studies exploring gene regulation during pre-implantation embryonic development in different species (Blakeley *et al.*, 2015; Cao *et al.*, 2014; Graf *et al.*, 2014; Jiang *et al.*, 2014; Xue *et al.*, 2013; Yan *et al.*, 2013). Several studies have shown that the timing of embryonic genome activation varies by species (Braude *et al.*, 1988; Cao *et al.*, 2014; Graf *et al.*, 2014; Hamatani *et al.*, 2004; Jiang *et al.*, 2014; Misirlioglu *et al.*, 2006; Wang *et al.*, 2004). The understanding of this variation may bring insights into embryonic developmental programming and species differences. Identifying the co-regulated gene clusters that are conserved across species is a key component in the understanding of this variation (Jiang *et al.*, 2014; Xue *et al.*, 2013). Such conserved gene clusters are likely involved in common biological processes that are fundamental to the embryonic development of mammals. However, due to the lack of effective computational methods, there has been limited understanding of the conservation of gene co-regulation during embryonic development.

In a typical study of mammalian embryonic development, expression levels of all genes are collected at multiple developmental milestones (stages), such as oocytes, 2-cell and 8-cell embryos. Gene expression data of different species are analyzed and compared to understand the similarities and variations in the embryonic development of the species. The most sophisticated method available so far for identifying conserved co-regulated genes consists of two steps in sequence (Jiang *et al.*, 2014; Xue *et al.*, 2013). First, co-regulated gene clusters are identified in each individual species by performing a cluster analysis of their gene expression data, usually by a hierarchical clustering method. Second, by computing overlaps among identified clusters in different species, co-regulated gene clusters that are conserved among species may be found. This two-step approach can be ineffective in two ways. There are innegligible noises in the expression data resulted from the complicated procedures in quantifying gene expression. The noises may prevent the detection of biologically meaningful and important gene clusters for each species (Jiang *et al.*, 2014). Moreover, the clusters identified in the first step may have no overlaps in the second step, thus unable to identify conserved gene clusters. In this paper, we address these two issues by proposing a novel cross-species bi-clustering approach.

A variety of methods have been proposed to reduce noise from a dataset, such as those for smoothing out noise, or identifying and removing outliers (Han *et al.*, 2011). However, a proper and effective noise reduction method is problem-specific. To identify *co-regulated* gene clusters, we search for genes that exhibit similar expression patterns over the embryonic developmental stages. We define that an *expression pattern* (or simply a *pattern*) is a specific series of high and low expression levels over a set of developmental stages. For example, in a study with three stages: oocytes, 2-cell and 4-cell embryos, the sequence of [*high, low, low*] is a pattern that a gene may follow, indicating that the gene has high expression level in oocytes, but low levels in the 2-cell and 4-cell stages. In order to reduce noise and focus on the biologically confirmed gene expression patterns, we propose to transform the raw gene expression data to reflect how closely the expression levels follow known patterns. In the new data matrix, rows represent genes and each column corresponds to a pattern in a pre-compiled list of patterns. Each gene is measured by the similarity between its gene expression path and each of the patterns in the list.

Instead of a separate cluster analysis for each species, we propose to integrate gene expression data of multiple species to search for confirmatory co-regulated gene clusters directly. This integrative way of data analysis allows the searching process to target at the gene clusters that show similar patterns across species. The multi-species joint cluster analysis corresponds to a machine learning principle: multi-view cluster analysis (Sun *et al.*, 2015), where the same set of subjects (i.e. genes here) is viewed in different input spaces, particularly here, in the developmental stages of different species. Further, we need to determine the expression patterns in each view (i.e. the columns in each data matrix) that are responsible for the grouping of subjects. Multi-view cluster analysis aims to group subjects into clusters in the same way no matter which view of data is used. However, most of the existing multi-view clustering methods assume that all columns in the data contribute equally in determining the clusters (Cai *et al.*, 2013; Chaudhuri *et al.*, 2009; Cheng *et al.*, 2013; Culp and Michailidis, 2009; Kumar and Daume, 2011; Langfelder and Horvath, 2008; Liu *et al.*, 2013). These methods cannot identify the specific patterns that the clustered genes actually follow. Even though a gene may follow multiple known patterns, the number of these patterns is much smaller than the total amount of pre-compiled biological patterns. Hence, these existing multi-view clustering methods are not suitable for solving our problem. We recently proposed two new multi-view bi-clustering methods (Sun *et al.*, 2014, 2015) that can identify consistent clusters across views and simultaneously specify a subset of variables in each view on which the genes in a cluster show high similarity. However, the algorithm developed in Sun *et al.* (2014), although is efficient, has not obtained a theoretical guarantee for convergence so far. The method in Sun *et al.* (2015) requires to predetermine the cluster size (i.e. the number of genes in a cluster) before the algorithm can be applied, which is obviously difficult to estimate for the gene co-regulation problem.

In this paper, we thus propose another new multi-view bi-clustering method that identifies both the gene clusters consistent across multiple species (views) and the expression patterns of the clustered genes for each species. By a sparse rank-one matrix factorization, this method decomposes a data matrix into a product of a sparse column vector and a sparse row vector. The non-zero entries of these vectors indicate the gene clusters and the selected expression patterns, respectively. We propose to use another sparse column vector to link the different data matrices. This column vector is used to enforce that the decomposed column vectors from every view correspond to the same subset of genes. The resultant optimization problem can be solved efficiently by developing an alternating optimization algorithm. Compared to the methods in Sun *et al.* (2014,2015), the proposed method is guaranteed to converge to a stationary point and does not require any prior knowledge of cluster size. We compared the proposed method in simulations to the traditional two-step approach, and several latest multi-view clustering methods developed by others, which demonstrated the superiority of our method. We then used the proposed approach to analyze the pre-implantation embryonic development datasets of the human and mouse. Across the two species, 22 co-regulated gene clusters were identified to be conserved. A gene ontology analysis of the identified genes showed that they are involved in many fundamental biological networks. The expression patterns associated with these clusters were compared between the human and mouse embryos, showing that there are both similarities and variations between the human and mouse in the gene activation timing during the early development.

We briefly introduce the notation used throughout this paper. We use a bold-font upper case letter such as $\mathbf{X}$ to represent a matrix, a bold-font lower case letter such as $\mathbf{v}$ to denote a column vector,

and a lower case letter such as a to represent a scalar. We denote the component of $\mathbf{X}$ at the ith row and jth column by $\mathbf{X}(i,j)$ or x_{ij}, and the ith row and jth column of $\mathbf{M}$, respectively, by $\mathbf{X}(i,\cdot)$, and $\mathbf{X}(\cdot,j)$. Similarly, we use $\mathbf{v}(i)$ to denote the ith component of $\mathbf{v}$. The Frobenius norm of a matrix $\mathbf{X}$ is denoted by $||\mathbf{X}||_F$ which is calculated as $\sqrt{\sum_i \sum_j |x_{ij}|^2}$. Further, the ℓ_1-norm of a vector $\mathbf{v}$ is denoted by $||\mathbf{v}||_1$ and calculated as $\sum_i |v_i|$, where v_i is its ith component. The operator $\mathbf{z} \odot \mathbf{u}$ is the element wise product of $\mathbf{z}$ and $\mathbf{u}$. We use an italic upper case letter as S to represent a set of elements.

