Li et al. BMC Genomics 2013, 14(Suppl 4):S2
http://www.biomedcentral.com/1471-2164/14/S4/S2

BMC Genomics

RESEARCH Open Access

MultiFacTV: module detection from higher-order time series biological data

Xutao Li[1,3], Yunming Ye[1,3], Michael Ng[2], Qingyao Wu[1,3*]

From IEEE International Conference on Bioinformatics and Biomedicine 2012
Philadelphia, PA, USA. 4-7 October 2012

Abstract

Background: Identifying modules from time series biological data helps us understand biological functionalities of a group of proteins/genes interacting together and how responses of these proteins/genes dynamically change with respect to time. With rapid acquisition of time series biological data from different laboratories or databases, new challenges are posed for the identification task and powerful methods which are able to detect modules with integrative analysis are urgently called for. To accomplish such integrative analysis, we assemble multiple time series biological data into a higher-order form, e.g., a gene × condition × time tensor. It is interesting and useful to develop methods to identify modules from this tensor.

Results: In this paper, we present MultiFacTV, a new method to find modules from higher-order time series biological data. This method employs a tensor factorization objective function where a time-related total variation regularization term is incorporated. According to factorization results, MultiFacTV extracts modules that are composed of some genes, conditions and time-points. We have performed MultiFacTV on synthetic datasets and the results have shown that MultiFacTV outperforms existing methods EDISA and Metafac. Moreover, we have applied MultiFacTV to Arabidopsis thaliana root(shoot) tissue dataset represented as a gene×condition×time tensor of size 2395 × 9 × 6(3454 × 8 × 6), to Yeast dataset and Homo sapiens dataset represented as tensors of sizes 4425 × 6 × 6 and 2920×14×9 respectively. The results have shown that MultiFacTV indeed identifies some interesting modules in these datasets, which have been validated and explained by Gene Ontology analysis with DAVID or other analysis.

Conclusion: Experimental results on both synthetic datasets and real datasets show that the proposed MultiFacTV is effective in identifying modules for higher-order time series biological data. It provides, compared to traditional non-integrative analysis methods, a more comprehensive and better view on biological process since modules composed of more than two types of biological variables could be identified and analyzed.

Background

Identification of biological modules plays a key role in bioinformatics because it can reveal interesting groups of proteins/genes having strong interactions, which may be related to some biological functionalities. In the literature, many methods have been proposed for this purpose. One popular way is to make use of clustering algorithms [1-4], which reveals module patterns by clustering proteins/genes into groups such that intragroup similarities are maximized while inter-group similarities are minimized. The performance of this type of methods relies significantly on the similarity function used during the clustering process. Due to this shortcoming, some researchers also tune to matrix factorization techniques for detecting biological modules. For example, in [5-7], singular value decomposition based methods have been studied and developed to detect modules from gene expression data. In [8-10], nonnegative matrix factorization related methods have been developed to cluster and explore biological data.

* Correspondence: wuqingyao.china@gmail.com
[1]Department of Computer Science, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, China
Full list of author information is available at the end of the article

Table 1 Experimental results on synthetic datasets.

Low noise level(0.005)								
	3-module NMI	dataset Fscore	5-module NMI	dataset Fscore	8-module NMI	dataset Fscore	10-module NMI	dataset Fscore
EDISA	0.5923	0.8273	0.4997	0.7554	0.6025	0.8100	0.4512	0.6709
MetaFac	0.9117	0.9787	0.8473	0.9472	0.6181	0.7717	0.5485	0.7160
MultiFacTV	**0.9874**	**0.9982**	**0.9936**	**0.9986**	**0.8273**	**0.9140**	**0.8035**	**0.8844**
Middle noise level(0.01)								
	3-module NMI	dataset Fscore	5-module NMI	dataset Fscore	8-module NMI	dataset Fscore	10-module NMI	dataset Fscore
EDISA	0.4200	0.7136	0.2907	0.6381	0.6670	0.8654	0.3142	0.6027
MetaFac	0.9312	0.9830	0.4710	0.6916	0.5444	0.7051	0.4146	0.6037
MultiFacTV	**0.9920**	**0.9987**	**0.9898**	**0.9978**	**0.8928**	**0.9552**	**0.7801**	**0.8678**
High noise level(0.02)								
	3-module NMI	dataset Fscore	5-module NMI	dataset Fscore	8-module NMI	dataset Fscore	10-module NMI	dataset Fscore
EDISA	0.4493	0.7189	0.2514	0.5923	0.2055	0.4355	0.1496	0.4411
MetaFac	0.9260	0.9793	0.5318	0.6804	0.2727	0.5222	0.2479	0.4233
MultiFacTV	**0.9914**	**0.9985**	**0.9656**	**0.9898**	**0.7757**	**0.8723**	**0.6565**	**0.7747**

Recently, CUR decomposition, a new method approximating original data matrix by selecting a set of columns and rows, has been applied to analyze microarray data and SNP data [11,12] because of its scalability and interpretability. This method may possibly be used to cluster large-scale biological data as well. However, all these methods are developed for analyzing biological data represented as matrix form, which models interactions between only two types of variables.

With rapid acquisition of biological experiments from different laboratories or studies based on different databases, many higher-order biological data representing interactions between more than two types of variables can be obtained. For instance, researchers in different laboratories may be interested in analysing gene co-expression networks under different stimulus, each of which is represented as a gene×gene matrix. Integrating these matrices results in a higher-order biological data, namely a gene×gene×stimulus tensor, and finding module patterns from such data tends to offer a better view of the underlying biological structures. Therefore powerful methods which are able to detect modules with integrative analysis are urgently called for.

In the literature, several integrative analysis methods have already been put forward. In [13], Li et al. developed a framework to find *recurrent heavy subgraphs* from multiple weighted networks represented as a 3D tensor, i.e., gene × gene × network. In the framework, a tensor objective function is proposed and solved, the solution of which helps to discovery a heavy subgraph. In [14], Omberg et al. employed higher-order singular value decomposition(HOSVD) to perform integrative analysis of multiple mircoarray data from different studies. Zhang et al. extended nonnegative matrix factorization method for exploring protein modules from multiple data sources [15]. In [16], a JointCluster algorithm was proposed to extract coherent clusters from multiple networks. However, all these methods are not suitable for analyzing time series data, which is also a task of particular importance in bioinformatics.

In this paper, we are interested in identifying biological modules from multiple time series data with integrative analysis. There are two ways to build up such data in general. One is to collect and accumulate from different time series data sources [17,18], and the other is to perform time series biological experiments under different stimulus/conditions [19,20]. The second way is usually more popular. For instance, in [20], researchers studied the time series expression profiles of genes in Arabidopsis thaliana under several abiotic stimulus; in [19], researchers studied time series gene expression of several sclerosis patients after IFN-β injection. Joining such data together, we can form a higher-order time series tensor, e.g., a gene×condition×time tensor. Identifying modules of genes, conditions and time-points from such tensor data could offer us a better understanding of the corresponding biological processes. For example, Supper et al. proposed EDISA algorithm by extending the 2D *iterative signature algorithm* to extract and analyze such modules [21].

