Front. Comput. Sci.
https://doi.org/10.1007/s11704-016-6287-7

RESEARCH ARTICLE

Identification and prioritization of differentially expressed genes for time-series gene expression data

Linlin XING, Maozu GUO (✉), Xiaoyan LIU, Chunyu WANG

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Abstract Identification of differentially expressed genes (DEGs) in time course studies is very useful for understanding gene function, and can help determine key genes during specific stages of plant development. A few existing methods focus on the detection of DEGs within a single biological group, enabling to study temporal changes in gene expression. To utilize a rapidly increasing amount of single-group time-series expression data, we propose a two-step method that integrates the temporal characteristics of time-series data to obtain a B-spline curve fit. Firstly, a flat gene filter based on the Ljung–Box test is used to filter out flat genes. Then, a B-spline model is used to identify DEGs. For use in biological experiments, these DEGs should be screened, to determine their biological importance. To identify high-confidence promising DEGs for specific biological processes, we propose a novel gene prioritization approach based on the partner evaluation principle. This novel gene prioritization approach utilizes existing co-expression information to rank DEGs that are likely to be involved in a specific biological process/condition. The proposed method is validated on the Arabidopsis thaliana seed germination dataset and on the rice anther development expression dataset.

Keywords time-series gene expression, flat gene filter, gene prioritization, co-expression, differentially expressed genes

Received May 31, 2016; accepted November 3, 2016

E-mail: maozuguo@hit.edu.cn

1 Introduction

Gene microarray analysis is the most popular approach to studying gene expression patterns. Gene expression arrays can be divided into two categories: static design arrays and dynamic design arrays. In general, static designs consider treatment-control pairs for analysis of gene expression patterns under different biological conditions, and do not involve time. On the other hand, dynamic designs are more suitable for longitudinal studies, which address temporal changes in gene expression. Identification of differentially expressed genes (DEGs) is important for understanding the biological mechanisms associated with gene expression. Here, we propose a new method for identification of DEGs in temporal studies of gene expression, within a single biological group.

Many methods have been proposed for detection of DEGs. Fold-change and statistical approaches, such as t-test and ANOVA, have been applied to analysis of gene expression data [1]. The SAM method [2] and the LIMMA package [3] are among the most widely used methods that utilize traditional statistical approaches. Different properties have been considered for improving the performance of statistical methods; for example, ElBakry et al. [4] proposed a novel approach that takes into account the temporal dependence of measurements to improve the statistical power of ANOVA.

Gene expression is a spatial and temporal specific process. Thus, gene expression characteristics have been increasingly analyzed using temporal data. Initially, clustering methods [5–7] were applied to time-series data; however, these methods cannot be used to directly discriminate DEGs. Existing

approaches were extended at different levels, to analyze temporal data. For example, the SAM software allows to treat data at different time points as different experiments, which in turn enables pairwise comparisons. The LIMMA package and other packages were also adapted to treat temporal data. These modifications helped to treat temporal data, but only to a certain extent. Existing approaches ignore the temporal structure of these data, and thus are not suitable for analysis of fast-growing time-series datasets. Bar-Joseph et al. published a review of existing statistical methods for identification of DEGs in time-series expression data [8]. Some new methods for analysis of time-series data have been developed recently. MaSigpro is a two-step regression method for identification of differentially expressed gene profiles in temporal studies; the method consists of the gene selection step and the variable selection step [9]. The EDGE package, proposed by Storey et al., is based on spline curves for identification of DEGs in time-series data [10]. Many other methods have been developed, which use state-of-the-art mathematical techniques [11–14] and address different problems [15–18], such as biological replicates.

However, detection of DEGs in single-group time-series data remains problematic. First, existing methods for detection of DEGs predominantly focus on the pairwise design. Very few methods focus on single-group time-series data. Second, existing methods often require expensive biological replicates. Third, temporal properties of time-series gene expression data are ignored. In this paper, we propose a novel two-step method that combines a flat gene filter and B-spline curve fitting, which is especially suitable for identification of DEGs in single-group time-series data. In addition, the proposed method does not require uniform sampling or biological replicates.

In general, DEGs are sorted according to their p values. This ordering of DEGs depends only on their relative statistical significance and does not reflect their biological significance. Identification of high-confidence DEGs that are promising for specific biological processes, from large (tens of thousands) sets of DEGs identified using these methods, can be both time-consuming and costly. Therefore, a novel approach to gene prioritization is required for reducing the cost of follow-up studies by efficient elimination of irrelevant candidates. Gene prioritization amounts to estimating the likelihood that candidate genes are involved in disease. Different computational methods that combine heterogeneous datasets, such as expression data, sequence information, functional annotation, and biomedical literature, have been developed for this purpose [19]; examples include GeneProspector [20], ToppGene [21], and PROSPECTR [22]. In this paper, gene prioritization refers to finding those genes among all the DEGs that are more important during a specific time window. Because co-expression is associated with co-regulation, information about co-expression can be used for identification and ranking of DEGs. Secondary databases of co-expression have been constructed [23], benefiting from an increasing amount of microarray data. The ATTED database is a pre-calculated co-expression database for plants, especially for Arabidopsis thaliana and rice [24]. In this paper, we address this problem by proposing a novel approach for gene prioritization that utilizes existing co-expression information to rank DEGs according to their importance in specific biological processes/conditions. This method is based on the partner evaluation principle and makes use of the existing co-expression information.

