SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2017, Vol. 60 012108:1–012108:16
doi: 10.1007/s11432-016-0030-6

Differential function analysis: identifying structure and activation variations in dysregulated pathways

Chuanchao ZHANG[1,2], Juan LIU[1*], Qianqian SHI[2], Tao ZENG[2*] & Luonan CHEN[2*]

[1]*State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan* 430072*, China;*
[2]*Key Laboratory of Systems Biology, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai* 200031*, China*

Received January 5, 2016; accepted June 13, 2016; published online November 18, 2016

Abstract Complex diseases are generally caused by the dysregulation of biological functions rather than individual molecules. Hence, a major challenge of the systematical study on complex diseases is how to capture the differentially regulated biological functions, e.g., pathways. The traditional differential expression analysis (DEA) usually considers the changed expression values of genes rather than functions. Meanwhile, the conventional function-based analysis (e.g., PEA: pathway enrichment analysis) mainly considers the varying activation of functions but disregards the structure change of genetic elements of functions. To achieve precision medicine against complex diseases, it is necessary to distinguish both the changes of functions and their elements from heterogeneous dysregulated pathways during the disease development and progression. In this work, in contrast to the traditional DEA, we developed a new computational framework, namely differential function analysis (DFA), to identify the changes of element-structure and expression-activation of biological functions, based on comparative non-negative matrix factorization (cNMF). To validate the effectiveness of our method, we tested DFA on various datasets, which shows that DFA is able to effectively recover the differential element-structure and differential activation-score of pre-set functional groups. In particular, the analysis of DFA on human gastric cancer dataset, not only capture the changed network-structure of pathways associated with gastric cancer, but also detect the differential activations of these pathways (i.e., significantly discriminating normal samples and disease samples), which is more effective than the state-of-the-art methods, such as GSVA and Pathifier. Totally, DFA is a general framework to capture the systematical changes of genes, networks and functions of complex diseases, which not only provides the new insight on the simultaneous alterations of pathway genes and pathway activations, but also opens a new way for the network-based functional analysis on heterogeneous diseases.

Keywords complex disease, biological function, non-negative matrix factorization, network structure, function activation

Citation Zhang C C, Liu J, Shi Q Q, et al. Differential function analysis: identifying structure and activation variations in dysregulated pathways. Sci China Inf Sci, 2017, 60(1): 012108, doi: 10.1007/s11432-016-0030-6

1 Introduction

The etiology of complex diseases involves numerous genes, environmental factors and their interactions [1,2], and thus the study of complex diseases is more complicated than expected. Traditional gene-based analysis explores the associations between individual genes and a disease, but only identifies a small proportion of the genetic variants related to a disease, which contribute to a limited understanding of complex

* Corresponding author (email: liujuan@whu.edu.cn, zengtao@sibs.ac.cn, lnchen@sibs.ac.cn)

diseases. There is a growing consensus that complex diseases are mostly contributed by multiple genes through their sophisticated interactions, rather than by the individual genes [3, 4]. Hence, the molecular network analysis of complex diseases could make us further interpret the molecular mechanisms of complex diseases [5] at a system level. However, the molecular network analysis could not directly elucidate the biological or functional roles of the excavated genes/interactions. To provide a comprehensive understanding of the molecular mechanisms causing complex diseases, it is necessary to develop function-based analysis based on the molecular network for complex diseases.

Actually, there are many studies on investigating the biological functions and constructing biological function databases, such as GO ontology database [6] and KEGG pathway database [7]. Based on these well-known databases, a few methods of function-based analysis have been proposed, such as Gene Ontology-based analysis [8, 9] and pathway-based analysis [10]. Especially, the biological pathways could provide the genetic regulated information of the biological functions, and thus the pathway-based analysis is specific and direct on biological functions.

The early developed pathway-based approaches were motivated to understand the biological roles of the excavated genes/interactions, such as the gene set enrichment analysis (GSEA) [10] and network ontology analysis (NOA) [11, 12]. Recently, some new methods have also been proposed to detect the dysregulated pathways by parsing the topological information of a pathway, e.g., signaling pathway impact analysis (SPIA) [13] and CliPPER [14]. Moreover, some other approaches have also been proposed to transform the genetic expression values into functional activation-scores and identify the differentially regulated pathways, such as pathway-based personalized analysis of cancer (Pathifier) [15] and GSVA [16]. Currently, those methods for function-based analysis have become an important way to understand the molecular mechanisms of complex diseases [17].

However, the topological information of the pathways deposited in databases is the assembled information, which is the combination of experimental results from the different labs under the different conditions. Therefore, the topological information could not represent the actual regulated relationships of pathway elements/genes in a specific condition, e.g., a specific disease. The conventional pathway-based analysis directly used such aggregated topological information, and thus considered the varying activation of pathways but disregarded the details of the structure change of pathways.

In this work, in contrast to the traditional DEA on individual molecules, we develop a novel computational framework, namely differential function analysis (DFA), to identify the changes of network-structure and expression-activation of the pathways at a network level, based on our new integration method, i.e., cNMF. By testing on various datasets, we show that DFA is able to efficiently recover the differential element-structure and differential activation-score of pre-set functional groups. Particularly, the analysis of DFA on human gastric cancer dataset, not only captures the changed network-structure of the pathways associated with gastric cancer, but also detects the differential activations of these pathways, which significantly distinguishes normal samples and disease samples and is also more effective than the state-of-the-art pathway-based methods, such as GSVA and Pathifier. Our analysis show that DFA is a general framework to detect the systematical changes of genes, networks and functions of complex diseases, which not only provides a new insight on the simultaneous alterations of pathway genes and pathway activations, but also opens a new way for the network-based functional analysis on heterogeneous diseases.

2 Methods and material

In this section, we describe the framework of DFA (Figure 1). We first introduce the problem, and then present the mathematical model of DFA. Next we describe the iterative multiplicative updating algorithm to solving the model.

2.1 The problem

We could measure the gene expression of samples in complex disease. Because some genes implement the same biological functions, the gene expression data actually have particular sub-structures, which include