We propose in this paper, MultiFacTV, a method to find modules from tensor time series data. This method employs a tensor factorization objective function and makes use of the decomposition results to identify

modules. As we consider time series data, the modules are expected to be as consecutive as possible in time dimension. Therefore we incorporate a time-related regularization term of total variation into the objective function. Different from the conference version [22], we have re-derived the factorization formulas and updated the algorithm because we do not assume that input biological tensor is nonnegative in this paper. We have compared MultiFacTV with EDISA [21] and MetaFac [23] on synthetic datasets, and the results have shown that MulitFacTV outperforms the other two algorithms. In addition, we have applied MultiFacTV to Arabidopsis thaliana root(shoot) tissue dataset, Yeast dataset and Homo Sapiens dataset, and the results have shown that MultiFacTV indeed identifies some interesting biological modules, most of which have not yet been reported in our conference version. These interesting findings have also been validated and explained by using Gene Ontology analysis with DAVID or other analysis.

Methods

Terminologies

A tensor refers to a multidimensional array or matrix. The order of a tensor is defined to be the number of dimensions, also known as modes, of the corresponding multidimensional array. For instance, given a $n_1 \times n_2 \times n_3$ tensor $\mathcal{A} = (a_{r,s,t})$, it is called a third-order tensor. The process of rearranging a tensor into a two-dimensional matrix is called unfolding. A n-th order tensor can be unfolded into n matrices in terms of each of its modes. For example, unfolding the tensor $\mathcal{A}$ in terms of mode 1, mode 2 and mode 3, we obtain three matrices $\mathbf{A}^{(1)}$, $\mathbf{A}^{(2)}$ and $\mathbf{A}^{(3)}$ of sizes $n_1 \times n_2 n_3$, $n_2 \times n_1 n_3$ and $n_3 \times n_1 n_2$ respectively. In this paper, we let $\mathbf{A}^{(p)}$ denote the unfolding matrix of tensor $\mathcal{A}$ in terms of mode p.

Let $\mathbf{I}_{n\times n}$ be the $n \times n$ identity matrix. Let $\mathbf{M}^T$ be the transpose of matrix $\mathbf{M}$. Given a $n_1 \times n_2$ matrix $\mathbf{M}$, we define $vec(\mathbf{M})$ to be a $n_1 n_2 \times 1$ vector that is obtained by stacking each column of $\mathbf{M}$. We define shrinkage$_{\alpha/\rho}$ $(\cdot)$ to be a shrinkage-thresholding operator for each entry of a matrix, i.e.,

$$\text{shrinkage}_{\alpha/\rho}(\mathbf{M})_{i,j} = m_{i,j} - \min(\alpha/\rho, |m_{i,j}|).\frac{m_{i,j}}{|m_{i,j}|},$$

where $\frac{m_{i,j}}{|m_{i,j}|}$ should be zero when $m_{i,j} = 0$.

Let $\otimes$ and $\circ$ be the Kronecker product operator and outer product operator. Given a n-dimensional vector $\mathbf{x} = [x_1, x_2, ..., x_n]^T$, let $\mathbf{x}^+ = \{x_i | x_i > 0, 1 \le i \le n\}$ and $\mathbf{x}^- = \{x_i | x_i < 0, 1 \le i \le n\}$ denote the sets of its positive entries and negative entries respectively. Besides, we define $\sum \mathbf{x}^+ = \sum_{\gamma \in \mathbf{x}^+} \gamma$ and $\sum \mathbf{x}^- = \sum_{\gamma \in \mathbf{x}^-} \gamma$. In this paper, max$(\cdot)$ and min$(\cdot)$ are functions used to find the maximum value and minimum value respectively.

MultiFacTV

Our idea to extract modules from higher-order time series biological data is using tensor factorization techniques. A higher-order time series biological data can be represented as a tensor. For example, a gene-condition-time interaction data is represented as a tensor in Figure 1(a). Factorizing this tensor with two decompositions for gene, condition and time-point respectively, we find two modules, i.e., the first module $m_1 = \{g_1, g_2, g_5, g_7, c_1, c_2, c_4, c_5, c_6, t_1, t_2\}$ and the second module $m_2 = \{g_3, g_4, g_6, c_3, t_2, t_3, t_4\}$, by using a threshold (say 4) to cut off the decompositions shown as in Figure 1(b). However, we may not be able to obtain good modules merely based on traditional tensor factorization techniques because

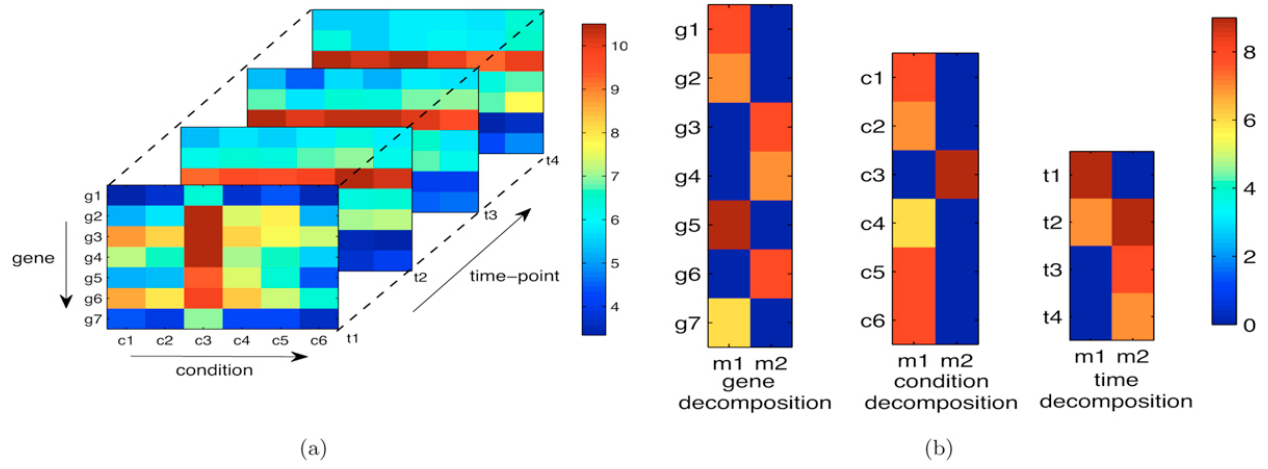

Figure 1 Illustration of module detection by using tensor factorization. (a) higher-order biological (tensor) data; (b) decomposition vectors obtained by factorizing the corresponding tensor.

the data we are considering includes time dimension. We need to make sure the modules exist consistently in some consecutive time periods, e.g., the time-points involved in module m_1/m_2 are expected to be as consecutive as possible. To achieve this property, some suitable constraints must be incorporated into the factorization process. Next we will formulate the proposed MultiFacTV method.

We assume that the higher-order biological data represents interactions between three types of variables, for example gene × condition × time data. We formulate the proposed MutliFacTV based on such data in this paper. However, it is remarkable that MultiFacTV is a general framework that can be derived similarly for biological data representing interactions more than three types of variables. Suppose we consider the genomic expression profiles of n_1 genes under n_2 conditions over n_3 time-points. The corresponding interactions can be represented as a $n_1 \times n_2 \times n_3$ tensor $\mathcal{A} = (a_{r,s,t})$, where $a_{r,s,t}$ is a value recording how the gene r responds to the condition s at the time-point t. We note that $a_{r,s,t}$ can be a positive or negative value, i.e., the input tensor $\mathcal{A}$ is not necessarily a nonnegative tensor.