2 Pattern preserving noise reduction

We start from introducing our noise reduction technique that aims to preserve the important expression patterns identified in the literature or in hypothesized biological processes. A list of patterns can be pre-compiled by collecting them from the current literature of embryonic development. Note that expression patterns can also be created by a biological hypothesis, and our algorithm will automatically evaluate if the patterns are useful for identifying conserved co-regulated genes. If a specific analysis is not interested in a known pattern, the pattern can be excluded from the list. Particularly in this paper, we have compiled 22 and 18 gene expression patterns, respectively, for the human and mouse pre-implantation embryonic developmental processes. (Readers can consult with Tables 2 and 3 in Section 5 for details.)

Although the actual gene expression data are continuous, the patterns are represented by discretized expression levels. For instance, if seven developmental stages: oocytes, pronucleus, zygote, 2-cell, 4-cell, 8-cell and morula, are considered, a gene is expressed high in oocytes, medium in pronucleus but low in the rest of the stages. This gene may be characterized by the following two patterns: a pattern with a *high* value in oocytes and a *low* value for all subsequent stages, or another pattern with a *high* value in both oocytes and pronucleus but a *low* value for the other stages. If we summarize all patterns using binary levels such as high and low, we can represent each of the two patterns by a 7-entry vector: [1, 0, 0, 0, 0, 0, 0] and [1, 1, 0, 0, 0, 0, 0] where 1 means *high* and 0 means *low*.

We transform a gene's expression levels at the different developmental stages into a vector of length that is equal to the number of pre-compiled patterns (e.g. d). Let p represent a pattern and its values at the different developmental stages form a vector denoted by $\mathbf{y}$. Let g represent the actual gene expression of a gene and its values at the different stages form another vector denoted by $\mathbf{x}$. The correlation between the two random variables p and g is computed as follows:

$$\mathrm{cor}(p,g) = \frac{\mathrm{cov}(p,g)}{\mathrm{var}(p)\mathrm{var}(g)},$$

where $\mathrm{cov}(p,g)$ is the sample covariance of p and g and calculated as:

$$\mathrm{cov}(p,g) = \sum_i y_i x_i - \frac{\sum_i y_i \sum_i x_i}{n},$$

$\mathrm{var}(p)$ and $\mathrm{var}(g)$ are sample variance and can be calculated as:

$$\mathrm{var}(p) = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n},$$

and

$$\mathrm{var}(g) = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n},$$

respectively. Given a threshold t, we can determine that the gene g follows the pattern p if $\mathrm{cor}(p,g) \geq t$. Hence, the expression levels of a specific gene g are converted into a binary vector of length d where a value of 1 indicates that the gene follows (is highly correlated to) the corresponding pattern, and a value of 0 means otherwise. The transformed data matrix for a species is an $n \times d$ matrix of binary values where n represents the number of genes.

3 Multi-view bi-clustering

3.1 Sparse rank-one matrix factorization

Given a data matrix $\mathbf{X}_{n\times d}$ of n genes and d variables, its rank-one matrix factorization can be represented by $\mathbf{u}\mathbf{v}^T$, where vector $\mathbf{u}$ is of length n and vector $\mathbf{v}$ is of length d. When we enforce $\mathbf{u}$ and $\mathbf{v}$ to be sparse, the optimal factorization captures the most prominent block structure in $\mathbf{X}$ because the rows and columns included in a block (as indicated by the non-zero entries of $\mathbf{u}$ and $\mathbf{v}$) naturally form row and column clusters, respectively. More precisely, the rows corresponding to non-zero values in $\mathbf{u}$ form a row (subject) cluster. The columns corresponding to non-zero values in $\mathbf{v}$ form a column (variable) cluster. This is illustrated in Figure 1, where darker color indicates a larger value at the corresponding position in $\mathbf{X}$ assuming all values in $\mathbf{X}$ are positive.

The optimal sparse rank-one matrix factorization of $\mathbf{X}$ can be found by solving the following optimization problem:

$$\min_{\mathbf{u},\mathbf{v}} ||\mathbf{X} - \mathbf{u}\mathbf{v}^T||_F^2 + \lambda_u||\mathbf{u}||_1 + \lambda_v||\mathbf{v}||_1. \quad (1)$$

The term $||\mathbf{X} - \mathbf{u}\mathbf{v}^T||_F^2$ is for achieving the closest approximation of $\mathbf{X}$, while $\lambda_u||\mathbf{u}||_1$ and $\lambda_v||\mathbf{v}||_1$ enforce the sparsity of $\mathbf{u}$ and $\mathbf{v}$. This optimization problem can be efficiently solved by alternatively solving two subproblems until convergence: (i) solving $\mathbf{v}$ while fixing $\mathbf{u}$, (ii) solving $\mathbf{u}$ while fixing $\mathbf{v}$. Both of the two subproblems have an analytical solution, which will be discussed in detail in Section 4. Problem (1) is different from sparse singular value decomposition as in Lee *et al.* (2010) because both $\mathbf{u}$ and $\mathbf{v}$ are not required to be unit vectors, and we do not have a scalar, i.e. the singular value in Lee *et al.* (2010), involved in Problem (1) as a variable. The bi-convexity of our formulation (which is convex in terms of u and v when one of them is fixed) ensures a better convergence property for the alternating algorithm. We will discuss this in more detail when the optimization algorithm is introduced in Section 4.

3.2 Multi-view sparse rank-one matrix factorization

We have discussed how we obtain gene clusters and their associated variables using the data matrix of one species in one view. Now we introduce the procedure to obtain consistent gene clusters across

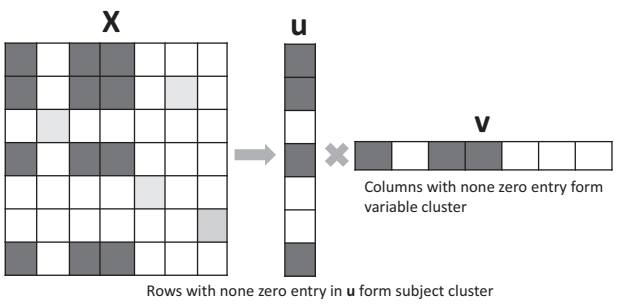


Fig. 1. Sparse rank-one matrix factorization of $\mathbf{X}$: $\mathbf{u}\mathbf{v}^T$. All values in $\mathbf{X}$ are assumed to be positive. Heavier color represents larger value at corresponding opposition in $\mathbf{X}$

multiple views and simultaneously identify their associated variables in each view. We propose to use a common vector $\mathbf{z}$ to link together the rank-one matrix factorization of multiple data matrices. Let m be the number of views, the proposed formulation is as follows:

$$\min_{\mathbf{z},\mathbf{u}_i,\mathbf{v}_i} \sum_{i=1}^{m} ||\mathbf{X}_i - (\mathbf{z} \odot \mathbf{u}_i)\mathbf{v}_i^T||_F^2 + \lambda_z ||\mathbf{z}||_1 + \sum_{i=1}^{m} \lambda_{u_i} ||\mathbf{u}_i||_1 + \sum_{i=1}^{m} \lambda_{v_i} ||\mathbf{v}_i||_1. \quad (2)$$

Here, we enforce $\mathbf{z}$ to be sparse for identifying common gene cluster across all views because when a component in $\mathbf{z}$ is zero, $\mathbf{u}$ will automatically have a value of zero at the corresponding position.