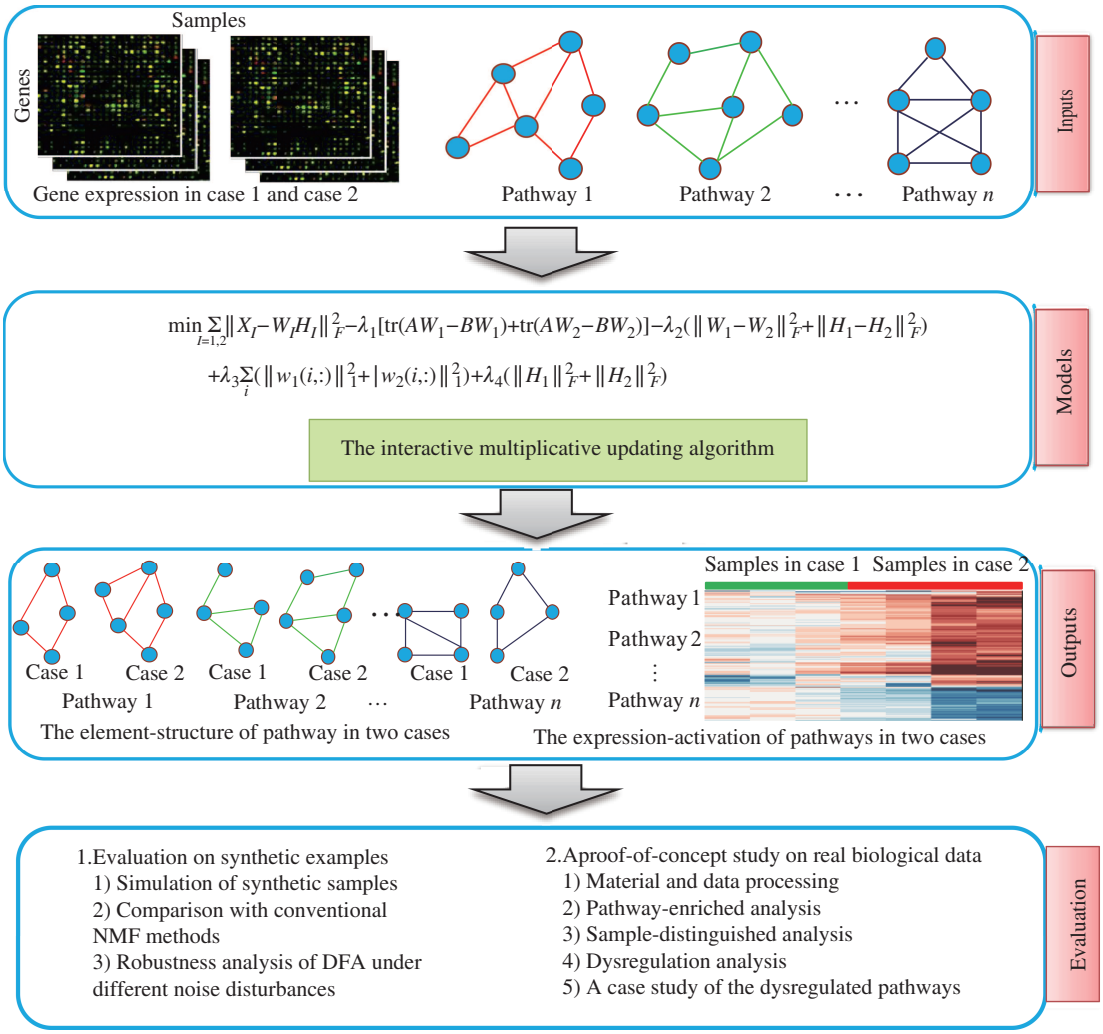

Figure 1 (Color online) Overview of DFA for identifying the dysregulated pathways in complex diseases.

the element component of the biological functions and the activation expression of the biological functions in samples. So, the gene expression (i.e., the matrix X) could decompose to the so-called basis matrix W and the coefficient matrix H. The basis matrix contains the element component information of each biological function, while the coefficient matrix represents the activation information of each biological function in samples. Such a problem of expression decomposition could be formulated as the non-negative matrix factorization (NMF) problem.

The non-negative matrix factorization could divide a matrix X into two non-negative matrices including a coefficient matrix H and a basis matrix W with a lower rank than matrix X [18, 19]. The solution of NMF can be used to easily identify sub-structures of the data [20, 21]. Especially, NMF can be applied to decompose the observed element expression matrix X of prior-known pathways into the actual element component matrix W and the activation matrix H of these pathways. Due to the effectiveness and the inherent advantages of NMF, there were plentiful applications of NMF and its variants in the analysis of large-scale gene expression datasets [22–25], the classification and clustering [26–28], and new class discovery [29, 30]. Besides, several variants of NMF have also been developed by incorporating various kinds of constraints: discriminative constraints [31], locality-preserving or network-regularized constraints [32–34], sparsely constraints [35–38], etc.

However, the main focus of this work is to identify the dysregulated biological functions in complex diseases. To interpret the dysregulation of pathways in complex diseases, one key is to find the differential elements/genes and the differential activations of corresponding biological function under different

conditions (e.g., normal or disease). This could be described in mathematical terms as follows. The input data are the elements of each pathway and the expression values of these elements in two sample groups (e.g., the matrices X_1 and X_2 for normal and disease sample groups respectively). A mathematical model is designed to decompose the matrices (X_1, X_2) into the basis matrices (W_1, W_2) and the coefficient matrices (H_1, H_2) correspondingly. Therefore, some joint-NMF model is needed. Recently, a number of joint-NMF models have been proposed [23, 39]. These methods usually assume that the coefficient matrix or the basis matrix of pathways would be identical in two sample groups, which could not detect the differentially regulated elements and the differentially activated pathways simultaneously. Thus, to address this issue, we proposed a new model for DFA, based on a novel technique, i.e., cNMF. This model is more general than the traditional joint-NMF, and makes that the coefficient matrix and the basis matrix of pathways would have restricted differences, simultaneously.

2.2 Comparative non-negative matrix fraction (cNMF)

The model of DFA is mainly based on cNMF, which includes the fitness function, the pathway-enriched constraint, the dysregulation constraint and the sparse constraint.

2.2.1 *Fitness function*

DFA does not require the same coefficient matrix or the same basis matrix of pathways in the two group samples, and thus cNMF could decompose the matrices (X_1, X_2) into the basis matrices (W_1, W_2) and the coefficient matrices (H_1, H_2) respectively. In other words, the basis matrices W_1 and W_2, and the coefficient matrices H_1 and H_2 could be different. Hence, the joint decomposition of the expression data for the two group samples can be derived by optimizing the following fitness function:

$$F(W_1, W_2, H_1, H_2) = \min \sum_{I=1,2} \|X_I - W_I H_I\|_F^2, \tag{1}$$

where X_1 and X_2 have the same dimensions $s \times m$; W_1 and W_2 have the same dimensions $s \times k$; H_1 and H_2 have the same dimensions $k \times m$. The parameter k is chosen prior to optimization, which is just the number of the analyzed pathways in this study.

2.2.2 *Pathway-enriched constraint*

The information on the known components or elements of pathways can be used to make each column of the basis matrix enriched on one corresponding pathway. That means, the basis matrix could contain the actual element information of each pathway where the value of each column represents the degree of an element belonging to a pathway. Thus, we have the following hypothesis: if one column of the basis matrix enriched on the corresponding pathway, the element values of this pathway should be very larger than those of the other elements. Then, the mathematical formalization of so-called pathway-enriched constraint could be derived by the following function:

$$\begin{aligned} O_1 &= \sum_i \sum_k (a(i,k)w_1(k,i) - b(i,k)w_1(k,i)) + \sum_i \sum_k (a(i,k)w_2(k,i) - b(i,k)w_2(k,i)) \\ &= \mathrm{tr}(AW_1 - BW_1) + \mathrm{tr}(AW_2 - BW_2), \end{aligned} \tag{2}$$

where the binary matrix A reflects the information of the known elements of pathways, and an element belongs to the corresponding pathway only when its value is one. The binary matrix B represents the information of the contrary elements, where an element does not belong to the corresponding pathway only when its value is one. Clearly the value of this constraint function is expected to be as large as possible, and thus the function (2) is actually taken as a soft constraint to add to the objective function (1).

2.2.3 *Dysregulation constraint*

In complex diseases, the elements/genes of pathways in two sample groups (i.e., normal and disease samples) may be different; meanwhile the activation scores of pathways in the samples may be also different. Hence, each column of the basis matrices W and each row of the coefficient matrix H in two samples groups could be different. We apply the Frobenius-norm on the basis matrix W and the coefficient matrix H to constrain the differential element-structures and the differential activation-scores of pathways respectively.

In details, the mathematical formalization of so-called dysregulation constraint could be derived by the following function:

$$G_1(W_1, W_2) = \sum_i \sum_k^s (w_1(k,i) - w_2(k,i))^2 = \|W_1 - W_2\|_F^2, \tag{3}$$

$$G_2(W_1, W_2) = \sum_i \sum_k^m (h_1(i,k) - h_2(i,k))^2 = \|H_1 - H_2\|_F^2. \tag{4}$$

Note that, the above function makes the model is suitable to handle the matched sample data, e.g., the tumor samples and the tumor-adjacent normal samples. All those constraints are soft constraints, which are added to the objective function (1).

2.2.4 *Sparsity constraint*

The sparse representations of NMF methods could discover the partial patterns [18], and several approaches have been proposed to obtain the sparse W and/or H factors [35,37,40,41]. Our cNMF of DFA similarly adopt the idea of imposing L_1-norm to make the sparsity of basis matrices W_1 and W_2 [41].