Assume we would like to find K modules. The following objective function is proposed to decompose the tensor $\mathcal{A}$ into three matrices **U**, **V** and **W**:

$$\min \left\| \mathcal{A} - \sum_{k=1}^{K} \mathbf{u}_k \circ \mathbf{v}_k \circ \mathbf{w}_k \right\|^2 + \alpha \sum_{k=1}^{K} ||\mathbf{B}\mathbf{w}_k||_1 \quad \text{s.t. } \mathbf{W} \geq 0, \quad \text{and } \mathbf{1}^T \cdot \mathbf{w}_k = 1 \text{ for } k = 1, 2, ..., K \tag{1}$$

where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_K]$, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K]$, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K]$ are three decomposition matrices regarding n_1 genes, n_2 conditions and n_3 time-points respectively; **B** is a $(n_3 - 1) \times n_3$ matrix satisfying

$$b_{i,j} = \begin{cases} 1 & i = j \\ -1 & i = j - 1 \\ 0 & \text{otherwise} \end{cases}$$

and $\alpha > 0$ is a regularization parameter. Clearly, $\alpha \sum_{k=1}^{K} ||\mathbf{B}\mathbf{w}_k||_1$ is a total variation constraint regarding the decomposition matrix of time. With this regularization term, we can control the modules identified such that they exist consistently in some consecutive time periods. Different from the conference version [22], the decomposition matrices **U** and **V** do not have nonnegative constraints because we allow negative entries in the tensor $\mathcal{A}$.

MultiFacTV seeks matrices **U**, **V** and **W** that minimize the objective function in (1). As there are three matrices unknown, we need to solve them in an iterative fashion, i.e., changing the optimization problem into three subproblems with one unknown matrix in each, and then solving them iteratively until it converges. Therefore we have three subproblems for MultiFacTV as follows.

Subproblem 1: Fix **V** and **W**, and solve **U** by minimizing the objective function in (1).

In this subproblem, the objective function is transferred into:

$$\min \left\| \mathbf{A}^{(1)} - \mathbf{U}\mathbf{F} \right\|^2 \tag{2}$$

where $\mathbf{F} = (\mathbf{W} \odot \mathbf{V})^T$. We have the following solution for **U**:

$$\mathbf{U} = \mathbf{A}^{(1)}\mathbf{F}^T\left(\mathbf{F}\mathbf{F}^T\right)^{-1} \tag{3}$$

Subproblem 2: Fix **U** and **W**, and solve **V** by minimizing the objective function in (1).

In this subproblem, the objective function is transferred into:

$$\min \left\| \mathbf{A}^{(2)} - \mathbf{V}\mathbf{F} \right\|^2 \tag{4}$$

where $\mathbf{F} = (\mathbf{W} \odot \mathbf{U})^{\mathbf{T}}$. We have the following solution for **V**:

$$\mathbf{V} = \mathbf{A}^{(2)}\mathbf{F}^T\left(\mathbf{F}\mathbf{F}^T\right)^{-1}. \tag{5}$$

Subproblem 3: Fix **U** and **V**, and solve **W** by minimizing the objective function in (1).

In this subproblem, the objective function is transferred into:

$$\min \left\| \mathbf{A}^{(3)} - \mathbf{W}\mathbf{F} \right\|^2 + \alpha \sum_{k=1}^{k} ||\mathbf{B}\mathbf{w}_k||_1 \quad \text{s.t. } 1^T \cdot w_k = 1 \text{ for } k = 1, 2, ..., K, \tag{6}$$

where $\mathbf{F} = (\mathbf{V} \odot \mathbf{U})^T$. In order to solve the matrix **W** in (6), we introduce two $(n_3 - 1) \times K$ auxiliary matrices **P** and **Q** and adopt the strategy of Alternating Direction Method of Multipliers (ADMM) [24,25]. As a result, three updating formulas are derived and obtained (see [22] for the detailed derivation):

$$vec(\mathbf{W}) = \left((\mathbf{F}\mathbf{F}^T \otimes 2\mathbf{I}_{n_3 \times n_3}) + (\rho \mathbf{I}_{K \times K} \otimes \mathbf{B}^T\mathbf{B})\right)^{-1} vec\left(\rho \mathbf{B}^T(\mathbf{P} - \mathbf{Q}/\rho) + 2\mathbf{A}^{(3)}\mathbf{F}^T\right) \tag{7}$$

$$\mathbf{P} = \text{shinkage}_{\alpha/\rho}(\mathbf{B}\mathbf{W} + \mathbf{Q}/\rho) \tag{8}$$

$$\mathbf{Q} = \mathbf{Q} + \rho\left(\mathbf{B}\mathbf{W} - \mathbf{P}\right) \tag{9}$$

Here ρ can be any positive number and we use $\rho = 1$ in our implementation. Clearly, we need to update matrices **W**, **P** and **Q** iteratively until it converges to solve this subproblem. Note that each column of **W** must be normalized after updating as equation (7) to guarantee its constraints in (1).

Iteratively solving these three subproblems leads to a local minimum of the MultiFacTV objective function in

(1) and the solutions for matrices **U**, **V** and **W** at the same time. Different from our conference version, the updating formulas for **U** and **V** in Subproblems 1 and 2 change here because we do not have nonnegative constraints on them. Next we summarize the proposed MultiFacTV method in Algorithm 1.

Algorithm 1 The MultiFacTV Algorithm

Input: a $n_1 \times n_2 \times n_3$ tensor $\mathcal{A}$, the number of modules K, parameter α, and thresholding parameters τ_1, τ_2, and τ_3

Output: K modules stored in $\Omega = \{\Omega_1, \Omega_2, ..., \Omega_K\}$

Procedure:

1: Randomly initialize matrices $\mathbf{U}_{(0)}$, $\mathbf{V}_{(0)}$ and $\mathbf{W}_{(0)}$, and set $t = 1$;

2: Compute $\mathbf{U}_{(t)} = \mathbf{A}^{(1)}\mathbf{F}^T (\mathbf{F}\mathbf{F}^T)^{-1}$ where $\mathbf{F} = (\mathbf{V}_{(t-1)} \odot \mathbf{U}_{(t-1)})^T$;

3: Compute $\mathbf{V}(t) = \mathbf{A}^{(2)}\mathbf{F}^T (\mathbf{F}\mathbf{F}^T)^{-1}$ where $\mathbf{F} = (\mathbf{W}_{(t-1)} \odot \mathbf{U}_{(t)})^T$;

4: Randomly initialize matrices $\mathbf{P}_{(0)}$ and $\mathbf{Q}_{(0)}$, and set $\mathbf{F} = (\mathbf{V}_{(t)} \odot \mathbf{U}_{(t)})^T$, s=1, ρ=1;

5: Iteratively update $\mathbf{W}_{(s)}$, $\mathbf{P}_{(s)}$, $\mathbf{Q}_{(s)}$ as follows:

$$vec(\mathbf{W}_{(s)}) = \left((\mathbf{F}\mathbf{F}^T \otimes 2\mathbf{I}_{n3\times n3}) + (\rho \mathbf{I}_{K\times K} \otimes \mathbf{B}^T\mathbf{B}) \right)^{-1} vec\left(\rho \mathbf{B}^T(\mathbf{P}_{(s-1)} - \mathbf{Q}_{(s-1)}/\rho) + 2\mathbf{A}^{(3)}\mathbf{F}^T \right),$$

$$\mathbf{P}_{(s)} = \text{shinkage}_{\alpha/\rho}(\mathbf{B}\mathbf{W}_{(s)} + \mathbf{Q}_{(s-1)}/\rho),$$

$$\mathbf{Q}_{(s)} = \mathbf{Q}_{(s-1)} + \rho(\mathbf{B}\mathbf{W}_{(s)} - \mathbf{P}_{(s)}),$$

until it converges;