Let $\widehat{\mathbf{z}}$, $\widehat{\mathbf{u}}_i$ and $\widehat{\mathbf{v}}_i$ be the optimal solution of Problem (2). There are two different approaches to obtaining gene clusters by inspecting $\widehat{\mathbf{z}}$ and $\widehat{\mathbf{u}}_i$. One way is to look for none zero entries in $\mathbf{z}$ and construct cluster by including all instances with none zero entry in $\mathbf{z}$. The other approach is to form cluster by including only subjects with non-zero entries in all $\widehat{\mathbf{u}}_i$. Let A and B be the two sets of subjects in the clusters defined by first and second approach, respectively. Let C_i be the set of subjects with none zero in $\widehat{\mathbf{u}}_i$. Since any subject with zero in $\widehat{\mathbf{z}}$ has zero in every $\widehat{\mathbf{u}}_i$, and also any subject with zero in all $\widehat{\mathbf{u}}_i$ has zero in $\mathbf{z}$, so we have $A = \cup C_i$. In addition, we have $B = \cap C_i$ by definition, so $A \supseteq B$. The choice between these two options depends on the nature of the problem being solved. In an application, such as identifying conserved co-regulated gene clusters, where tight clusters from the angle of each view are required, the latter approach is more favorable. While for applications where the objective is to find latent structures among subjects, such as a disease subtyping study with data from both phenotypic and genotypic views (Sun *et al.*, 2014), the first approach may be used.

The optimal solution of Problem (2) leads to the identification of a gene cluster and its associated variables in each view. When multiple clusters are needed, we can obtain the subsequent gene clusters by repeatedly solving Problem (2) with $\mathbf{X}_i$ replaced by a residual matrix $\bar{\mathbf{X}}_i$. There are two ways to create $\bar{\mathbf{X}}_i$ from $\mathbf{X}_i$ and the sparse rank-one approximation $\widehat{\mathbf{u}}_i\widehat{\mathbf{v}}_i^T$ of $\mathbf{X}_i$. One way is to calculate the difference between $\mathbf{X}_i$ and $\widehat{\mathbf{u}}_i\widehat{\mathbf{v}}_i^T$, i.e. $\bar{\mathbf{X}} = \mathbf{X}_i - \widehat{\mathbf{u}}_i\widehat{\mathbf{v}}_i^T$. The other way is to exclude the rows corresponding to all the subjects in the identified cluster from $\mathbf{X}_i$. The first approach may lead to a cluster solution that assigns a subject to more than one cluster whereas the clusters resulted from the second way are always mutually exclusive. The second approach was used in our experiment.

4 Optimization

In this section, we propose a computational algorithm to solve Problem (2) by following the block coordinate decent (BCD) framework (Tseng, 2001). We start with a brief introduction of soft-thresholding rule for solving the minimization problem bellow, as it is used frequently in our algorithm.

$$\min_x x^2 - 2\alpha x + 2\beta|x|, \quad (3)$$

where α and $\beta > 0$ are two constants. Let $f(x) = x^2 - 2\alpha x + 2\beta|x|$, we have:

$$f(x) = \begin{cases} (x - (\alpha - \beta))^2 - (\alpha - \beta)^2 & x > 0 \\ 0 & x = 0 \\ (x - (\alpha + \beta))^2 - (\alpha + \beta)^2 & x < 0. \end{cases}$$

When $\alpha > \beta$, $(\alpha - \beta)$ minimizes $f(x)$ when $x > 0$ with minimum $-(\alpha - \beta)^2$ and 0 minimizes $f(x)$ when $x \leq 0$ with 0 being the minimum. Obviously, $-(\alpha - \beta)^2 < 0$, so $(\alpha - \beta)$ is the overall minimizer when $\alpha > \beta$. Similarly, when $\alpha < -\beta$, $(\alpha + \beta)$ minimizes $f(x)$ with $-(\alpha + \beta)^2$ being the minimum; and when $|\alpha| < \beta$, 0 minimizes $f(x)$ with minimum 0. Collectively, Problem (3) has an analytical solution that can be summarized as follows:

$$\widehat{x} = \begin{cases} \alpha - \beta & \alpha > \beta \\ 0 & |\alpha| \leq \beta \\ \alpha + \beta & \alpha < -\beta. \end{cases} \quad (4)$$

This is the so called soft-thresholding rule for solving Problem (3).

In our algorithm, we iteratively search for the optimal $\mathbf{z}$, $\mathbf{u}_i$'s and $\mathbf{v}_i$'s. In each iteration, we alternatively search for optimal $\mathbf{z}$, $\mathbf{u}_i$'s and $\mathbf{v}_i$'s in sequence by solving one with fixing the other two. When $\mathbf{z}$ is fixed, both the two subproblems of finding optimal $\mathbf{u}_i$ with fixed $\mathbf{v}_i$ and finding optimal $\mathbf{v}_i$ with fixed $\mathbf{u}_i$ are independent among views, thus can be solved separately for each view and in parallel.

(a) Solving for $\mathbf{u}_i$ *when* $\mathbf{z}$ *and* $\mathbf{v}_i$ *are fixed*

When $\mathbf{z}$ and $\mathbf{v}_i$ are fixed, and $\mathbf{u}_i$ remains as the only variable, Problem (2) is reduced to:

$$\min_{\mathbf{u}_i} ||\mathbf{X}_i - (\tilde{\mathbf{z}} \odot \mathbf{u}_i)\tilde{\mathbf{v}}_i^T||_F^2 + \lambda_{u_i}||\mathbf{u}_i||_1, \quad (5)$$

where $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{v}}_i$ are constant. By expanding both the Frobenius norm and ℓ_1-norm, this sub-problem can be transformed to:

$$\min_{\mathbf{u}_i} \sum_{j,k} (\mathbf{X}_i(j,k) - \tilde{\mathbf{z}}(j)\tilde{\mathbf{v}}_i(k)\mathbf{u}_i(j))^2 + \sum_j \lambda_{u_i}|\mathbf{u}_i(j)|.$$

Since there is no interacting terms among components of $\mathbf{u}_i$, each component $\mathbf{u}_i(j)$ can be solved independently. After excluding all constant terms, the optimal $\mathbf{u}_i(j)$ can be found by optimizing:

$$\min_{\mathbf{u}_i(j)} \mathbf{u}_i(j)^2 - 2\frac{\mathbf{X}_i(j,\cdot)\tilde{\mathbf{v}}_i}{\tilde{\mathbf{z}}(j)||\tilde{\mathbf{v}}_i||_2^2}\mathbf{u}_i(j) + \frac{\lambda_{u_i}}{\tilde{\mathbf{z}}(j)^2||\tilde{\mathbf{v}}_i||_2^2}|\mathbf{u}_i(j)|.$$

Let

$$\alpha_{\mathbf{u}_i(j)} = \frac{\mathbf{X}_i(j,\cdot)\tilde{\mathbf{v}}_i}{\tilde{\mathbf{z}}(j)||\tilde{\mathbf{v}}_i||_2^2}, \quad \beta_{\mathbf{u}_i(j)} = \frac{\lambda_{u_i}}{2\tilde{\mathbf{z}}(j)^2||\tilde{\mathbf{v}}_i||_2^2},$$

and the soft-thresholding rule as in Eq. (4) can be applied by setting $\alpha = \alpha_{\mathbf{u}_i(j)}$ and $\beta = \beta_{\mathbf{u}_i(j)}$ to obtain optimal $\mathbf{u}_i(j)$ as follows:

$$\widehat{\mathbf{u}}_i(j) = \begin{cases} \alpha_{\mathbf{u}_i(j)} - \beta_{\mathbf{u}_i(j)} & \alpha_{\mathbf{u}_i(j)} > \beta_{\mathbf{u}_i(j)} \\ 0 & |\alpha_{\mathbf{u}_i(j)}| \leq \beta_{\mathbf{u}_i(j)} \\ \alpha_{\mathbf{u}_i(j)} + \beta_{\mathbf{u}_i(j)} & \alpha_{\mathbf{u}_i(j)} < -\beta_{\mathbf{u}_i(j)}. \end{cases} \quad (6)$$