Finally, by adding all of those soft constraints, the extended objective function of cNMF is defined as follow:

$$\begin{aligned} \min \quad & \sum_{I=1,2} \|X_I - W_I H_I\|_F^2 - \lambda_1[\mathrm{tr}(AW_1 - BW_1) + \mathrm{tr}(AW_2 - BW_2)] \\ & - \lambda_2(\|W_1 - W_2\|_F^2 + \|H_1 - H_2\|_F^2) + \lambda_3 \sum_i (\|w_1(i,:)\|_1^2 + \|w_2(i,:)\|_1^2) \\ & + \lambda_4(\|H_1\|_F^2 + \|H_2\|_F^2), \end{aligned} \tag{5}$$

where the term $\lambda_3 \sum_i (\|w_1(i,:)\|_1^2 + \|w_2(i,:)\|_1^2)$ encourages the sparsity of the matrix W, while $\lambda_4(\|H_1\|_F^2 + \|H_2\|_F^2)$ limits the growth of the matrix H.

2.3 Differential function analysis: a new model of function analysis

2.3.1 *Solving DFA by cNMF*

Obviously, cNMF is not convex in W_1,W_2, H_1 and H_2. Therefore, it is unrealistic to find the global minimum. Similar to the classical NMF algorithms [19,23,42], we have developed the iterative algorithm of cNMF to solve the DFA model as the following algorithm, which efficiently converges to a local minimum by iteratively updating the matrix decomposition. The updating rules and proof of this algorithm are provided in supplementary information (SI).

2.3.2 *Pathway-enriched analysis of DFA by pathway remodeling on basis matrices*

Each column of the basis matrix is expected to represent a particular pathway. Thus, the enrichment significant score (ES-score) of each pathway is designed to evaluated whether the estimated basis matrix of DFA could recover the element-structure of the analyzed pathways, and ES-score is calculated on one column of the basis matrix as bellows.

The elements of one column of the basis matrix are ranked by their values in a descending order, and the first N elements are selected, where N denotes the element number of the particular pathway of

Algorithm 1 Algorithmic framework for DFA

Step 1 Initialize W_1, W_2, H_1 and H_2 with non-negative values, and set the iteration index $t = 0$.

Step 2 Fix H_1 and H_2, solve the constrained problem

$$\min \sum_{I=1,2} \|X_I - W_I H_I\|_F^2 - \lambda_1[\mathrm{tr}(AW_1 - BW_1) + \mathrm{tr}(AW_2 - BW_2)] - \lambda_2\|W_1 - W_2\|_F^2 + \lambda_3 \textstyle\sum_i (\|w_1(i,:)\|_1^2 + \|w_2(i,:)\|_1^2).$$

That is, update W_1, W_2 with

$$w_{ij}^1 \leftarrow w_{ij}^1 \frac{(2X_1H_1^{\mathrm{T}} + \lambda_1 A^{\mathrm{T}} + 2\lambda_2 W_1)_{ij}}{(2W_1H_1H_1^{\mathrm{T}} + \lambda_1 B^{\mathrm{T}} + 2\lambda_3 W_1 e_{k\times k} + 2\lambda_2 W_2)_{ij}}.$$

$$w_{ij}^2 \leftarrow w_{ij}^2 \frac{(2X_2H_2^{\mathrm{T}} + \lambda_1 A^{\mathrm{T}} + 2\lambda_2 W_2)_{ij}}{(2W_2H_2H_2^{\mathrm{T}} + \lambda_1 B^{\mathrm{T}} + 2\lambda_3 W_2 e_{k\times k} + 2\lambda_2 W_1)_{ij}}.$$

Step 3 Fix W_1 and W_2, solve the constrained problem

$$\min \sum_{I=1,2} \|X_I - W_I H_I\|_F^2 - \lambda_2\|H_1 - H_2\|_F^2 + \lambda_4(\|H_1\|_F^2 + \|H_2\|_F^2).$$

That is, update H_1, H_2 with

$$h_{ij}^1 \leftarrow h_{ij}^1 \frac{(2W_1^{\mathrm{T}}X_1 + 2\lambda_2 H_1)_{ij}}{(2W_1^{\mathrm{T}}W_1H_1 + 2\lambda_4 H_1 + 2\lambda_2 H_2)_{ij}}.$$

$$h_{ij}^2 \leftarrow h_{ij}^2 \frac{(2W_2^{\mathrm{T}}X_2 + 2\lambda_2 H_2)_{ij}}{(2W_2^{\mathrm{T}}W_2H_2 + 2\lambda_4 H_2 + 2\lambda_2 H_1)_{ij}}.$$

Step 4 let $t \leftarrow t + 1$, repeat Steps 2 and 3 until convergence criteria are satisfied.

this column. (i) For the synthetic examples, the ES-score of each pathway in one column directly uses the percentage of pathway elements in the Top-N. (ii) For the real biological samples, the ES-score is the P-value of a hypergeometric test on the enrichment of pathway elements in the Top-N, where the hypergeometric test is introduced in SI.

Note that there are two basis matrices in cNMF, and thus the same columns in the two matrices are considered to play the same biological role, e.g., the same pathway enriched. The moderate significant value, which is the minimum score of the two ES-scores of one pathway for the synthetic examples and is the maximum score of the two ES-scores of one pathway for the real biological samples, is the final ES-score of this pathway on this column in both basis matrices.

The ES-score of the pathway corresponding to one column can represent the pathway-enriched significance of this column. The pathway-enriched significance of all columns consist of the pathway-enriched significance vector, and their average value is the pathway-enriched significant score (PE-score) of the basis matrices. If the pathway-enriched significance of one column is larger than ES-scores of other pathways, this column is called as the pathway-identified column. The pathway-identified ratio is further measured by the percentage of the pathway-identified columns in all.

The pearson correlation between the same column from the calculated and pre-set basis matrices is calculated. And in cNMF, the minimum value of the two correlation values for the same column from the two calculated basis matrices is defined as the final correlation value (P-score) of this column. The P-score of one column can represent the significance of the element weights of this column consistent with the pre-set element weights. The P-scores for all columns consist of a new correlation vector, and the average value of this vector is called as pathway remodeling score (PR-score). PR-score represents the pathway recovery on element weights of the basis matrices.

Especially, for the synthetic example, some additional measurements are designed for evaluation. Two measurements are used to evaluate the performance of the basis matrix recovery on synthetic samples. One is the Euclidean distance (E_{ES}) between the PE-scores from the calculated vector and the prior-known vector where each column is the pathway-identified column and its ES-score is one. The other one is the Euclidean distance (E_{PS}) between P-scores from the calculated correlation vector and the prior-known correlation vector where the element weights of each column is consistent with pre-set basis matrices and its P-score is one. Obviously, E_{ES} indicates the pathway recovery on element number and E_{PS} reflects the pathway recovery on element weights.

Noted, the basis matrix is considered as the element-structure of the computed pathways. In order to achieve this purpose, a pathway-enriched constraint is designed and actually taken as a soft constraint to add to the objective function. To deal with the tradeoff of optimization with such objective function, the ES-score is further designed to evaluate whether the estimated basis matrix of DFA could recover

the element-structure of the analyzed pathways. Thus, ES-score could be considered as a criterion for judging DFA in order to select the optimal parameters of DFA.

2.3.3 *Sample-distinguished analysis of DFA by reclassifying samples based on coefficient matrices*

Each row of the coefficient matrix represents the activation scores of one pathway in the samples. The K-means algorithm is applied for each row of the coefficient matrix, and the percentage of the correctly distinguished samples in all samples is used as the sample-distinguished score (SD-score) of the corresponding pathway. The SD-scores for all rows consist of the SD-score vector. And the average value of this vector can measure the global sample-distinguished score (GSD-score) of coefficient matrices.

For the synthetic examples, the true sample classification is prior known. Thus, the Euclidean distance (E_{SD}) is used to evaluate the differences between the two SD-score vector from the calculated coefficient matrix and the pre-set coefficient matrix. Such Euclidean distance is considered to represent the ability to reclassify samples by DFA, which is expected to be as minimal as best.