6: Set $\mathbf{W}_{(t)} = \mathbf{W}_{(s)}$;

7: If $||\mathbf{U}_{(t)} - \mathbf{U}_{(t-1)}||^2 + ||\mathbf{V}_{(t)} - \mathbf{V}_{(t-1)}||^2 + ||\mathbf{W}_{(t)} - \mathbf{W}_{(t-1)}||^2 > 0.001$, set $t = t + 1$ and goto Step 2;

otherwise, goto Step 8;

8: For k = 1 to K

Set $\Omega_k = \emptyset$

If $\sum \mathbf{u}_k^+ < -\sum \mathbf{u}_k^-$, set $\mathbf{u}_k = -\mathbf{u}_k$;

For r = 1 to n_1

If $u_{r,k} >= 0.5 * \tau_1 * \left((\max(\mathbf{u}_k^+) + \min(\mathbf{u}_k^+)\right)$, set $\Omega_k = \Omega_k \cup \{\text{gene } r\}$;

If $\sum \mathbf{v}_k^+ < -\sum \mathbf{v}_k^-$, set $\mathbf{v}_k = -\mathbf{v}_k$;

For s = 1 to n_2

If $v_{s,k} >= 0.5 * \tau_2 * \left((\max(\mathbf{v}_k^+) + \min(\mathbf{v}_k^+)\right)$, set $\Omega_k = \Omega_k \cup \{\text{condition } s\}$;

For t = 1 to n_3

If $w_{t,k} >= 0.5 * \tau_3 * ((\max(\mathbf{w}_k) + \min(\mathbf{w}_k))$, set $\Omega_k = \Omega_k \cup \{\text{time point } t\}$;

9: Return $\Omega = \{\Omega_1, \Omega_2, ..., \Omega_K\}$.

In this algorithm, we need to input a tensor and five parameters. At the beginning, the algorithm randomly initializes matrices **U**, **V** and **W** in step 1, and then it updates them iteratively from steps 2 to 7. We note that there is an inner loop in step 5 in order to update **W**. When finishing the computation of **U**, **V** and **W**, the algorithm outputs K modules in step 8 by cutting off each column of **U**, **V** and **W** with thresholding parameters τ_1, τ_2, and τ_3 respectively. Since the decomposition matrices **U** and **V** are not necessarily nonnegative, the module extraction in step 8 is also different from the conference version.

Results

In this section, we run MultiFacTV on synthetic datasets, Arabidopsis thaliana dataset, Yeast dataset and Homo sapiens dataset to test its performance and usefulness. The synthetic datasets are generated artificially and the other three real datasets can be found on http://www.ra.cs.uni-tuebingen.de/software/EDISA/downloads/index.htm.

Results on synthetic datasets

In this experiment, we generated gene×condition×time tensor data to test the effectiveness of MultiFacTV. In the synthetic datasets, some "ground-truth" modules containing a set of genes, conditions and consecutive time intervals were generated. There were 400 genes, 400 conditions and 50 time-points. Based on the number of modules included, the datasets were categorized into four types, 3-module dataset, 5-module dataset, 8-module dataset and 10-module dataset. To test the robustness of MultiFacTV, we added different level of noise in the corresponding tensors, i.e., using 0.005, 0.01 and 0.02 as densities to add noise into the tensors respectively. Our objective was to identify the "ground-truth" modules accurately.

As for a comparison, we performed EDISA and MetaFac as well. For MetaFac and MultFacTV, we set K to be the number of modules in the dataset. For EDISA, the sample size was set to be 20 and the iteration number was set to be 50. The parameters τ_g and τ_c were turned in the interval [0,1] with 0.1 as increasing step and then the best parameter values were to produce final result. For MultiFacTV, we used $\tau_1 = \tau_2 = 1.0$ and $\tau_3 = 0.75$. All results were evaluated based on the Fscore and NMI (Normalized Mutual Information) by considering the discovered modules and the "ground truth" modules.

Before comparing the performance of MultiFacTV and the other two algorithms, we first demonstrate the convergence of the proposed MultiFacTV and how its performance changes against the tuning of parameter α. In Figure 2, we show the convergence of MultiFacTV based on one synthetic dataset. We see from this figure that the objective function value is decreasing as the number of iterations increases, and after 40 iterations the change is very little and the algorithm is stopped. In Figure 3, we show how the performance of MutliFacTV changes with respect to the tuning of α. We see from this figure that its performance does not change significantly as parameter α changes from 1 to 20, and the best result is yielded when $\alpha = 10$. Therefore we used this value for α in the experiments. Table 1 shows the results of EDISA, MetaFac and the proposed algorithm on these synthetic datasets. We see from the table that MultiFacTV algorithm outperforms the other two comparison algorithms.

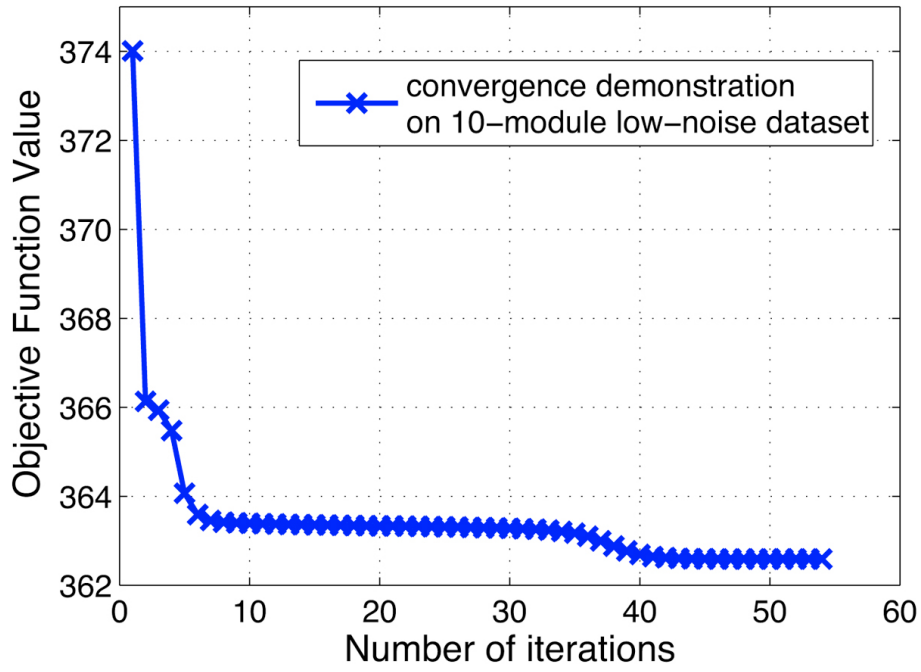

Figure 2 The demonstration of convergence for a particular run of MultiFacTV on synthetic dataset.