(b) Solving for $\mathbf{v}_i$ *when* $\mathbf{z}$ *and* $\mathbf{u}_i$ *are fixed*

When $\mathbf{z}$ and $\mathbf{u}_i$ are fixed to $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{u}}_i$, respectively, the subproblem of Problem (2) with $\mathbf{v}_i$ being the only variable can be written as:

$$\min_{\mathbf{v}_i} ||\mathbf{X}_i - (\tilde{\mathbf{z}} \odot \tilde{\mathbf{u}}_i)\mathbf{v}_i^T||_F^2 + \lambda_{v_i}||\mathbf{v}_i||_1. \quad (7)$$

By expanding the Frobenius norm and ℓ_1-norm, this sub-problem can be transformed to:

$$\min_{\mathbf{v}_i} \sum_{j,k} (\mathbf{X}_i(j,k) - \tilde{\mathbf{z}}(j)\tilde{\mathbf{u}}_i(j)\mathbf{v}_i(k))^2 + \lambda_{v_i}\sum_k |\mathbf{v}_i(k)|.$$

Similar to sub-problem (5), here we also have no interacting terms among components of $\mathbf{v}_i$, so each of its components $\mathbf{v}_i(k)$ can also be solved independently. The sub-problem for solving $\mathbf{v}_i(k)$ is as follows:

$$\min_{\mathbf{v}_i(k)} \mathbf{v}_i(k)^2 - 2\frac{(\tilde{\mathbf{z}} \odot \tilde{\mathbf{u}}_i)^T \mathbf{X}_i(\cdot,k)}{||\tilde{\mathbf{z}} \odot \tilde{\mathbf{u}}_i||_2^2}\mathbf{v}_i(k) + \frac{\lambda_{v_i}}{||\tilde{\mathbf{z}} \odot \tilde{\mathbf{u}}_i||_2^2}|\mathbf{v}_i(k)|.$$

Let

$$\alpha_{\mathbf{v}_i(k)} = \frac{(\tilde{\mathbf{z}} \odot \tilde{\mathbf{u}}_i)^T \mathbf{X}_i(\cdot,k)}{||\tilde{\mathbf{z}} \odot \tilde{\mathbf{u}}_i||_2^2}, \quad \beta_{\mathbf{v}_i(k)} = \frac{\lambda_{v_i}}{2||\tilde{\mathbf{z}} \odot \tilde{\mathbf{u}}_i||_2^2},$$

this problem can also be solved by applying the soft-thresholding rule. The optimal $\mathbf{v}_i(k)$ is calculated as:

$$\widehat{\mathbf{v}}_i(k) = \begin{cases} \alpha_{\mathbf{v}_i(k)} - \beta_{\mathbf{v}_i(k)} & \alpha_{\mathbf{v}_i(k)} > \beta_{\mathbf{v}_i(k)} \\ 0 & |\alpha_{\mathbf{v}_i(j)}| \leq \beta_{\mathbf{v}_i(j)} \\ \alpha_{\mathbf{v}_i(k)} + \beta_{\mathbf{v}_i(k)} & \alpha_{\mathbf{v}_i(k)} < -\beta_{\mathbf{v}_i(k)}. \end{cases} \tag{8}$$

(c) Solving for $\mathbf{z}$ *while* $\mathbf{u}_i$ *and* $\mathbf{v}_i$ *are fixed*

When $\mathbf{u}_i$ and $\mathbf{v}_i$ are fixed to $\tilde{\mathbf{u}}_i$ and $\tilde{\mathbf{v}}_i$, Problem (2) is reduced to:

$$\min_{\mathbf{z}} \sum_i ||\mathbf{X}_i - (\mathbf{z} \odot \tilde{\mathbf{u}}_i)\tilde{\mathbf{v}}_i^T||_F^2 + \lambda_z||\mathbf{z}||_1. \tag{9}$$

As in both (a) and (b), it can be shown that each component of $\mathbf{z}$ can be solved independently. Let

$$\mathbf{M} = [\mathbf{X}_1, \cdots \mathbf{X}_m], \quad \mathbf{E} = [\tilde{\mathbf{u}}_1\tilde{\mathbf{v}}_1^T, \cdots, \tilde{\mathbf{u}}_m\tilde{\mathbf{v}}_m^T],$$

the problem for solving each $\mathbf{z}(j)$ can be written as:

$$\min_{\mathbf{z}(j)} \mathbf{z}(j)^2 - 2\frac{\mathbf{E}(j,:)\mathbf{M}(j,:)^T}{||\mathbf{E}(j,:)||_2^2}\mathbf{z}(j) + \frac{\lambda_z}{||\mathbf{E}(j,:)||_2^2}|\mathbf{z}(j)|.$$

Let

$$\alpha_{\mathbf{z}(j)} = \frac{\mathbf{E}(j,:)\mathbf{M}(j,:)^T}{||\mathbf{E}(j,:)||_2^2}, \quad \beta_{\mathbf{z}(j)} = \frac{\lambda_z}{2||\mathbf{E}(j,:)||_2^2},$$

and apply the soft-thresholding rule, the optimal $\tilde{\mathbf{z}}(j)$ is calculated as:

$$\widehat{\mathbf{z}}(j) = \begin{cases} \alpha_{\mathbf{z}(j)} - \beta_{\mathbf{z}(j)} & \alpha_{\mathbf{z}(j)} > \beta_{\mathbf{z}(j)} \\ 0 & |\alpha_{\mathbf{z}(j)}| \leq \beta_{\mathbf{z}(j)} \\ \alpha_{\mathbf{z}(j)} + \beta_{\mathbf{z}(j)} & \alpha_{\mathbf{z}(j)} < -\beta_{\mathbf{z}(j)}. \end{cases} \tag{10}$$

We summarize our algorithm in Algorithm (1).

Algorithm 1. Multi-view Sparse Vector Decomposition

Input: $\mathbf{X}_i$, λ_z, λ_{u_i} and λ_{v_i} for $i = 1, \cdots, m$
Output: $\mathbf{z}$, $\mathbf{u}_i$ and $\mathbf{v}_i$ for $i = 1, \cdots, m$
1. Initialize $\mathbf{z}$ with a vector of all ones.
2. Initialize each $\mathbf{v}_i$ using $\sqrt{\sigma_i}\bar{\mathbf{v}}_i$, where σ_i and $\bar{\mathbf{v}}_i$ are the first largest singular vector of $\mathbf{X}_i$.
3. For $i = 1, \cdots, m$,
   Update $\mathbf{u}_i$ according to Eq. (6).
   Update $\mathbf{v}_i$ according to Eq. (8).
4. Update $\mathbf{z}$ according to Eq. (10).
Repeat Steps 3 and 4 until convergence.

4.1 Convergence analysis

Given a function $f(\mathbf{x})$, its directional derivative at a point $\mathbf{z}$ in its domain along a direction $\mathbf{d}$ is calculated as:

$$f'(\mathbf{z};\mathbf{d}) = \lim_{s \to 0} \frac{f(\mathbf{z} + s\mathbf{d}) - f(\mathbf{z})}{s}.$$

We say f is Gâteaux differentiable at $\mathbf{z}$, if $f'(\mathbf{z};\mathbf{d})$ is well defined for all $\mathbf{d}$. In addition, when

$$f'(\mathbf{z};\mathbf{d}) \geq 0, \quad \forall d,$$

we say $\mathbf{z}$ is a stationary point of f.