3 Result and discussion

3.1 Evaluation on synthetic examples

To illustrate DFA for analysis of the dysregulation of pathways between the two sample groups, a large number of numeric examples have been produced by the randomly generated datasets. DFA has been applied on these examples and compared with the conventional NMF approaches. In addition, the robustness of DFA has also been evaluated.

3.1.1 *Simulation of synthetic samples*

In order to completely evaluate DFA, the numeric examples include nine categories, which represent nine different kinds of pathway alterations. Generally, the dysregulated pathways have three types: the pathways with differential elements and differential activation (dWdH); the pathways with only differential element (dW); and the pathways with only differential activation (dH). And three different fractions (100%, 90% and 80%) are also used to represent the degree of the differences/dysregulations. For example, given the dysregulated pathways with differential element and differential activation, three datasets were randomly generated corresponding to three different fractions (dWdH_0, dWdH_1 and dWdH_2), where the difference degree of element-structure and activation-score are 100%, 90% and 80% respectively. Thus, the nine category datasets are denoted by dWdH_0, dWdH_1, dWdH_2, dW_0, dW_1, dW_2, dH_0, dH_1 and dH_0. For each dataset, there are 10 pathways and each pathway contains 10 elements. Hence, the element expression data of these pathways have 100 elements and 50 samples in each sample group. More details of the synthetic examples are included in SI.

3.1.2 *Comparison with conventional NMF methods*

DFA was applied in these simulated datasets, and evaluated its performance with other conventional NMF methods. These conventional methods include two individual NMF model with no joint-constraint (NMF_t) and two joint NMF model (jNMF). jNMF model also contains the joint-NMF with fixed matrix W (jNMF_w) and the joint-NMF with fixed matrix H (jNMF_h). To show our method for recovering the differential element-structures and differential activation-scores of pathways, we give an overview of the performance of DFA and the conventional jNMF method (e.g., jNMF_h) on the simulated pathways with differential elements and differential activation (Figure 2).

The four measurements are proposed to evaluate such four methods completely, including the PR-score of the basis matrices, the PE-score of the basis matrices, the GSD-score of the coefficient matrices, and the comprehensive score which is the average value of E_{ES} of the basis matrices, E_{PS} of the basis matrices and E_{SD} of the coefficient matrices, and represents the average performance of our method for the pathway recovery.

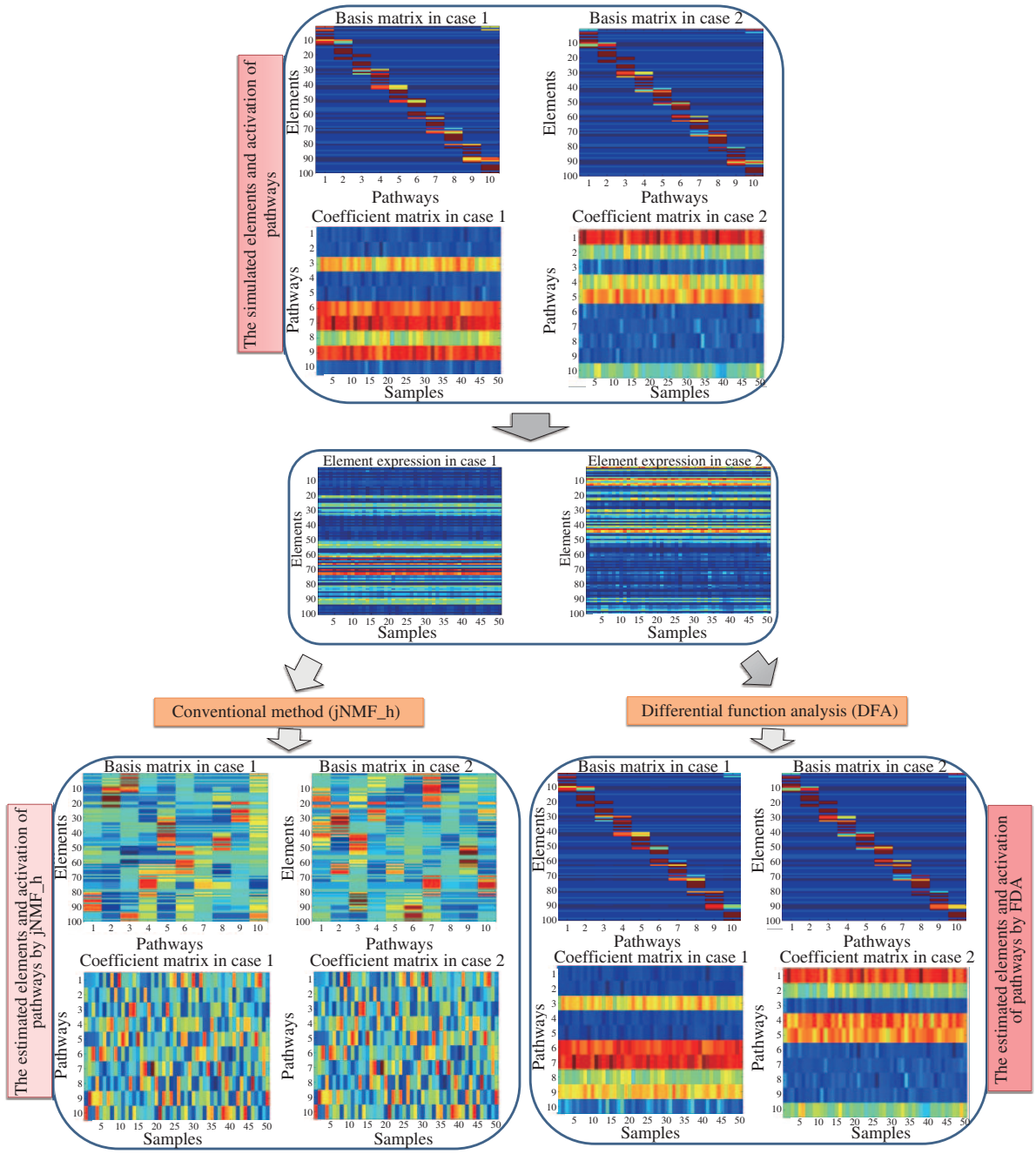

Figure 2 (Color online) The overview of the performance of our method DFA and the conventional jNMF method on the simulated pathways with differential elements and differential activation. The basis matrix contains the information for elements of each pathway, and the coefficient matrix contains the activation information of each pathway in each sample.

The comparison results are shown in Figure 3. According to the above measurements, DFA actually displays the best ability to recover the differential element-structures and differential activation-scores of the pre-set functional groups among all the compared methods/strategies.

3.1.3 *Robustness analysis of DFA under different noise disturbances*

Moreover, the above nine simulated datasets have additional noise disturbances including 0%, 0.0001%, 0.001%, 0.01%, 0.1%, 1%, 10%, 20%, 30% and 50% noise, respectively. DFA has been carried on these additional noisy simulated datasets. As shown in Figure 4, DFA indeed is robust for noisy data.