Results on Arabidopsis thaliana datasets

In this experiment, the MultiFacTV was applied to Arabidopsis thaliana data to explore biological module patterns therein. The data recorded the time-series genomic expression of the root/shoot tissue in Arabidopsis thaliana when different abiotic stresses were considered. For the genomic expression data of root tissue, we constructed a gene×condition×time tensor $\mathcal{A}$ of size 2395 × 9 × 6. For the genomic expression data of shoot tissue, we constructed a gene×condition×time

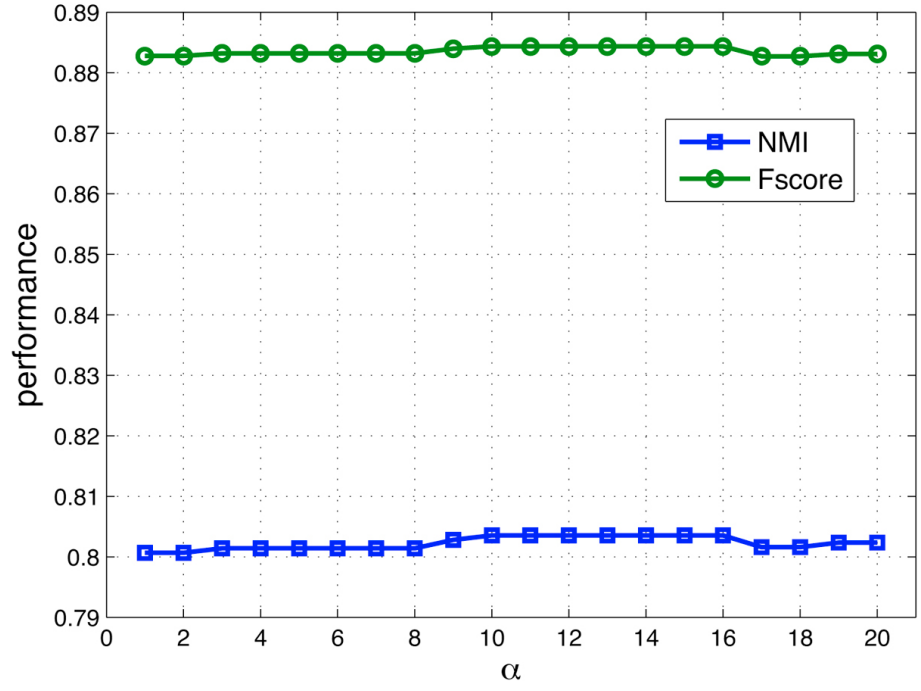

Figure 3 The change of performance for MultiFacTV with respect to the tuning of parameter α on 10-module low noise dataset.

tensor $\mathcal{A}$ of size 3454 × 8 × 6. Both tensors were nonnegative. We run the MultiFacTV method with $K = 40$, $\alpha = 10$, $\tau_1 = \tau_2 = 1.0$ and $\tau_3 = 0.75$ on each tensor.

Next we present some biological modules discovered from each of these tensors(data). To validate these modules, we associate them to some functional annotation terms with DAVID analysis [26]. Besides, the corresponding p-values are also given to demonstrate the statistical significance of these functional terms.

Interesting genomic modules in root tissue: Some interesting biological modules detected from root tissue by MultiFacTV are given in the following.

1. Cold-osmotic modules. In [27], it has been manifested that a large portion of the Arabidopsis genes are sensitive to cold and osmotic stress stimulus. In our results, we found several modules participating in the response to both stresses. We present two of such modules here and their genomic expression profiles are shown in Figure 4. We observe from Figures 4(a) and 4(b) there are distinct expression shapes, where the shapes for cold and osmotic conditions are quite similar. This observation indicates the genes in these two modules co-regulate under these two conditions, suggesting that Arabidopsis may not distinguish between cold and osmotic stresses. The first module is associated to functional terms like "response to water deprivation", "cold acclimation" and "response to cold" (p-values: 3.9×10^{-4}, 7.1×10^{-4} and 1.2×10^{-3} respectively) by using DAVID, and the second module is associated to "response to osmotic stress", "response to temperature stimulus" and "response to cold" (p-values: 9.1×10^{-4}, 8.9×10^{-6} and 1.4×10^{-2} respectively). These facts confirm that both modules play key roles in the response to cold and osmotic stresses.
2. Salt module. In Figure 5(a), we show a module detected by MultiFacTV that responds to salt stress. Apparently, this module has quite different expression shapes under salt stress compared to under the other stresses. Moreover, the terms like "response to water deprivation" and "response to salt stress" (p-values: 2.6×10^{-9} and 5.0×10^{-3} respectively) are mapped to it, which manifests this module indeed functions under salt stress.
3. Heat module. We obtained a module participating in the response to heat shock, shown as in Figure 5 (b). Clearly, it has quite distinct expression shapes under heat condition. With DAVID, the genes in this module are mapped to "response to heat" and "response to temperature stimulus" (p-values: 1.1×10^{-55} and 1.3×10^{-43} respectively).
4. Uvb-wound modules. We obtained two modules responding to uvb light and wound stresses, see Figure 6. In Figure 6(a), we observe that the module 1 down-regulates slightly from 0.5h to 12h and up-regulates from 12h to 24h. This module is significant for "photosynthesis, light harvesting", "response to light stimulus" and "defense response" (p-values: 1.4×10^{-8}, 2.8×10^{-2} and 1.2×10^{-2} respectively). It can be observed from Figure 6(b) that the module 2 has different genomic expression profiles for uvb and wound stresses in comparison with the other stresses. The module is pronounced under "response to light stimulus", "response to UV" and "response to wounding" (p-values: 1.4×10^{-4}, 1.4×10^{-4} and 8.7×10^{-3}). Clearly, both modules indeed participate in the response to uvb light and wound stresses.

Interesting genomic modules in shoot tissue: In the following, we show two interesting genomic modules in shoot tissue output by the proposed MultiFacTV.

1. Salt-oxidative-drought module. We found a module participating in the response to salt, oxidative and drought stresses, see Figure 7(a). We observe that the module has similar genomic expression profiles for salt, oxidative and drought stresses. It is annotated to functional terms like "response to salt stress", "oxidoreductase", "oxidation reduction" (p-values: 7.8×10^{-2}, 3.1×10^{-2} and 4.1×10^{-2} respectively). This suggests that the module is significant and indeed has biological functionalities related to salt and oxidative stresses.
2. Wound module. We obtained a module participating in the response to wound stress, see Figure 7 (b). It can be observed that the module first up-regulates and then down-regulates from 1h to 12h under wound stress, and its genomic expression shapes are quite different in comparison with the ones for the other stresses. By using DAVID, the module is annotated to functional terms like "defense response" and "response to wounding" (p-values: 2.4×10^{-4} and 2.5×10^{-2} respectively). This suggests the module identified by MultiFacTV indeed has wound-related biological functionalities.

Results on yeast dataset

We performed the proposed MultiFacTV on Yeast dataset to explore interesting module patterns. This dataset recorded multiple time series genomic expression of yeast *Saccharomyces cerevisiae* regarding to different environmental changes [28]. We considered six environmental stresses in this dataset, including heat shock, 0.32mM H_2O_2, 1mM menadione, 2.5mM DTT(dithiothreitol), 1.5mM diamide and 1M sorbitol. Since different time-points were adopted to record the expression

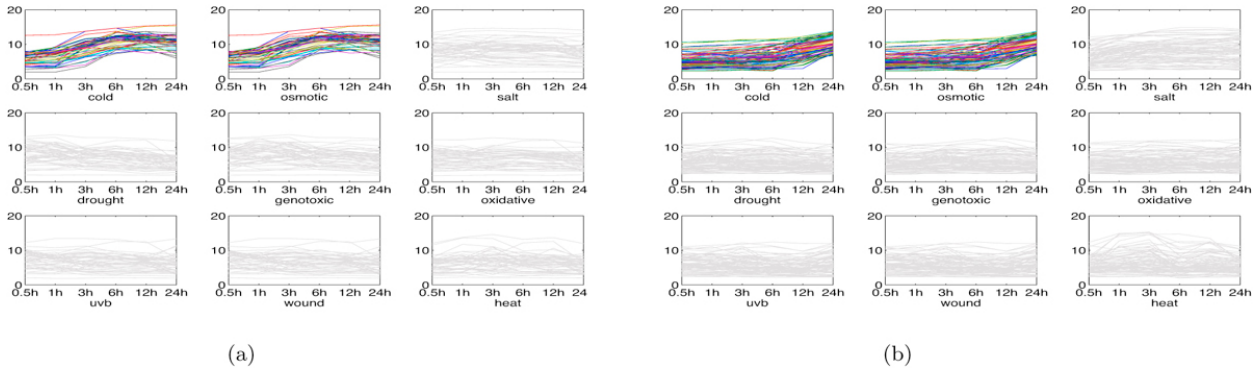

Figure 4 Two cold-osmotic modules in root tissue (Both figures come from [22]**)**. (a) genomic expression of module 1; (b) genomic expression of module 2.