For simplifying the presentation, we use $f(\mathbf{z}, \mathbf{u}_i, \mathbf{v}_i)$ to represent the objective function of Problem (2). Let

$$f_0(\mathbf{z}, \mathbf{u}_i, \mathbf{v}_i) = \sum_{i=1}^m ||\mathbf{X}_i - (\mathbf{z} \odot \mathbf{u}_i)\mathbf{v}_i^T||_F^2, \quad f_{\mathbf{z}}(\mathbf{z}) = \lambda_z||\mathbf{z}||_1,$$
$$f_{\mathbf{u}_i}(\mathbf{u}_i) = \sum_{i=1}^m \lambda_{u_i}||\mathbf{u}_i||_1, \quad f_{\mathbf{v}_i}(\mathbf{v}_i) = \sum_{i=1}^m \lambda_{v_i}||\mathbf{v}_i||_1,$$

then we have:

$$f(\mathbf{z}, \mathbf{u}_i, \mathbf{v}_i) = f_0(\mathbf{z}, \mathbf{u}_i, \mathbf{v}_i) + f_{\mathbf{z}}(\mathbf{z}) + f_{\mathbf{u}_i}(\mathbf{u}_i) + f_{\mathbf{v}_i}(\mathbf{v}_i). \tag{11}$$

Theorem 1: Let $\{(\mathbf{z}, \mathbf{u}_i, \mathbf{v}_i)^r\}$ be a sequence generated by Algorithm 1, every limit point of $\{(\mathbf{z}, \mathbf{u}_i, \mathbf{v}_i)^r\}$ is a stationary point of $f(\mathbf{z}, \mathbf{u}_i, \mathbf{v}_i)$.

Proof: First, the overall function $f(\mathbf{z}, \mathbf{u}_i, \mathbf{v}_i)$ is continuous on its entire domain R^p, here p is the total number of components of $\mathbf{z}$, $\mathbf{u}_i$ and $\mathbf{v}_i$ combined. Second, it can be easily shown that $f_0(\mathbf{z}, \mathbf{u}_i, \mathbf{v}_i)$ is Gâteaux differentiable with respect to all the variables: $\mathbf{z}$, $\mathbf{u}_i$ and $\mathbf{v}_i$. Third, all the three sub-problems, i.e. Problem (6), (8) and (10) have one unique optimal solution, which can be found analytically as in Eq. (6), (8) and (10). According to Theorem (4.1) in Tseng (2001), for an optimization problem as shown in Eq. (2) with its objective function generally formatted as in Eq. (11), when f is a continuous, f_0 is Gâteaux is differentiable and has open domain, and all the sub-problems, i.e. problems that are solved for variables in one block while fixing those in all others, have unique solution, every limit point generated by a block coordinate decent (BCD) algorithm, such as our Algorithm (1), is a stationary point of f. This leads to our conclusion. □

5 Results

We first evaluated the effectiveness of the proposed method using synthetic data, and subsequently applied it to two real world datasets of pre-implantation embryonic development in the human and mouse. To demonstrate its advantage, we compared the proposed multi-view bi-clustering method with several existing approaches using synthetic data where we know the ground truth. The compared methods include both base line approaches and advanced multi-view clustering methods that are recognized as the state of art in the machine learning field. These methods are briefly describes as follows:

- **Single view overlap:** This is the traditionally and commonly used two-step approach, i.e. clustering analysis in each view separately followed by the computation of overlaps among clusters from different views (Jiang *et al.*, 2014; Xue *et al.*, 2013). We ran this two-step approach with both the hierarchical clustering as implemented in tool WGCNA (Langfelder and Horvath, 2008) and the bi-clustering via sparse rank-one matrix factorization as the clustering method on each view.

- **Kernel addition/product**: Radial basis function (RBF) kernels of all views are combined via addition or component wise product. Spectral clustering was subsequently applied to the combined kernel to obtain clusters.
- **Feature concatenation**: Data from all views were simply arranged together by feature concatenation and the RBF kernel of this combined data was calculated and used in spectral clustering to obtain clusters.
- **Co-trained spectral**: Homogeneous kernels among views are sought via iterative search. In each iteration, the kernel of one view is updated with information from the remaining views. Spectral clustering was subsequently used with these homogeneous kernels to obtain clusters (Kumar and Daume, 2011).
- **Co-regularized spectral**: This method also performs joint spectral clustering (Kumar *et al.*, 2011). The eigendecomposition of the graph Laplacian of all views is linked to obtain homogeneous eigenvectors that are used subsequently in *k*-means to obtain clusters.

5.1 Simulation study

We simulated datasets with implanted block structures that give both clusters of subjects and variables by mimicking datasets from a real study in which genes are characterized with expression patterns. Two views of data for 1000 subjects were created. There were 12 variables in view 1, and 15 variables in view 2. The data matrix of each view is created by randomly setting 0 or 1 to each entry with varying probability that is determined according to prefixed block structures projected in the data. More specifically, we start from a data matrix filled with all 0. Then we reset data entries inside and outside the blocks to 1 with probability 0.9 and 0.1, respectively. For simplifying the process and easy presentation, we had subjects in the two datasets well aligned and indexed from 1 to 1000; and variables were also indexed using consecutive number starting from 1. View 1 was designed to have two blocks. The first block consists of subjects from 1 to 400 and variables from 1 to 3. The second includes the 200 subjects indexed from 481 to 680 and variables from 4 to 6. Three blocks were included in view 2. The first block contains subjects from 1 to 240 and the first three variables. The second block consists of subjects from 241 to 480 and variable 4, 5 and 6. The last block includes 320 subjects indexed from 481 to 800 and variables from 7 to 9. By comparing blocks of the two views, it is obvious that there are three consistent blocks (i.e. containing same subjects) between the two views. Variables of each view and number of subjects in these blocks are provided in Table 1. Block 1 consists of 240 subjects and contains variables from 1 to 3 in both view. There are 200 subjects in block 2. The corresponding variables are 4, 5 and 6 in view 1 and 7, 8 and 9 in view 2. Block 3 consists of 160 subjects and contains variables from 1 to 3 in view 1 and variables from 4 to 6 in view 2.

We randomly generated six datasets using the settings as described above. For each dataset, all compared methods were run to obtain four clusters. Three out of the four clusters correspond to the three consistent blocks, respectively, in the data; and the remaining one corresponds to the set including all other subjects. The normalized mutual information (NMI) by comparing the cluster solution resulted from each method with the true solution (blocks) is calculated to measure their performance. It ranges from 0 to 1. A higher value indicates stronger consistency between the two compared cluster solutions.

Table 1. Variables and number of subjects in the three true consistent blocks between the two views of the synthetic datasets

		Block 1	Block 2	Block 3
Variables	*view 1*	1–3	4–6	1–3
	view 2	1–3	7–9	4–6
Number of subjects		240	200	160

The variable set is represented by *i–j*, which includes variables indexed from *i* through *j* (with both *i* and *j* included).

The mean and standard deviation of NMIs obtained by all compared methods on the six synthetic datasets are presented in Figure 2. For single view overlap, only the results obtained when bi-clustering via sparse rank-one matrix factorization was used as the clustering method are reported, as they are better than that when hierarchical clustering was used. The proposed multi-view bi-clustering method is labeled with MVBC. It has the highest mean NMI 0.8576 with standard deviation 0.0135, which is significantly higher than that of all other compared methods, and thus has the best performance. In order to have a better idea on what the consistent blocks identified by each method look like, we draw data matrix plots in Figure 3 with subjects arranged according to their block assignments determined by each method on one of the six synthetic datasets. That is data points from the same identified block are plotted together. These data matrix plots also demonstrate the advantage of MVBC by showing that it uncovers the true blocks with minor and the least mismatching when comparing to the others. The superior performance of MVBC over the traditional two-step approach demonstrates the improved power of joint multi-view analysis in identifying consistent clusters. The observation that it outperforms all other compared multi-view clustering methods shows the advantage of performing subspace space searching in the situation where consistent clusters are determined by only subset of variables in the data.