2 Methods

This section describes the details of the proposed method, which is schematically shown in Fig. 1. The microarray data are normalized using the quantile normalization method before starting the identification process. The proposed method is a two-step method: in the first step, a flat gene filter is applied based on the Ljung–Box test; in the second step a B-spline model is fitted, to identify DEGs. Finally, a novel approach to gene prioritization is used, in which the identified DEGs are ranked according to the new ranking strategy based on the partner evaluation principle.

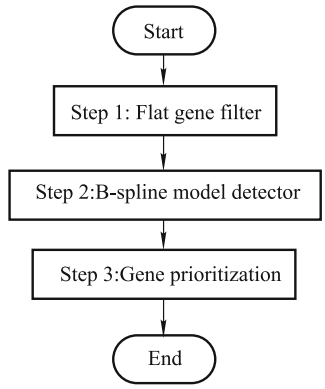


Fig. 1 Flow chart of the proposed method

2.1 Flat gene filter

Based on the temporal and spatial characteristics of gene expression, specific genes are expressed in specific developmental stages/tissues. Moreover, some important genes, which are necessary for ensuring some basic functionality

of organisms, are expressed stably. In the flat gene filter step, a statistical test method is applied to each gene for identification of flat genes. For each gene, the expression data are organized into a short time-series vector. Therefore, the expression patterns of stably expressed and unexpressed genes are similar to white noise sequences. The autocorrelation coefficient captures the correlation of the process with itself at different time points, as shown in Eq. (1). Let k be the time lag for autocorrelation calculations. Further, let $g_i = \{x_0, x_1, \ldots, x_n\}$ be the expression vector of gene i, and let t denote time. Then

$$\rho(k) = \frac{E[(x_t - u)(x_{t+k} - u)]}{\sigma^2}, \tag{1}$$

where $\rho(k)$ is the delay k autocorrelation coefficient, and μ, σ^2 are, respectively, the mean and variance of the expression vector.

Quite generally, for autocorrelation coefficients under 0.3, the sequence is considered to be non-correlated. For autocorrelation coefficients from 0.3 to 0.5, the sequence is weakly correlated. For autocorrelation coefficients from 0.5 to 0.8, the sequence is considered to be correlated. For values above 0.8, the sequence is significantly correlated. The autocorrelation coefficient reflects the correlation at a specific time lag.

Here, the Ljung–Box test was introduced to accomplish the detection of white noise. Instead of testing randomness at each time lag, the Ljung–Box tests the extent of overall randomness based on a number of time lags. The null hypothesis of this test is that the analyzed sequence is completely independent similar to a white noise process.

The null hypothesis and the alternative hypothesis are

$$H0 : \rho(1) = \rho(2) = \cdots = \rho(k) = 0.$$
$$H1 : \exists k \leqslant m, \rho(k) \neq 0.$$

The Q statistic is defined by Eq. (2), as follows:

$$Q_{lb} = n(n+2) \sum_{k=1}^{m} \left(\frac{\rho_k^2}{n-k} \right) \sim \chi^2(m), \tag{2}$$

where n is the sample size, ρ_k is the delay k autocorrelation coefficients, and m is the number of time lags in the test. The Q statistic is described by the chi-square distribution. In single-side hypothesis testing, if $Q_{lb} > \chi^2_{1-\alpha}(m)$, then the null hypothesis is rejected. Rejection of the null hypothesis will imply that the analyzed gene expression vector is not noise. Noise can significantly affect expression data. Therefore, a loose alpha value is chosen to reduce the false detection rate of flat genes. In practice, a sharp peak detection procedure is incorporated into the program to avoid filtering out the genes with only one sharp peak. Simply put, if the extremum of a series is at least twofold stronger or weaker than the average over the rest of the sequence's points, the sequence is concluded to have a sharp peak.

2.2 B-spline model detector

For each gene, its expression data can be described as a vector $g_i = \{x_0, x_1, \ldots, x_n\}$. Then, the process of the detection of DEGs uses the goodness of fit criterion to determine whether the analyzed gene is differentially expressed. Given this set of expression data points $\{x_0, x_1, \ldots, x_n\}$, a degree p, and a number h, where $n > h \geqslant p \geqslant 1$, are determined, and the goal is to find a B-spline curve of degree p that is defined by $h + 1$ control points. This curve must pass through the first and last data points (x_0, x_n), and it should approximate the data polygon in the sense of least squares, as shown in Fig. 2. The B-spline model of degree p can be described using Eq. (3):

$$g(x) = \sum_{i=0}^{h} N_{i,p}(x) P_i. \tag{3}$$

In Eq. (3), g is the gene expression vector. $N_{i,p}(x)$ is a polynomial of degree p, defined in in $[x_i, x_{i+p+1})$. $N_{i,p}$ is the matrix of basis function coefficients. P_i are the control points.

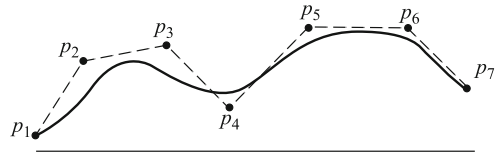


Fig. 2 B-spline model of seven control points

As described above, let $D_0 = P_0, D_n = P_n$ to force the curve to pass through the first and last data points. Then, Eq. (3) can be rewritten as Eq. (4):

$$g(x) = N_{0,p} D_0 + \sum_{i=1}^{h} N_{i,p}(x) P_i + N_{i,p} Dn. \tag{4}$$

As a constraint, parameters n, h, p must satisfy $n > h \geqslant p \geqslant 1$. The knots of the B-spline curve were generated using "average" the parameters method. Actually, the first and last p knots were set to the minimal and maximal parameters, to ensure that the curve passes through the first and last points. Other knots were calculated as the average over contiguous p parameters. For each gene, a global approximation was applied to determine whether it is differentially expressed. In the B-spline model, the degree and the number of control points both affect the shape of the approximating curve. A lower degree or a smaller number of control points in general will