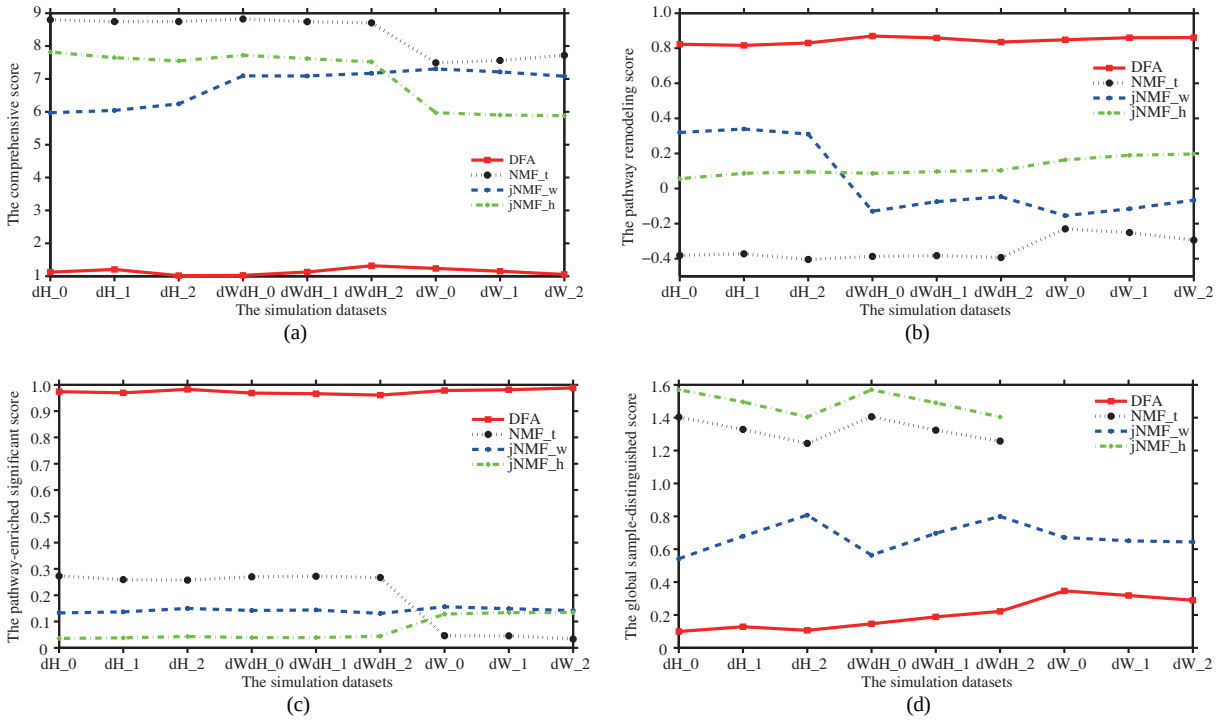

Figure 3 (Color online) The performance of DFA and conventional methods. They are the average performances from random 1000 times. (a) The comprehensive score of DFA and conventional methods, and the minimal score is the best result; (b) the PR-score of the basis matrices, and the maximal score is the best result; (c) the PE-score of the basis matrices, and the maximal score is the best result; (d) the GSD-score of the coefficient matrices, and the minimal score is the best result. Note that NMF_t and jNMF_h cannot provide non-trivial results on the datasets with dW type.

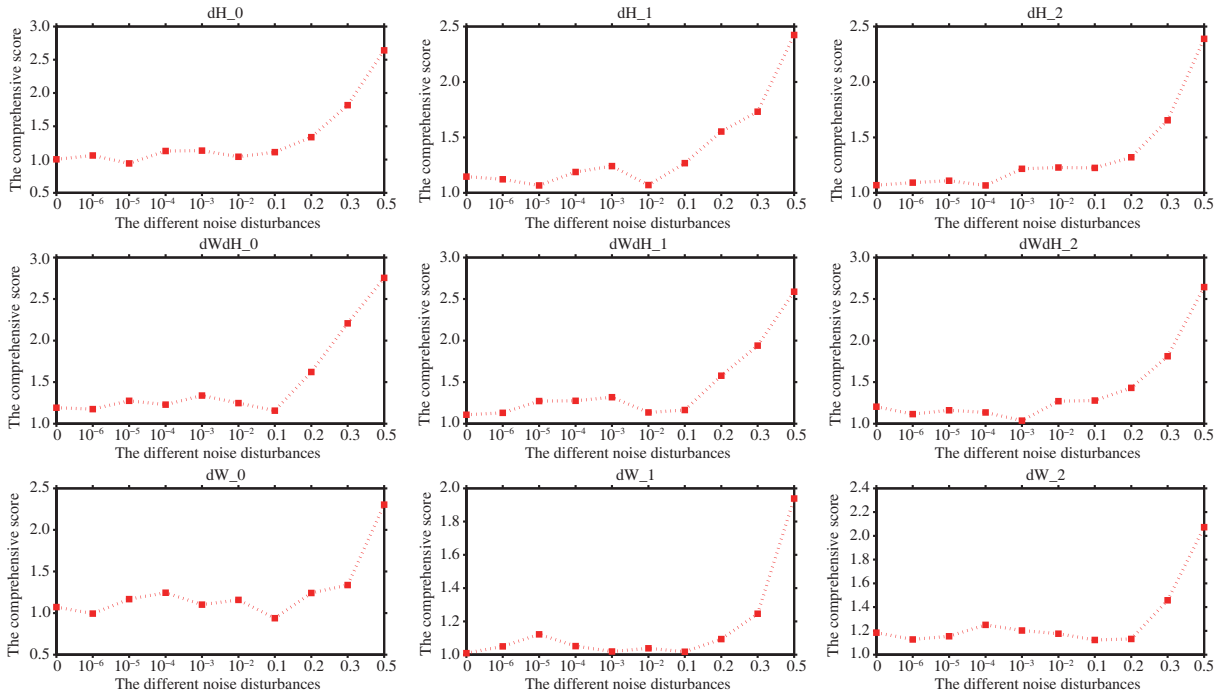

Figure 4 (Color online) The robustness evaluation of DFA under different noise disturbances. The nine sub-figures show the robustness performance of DFA in nine groups with different noise levels respectively. Each sub-figure displays the comprehensive score of DFA output under different noise disturbances.

3.2 A proof-of-concept study on real biological data

3.2.1 *Material and data processing*

The gene expression profiles (GSE27342) of 160 paired gastric cancer samples were downloaded, including 80 tumor samples, and 80 tumor-adjacent normal samples [43].

As a biological case analysis of our methodology, the KEGG pathways [7] are used as input, which would be comparable with previous studies.

Particularly on the analysis of gastric cancer, the KEGG pathways related to gastric cancer were selected by the prior knowledge. In details, the pathways related to gastric cancer were defined as the pathways which have significant enrichments on the disease genes and the differential expressed genes, which include several steps as follows:

(1) the genes related with gastric cancer are obtained from the GeneCard (http://www.genecards.org/);

(2) the differential expressed genes are obtained by student's T-test on normal and tumor samples;

(3) the enrichment P-value of each pathway on disease genes and differential expressed genes is calculated by hypergeometric test;

(4) last, there are 53 KEGG pathways with significant enrichment (P-value<0.05), and were used in the following analysis on human gastric cancer.

Based on the above selected pathways and the parameter k is just the number of the analyzed pathways as 53, the all parameters of DFA are tuned in a reasonable scale and chosen according to the sum of the optimal pathway-enriched score and SD-score in the following ways.

(1) The basis matrix of DFA is considered as the element-structure of the analyzed pathways, where each column is expected to represent a particular pathway. So, we design the pathway-enriched score to evaluate the accuracy of DFA by requiring each column of the calculated basis matrix enriched in corresponding particular pathway. And the larger pathway-enriched score is, the better DFA is.

(2) Meanwhile, the coefficient matrix is considered as the activation-score of the analyzed pathways, where the values of the coefficient matrix are required to distinguish normal and tumor samples. Therefore, we design the SD-score to evaluate the accuracy of DFA by clearly grouping normal and tumor samples. And the larger SD-score is, the better DFA is.

Totally, the sum of the pathway-enriched score and the SD-score could be used together as the criteria for evaluating the performance of DFA, by which we can select the optimal parameters for DFA.

Note that the main elements of pathways were the topological information. Not only the genes involved in pathways but also the edges/gene-pairs of pathways could provide valuable information to understand complex diseases. Thus, in this work, we used the edges of pathways to analyze the dysregulation of pathways from the perspective of edges (i.e., network). That means, the input data of DFA was the quantified element score of gene-pairs rather than original expression value of genes in pathways (see more details in SI).