5.2 Case study: the human and mouse embryonic development

We applied the proposed method to two datasets that were collected respectively for the human and mouse embryonic development. The two datasets were downloaded from www.ncbi.nlm.nih.gov/geo with accessing number GSE44183 and have been used in previous studies (Jiang *et al.*, 2014; Xue *et al.*, 2013). Both gene expression datasets were obtained from single cell RNA sequencing. There are

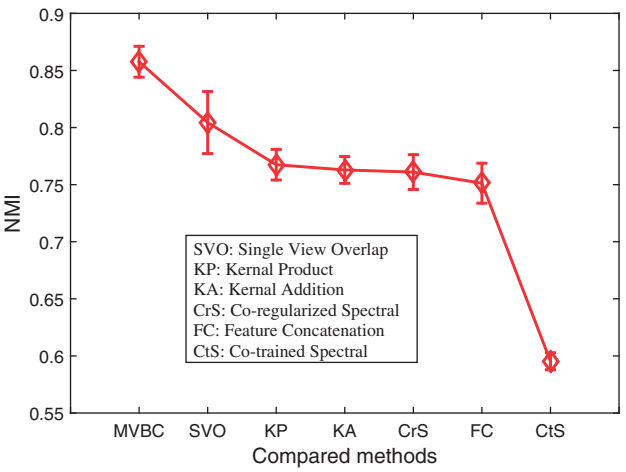


Fig. 2. Plot of mean and standard deviation of NMIs obtained by each compared method on the six synthetic datasets. The proposed method is labeled with MVBC

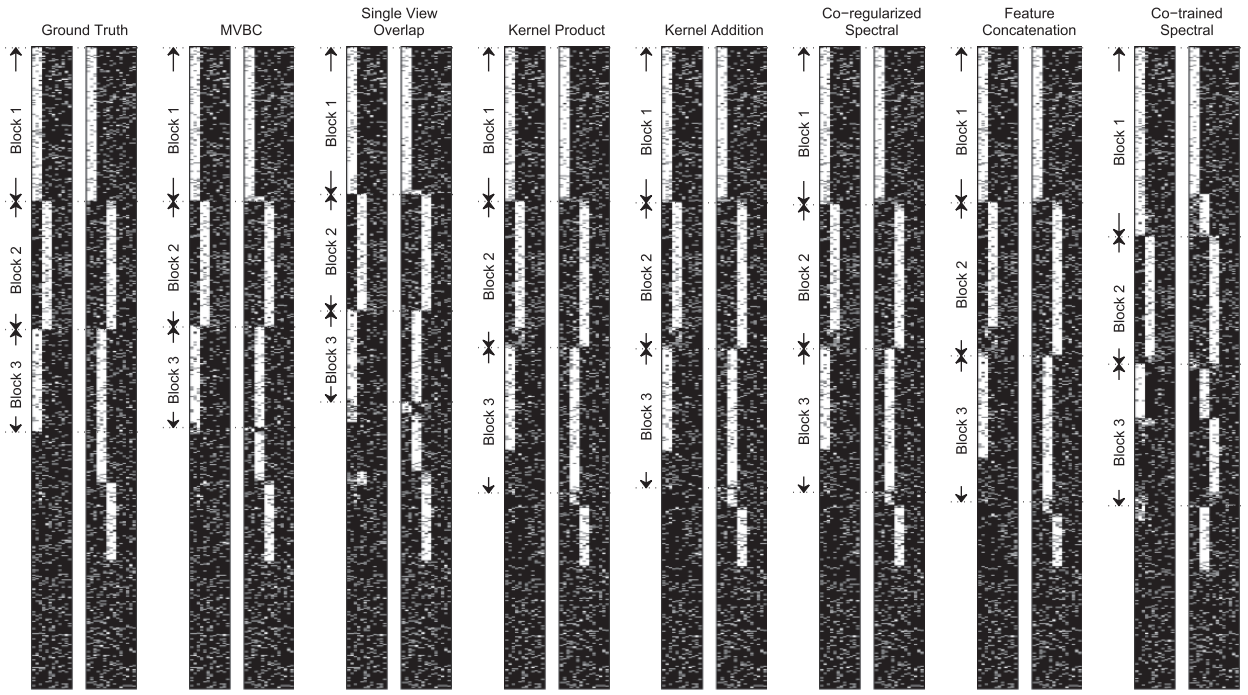


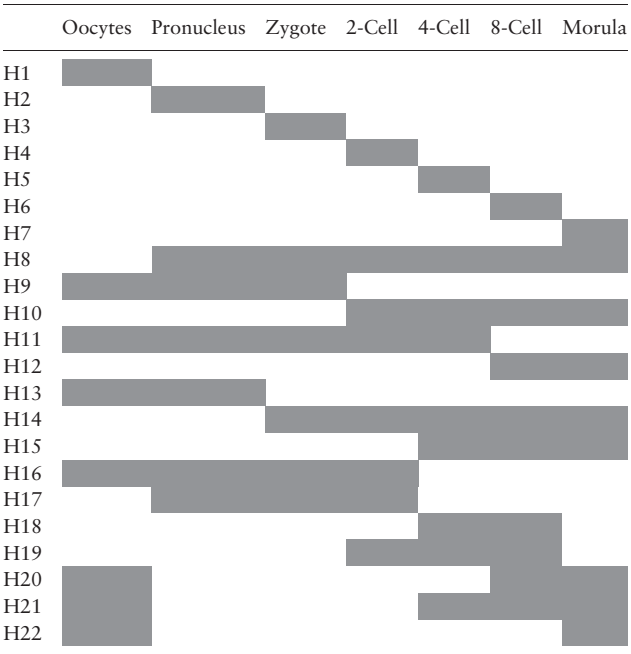
Fig. 3. Consistent blocks identified by all compaired methods on one of the six synthetic datasets. The proposed method is labeled with MVBC. Data matrixes are plotted with black spot indicating 0 and white spot indicating 1. Subjects in the plot are arranged according to the consistent blocks identified by each method. Two matrixes are plotted for each method, i.e. one per each view. The left most set of two matrix plots indicates the true consistent blocks in the data. See Table 1 for details of these three blocks

14 766 genes and seven embryonic development stages, oocytes, pronucleus, zygote, 2-cell, 4-cell, 8-cell and morula, in the human dataset. For the mouse, gene expression levels of 13 879 genes at six embryonic development stages, oocytes, pronucleus, 2-cell, 4-cell, 8-cell and morula, are available. Because we aimed to identify co-regulated gene clusters conserved during the human and mouse embryo development, the 11 018 common genes in both datasets were included in the analysis.

Gene expression patterns used in our analysis consisted of both those identified by existing works (Jiang *et al.*, 2014; Xue *et al.*, 2013)and those that might be present in the embryonic development indicated in literatures (Blakeley *et al.*, 2015; Cao *et al.*, 2014; Graf *et al.*, 2014; Hamatani *et al.*, 2004; Wang *et al.*, 2004; Yan *et al.*, 2013; Zeng *et al.*, 2004). For humans, we aggregated 22 gene expression patterns as listed in Table 2. We compiled a list of 18 patterns for the mouse, which are listed in Table 3.