not be able to approximate the data polygon satisfactorily. However, the more control points the curve has, the closer it becomes to the interpolation curve. These parameters affect the generalization ability of the model. We set degree p to be constant across all of the analyzed genes, to avoid inconsistency and overfitting. The approach in this work was different from that in Storey's work [10], in which a fixed cubic spline fitting was used. The degree of the B-spline model is set by the user, with the recommended degree ranging from three to five, for flexibility. In addition, the control points can be set according to the time points, e.g., $n-2 \sim n-3$. Owing to the advantage of the flat gene filter, allowing the degree to be user-defined makes the method more flexible and helps to avoid overfitting to a certain extent.

Based on the B-spline model, we use hypothesis testing to determine whether or not a gene is differentially expressed. At first, the fitting parameters of the null hypothesis and the alternative hypothesis are calculated, and then the fitting residuals of the two models are calculated. The goodness of fit of the model is measured in terms of the residual sum of squares, and the F statistic is constructed to compare the two models. The null hypothesis is that the analyzed gene is not differentially expressed. Obviously, a stably expressed gene will consistently collapse to the mean curve. Therefore, the mean curve is treated as the null hypothesis: $g(u) = \bar{g}$. To calculate the statistical significance, the F statistic is defined by Eq. (5):

$$F = \frac{SSE_0 - SSE_1}{SSE_0}, \tag{5}$$

where SSE_0, SSE_1 are the sums of the fit residuals for the null and alternative hypotheses, respectively. It is clear that the F values for differentially expressed genes will be higher than for other genes. The p-values are calculated using the resampling approach [25]. The fit residuals of the B-spline model are resampled to form a new vector with the mean value. This new vector is used to yield F_i^0 for gene i, as shown in Eq. (6):

$$F_i^0 = \frac{SSE_0 - SSE_1^*}{SSE_0}, \tag{6}$$

where $SSE_1^*$is the sum of the fit residuals for the new data vector. The procedure is iterated M times over all of the analyzed genes, to generate their p-values. In practice, M can be set to 10,000. Thus, the p-value for gene i is calculated from Eq. (7):

$$p_i = \frac{count(F_i^{0k} > F_i)}{M}, k = 1 \cdots M. \tag{7}$$

Up to this point, we have detailed the algorithm using the Ljung–Box test and the B-spline model (LBS) for identification of DEGs. The overall algorithm is listed in the following.

Algorithm LBS

Input: gene expression dataset $D_{\exp}$, gene set $G = \{g_1, g_2, \ldots, g_n\}$, degree d, control points h

Output: DEGs set S_{DEG}

1. Normalize the gene expression dataset $D_{\exp}$
2. $S_{DEG} \leftarrow \emptyset$
3. **FOR** $i = 1 : n$
4. Get expression value vector $g_i = \{x_0, x_1, \ldots, x_n\}$
5. Compute autocorrelation coefficient $\rho(k)$ for each time lag according to Eq. (1)
6. Construct the Q statistic according to Eq. (2)
7. **IF** $(Q_{lb} > \chi^2_{1-\alpha}(m))$
8. flatmark =true;
9. **ELSE**
10. flatmark =false;
11. **END IF**
12. Perform B-spline curve fitting under alternative hypothesis $H1$ according to Eq. (4)
13. Calculate F-Statistics F_i according to Eq. (5)
14. Perform B-spline curve fitting under null hypothesis $H0$
15. Resample the residuals and calculateF_i^0 according to Eq. (6)
16. Calculate p-value p_i according to Eq. (7)
17. **IF** (! flatmark && $p_i < 0.05$)
18. $S_{DEG} \leftarrow S_{DEG} \cup g_i$
19. **END IF**
20. **END FOR**

2.3 Gene prioritization approach using co-expression information based on the partner evaluation principle

DEG identification methods always generate tens of thousands of DEGs, which is not conducive to subsequent experimental analysis. Traditionally, the DEGs are ranked according to some statistics (e.g., p-values or Q-values). However, this order of DEGs depends on their statistical significance only, and does not account for their biological significance. Therefore, identification of high-confidence promising DEGs for specific biological processes can be costly and time-consuming. Prior biological knowledge is helpful for determining the importance of genes, especially their abundant co-expression information. Genes are not independent, and genes with similar functions are expressed in the same biological time-window. Similar genes have similar expression curves. Therefore, inspired by the partner evaluation principle, a novel approach to gene prioritization based on the partner evaluation principle is proposed here, in which co-expression information is used as partner evaluation information. Here, for each gene its partners are its co-expressed genes. In this paper, the ATTED database was selected as a data source for co-expression; this database contains the pre-calculated co-expression data for some plants (e.g., arabidop-

sis and rice). The mutual rank (MR) value is a measure used in the ATTED database, which reflects the strength of an association between two genes. The MR value, as the partner evaluation information, is used to construct the novel prioritization criterion of DEGs, which is termed the partner evaluation score (PES). PES reflects the behavioral consistency of a gene with its partners. If the partners of a certain gene are all differentially expressed, the gene should have higher priority. If the partners of a certain gene are not differentially expressed, the gene may be of lower priority. High-scoring genes may play a more important role in specific gene clusters or during specific developmental stages.