3.2.2 *Pathway-enriched analysis reveals significant pathway recovered by DFA*

Based on the estimated basis matrix, the pathway-enriched significance of all column/pathways is calculated (see Section 2). The pathway-enriched significance of most columns is small, and thus the most columns are the identified pathways. With different thresholds of significance, the columns/pathways with less pathway-enriched score than the thresholds are counted, and the percentage of these enriched columns/pathways are shown in Figure 5. Obviously, DFA is very effective to recover pathways even when the significance threshold is strict. The pathway-identified columns are also counted and the pathway-identified ratio is 96% (51/53), which means that the differential information of different pathways are well decomposed.

3.2.3 *Sample-distinguished analysis reveals accurate sample discrimination achieved by DFA*

The activation expression of the analyzed pathways can be used to cluster samples by K-means algorithm, and the SD-score of each pathway was calculated (see Section 2). To illustrate reasonability of the DFA

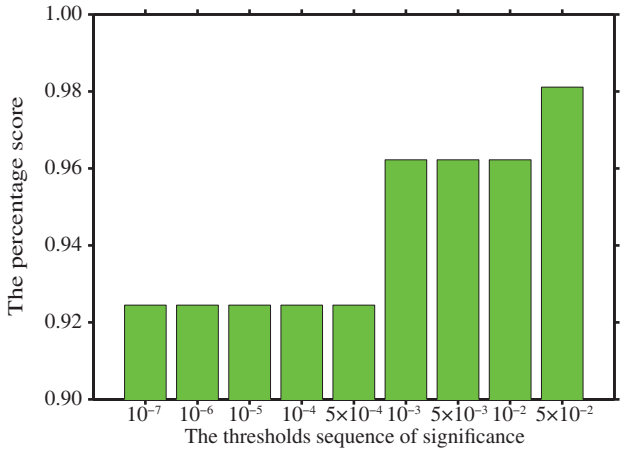

Figure 5 (Color online) Performance of the pathway-enriched significance score under the different thresholds of significance.

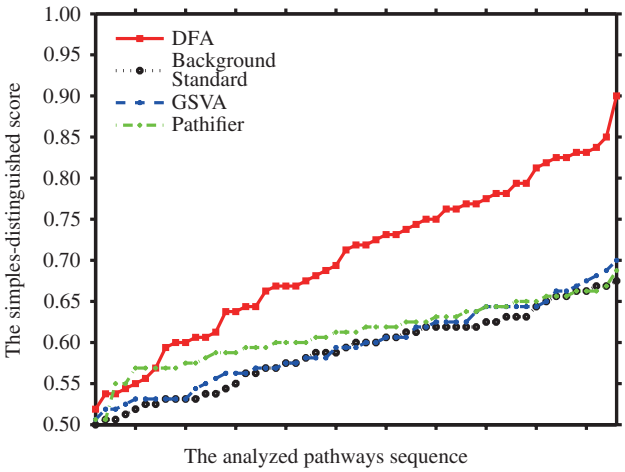

Figure 6 (Color online) Sample-distinguished evaluation of DFA and the conventional methods. The x-axis represents the analyzed pathways in a sequence. The y-axis represents the SD-score of the analyzed pathways. Note that, the SD-score of the analyzed pathways in the figure has been sorted for convenient visualization.

Table 1 The average GSD-score of DFA and the conventional methods

Method	Background standard	GSVA	Pathifier	DFA
GSD-score	0.594	0.596	0.611	0.7026
Increased ratio (%)		0.42	2.8	18.3

consideration on the change of element-structure, DFA and some conventional methods were compared by evaluating the SD-score, which support that the decomposed pathway structure-change can improve the pathway activation-score to distinguish normal and tumor samples. The conventional methods (such as GSVA and Pathifier) only obtained the activation expression of pathways but not the change of element-structure. As an experiment control, the SD-score of each analyzed pathway was also directly calculated from the original gene expression level, which was regarded as the Background Standard.

As shown in Figure 6, GSVA has performance close to control; Pathifier is slight better; and DFA is much better than all compared approaches for most pathways. The performance values of these methods were shown in Table 1, and the increased ratio indicates the improvement of particular method compared to control. Clearly, these assessments strongly support the necessary of investigating the changes of element-structure, and our DFA can address this serious issue well and much better than the state-of-the-art methods.

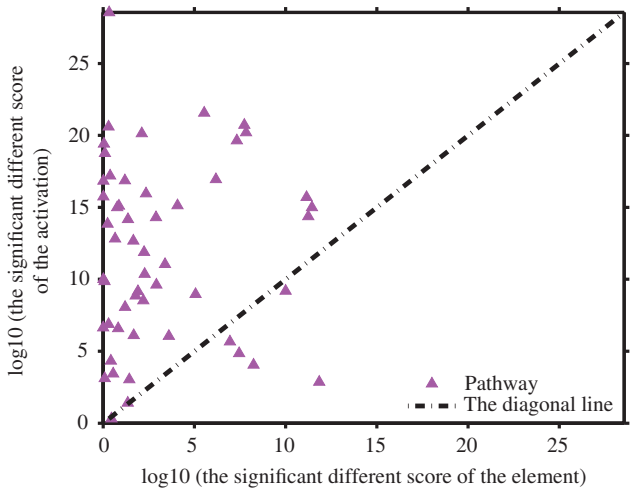

Figure 7 (Color online) The illustration of the dysregulation of pathways in the differential elements and the differential activation.

Table 2 The significant pathways regarding to the element-structure[a)]

Pathway name	P-values_E	P-values_A	R
Pyrimidine metabolism	1.43×10^{-12}	0.001369	N
Basal cell carcinoma	3.58×10^{-12}	9.99×10^{-16}	N
Dopaminergic synapse	5.61×10^{-12}	4.24×10^{-15}	N
Purine metabolism	6.99×10^{-12}	1.96×10^{-16}	N
Small cell lung cancer	9.85×10^{-11}	6.38×10^{-10}	Y
Cell cycle	5.71×10^{-9}	8.81×10^{-5}	Y
Gastric acid secretion	1.45×10^{-8}	6.21×10^{-21}	Y
Phenylalanine metabolism	1.8×10^{-8}	1.88×10^{-21}	N
Metabolism of xenobiotics by cytochrome P450	3.48×10^{-8}	1.4×10^{-5}	Y
Pathways in cancer	4.62×10^{-8}	2.25×10^{-20}	Y
Epstein-Barr virus infection	1.11×10^{-7}	2.14×10^{-6}	Y
Systemic lupus erythematosus	6.53×10^{-7}	1.13×10^{-17}	Y
Proteoglycans in cancer	2.86×10^{-6}	2.68×10^{-22}	Y
HTLV-I infection	8.58×10^{-6}	1.08×10^{-9}	N
Melanogenesis	8.44×10^{-5}	7.29×10^{-16}	Y

a) P-values_E evaluates the different significance of the elements; P-values_A evaluates the different significance of the activation; R represents whether one pathway is known related with the gastric cancer; Y represents the pathway is known associated with the gastric cancer; N represents the pathway is unclearly associated with the gastric cancer

3.2.4 *Dysregulation analysis reveals the simultaneous changes of pathway element and activation by DFA*

As mentioned in this paper, identifying the dysregulations of pathways need to consider the changes of elements and activation. The significant different score of each analyzed pathway between the two sample groups was calculated by student's T-test in two aspects as the elements and the activation, which are shown in Figure 7. Obviously, DFA not only considers the changes of expression-activation, but also detects the changes of element-structure which are usually disregarded by the conventional methods.

We list the most significant 15 pathways in Table 2. There are 9 pathways related to the gastric cancer as reported in literatures. These pathways contained: the outcome of gastric cancer (e.g., Epstein-Barr virus infection), the common pathways in cancer (e.g., Pathways in cancer, Cell cycle and Proteoglycans in cancer), and the pathways synchronously occurred with gastric cancer (e.g., Small cell lung cancer [44], Metabolism of xenobiotics by cytochrome P450 [45], Systemic lupus erythematosus [46] and Gastric acid secretion [47] and Melanogenesis [48, 49]).