We first reformatted the raw data of gene expression levels of genes, so gene regulations are directly characterized by the expression patterns included. We used 0.75 as the cutoff threshold: t (as described in Section 2) while performing the reformatting. Then we ran the proposed multi-view bi-clustering method with the two reformatted datasets to identify conserved co-regulated gene clusters between the two species. As we know, co-regulated genes suggest their involvement in a common network of biological processes and functions. Moreover, conservation of co-regulations among different species implies that the corresponding biological processes and functions are fundamental to all species studied. For further understanding of the conserved co-regulated gene clusters obtained by running our approach, we performed gene ontology (GO) analysis using DAVID (Huang da *et al.*, 2009) for all clusters. Lastly, we compared the expression patterns that are associated with the same clusters in both species to reveal the similarities and differences in developmental programing.

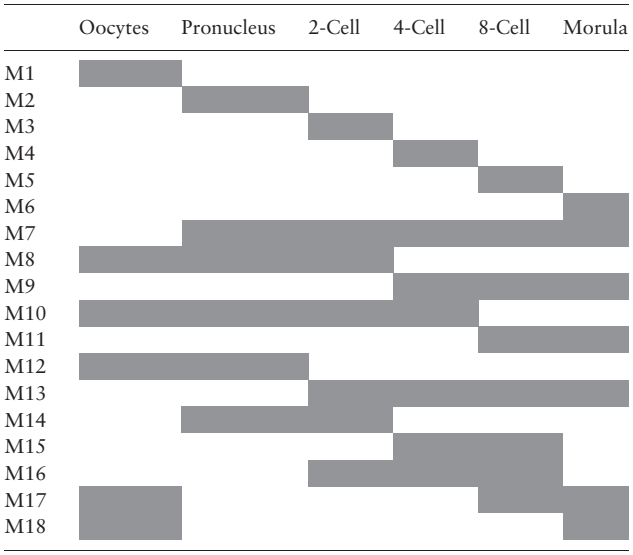
Table 2. The 22 gene expression patterns included in our analysis for characterizing gene regulation in the human pre-implantation embryonic development



Dark (white) color indicates high (low) expression level.

In total, 22 co-regulated gene clusters that were conserved between mice and humans were identified in our analysis. The results are summarized in Table 4 including: the size of each cluster, the patterns with the strongest association to each cluster, and the top GO terms that are significantly associated to genes in these clusters

Table 3. The 18 gene expression patterns included in our analysis for characterizing gene regulation in mouse pre-implantation embryonic development



Dark (white) color indicates high (low) expression level.

(with P value ≤ 0.05). In the table, expression patters are represented by a sequence of 0 and 1, with 0 denoting low level or no expression and 1 indicating high level expression.

Out of the 22 clusters, seven were relatively large, i.e. clusters C1–2, C4, C5, C7 and C9–10, with more than 100 genes in each. Analysis of the functions of genes in these clusters revealed that they are engaged in fundamental biological processes. More specifically, the 1042 co-regulated genes in C1 are involved in cell death and survival; the 1510 co-regulated genes in C2 are engaged in RNA post-transcriptional modification, protein synthesis, cellular growth and proliferation; and the 765 co-regulated genes in C4 are involved in cell cycle, gene expression and cellular assembly and organization. Moreover, genes engaged in carbohydrate and lipid metabolism, DNA replication, embryonic development and cellular function and maintenance are also co-regulated with many others in both species as indicated by clusters C5, C7 and C9–10. The remaining 15 clusters (i.e. C3, C6, C8 and C11–22) are small. The GO analysis of genes in these clusters shows significant over-representation of genes involved in transcription, translation, reproduction, sex differentiation, mitochondrial functions and stem cell maintenance, which implies that genes involved in these biological functions are co-regulated in humans and mice in a conserved fashion.

Intriguingly, several clusters contain the co-regulated genes that follow similar expression patterns between the human and mouse embryos. Of note, genes in clusters C2, C4 and C6 shows similar

Table 4. Conserved co-regulated gene clusters identified by our proposed method during the human and mouse pre-implantation embryonic development

Co-regulated gene cluster	No. of genes	Mouse (Ooc,Pr,2c,4c,8c,M)	Human (Ooc,Pr,Zy,2c,4c,8c,M)	Gene Ontology
C1	1042	M12 (1,1,0,0,0,0)	H12 (0,0,0,0,0,1,1)	Cell death and survival, cancer
C2	1510	M9 (0,0,0,1,1,1)	H12 (0,0,0,0,0,1,1)	RNA post-transcriptional modification, protein synthesis, cellular growth and proliferation genes
C4	765	M12 (1,1,0,0,0,0)	H11 (1,1,1,1,1,0,0)	Cell cycle, gene expression, cellular assembly and organization
C9	207	M9 (0,0,0,1,1,1)	H11 (1,1,1,1,1,0,0)	Cancer, cell cycle, carbohydrate metabolism, lipid metabolism, small molecule biochemistry
C7	179	M9 (0,0,0,1,1,1)	H1 (1,0,0,0,0,0,0)	DNA replication, recombination and repair, cell cycle
C10	158	M9 (0,0,0,1,1,1)	H5 (0,0,0,0,1,0,0)	Cellular function and maintenance, cell cycle, reproductive system development and function
C5	143	M9 (0,0,0,1,1,1)	H7 (0,0,0,0,0,0,1)	Embryonic development,
C6	54	M9 (0,0,0,1,1,1)	H7 (0,0,0,0,0,0,1)	Cellular growth and proliferation
C8	53	M12 (1,1,0,0,0,0)	H7 (0,0,0,0,0,0,1)	Amino acid Metabolism, small molecule biochemistry, carbohydrate metabolism, small molecule biochemistry
C3	51	M12 (1,1,0,0,0,0)	H6 (0,0,0,0,0,1,0)	Hereditary disorder, neurological disease, cell-to-cell signaling and interaction, cell morphology
C13	38	M3 (0,0,1,0,0,0)	H12 (0,0,0,0,0,1,1)	RNA processing
C14	34	M3 (0,0,1,0,0,0)	H2 (0,1,0,0,0,0,0)	Organic alcohol transport
C11	33	M16 (0,0,1,1,1,0)	H5 (0,0,0,0,1,0,0)	Sex differentiation, stem cell maintenance
C12	20	M6 (0,0,0,0,0,1)	H11 (1,1,1,1,1,0,0)	Regulation of muscle cell differentiation, cell motion
C20	18	M4 (0,0,0,1,0,0)	H22 (1,0,0,0,0,0,1)	Mitochondrial
C16	17	M3 (0,0,1,0,0,0)	H9 (1,1,1,0,0,0,0)	Gene silencing by RNA, DNA metabolic process
C15	12	M2 (0,1,0,0,0,0)	H12 (0,0,0,0,0,1,1)	Cellular amino acid derivative metabolic process
C21	12	M5 (0,0,0,0,1,0)	H17 (0,1,1,1,0,0,0)	mRNA metabolic process
C17	11	M1 (1,0,0,0,0,0)	H18 (0,0,0,0,1,1,0)	Transcription
C18	11	M4 (0,0,0,1,0,0)	H2 (0,1,0,0,0,0,0)	Translation, protein transport
C19	10	M5 (0,0,0,0,1,0)	H2 (0,1,0,0,0,0,0)	Reproduction
C22	8	M8 (1,1,1,0,0,0)	H17 (0,1,1,1,0,0,0)	Mitosis II

The size of each cluster, the patterns for both the human and mouse that have the strongest association with genes in each cluster as indicated by the component with the largest value in vector $\mathbf{v}_i$ of Problem (2), and the top GO terms that are significantly associated to genes in these clusters (with P value ≤ 0.05) are provided. Expression patterns are represented by a sequence of 0 and 1, with 0 denoting low level or no expression and 1 indicating high level expression.