The novel PES score for gene i is given by Eq. (8):

$$PES_i = \frac{\sum_{k=1}^{m} 1/MR_{ik}}{\sum_{j=1}^{100} 1/MR_{ij}}. \tag{8}$$

For gene i, we choose the first 100 co-expressed genes as its partners. The numerator is the reciprocal of the MR value of the 100 partners. Suppose m genes are differentially expressed among the 100 partner genes in this dataset. Then, the denominator is the reciprocal of MR values of its m co-expressed partners.

By performing statistical analysis of co-expression lists of all genes, we find that 100 partners will include all of the genes that considerably affect the target gene, thus reducing the calculation complexity. Hence, only 100 co-expressed genes are considered in PES, to balance the computational complexity and accuracy.

PES reflects the approval of its partners in a specific condition. High-scoring genes are likely to play a more important role. Owing to the existence of different microarray platforms, the ATTED database does not contain all genes. For those genes not in the co-expression table, only DEGs are identified. In the Arabidopsis Thaliana dataset, the microarray platform is GPL198, and the co-expression table covers about 91.4% (20,837/22,810) of genes. In the case of rice, the co-expression table covers about 40% (20,626/57,381) genes for the platform GPL2025. However, most of the differentially expressed genes for the data used in this paper were in the co-expression table.

3 Results and discussion

In this section, we validate the proposed method on two datasets: one is the expression data of the seed germination process of Arabidopsis Thaliana, the other is the data for rice anther development. For the identification of DEGs, we choose the SAM and EDGE methods as reference methods. Mev4.9 [26] was used to complete the SAM method. EDGE is a spline-based method that is applicable to single-group data. All parameters were set to their default values. For the evaluation of PES, GO enrichment analysis is accomplished by agriGO [27]. Then, the genes with high PES scores were further validated in terms of their average degrees and average clustering coefficients in the co-expression network.

3.1 Case I: Seed germination process of Arabidopsis Thaliana

Narsai et al. [28] published this dataset with index number GSE30223 in GEO. Data for ten time-points during germination of Arabidopsis seeds were collected; the goal of the study was to investigate the transcriptome during the germination of Arabidopsis seeds. This dataset provides significant information for in-depth investigations of seeds' germination, which consists of the stages of maturation, stratification, germination, and post-germination of seeds. The method used in [28] was a pairwise comparison of a sample of freshly harvested seeds with samples of previously harvested seeds. In our experiment, we did not analyze the sample of freshly harvested seeds. The other nine time-points were treated as a seed germination procedure, from dry seeds to germens.

In the flat gene filter step, 5,713 genes were filtered out. These genes showed no differential expression patterns. The results of the clustering procedure are shown in Fig. 3. As a result of figure of merit [29], six clusters were detected using the k-means method. The clusters single out the more stable genes, that exhibit very small fluctuations of expression. Compared with the expression cluster results for the identified DEGs in Fig. 4, the difference between the groups is clearly seen.

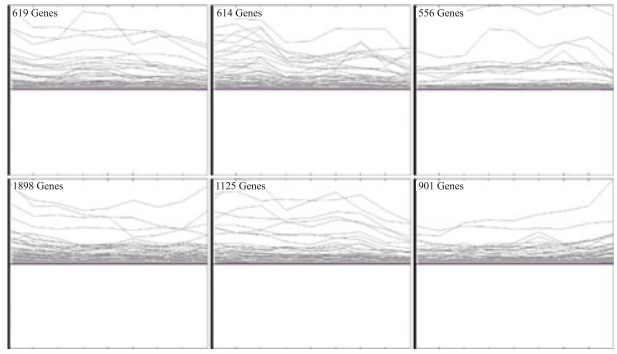


Fig. 3 Cluster analysis of filtered flat genes in Case I (x-axis shows the nine time-points, y-axis shows the corresponding expression values of different genes)

As shown in Table 1, when the degree and the number

of control points are set to four and seven, the number of differentially expressed genes identified using our method is ~15,000. This is inconsistent with conclusions in the original paper [28]. The SAM analysis results were generated for default settings. The first time-point was selected as the base sample. The second and the last time-points were compared with the base sample. The EDGE software was set to the time-course configuration. The two results in the table were generated with parameters set to 3 and 7. Nevertheless, for Q = 0.05, nearly all of the genes (21,168) were identified as differentially expressed.

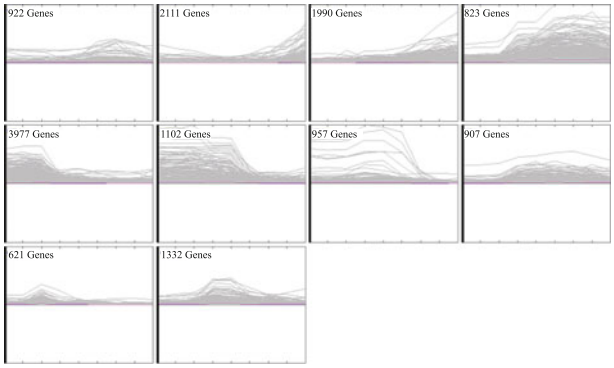


Fig. 4 Cluster analysis of DEGs in Case I (*x*-axis shows the nine time-points, *y*-axis shows the corresponding expression values of different genes)

Table 1 Comparison of the number of identified DEGs obtained using three methods in Case I

	LBS[a]	EDGE		SAM	
Parameter	degree=4 c.p.[b] =7	degree=7	degree=3	0vs2	0vs9
Threshold	p=0.05	Q=8.2e–7	Q=5.5e–10		
Gene No.	14,963	12,251	10,357	14	16

Note: [a]LBS: the proposed method that uses the Ljung–Box test and the B-spline model. [b]c.p.: the number of control points in the B-spline model

As a result of the figure of merit, ten clusters were detected using the k-means method applied to the differentially expressed genes. Figure 4 visualizes the clusters. All of the four patterns described in [28] are captured by our method, without the need for control design and pairwise comparison of each time-point.