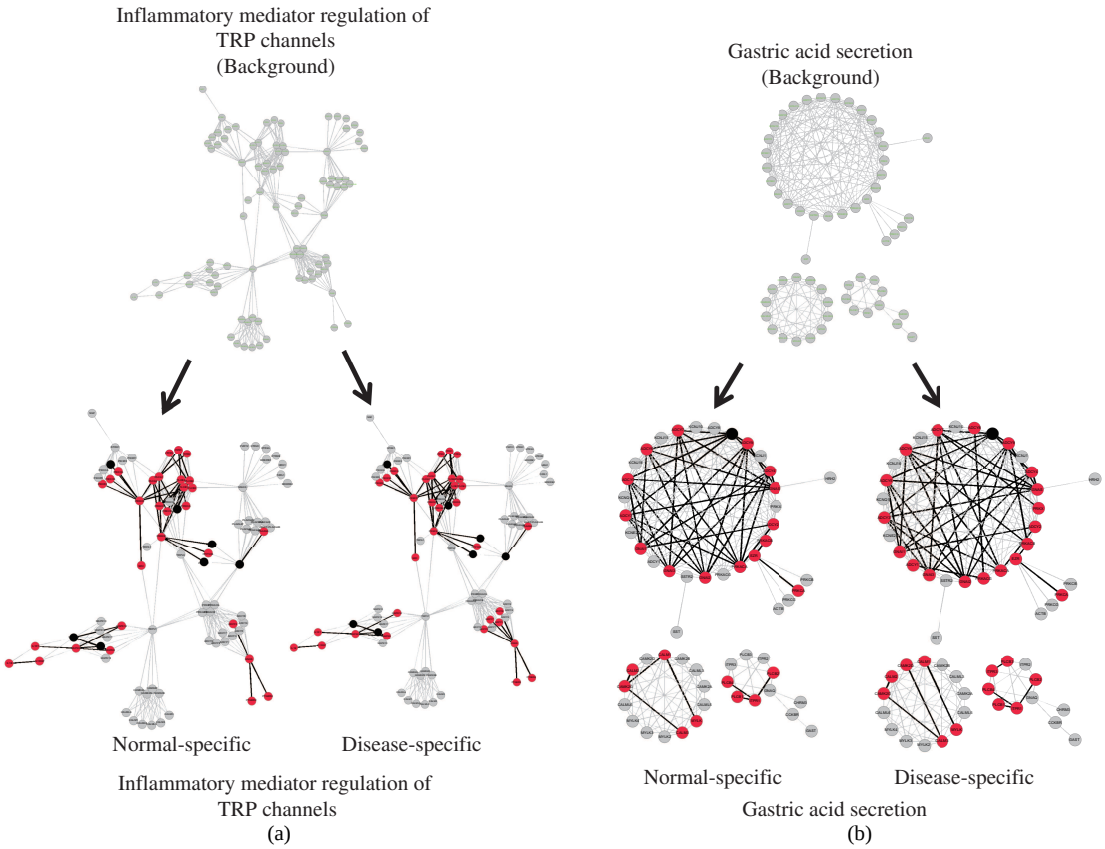

Figure 8 (Color online) The case studies of the dysregulated pathways. (a) shows the Inflammatory mediator regulation of TRP channels, where its normal-specific network and disease-specific network were shown; (b) shows Gastric acid secretion, where its normal-specific network and disease-specific network were shown. The red gene represents the disease-related gene, the black gene represents the actual gene and the gay gene represents visional gene.

3.2.5 *A case study of the dysregulated pathways with different changes on element-structure and expression-activation*

Finally, as a real example to display the topological characteristics of two kinds of dysregulated pathways, Gastric acid secretion and the inflammatory mediator regulation of TRP channels were used in the case studies, which correspond to pathways with different element-structure and different expression-activation, respectively.

For the inflammatory mediator regulation of TRP channels, it was related with gastric cancer [49], and its normal-specific topological information and disease-specific topological information were shown in Figure 8(a). The significant different score of this pathway on the elements and activation are 0.02 and 2.11×10^{-13}, respectively. Thus, this pathway tends to have only big changes on its activations.

For Gastric acid secretion, it was also reported to be related with gastric cancer [47], and its normal-specific topological information and disease-specific topological information were shown in Figure 8(b). The significant different score of this pathway on the elements and activation are 1.45×10^{-8} and 6.21×10^{-21}, respectively. Thus, this pathway would have both significant changes on pathway structure and pathway activation.

4 Conclusion

The molecular network analysis of the complex disease is a powerful way to interpret the molecular mechanisms of complex diseases. But, conventional molecular-based/gene-based analysis could not directly recover the biological roles of the excavated genes/interactions. Hence, the function analysis based on

network has increasingly attracted the attention from the communities of biological and medical sciences.

However, the conventional methods for function methods mainly consider the varying activation of pathways but disregard the details of the topological change of pathways. Therefore, it is difficult to precisely indicate the dysregulation of biological functions by these methods. In this work, we present an elaborate computational framework DFA to investigate the changes of network-structure and expression-activation of the pathways, especially for complex diseases. DFA has been carried on various synthetic datasets and some real disease datasets as gastric cancer. The results indeed show the DFA has the superior ability to quantify the differential element-structure and differential activation-score of dysregulated pathways. Particularly, the analysis of DFA in human gastric cancer actually detects the significantly differential activations of the pathways associated with gastric cancer including the changed network-structure and pathway activation, which is more effective than the state-of-the-art as GSVA and Pathifier.

Besides, we also discussed the influence of different pathway sources for DFA and conventional methods on human gastric cancer dataset. Similar to the usage of KEGG, we used Reactome database [50] and BioCarta database (http://www.biocarta.com) as the prior knowledge of DFA respectively. We evaluated and compared these estimated results of DFA with different pathway databases by the global sample-distinguished score (GSD-score). Moreover, we also integrated the pathways of the KEGG, BioCarta and Reactome to form the integrated set of all pathways and discussed the performance of DFA with such integrated pathways as prior knowledge (see more details in SI). Based on these additional results, we could find that DFA always has the best performance than conventional methods under different settings of pathway sources, which support DFA is actually more effective than the state-of-the-art pathway-based methods. Meanwhile, the performance of DFA is further improved when the integrated pathways are used, which implies that the integration of pathways maybe a way to enhance the accuracy of DFA which is worth of studying in future.

In future, we will develop new dysregualtion constraint function to further extend DFA for wide applications including paired samples or unpaired samples. Also it is important to consider direct or causal associations in networks (i.e., by using partial correlation for linear systems or by part mutual information for nonlinear systems) [51, 52], and further consider dynamics of living organisms (i.e., by dynamical network biomarkers) [53–56] and modularity of interactome (i.e., by edge or module network [57–60] for the analysis of biological functions).