Note: C5 and C6 are two distinct clusters, as besides the pattern with strongest support from genes in the cluster (data shown), there are other associated patterns that are distinct between these two clusters (data not shown).

expression pattern and are highly expressed at morula in both species. These genes are involved in RNA post-transcriptional modification, embryonic development and cellular growth and proliferation. Similarly, genes in clusters C4 and C22 also exhibit similar expression pattern between humans and mice, with high expression levels at the zygote and 2-cell stages. GO analysis of these genes indicates significant over-representation of cell cycle and mitosis II. Together, these results suggest that humans and mice share many core transcriptional programming in their pre-implantation embryonic development. In the contrary, there are also clusters of co-regulated genes that show completely reverse expression patterns between the two species. For example, co-regulated genes in clusters C1, C3, C8, C15 and C17 were expressed highly by mouse oocytes and/or prenuclear embryos, but highly enriched in the human 8-cell and morula stages. Also genes in clusters C7, C9 and C12 were high from 4-cell to morula in the mouse but low in corresponding stages in the human. Interestingly, most genes with the reserved expression patterns are involved in cell death and survival, cancer, metabolism and recombination and repair. These results suggest that the mouse and human early embryos employ very different pathways to prepare themselves for the upcoming processes of implantation. In addition, comparing patterns associated to clusters C13–14, C16 and C18–21 between the two species shows variations in the timing of activation of genes included, suggesting the potential mechanism of embryonic developmental speed varies between humans and mice.

Collectively, our results here show that genes involved in many fundamental biological networks during pre-implantation embryonic development are regulated in a conserved fashion between humans and mice. There are both similarities and differences in the activation timing of the co-regulated genes between the two species. For example, genes engaged in networks such as mitosis II and proliferation show the same activation timing; while genes involved in biological processes such as cell death and survival show completely reversed activation timing; and genes with roles in networks such as mRNA metabolic process show delayed or advanced activation. These cluster results bring unique insights to the little-known developmental programming of mammalian pre-implantation embryos.

6 Discussion

We have developed a new approach that can be used to identify co-regulated gene clusters that are conserved among multiple species using samples collected at a series of different time points such as during pre-implantation embryonic development. The proposed approach consists of two components: pattern preserving noise reduction and multi-view bi-clustering via sparse rank-one matrix factorization. We have developed an efficient algorithm that is guaranteed to converge for solving the optimization problem in the proposed multi-view bi-clustering. Compared to the commonly used two-step approach (Blakeley *et al.*, 2015; Jiang *et al.*, 2014; Xue *et al.*, 2013), our approach is less vulnerable to noise in the gene expression data and has the advantage of identifying conserved co-regulated gene clusters among species. In this study, we did not attempt to normalize data between species because in real world situations, direct comparisons in gene expression levels among species may not be necessary. However, such normalization is intriguing and new strategies should be developed when a need is presented.

We have succeeded in identifying conserved co-regulated gene clusters between the human and mouse in their pre-implantation embryos by applying the proposed approach. The clusters not only represent functional gene networks that conserved in embryogenesis between the two species, but reveal similarities and differences in progression of developmental programming of embryos across species. The identification of these orchestrated functional changes is among the first step to unveil the little-known embryonic programming, and provide directions of future research in embryogenesis.

Even though the development of the proposed method is motivated by studying the pre-implantation embryonic development of multiple mammalian species, it can certainly be applied to many other similar situations. The approach that we have proposed for cleaning the data can be employed to denoise other similar datasets when gene expression patterns are the focus of the study. The proposed multi-view bi-clustering method is a general clustering approach and can be used in any multi-view setting, especially in situations where consistent gene clusters across views only exist in the subspaces of the variables in the views.

Because expression patterns used here in noise reduction are essentially variables that groups genes in the subsequent cluster analysis. The success of the method can be limited by the expression patterns that are used. We suggest using all patterns that potentially make biologic sense. The method is flexible in that the patterns can be modified when new biological questions arise. When combined with the traditional two-step clustering approach, our method is a great tool to obtain more information from the same dataset.

Funding

This work was supported by National Science Foundation (NSF) grants DBI-1356655 and IIS-1320586, USDA-ARS (1265-31000-091-02S) and USDA regional collaboration project W3171. Jinbo Bi was also supported by NSF grants CCF-1514357 and IIS-1447711.

Conflict of Interest: none declared.

References

Blakeley,P. *et al.* (2015) Defining the three cell lineages of the human blastocyst by single-cell rna-seq. *Development*, **142**, 3151–3165.

Braude,P. *et al.* (1988) Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature*, **332**, 459–461.

Cai,X. *et al.* (2013) Multi-view k-means clustering on big data. In: *International joint conference on Artificial Intelligence*, pp. 2598–2604.

Cao,S. *et al.* (2014) Specific gene-regulation networks during the pre-implantation development of the pig embryo as revealed by deep sequencing. *BMC Genomics*, **15**, 4.

Chaudhuri,K. *et al.* (2009) Multi-view clustering via canonical correlation analysis. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, ACM. pp. 129–136, New York, NY, USA.

Cheng,W. *et al.* (2013) Flexible and robust co-regularized multi-domain graph clustering. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, vol. 1, pp. 320–328.

Culp,M. and Michailidis, G. (2009) A co-training algorithm for multi-view data with applications in data fusion. *J. Chemometr.*, **23**, 294–303.

Graf,A. *et al.* (2014) Fine mapping of genome activation in bovine embryos by RNA sequencing. *Proc. Natl. Acad. Sci. USA*, **111**, 4139–4144.

Hamatani,T. *et al.* (2004) Dynamics of global gene expression changes during mouse preimplantation development. *Dev. Cell*, **6**, 117–131.

Han,J. *et al.* (2011) *Data Mining: Concepts and Techniques*, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Huang,daW. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Jiang,Z. *et al.* (2014) Transcriptional profiles of bovine in vivo pre-implantation development. *BMC Genomics*, **15**, 756.

Kumar,A. and Daume,III,H. (2011) A co-training approach for multi-view spectral clustering. In: Getoor,L. and Scheffer,T. (eds.) *Proceedings of the*

28th International Conference on Machine Learning, ACM, New York, NY, USA. pp. 393–400.

Kumar,A. *et al*. (2011) Co-regularized multi-view spectral clustering. In: Shawe-Taylor,J. *et al*. (eds.) *Advances in Neural Information Processing Systems 24*, pp. 1413–1421.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Lee,M. *et al*. (2010) Biclustering via sparse singular value decomposition. *Biometrics*, **66**, 1087–1095.

Liu,J. *et al*. (2013) Multi-view clustering via joint nonnegative matrix factorization. In: *Proceedings of the SIAM International Conference on Data Mining*. vol. 13, pp. 252–260.

Misirlioglu,M. *et al*. (2006) Dynamics of global transcriptome in bovine matured oocytes and preimplantation embryos. *Proc. Natl. Acad. Sci. USA*, **103**, 18905–18910.

Sun,J. *et al*. (2014) Multi-view singular value decomposition for disease subtyping and genetic associations. *BMC Genet*., **15**, 73.

Sun,J. *et al*. (2015) Multi-view sparse co-clustering via proximal alternating linearized minimization. In: *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37 of *JMLR Proceedings*, pp. 757–766. JMLR.org.

Tseng,P. (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl*., **109**, 475–494.

Wang,Q.T. *et al*. (2004) A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Dev. Cell*, **6**, 133–144.

Xue,Z. *et al*. (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, **500**, 593–597.

Yan,L. *et al*. (2013) Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol*., **20**, 1131–1139.

Zeng,F. *et al*. (2004) Transcript profiling during preimplantation mouse development. *Dev. Biol*., **272**, 483–496.