To identify high-confidence candidate genes, the DEGs were then ranked using the gene prioritization method based on the PES score. Table 2 shows the number of DEGs for different PES scores. Figure 5 shows the trends of average degree and average cluster coefficient, for genes with different scores. As shown by Table 2, the new ranking strategy exhibits good filterability. Moreover, when the PES continues to rise, the average degree and average clustering coefficient of the selected genes gradually increase, as shown in Fig. 5. Conversely, in a random selection process, the average degree and average clustering coefficient do not increase with increasing the score. A GO enrichment analysis of a gene set with score >= 0.9 was applied. The GO enrichment results for Biological Process showed that genes are enriched at "embryonic development ending in seed dormancy", "translation", "post-embryonic development", and "nitrogen compound metabolic process" in the seed germination process, which is very consistent with the conclusion of the original paper.

Table 2 The number of DEGs for different PES scores in Case I

PES	–	> 0.1	> 0.2	> 0.3	> 0.4
Gene No.	14,841	13,824	13,570	13,323	13,032
PES	> 0.5	> 0.6	> 0.7	> 0.8	> 0.9
Gene No.	12,696	12,244	11,537	10,181	7,188

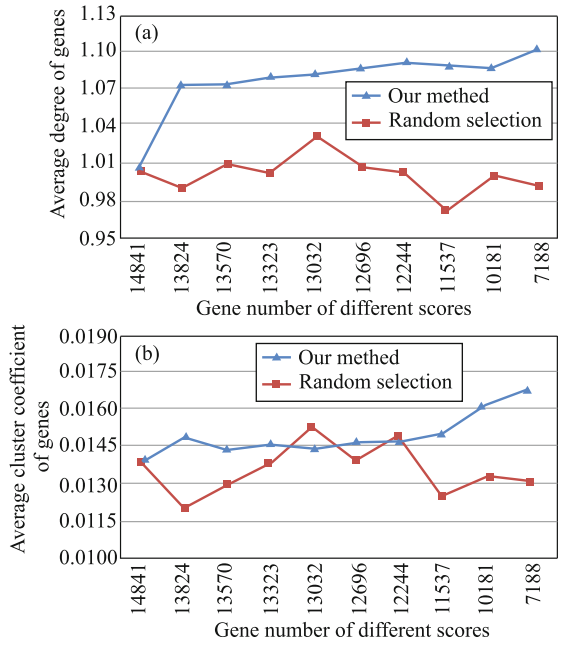


Fig. 5 (a) Average degree and (b) average cluster coefficient of genes corresponding to different scores

As expected, the results of the ranking procedure show that the high-scoring genes play an important role in the co-expression network. The increasing degree illustrates that the genes with higher PES scores may be the hubs of the gene network. These play a more important role in the seed germination stage.

3.2 Case II: Rice anther development

This dataset numbered GSE13988 contains information about anther development of rice [30]. The original data can be downloaded from the GEO database. The experimental

data covers the time from the formation of hypodermal archesporial cells to tri-cellular mature pollens. The entire data are divided into eight stages (eight time-points). Each time-point contains three or four biological replicates. First, all the data were processed by the flat gene filter. In this dataset, 17,132 of the 57,381 probes were filtered out. These genes were considered as unexpressed/ constant expressed genes. The gene filter results are shown in Fig. 6. Cluster analysis reveals that the majority of genes are not differentially expressed. Their expression curves are flat or fluctuate somewhat. The result shows that our filtering method is able to filter out flat genes.

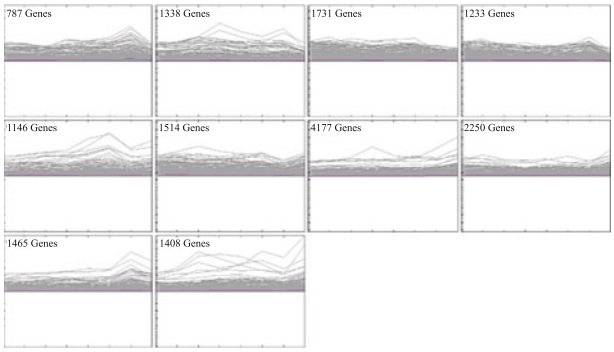


Fig. 6 Cluster analysis of filtered flat genes in Case II (*x*-axis shows the eight time-points, *y*-axis shows the corresponding expression values of different genes)

Differentially expressed genes identified by the B-spline model are shown in Table 3. In this table, the EDGE result and the SAM result are also included for comparisons. The second column in Table 3 shows the differentially expressed genes identified by our method. In addition, the results obtained using the comparison methods are listed respectively in columns 3 and 4. The Q-value threshold is the default for the EDGE software. Nevertheless, when the Q-value threshold was set to an empirical value of 0.05, the EDGE software generated a meaningless result according to which all of the analyzed genes were identified as differentially expressed. Clearly, the proposed method is able to detect more differentially expressed genes than other methods. The corresponding clustering results for DEGs are shown in Fig. 7. Ten clusters were detected using the k-means method. Obviously, the differential expression patterns of the DEGs in the cluster diagram show that these genes are truly differentially expressed during the anther development stage. In addition, an advantage of the Ljung–Box filter is its robustness with respect to different thresholds, and this method ensures that most of the false positive results are filtered out.