Acknowledgements This paper was supported by Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (Grant No. XDB13040700), and National Natural Science Foundation of China (Grant Nos. 61272274, 60970063, 61134013, 91439103, 91529303, 31200987 and 81471047). It was also partially supported by JSPS KAKENHI (Grant No. 15H05707).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Jin L, Zuo X Y, Su W Y, et al. Pathway-based analysis tools for complex diseases: a review. Genom Proteom Bioinform, 2014, 12: 210–220
2 Panoutsopoulou K, Zeggini E. Finding common susceptibility variants for complex disease: past, present and future. Brief Funct Genom Proteom, 2009, 8: 345–352
3 Freimer N B, Sabatti C. Human genetics: variants in common diseases. Nature, 2007, 445: 828–830
4 Thomas D. Gene-environment-wide association studies: emerging approaches. Nat Rev Genet, 2010, 11: 259–272
5 Cordell H J. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet, 2009, 10: 392–404
6 Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, 25: 25–29
7 Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000, 28: 27–30
8 Holmans P, Green E K, Pahwa J S, et al. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet, 2009, 85: 13–24
9 Zhang C C, Liu J, Shi Q Q, et al. Identification of phenotypic networks based on whole transcriptome by comparative network decomposition. In: Proceedings of Bioinformatics and Biomedicine (BIBM), Washington, 2015. 189–194

10 Subramanian A, Tamayo P, Mootha V K, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Nat Acad Sci, 2005, 102: 15545–15550
11 Wang J, Huang Q, Liu Z P, et al. NOA: a novel network ontology analysis method. Nucleic Acids Res, 2011, 39: e87
12 Zhang C, Wang J, Hanspers K, et al. NOA: a cytoscape plugin for network ontology analysis. Bioinformatics, 2013, 29: 2066–2067
13 Tarca A L, Draghici S, Khatri P, et al. A novel signaling pathway impact analysis. Bioinformatics, 2009, 25: 75–82
14 Martini P, Sales G, Massa M S, et al. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res, 2013, 41: 218–225
15 Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. Proc Nat Acad Sci, 2013, 110: 6388–6393
16 Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinform, 2013, 14: 1–15
17 Khatri P, Sirota M, Butte A J. Ten years of pathway analysis: current approaches and outstanding challenges. Plos Comput Biol, 2012, 8: 1454–1459
18 Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401: 788–791
19 Lee D D, Seung H S. Algorithms for non-negative matrix factorization. Adv Neural Inform Proc Syst, 2001, 13: 556–562
20 Wang Y X, Zhang Y J. Nonnegative matrix factorization: a comprehensive review. IEEE Trans Knowl Data Eng, 2013, 25: 1336–1353
21 Jia Z L, Zhang X, Guan N Y, et al. Gene ranking of RNA-seq data via discriminant non-negative matrix factorization. Plos One, 2015, 10: e0137782
22 Zhang X, Guan N Y, Jia Z L, et al. Semi-supervised projective non-negative matrix factorization for cancer classification. Plos One, 2015, 10: e0138814
23 Zhang S H, Li Q J, Liu J, et al. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics, 2011, 27: i401–i409
24 Leo T, Bjorn N. A framework for regularized non-negative matrix factorization, with Application to the analysis of gene expression data. Plos One, 2012, 7: e46331
25 Lee C M, Mudaliar M A V, Haggart D R, et al. Simultaneous non-negative matrix factorization for multiple large scale gene expression datasets in toxicology. Plos One, 2012, 7: 1411
26 Ma H, Jia M, Shi Y K, et al. Semi-supervised nonnegative matrix factorization for microblog clustering based on term correlation. Web Technol Appl, 2014, 8709: 511–516
27 Seichepine N, Essid S, Fevotte C, et al. Soft nonnegative matrix co-factorization. IEEE Trans Signal Process, 2014, 22: 5940–5949
28 Liu H F, Wu Z H, Li X L, et al. Constrained nonnegative matrix factorization for image representation. IEEE Trans Patt Anal Mach Intell, 2012, 34: 1299–1311
29 Wu Q Y, Wang Z Y, Li C S, et al. Protein functional properties prediction in sparsely-label PPI networks through regularized non-negative matrix factorization. BMC Syst Biology, 2015, 9: 1–14
30 Fogel P, Young S S, Hawkins D M, et al. Inferential, robust non-negative matrix factorization analysis of microarray data. Bioinformatics, 2007, 23: 44–49
31 Zafeiriou S, Tefas A, Buciu I, et al. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. IEEE Trans Neural Netw, 2006, 17: 683–695
32 Jiang J J, Zhang H B, Xue Y. Fast local learning regularized nonnegative matrix factorization. Adv Comput Environm Sci, 2012, 142: 67–75
33 Gu Q Q, Zhou J. Local learning regularized nonnegative matrix factorization. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, 2009. 1046–1051
34 Cai D, He X F, Wu X Y, et al. Non-negative matrix factorization on manifold. In: Proceedings of IEEE International Conference on Data Mining, Pisa, 2008. 63–72
35 Liu Y L, Du J L, Wang F. Non-negative matrix factorization with sparseness constraints for credit risk assessment. In: Proceedings of IEEE International Conference on Grey Systems and Intelligent Services, Macau, 2013. 211–214
36 Liu C L, Ma J W. Automatic non-negative matrix factorization clustering with competitive sparseness constraints. Intell Comput Methodol, 2014, 8589: 118–125
37 Hoyer P O. Non-negative matrix factorization with sparseness constraints. J Mach Learn Res, 2004, 5: 1457–1469
38 Canadas-Quesada F J, Vera-Candeas P, Ruiz-Reyes N, et al. Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints. Eur J Audio Speech Music Proc, 2014, 2014: 1–17
39 Zhang S, Liu C C, Li W, et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res, 2012, 40: 9379–9391
40 Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics, 2005, 21: 3970–3975
41 Kim H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics, 2007, 23: 1495–1502
42 Peng C, Wong K C, Rockwood A, et al. Multiplicative algorithms for constrained non-negative matrix factorization. In: Proceedings of IEEE International Conference on Data Mining, Brussels, 2012. 1068–1073
43 Cui J, Li F, Wang G Q, et al. Gene-expression signatures can distinguish gastric cancer grades and stages. Plos One, 2011, 6: 1387

44 Frances N, Zeichner S B, Francavilla M, et al. Gastric small-cell carcinoma found on esophagogastroduodenoscopy: a case report and literature review. Case Rep Oncol Med, 2013, 2013: 475961

45 Hu K W, Chen F H. Identification of significant pathways in gastric cancer based on protein-protein interaction networks and cluster analysis. Genet Mol Biol, 2012, 35: 701–708

46 Shimoda T, Matsutani T, Yoshida H, et al. A case of gastric cancer associated with systemic lupus erythematosus and nephrotic syndrome. Nihon Shokakibyo Gakkai Zasshi, 2013, 110: 1797–1803

47 Axon A T. Relationship between Helicobacter pylori gastritis, gastric cancer and gastric acid secretion. Adv Med Sci, 2007, 52: 55–60

48 Lee J, Jung K, Kim Y S, et al. Diosgenin inhibits melanogenesis through the activation of phosphatidylinositol-3-kinase pathway (PI3K) signaling. Life Sci, 2007, 81: 249–254

49 Rappaport N, Nativ N, Stelzer G, et al. MalaCards: an integrated compendium for diseases and their annotation. Datab J Biolog Datab Curat, 2013, 2013: 1429–1438

50 Croft D, O'Kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res, 2011. 39(Database issue): D691–D697

51 Zhao J, Zhou Y W, Zhang X J, et al. Part mutual information for quantifying direct associations in networks. Proc Nat Acad Sci, 2016, 113: 5130–5135

52 Zhang X J, Liu K Q, Liu Z P, et al. NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. Bioinformatics, 2013, 29: 106–113

53 Chen L N, Liu R, Liu Z P, et al. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Sci Rep, 2012, 2: 342

54 Liu R, Wang X D, Aihara K, et al. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. Med Res Rev, 2013, 34: 455–478

55 Liu R, Chen P, Aihara K, et al. Identifying early-warning signals of critical transitions with strong noise by dynamical network markers. Sci Rep, 2015, 5: 17501

56 Zeng T, Zhang C C, Zhang W W, et al. Deciphering early development of complex diseases by progressive module network. Methods, 2014, 67: 334–343

57 Yu X T, Li G J, Chen L N. Prediction and early diagnosis of complex diseases by edge-network. Bioinformatics, 2014, 30: 852–859

58 Yu X T, Zeng T, Wang X D, et al. Unravelling personalized dysfunctional gene network of complex diseases based on differential network model. J Transl Med, 2015, 13: 1–13

59 Zeng T, Wang D C, Wang X D, et al. Prediction of dynamical drug sensitivity and resistance by module network rewiring-analysis based on transcriptional profiling. Drug Resist Update, 2014, 17: 64–76

60 Zeng T, Zhang W W, Yu X T, et al. Big-data-based edge biomarkers: study on dynamical drug sensitivity and resistance in individuals. Brief Bioinform, 2015, 21: 863–874