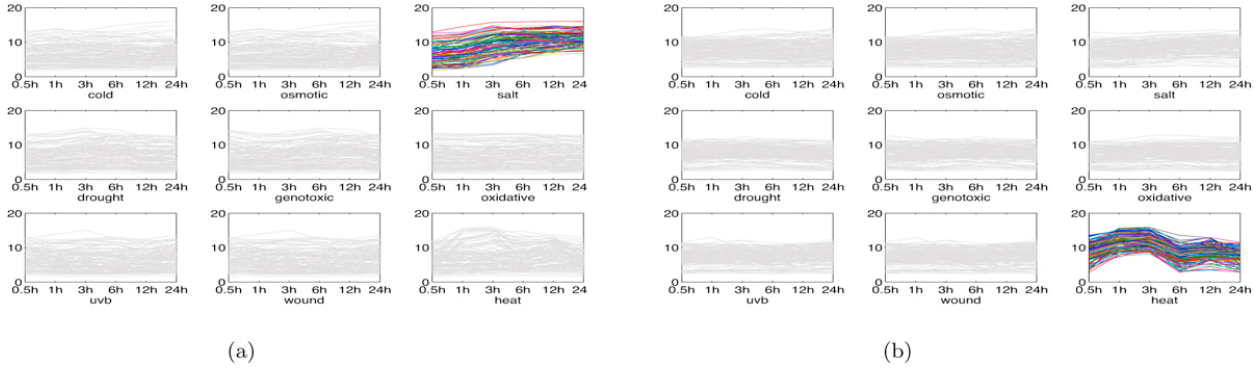

Figure 5 Salt module and heat module in root tissue (Figure (b) comes from [22]**)**. (a) genomic expression of salt module; (b) genomic expression of heat module.

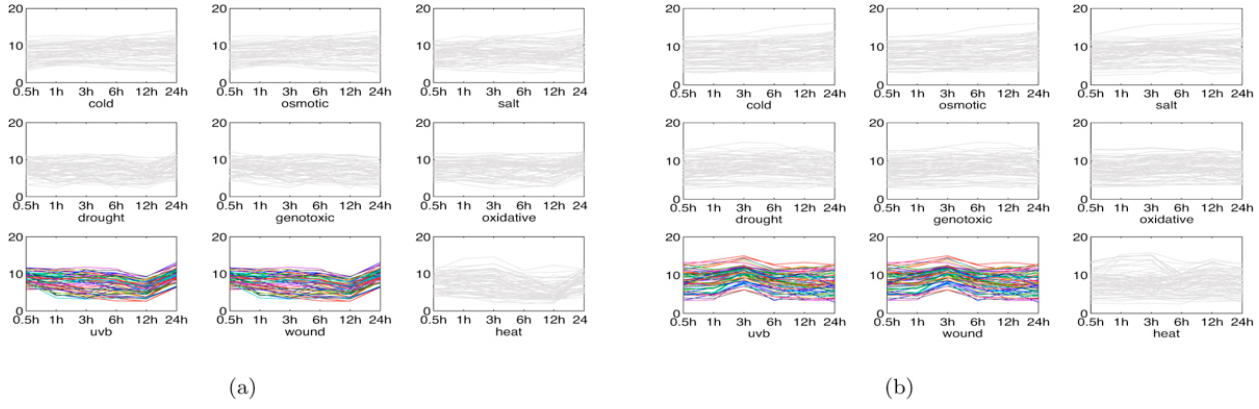

Figure 6 Two Uvb-wound modules in root tissue. (a) genomic expression of module 1; (b) genomic expression of module 2.

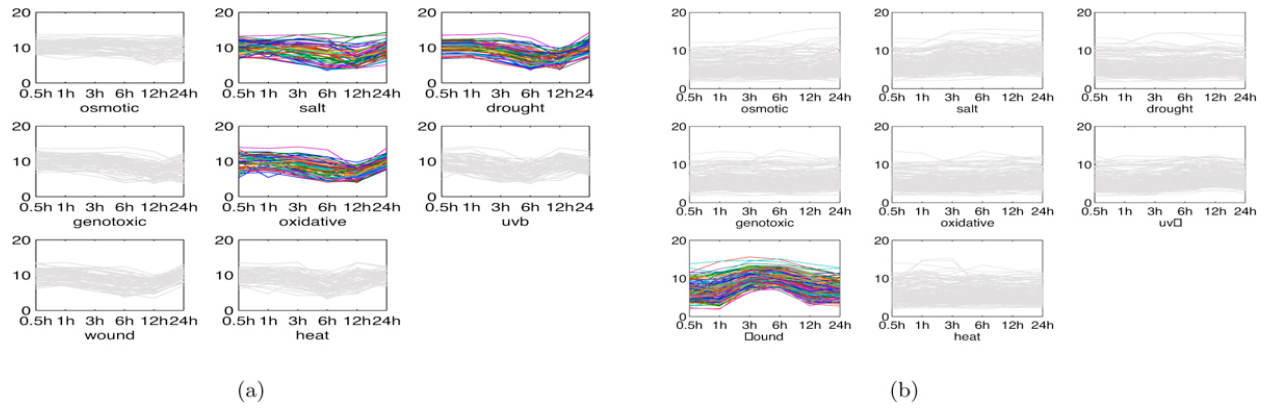

Figure 7 Salt-oxidative-drought module and wound module in shoot tissue. (a) genomic expression of salt-oxidative-drought module; (b) genomic expression of wound module.

under different environmental stresses in the original data, we preprocessed this data by selecting 6 time-points, i.e., 10min, 20min, 30min, 40min, 60min and 80min. The missing time-point was handled by using a linear interpolation of two closest time-points available. Other missing values were replaced with the average expression value at the corresponding time-point. As a result, we constructed a gene×condition×time tensor $\mathcal{A}$ of size 4425 × 6 × 6, i.e., there were 4425 genes, 6 stresses and 6 time-points. This tensor was not nonnegative because the genomic expression data included negative values. The MultiFacTV algorithm was performed with $K = 20$, $\alpha = 10$, $\tau_1 = \tau_3 = 0.75$ and $\tau_2 = 0.85$.

Next we present and analyze some interesting module patterns identified by the proposed MultiFacTV.