In the GO analysis, as expected, we found that gene expression is enriched in the nucleus, various vesicles, the membrane system, the chromosome, and the plasma membrane, as shown in Fig. 8. This is consistent with the anther development process in which cells undergo division and differentiation. All of the GO enrichment results can be recall using our method and the agriGO website.

Table 3 Comparison of the number of identified DEGs obtained using three methods in Case II

	LBS[a]	EDGE		SAM	
Parameter	degree=3 c.p.[b] =6	degree =7	degree =3	0vs2	0vs9
Threshold	p=0.05	Q=7.78–7	Q=2.8e–10		
Gene No.	11,249	4,061	7,958	39	959

Note: [a]LBS: the proposed method that uses the Ljung–Box test and the B-spline model. [b]c.p.: the number of control points in the B-spline model

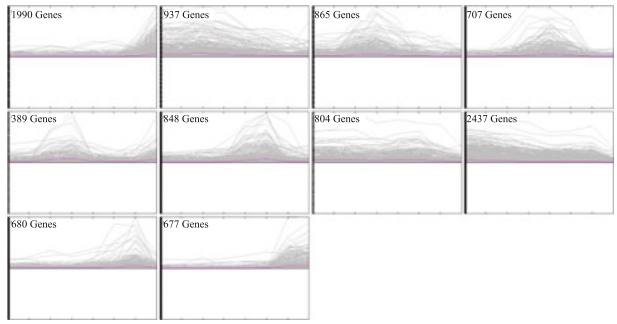


Fig. 7 Cluster analysis of DEGs in Case II (*x*-axis shows the eight time-points, *y*-axis shows the corresponding expression values of different genes)

As mentioned earlier, the co-expression table does not cover all the genes in the microarray dataset. Here, 76.9% (8,646 out of 11,249) of the DEGs were covered. The proportion is meaningful for a proof-of-concept test. We used different scores to filter the genes in the ranking list, as shown in Table 4.

Table 4 The number of DEGs for different PES scores in Case II

PES	–	> 0.1	> 0.2	> 0.3	> 0.4
Gene No.	8,646	8,150	7,588	6,830	5,878
PES	> 0.5	> 0.6	> 0.7	> 0.8	> 0.9
Gene No.	4,834	3,727	2,605	1,639	675

We analyzed the genes with different scores, and found that the average degree and the average cluster coefficient increase with increasing the PES score, as shown in Fig. 9. As the PES score continues to increase, non-important genes are gradually filtered out. The retained genes exert a much stronger influence in this biological process. This method allows to reduce the computational complexity associated with finding key genes and reconstructing gene regulatory networks. Conversely, in a random selection process, the average degree and

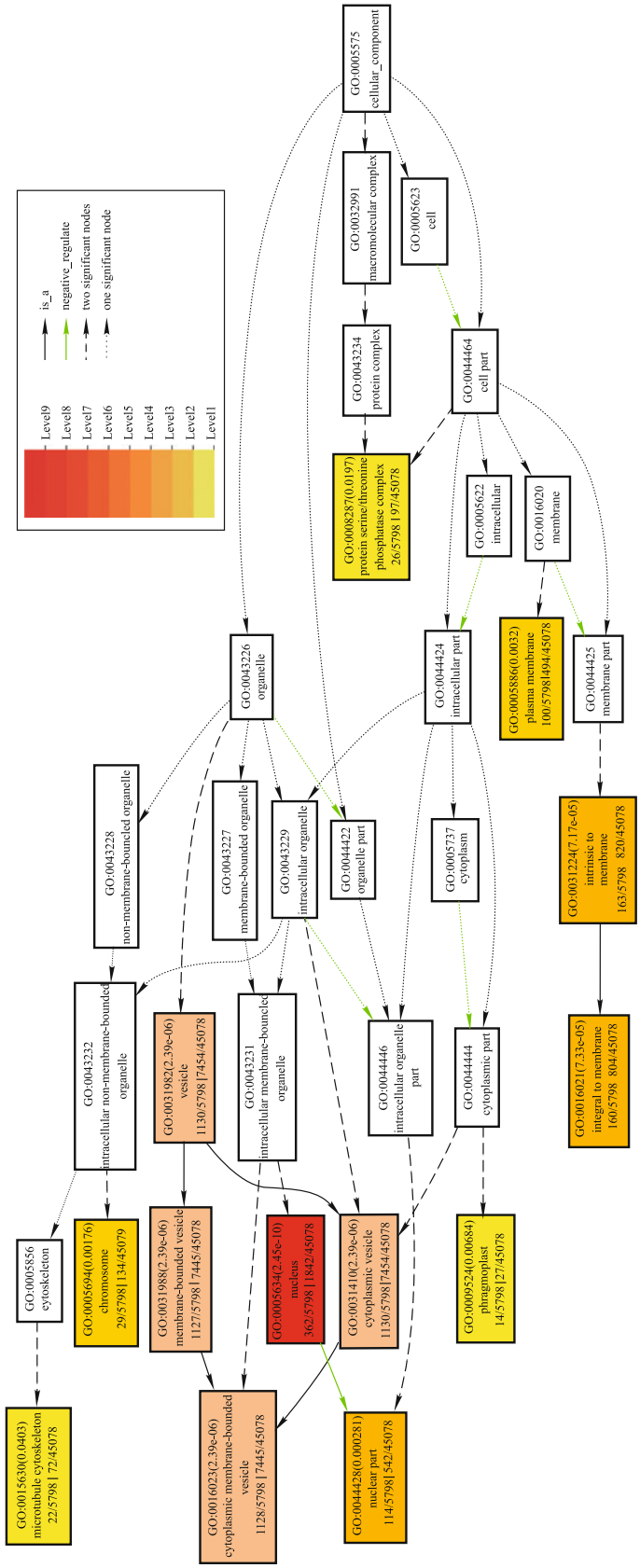


Fig. 8 GO enrichment analysis of in subcategory of cell components

the average cluster coefficient of the genes do not increase concurrently.