1. H_2O_2-menadione modules. In [28], it has been shown that a large portion of genes in yeast co-regulate under H_2O_2 stress and menadione stress despite that they are supposed to result in different reactive oxygen species. The MultiFacTV obtains similar findings and we present the genomic expression of two modules of such kind, see Figures 8(a) and 8(b). We observe that the module 1 up-regulates from 30min to 40min and down-regulates from 40min to 60min under both stresses, while the module 2 down-regulates from 10min to 20min and up-regulates from 20min to 30min. The analysis with DAVID have shown that the module 1 is functionally related to "reproduction of a single-celled organism", "mating projection tip" and "cell budding" (p-values: 1.9×10^{-2}, 7.8×10^{-2} and 6.1×10^{-2} respectively), and the module 2 is functionally associated to "glucose catabolic process", "hexose catabolic process" and "monosaccharide catabolic process" (p-values: 4.2×10^{-2}, 5.2×10^{-2} and 5.8×10^{-2} respectively). All these terms may be related to some biological process induced by the oxidative and reductive reactions taking place in the cells.
2. Heat shock modules. We obtained two interesting modules responding to heat shock in yeast. The genomic expression of both modules are shown as in Figures 9(a) and 9(b). We see that these two modules have opposite expression trends after heat stress where the module 1 down-regulates while the

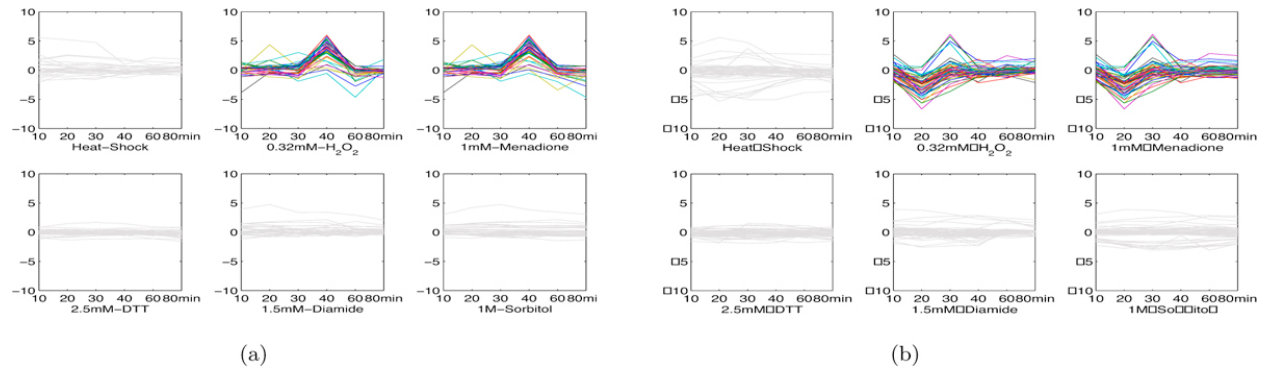

Figure 8 Two H_2O_2-menadione modules in yeast. (a) genomic expression of module 1; (b) genomic expression of module 2.

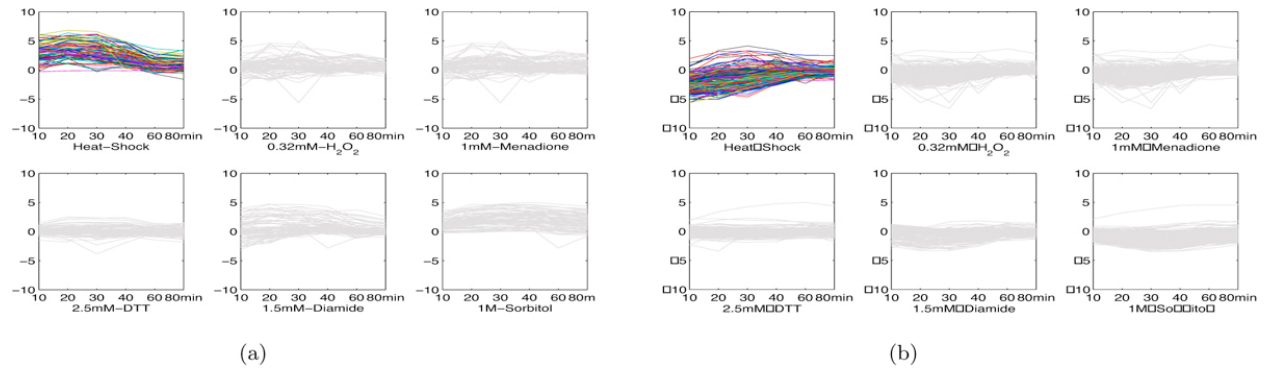

Figure 9 Two heat shock modules in yeast. (a) genomic expression of module 1; (b) genomic expression of module 2.

module 2 up-regulates. The analysis with DAVID have shown that the module 1 indeed takes part in the response to heat and temperature stimulus (p-values: 2.5×10^{-15} and 3.5×10^{-23} respectively). Moreover, we find this module is annotated to functional terms like "protein catabolic process" and "cellular macromolecule catabolic process" (p-values: 5.2×10^{-7} and 2.5×10^{-5} respectively). This can be interpreted by the fact that heat shock usually leads to protein unfolding [28]. The module 2 is annotated to functional terms like "ribonucleoprotein complex biogenesis" and "RNA binding" (p-values: 8.6×10^{-24} and 7.0×10^{-7}). This may be because the protein unfolding induces the concurrent ribonucleoprotein complex biogenesis.

Results on Homo Sapiens dataset

We applied the proposed MultiFacTV to Homo Sapiens dataset for exploring biological modules. It was a higher-order time series dataset about genomic expression of multiple sclerosis patients after IFN-β injection treatment. We represented this data as a nonnegative gene×patient×time tensor of size 2920 ×14×9, i.e, there were 2920 genes, 14 patients {A, B, C, D, E, F, G, H, I, J, K, L, M, N} and 9 time-points. The MultiFacTV was performed with $\tau_1 = \tau_2 = 0.5$, $\tau_3 = 0.75$, $\alpha = 10$ and $K = 40$. As a result, we found many interesting modules responding to IFN-β treatment similar to [21]. To exploit the usefulness of those modules from a different view, we made use of them to help us group the patients.

With the 40 modules identified, we constructed a binary matrix $\mathbf{M}$ of size 14×40 representing the membership of each patient to the modules, where $m_{i,j} = 1$ if the i-th patient was associated to the j-th module, otherwise $m_{i,j} = 0$. In such case, each of the 14 patients was represented as a 1×40 binary vector. Subsequently, we clustered the 14 patients by using k-means algorithm and the clustering results were {A, B, C, D}, {E, F, G, H}, {J, K, L, M }, {I, N }. This grouping result may suggest some differences of patients in their disease histories or progressions. We believe that this result will be beneficial to the designation of personalized medicine for the patients with multiple sclerosis [29].

Conclusions

As more and more time series biological data are being accumulated from different laboratories or databases, identification of modules with integrative analysis become an important and urgent task. One way to accomplish such integrative analysis is assembling multiple time series biological data into a tensor form. In this paper, we have proposed the MultiFacTV method, which extends the tensor factorization objective by introducing a time-related regularization term of total variation, to detect modules from such higher-order time series biological data. We have performed the MultiFacTV method on synthetic datasets, Arabidopsis dataset, Yeast dataset and Homo sapiens dataset to test its performance. The results have shown that the proposed MultiFacTV indeed reveals some interesting module patterns. We have shown and validated these interesting findings with DAVID analysis or other analysis.

In this paper, we assume that the multiple time series genomic expression data have the same size, i.e., the same number of genes and the same number of time-points, so that they can be joined into a tensor. In some cases, the data may be in different sizes. For example, the original Yeast dataset [28] has different number of time-points for different environmental stresses. In the future, it would be interesting to extend the tensor factorization objective function of Tucker1 or Tucker2 [30] in a similar way to perform integrative module detection for such data.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
XL, MN and YY participated in designing the algorithm, drafting and revising the manuscript. XL participated in implementing the algorithm and performing experiments. QW participated in the discussions of experimental results. All authors have read and approved the final version of this manuscript.