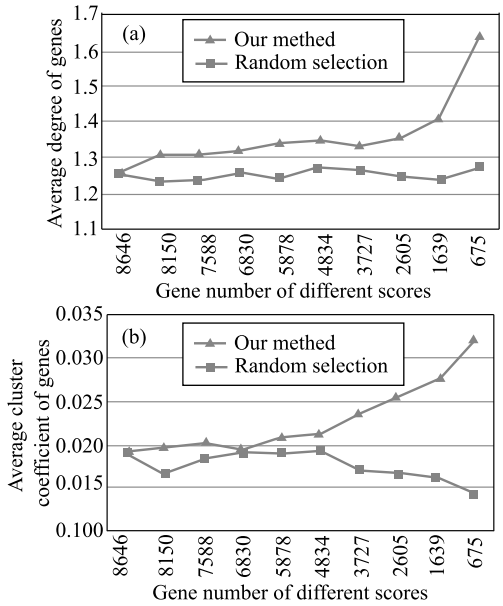


Fig. 9 (a) Average degree and (b) average cluster coefficient of genes corresponding to different scores

As can be seen from the above tables, our ranking strategy is consistent with the biological facts that if a gene's partners are differentially expressed, this gene plays a more important role in its network. Our ranking score can indicate the importance of specific genes in specific biological processes. In conclusion, the results of our studies show that the proposed prioritization approach is effective.

4 Conclusion

In this paper, we proposed a novel combined method for identification of DEGs in single-group time-series datasets and a novel approach to gene prioritization, for re-ranking the identified DEGs according to their biological importance. The DEGs identification method is a two-step method that considers the temporal characteristics of expression and does not rely on replicates. The first step is filtering out flat genes according to the characteristics of time series. Then, a parameterized B-spline model is applied to identify statistically significant genes. The gene prioritization approach based on the partner evaluation principle uses the co-expression information as the evaluation index to re-rank the DEGs that are likely related to specific biological processes/conditions.

Using the flat gene filter with the B-spline method combines the advantages of both methods, and makes up some defects. First, the flat gene filter makes full use of temporal characteristics of gene expression and limits the excessive flexibility of the spline method. For each gene, the expression vector is actually a time sequence and is very short, for cost reasons. Currently, statistical methods, such as the ARMA model, are not applicable for very short time series. Here, we introduced the Ljung–Box test, to determine whether the gene vector is white noise or not. In addition, at the same time, a loose alpha threshold and a sharp peak detection step were incorporated into the testing procedure to make sure no DEGs are missing. Second, the B-spline model brings sufficient capability to the method. The B-spline model is defined by the degree of the basis and the number of control points. The two parameters of the B-spline model can be adjusted to fit different data sets and models with different number of degrees of freedom. Different configurations may yield different results. However, using the flat gene filter ensures that truly flat genes are not selected. As a result, there are only a few false positives and misses.

After the identification of DEGs, they are re-ranked according to the PES score based on the partner evaluation principle. Inspired by the partner evaluation principle, we propose this gene prioritization approach for ranking DEGs that would be promising for follow-up studies. The co-expression information extracted from the ATTED database is considered for evaluation of partners. By performing statistical analysis, a suitable number of partners is selected to balance the accuracy and the computational complexity.

The results presented in this paper suggest that our method is effective for identification of DEGs in single-group time-series expression data. Moreover, the proposed approach for gene prioritization yields meaningful ranking for understanding the analyzed and ranked DEGs. The usage of the flat gene filter and gene prioritization approach based on the partner evaluation principle can be easily extended to other domains. In addition, in the future, this procedure can be extended to analyze experiments with balance controls and cover a larger number of genes in datasets. The proposed method based on the gene expression matrix can be easily adapted for processing RNA-seq data.

Acknowledgements This paper was supported by the National Natural Science Foundation of China (Grant Nos. 61271346, 61571163, 61532014, 91335112, 61671189 and 61402132).

References

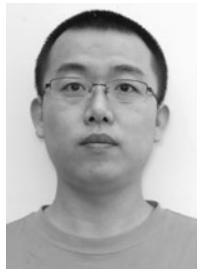
1. Dudoit S, Yang Y H, Callow M J, Speed T P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 2002, 12(1): 111–139

2. Tusher V G, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(9): 5116–5121
3. Smyth G K. Limma: linear models for microarray data. In: Gentleman R, Carey V J, Huber W, et al, eds. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. New York: Springer, 2005, 397–420
4. ElBakry O, Ahmad M O, Swamy M N. Identification of differentially expressed genes for time-course microarray data based on modified RM ANOVA. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(2): 451–466
5. Bar-Joseph Z. Analyzing time series gene expression data. Bioinformatics, 2004, 20(16): 2493–2503
6. Ernst J, Nau G J, Bar-Joseph Z. Clustering short time series gene expression data. Bioinformatics, 2005, 21(suppl_1): 159–168
7. Chaiboonchoe A, Samarasinghe S, Kulasiri G D. Using emergent clustering methods to analyse short time series gene expression data from childhood leukemia treated with glucocorticoids. In: Proceedings of the 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. 2009, 741–747
8. Bar-Joseph Z, Gerber G, Simon L, Gifford D K, Jaakkola T S. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proceedings of the National Academy of Sciences of the United States of America, 2003, 100(18): 10146–10151
9. Conesa A, Nueda M J, Ferrer A, Talon M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. Bioinformatics, 2006, 22(9): 1096–1102
10. Storey J D, Xiao W Z, Leek J T, Tompkins R G, Davis R W. Significance analysis of time course microarray experiments. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(36): 12837–12842
11. Kim J, Ogden R, Kim H. A method to identify differential expression profiles of time-course gene data with Fourier transformation. BMC Bioinformatics, 2013, 14(1): 310
12. Han X U, Sung W-K, Feng L I N. Identifying differentially expressed genes in time-course microarray experiment without replicate. Journal of Bioinformatics and Computational Biology, 2007, 5(02a): 281–296
13. Angelini C, Cutillo L, De Canditiis D, Mutarelli M, Pensky M. BATS: a Bayesian user-friendly software for analyzing time series microarray experiments. BMC Bioinformatics, 2008, 9: 415
14. Wu S, Wu H L. More powerful significant testing for time course gene expression data using functional principal component analysis approaches. BMC Bioinformatics, 2013, 14(1): 6
15. Yang E W, Girke T, Jiang T. Differential gene expression analysis using coexpression and RNA-Seq data. Bioinformatics, 2013, 29(17): 2153–2161
16. Pan J B, Hu S C, Wang H, Zou Q, Ji Z L. PaGeFinder: quantitative identification of spatiotemporal pattern genes. Bioinformatics, 2012, 28(11): 1544–1545
17. Xiao S J, Zhang C, Zou Q, Ji Z L. TiSGeD: a database for tissue-specific genes. Bioinformatics, 2010, 26(9): 1273–1275
18. Pan J B, Hu S C, Shi D, Cai M C, Li Y B, Zou Q, Ji Z L. PaGenBase: a pattern gene database for the global and dynamic understanding of gene function. PloS One, 2013, 8(12): E80747
19. Moreau Y, Tranchevent L C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics, 2012, 13(8): 523–536
20. Yu W, Wulf A, Liu T B, Khoury M J, Gwinn M. Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. BMC Bioinformatics, 2008, 9(1): 528
21. Chen J, Bardes E E, Aronow B J, Jegga A G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Research, 2009, 37(suppl_2): W305–W311
22. Adie E A, Adams R R, Evans K L, Porteous D J, Pickard B S. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics, 2005, 6(1): 55
23. Usadel B, Obayashi T, Mutwil M, Giorgi F M, Bassel G W, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart N J. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ, 2009, 32(12): 1633–1651
24. Obayashi T, Okamura Y, Ito S, Tadaka S, Aoki Y, Shirota M, Kinoshita K. ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. Plant and Cell Physiology, 2014, 55(1): e6
25. Storey J D, Tibshirani R. Statistical significance for genome wide studies. Proceedings of the National Academy of Sciences of the United States of America, 2003, 100(16): 9440–9445
26. Howe E, Holton K, Nair S, Schlauch D, Sinha R, Quackenbush J. MeV: multiexperiment viewer. In: Ochs M F, Casagrande J T, Davuluri R V, eds. Biomedical Informatics for Cancer Research. Springer US, 2010, 267–277
27. Du Z, Zhou X, Ling Y, Zhang Z H, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Research, 2010, 38(suppl_2): W64–W70
28. Narsai R, Law S R, Carrie C, Xu L, Whelan J. In-depth temporal transcriptome profiling reveals a crucial developmental switch with roles for RNA processing and organelle metabolism that are essential for germination in Arabidopsis. Plant Physiology, 2011, 157(3): 1342–1362
29. Yeung K Y, Haynor D R, Ruzzo W L. Validating clustering for gene expression data. Bioinformatics, 2001, 17(4): 309–318
30. Fujita M, Horiuchi Y, Ueda Y, Mizuta Y, Kubo T, Yano K, Yamaki S, Tsuda K, Nagata T, Niihama M, Kato H, Kikuchi S, Hamada K, Mochizuki T, Ishimizu T, Iwai H, Tsutsumi N, Kurata N. Rice expression atlas in reproductive development. Plant and Cell Physiology, 2010, 51(12): 2060–2081

Linlin Xing received his MS degree in computer science form Harbin Institute of Technology (HIT), China in 2012. He is currently a PhD candidate under the supervision of Professor Maozu Guo in the School of Computer Science and Technology, HIT. His research interests include gene expression data analysis and biological network construction.

Maozu Guo received his BS and MS degrees from Harbin Engineering University, China in 1988 and 1991 respectively, and PhD degree from Harbin Institute of Technology (HIT), China in 1998, all in computer science. He is currently a professor in the School of Computer Science and Technology, HIT. His research interests include Bioinformatics and machine learning.

Xiaoyan Liu received her BS and MS degrees in computer science from Harbin Engineering University, China, and PhD degree in Engineering Mechanics from Harbin Institute of Technology (HIT), China. She is currently an associate professor in School of Computer Science and Technology at HIT. Her research interests include Bioinformatics and knowledge-based systems.



Chunyu Wang received his BS, MS and PhD degrees in computer science from Harbin Institute of Technology (HIT), China. Now he is an associate professor in computer science and technology at HIT. His research interests include bioinformatics and machine learning.