Acknowledgements
Based on "MultiFacTV: finding modules from higher-order gene expression profiles with time dimension", by Xutao Li, Yunming Ye, Qingyao Wu and Michael Ng which appeared in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*. ©2012 IEEE [22].
X. Li's research was supported in part by NSFC under Grant No.61100190. Y. Ye's research was supported in part by NSFC under Grant No. 61272538, National Key Technology R&D Program of MOST China under Grant No. 2012BAK17B08, Shenzhen Science and Technology Program under Grant No. CXY201107010206A, and Shenzhen Strategic Emerging Industries Program under Grant No. ZDSY20120613125016389. M. Ng's research was supported in part by Centre for Mathematical Imaging and Vision, HKRGC Grant No. 201812.

Declarations
The publication costs for this article were funded by Yunming Ye.
This article has been published as part of *BMC Genomics* Volume 14 Supplement S4, 2013: Selected articles from the IEEE International Conference on Bioinformatics and Biomedicine 2012: Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/14/S4.

Authors' details
[1]Department of Computer Science, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, China. [2]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China. [3]Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen, 518055, China.

Published: 1 October 2013

References

1. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM, *et al*: **Systematic determination of genetic network architecture.** *Nature genetics* 1999, **22**:281-285.
2. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proceedings of the National Academy of Sciences* 1998, **95(25)**:14863-14868.
3. Pentney W, Meila M: **Spectral clustering of biological sequence data.** *Proceedings of the national conference on artificial intelligence* 2005, **20**:845-850.
4. Cheng Y, Church GM: **Biclustering of expression data.** *Proceedings of the eighth international conference on intelligent systems for molecular biology* 2000, **1**:93-103.
5. Wall M, Rechtsteiner A, Rocha L: **Singular value decomposition and principal component analysis.** *A practical approach to microarray data analysis* 2003, 91-109.
6. Lee M, Shen H, Huang JZ, Marron JS: **Biclustering via sparse singular value decomposition.** *Biometrics* 2010, **66(4)**:1087-1095.
7. Sill M, Kaiser S, Benner A, Kopp-Schneider A: **Robust biclustering by sparse singular value decomposition incorporating stability selection.** *Bioinformatics* 2011, **27(15)**:2089-2097.
8. Jung I, Kim D: **Linknmf: Identification of histone modification modules in the human genome using nonnegative matrix factorization.** *Gene* 2012, **518(1)**:215-221.
9. Carmona-Saez P, Pascual-Marqui R, Tirado F, Carazo J, Pascual-Montano A: **Biclustering of gene expression data by non-smooth non-negative matrix factorization.** *BMC bioinformatics* 2006, **7(1)**:78.
10. Mejía-Roa E, Carmona-Sáez P, Nogales R, Vicente C, Vazquez M, Yang XY, Garcia C, Tirado F, Pascual-Montano A: **Bionmf: a web-based tool for nonnegative matrix factorization in biology.** *Nucleic acids research* 2008, **36(suppl 2)**:523-528.
11. Mahoney M, Drineas P: **Cur matrix decompositions for improved data analysis.** *Proceedings of the National Academy of Sciences* 2009, **106(3)**:697-702.
12. Paschou P, Mahoney M, Javed A, Kidd JR, Pakstis AJ, Gu S, Kidd KK, Drineas P: **Intra-and interpopulation genotype reconstruction from tagging snps.** *Genome Research* 2007, **17(1)**:96-107.
13. Li W, Liu CC, Zhang T, Li H, Waterman MS, Zhou XJ: **Integrative analysis of many weighted co-expression networks using tensor computation.** *PLoS Computational Biology* 2011, **7(6)**:e1001106.
14. Omberg L, Golub GH, Alter O: **A tensor higher-order singular value decomposition for integrative analysis of dna microarray data from different studies.** *Proceedings of the National Academy of Sciences* 2007, **104(47)**:18371-18376.
15. Zhang Y, Du N, Ge L, Jia K, Zhang AD: **A collective nmf method for detecting protein functional module from multiple data sources.** *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine* 2012, 655-660.
16. Narayanan M, Vetta A, Schadt EE, Zhu J: **Simultaneous clustering of multiple gene expression and physical interaction datasets.** *PLoS computational biology* 2010, **6(4)**:e1000742.
17. Garcia-Hernandez M, Berardini T, Chen G, Crist D, Doyle A, Huala E, Knee E, Lambrecht M, Miller N, Mueller LA, *et al*: **Tair: a resource for integrated arabidopsis data.** *Functional integrative genomics* 2002, **2(6)**:239-253.
18. Edgar R, Domrachev M, Lash AE: **Gene expression omnibus: Ncbi gene expression and hybridization array data repository.** *Nucleic acids research* 2002, **30(1)**:207-210.
19. Weinstock-Guttman B, Badgett D, Patrick K, Hartrich L, Santos R, Hall D, Baier M, Feichter J, Ramanathan M: **Genomic effects of ifn-β in multiple sclerosis patients.** *The Journal of Immunology* 2003, **171(5)**:2694-2702.
20. Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, Angelo C, Bornberg-Bauer E, Kudla J, Harter K: **The atgenexpress global stress expression data set: protocols, evaluation and model data analysis of uv-b light, drought and cold stress responses.** *The Plant Journal* 2007, **50(2)**:347-363.
21. Supper J, Strauch M, Wanke D, Harter K, Zell A: **Edisa: extracting biclusters from multiple time-series of gene expression profiles.** *BMC bioinformatics* 2007, **8(1)**:334-348.
22. Li XT, Ye YM, Wu QY, Ng MK: **Multifactv: Finding modules from higher-order gene expression profiles with time dimension.** *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on: 4-7 October 2012* 2012, 1-6.
23. Lin YR, Sun J, Castro P, Konuru R, Sundaram H, Kelliher A: **Metafac: community discovery via relational hypergraph factorization.** *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* 2009, 527-536.
24. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J: **Distributed optimization and statistical learning via the alternating direction method of multipliers.** *Now Publishers* 2011.
25. Ng MK, Weiss P, Yuan X: **Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods.** *SIAM Journal on Scientific Computing* 2010, **32(5)**:2710-2736.
26. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA, *et al*: **David: database for annotation, visualization, and integrated discovery.** *Genome Biol* 2003, **4(5)**:P3.
27. Kreps JA, Wu YJ, Chang HS, Zhu T, Wang X, Harper J: **Transcriptome changes for arabidopsis in response to salt, osmotic, and cold stress.** *Plant Physiology* 2002, **130(4)**:2129-2141.
28. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Science's STKE* 2000, **11(12)**:4241-4257.
29. Vosslamber S, Baarsen L, Verweij CL: **Pharmacogenomics of ifn-β in multiple sclerosis: towards a personalized medicine approach.** *Pharmacogenomics* 2009, **10(1)**:97-108.
30. Acar E, Yener B: **Unsupervised multiway data analysis: A literature survey.** *IEEE Transactions on Knowledge and Data Engineering* 2009, **21(1)**:6-20.

doi:10.1186/1471-2164-14-S4-S2
Cite this article as: Li *et al.*: MultiFacTV: module detection from higher-order time series biological data. *BMC Genomics* 2013 **14**(Suppl 4):S2.