Mathematical Biosciences 281 (2016) 74–81

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Analyzing gene expression time-courses based on multi-resolution shape mixture model

Ying Li[a,b], Ye He[a,b], Yu Zhang[a,b,*]

[a] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[b] Key Laboratory of Symbolic Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun 130012, China

ARTICLE INFO

Article history:
Received 16 April 2016
Revised 23 August 2016
Accepted 31 August 2016
Available online 10 September 2016

Keywords:
Bayesian information criterion
Global fractal scale
Local fractal scale
Mixture model clustering
Multi-resolution fractal feature

ABSTRACT

Objective: Biological processes actually are a dynamic molecular process over time. Time course gene expression experiments provide opportunities to explore patterns of gene expression change over a time and understand the dynamic behavior of gene expression, which is crucial for study on development and progression of biology and disease. Analysis of the gene expression time-course profiles has not been fully exploited so far. It is still a challenge problem. We propose a novel shape-based mixture model clustering method for gene expression time-course profiles to explore the significant gene groups.

Results: Based on multi-resolution fractal features and mixture clustering model, we proposed a multi-resolution shape mixture model algorithm. Multi-resolution fractal features is computed by wavelet decomposition, which explore patterns of change over time of gene expression at different resolution. Our proposed multi-resolution shape mixture model algorithm is a probabilistic framework which offers a more natural and robust way of clustering time-course gene expression. We assessed the performance of our proposed algorithm using yeast time-course gene expression profiles compared with several popular clustering methods for gene expression profiles. The grouped genes identified by different methods are evaluated by enrichment analysis of biological pathways and known protein–protein interactions from experiment evidence. The grouped genes identified by our proposed algorithm have more strong biological significance.

Conclusion: A novel multi-resolution shape mixture model algorithm based on multi-resolution fractal features is proposed. Our proposed model provides a novel horizons and an alternative tool for visualization and analysis of time-course gene expression profiles.

Availability: The R and Matlab program is available upon the request.

1. Introduction

The real biological systems are complex and dynamic. Recently, microarray experiments emerge in large numbers to evaluate changes in gene expression over time. The gene expression time-course profiles provide opportunities to understand and explore the complex dynamic mechanisms of gene regulation, cell-cycle, cell development, external stimuli (e.g., drugs and stress) and disease progression.

The one of main difficulties for analysis of gene expression is the affection of high noise and the feature of small sample and high dimension.

Some popular clustering methods such as *k*-means clustering [1], self-organizing maps (SOM) [2], hierarchical clustering [3], mixture clustering model [4] affinity propagation-based clustering (APcluster) [5,6] are all applied to time-course gene expression. In addition, qualitative bicluster algorithm (QUBIC) [7] is also a better choice for time-course gene expression. For analysis of gene expression time-course profiles, the feature of shape change is an important issue. The methods for gene expression analysis from the point of view on shape-based similarity measure have been studied. Wen et al. [8] proposed a shape-based similarity measure, which compares two gene expression profiles based on the qualitative changes of expression values. Thus, two profiles are considered as similar if they increase and decrease more or less simultaneously. However, in this measure all time points are taken into consideration. The main drawback of the method is a risk of missing interesting (local) relationships [8]. Kwon et al. [9] suggested an 'event-based' edge detection method. The raw gene expression

* Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun, China.
E-mail address: zy26@jlu.edu.cn (Y. Zhang).

http://dx.doi.org/10.1016/j.mbs.2016.08.012
0025-5564/

profiles are converted to a string of events, and the event strings are aligned by the Needleman–Wunsch algorithm. An event in a specific time interval is considered as the directional change of the gene expression curve at that instant. Filkov et al. [10] proposed an 'edge detection' method for periodic data sets with small sequences. Futschik and Carlisle [11] designed a fuzzy *c*-means algorithm for clustering of time-course gene expression and a R package termed Mfuzz is provided [12]. Willbrand et al. [13] applied up-down analysis to microarray times series. A local shape-based similarity measure based on Spearman rank correlation was introduced [14]. A novel alignment method based on hidden Markov models (HMMs) was proposed to analyze the time-course gene expression [15]. Chiu et al. [16] proposed an affinity propagation-based clustering algorithm for time-series gene expression data, where a sliding-window mechanism was applied to extract a large number of features to explore the relationship between genes. However these methods tend to oversimplify the original gene expression data, which further loses a lot of information contained in the original time series.

In this paper, we propose a novel multi-resolution shape mixture model algorithm based on multi-resolution fractal features to cluster the time-course gene expression profiles. The multi-resolution fractal feature is a more exact description of change degree of signal from global to local range at different multi-resolutions. The multi-resolution fractal feature indicates the meaningful shape information of signal, which is very beneficial for signal classification and identification. Firstly, the multi-resolution fractal features of time-course gene expression profiles are extracted. Then the mixture model cluster algorithm is applied to genes based on the multi-resolution fractal features. Mixture model cluster algorithm [17] belongs to clustering based probability, where the number of clusters can be well inferred by Bayesian information criterion [18]. Compared with several popular clustering methods including *k*-means, mixture clustering model, APcluster, bicluster and Mfuzz, we evaluated the performance of our proposed algorithm using yeast time-course gene expression profiles methods for gene expression profiles. The grouped genes identified by different methods are validated by enrichment analysis of biological pathways and known protein–protein interactions from experiment evidence. The results shows that our proposed algorithm is effective and can explore the gene set with strong biological significance.

1.1. Multi-resolution fractal feature extraction of time series gene expression profiles

In this section, we give an algorithm to extract multi-resolution fractal feature of time series gene expression profile. For simplification the time series gene expression profile is also called as biological signal.

1.2. Decomposition with Mallat algorithm

Suppose $\varphi(x)$ is the scaling function, $\psi(x)$ is the wavelet function with compactly support. Let $\varphi_{j,k} := 2^{j/2}\varphi(2^j x - k)$, and $\psi_{j,k} := 2^{j/2}\psi(2^j x - k)$. For any continuous signal $f(x)$, the corresponding scaling coefficient c_k^j and wavelet coefficient d_k^j can be computed as follows:

$$c_k^j := \langle f, \varphi_{j,k}\rangle, d_k^j := \langle f, \psi_{j,k}\rangle, k \in Z.$$

For the sample signal, it is easy to extend the number of sample to be 2^N, which usually includes period extend or zero extend. In fact, $\{c_l^N\}_{l=0}^{2^N-1}$ is always taken as the 1-D original sample signal. Suppose$\{h_l\}_l$ and $\{g_l\}_l$ are finite low-pass and high-pass filters. Decompose the signal completely with Mallat algorithm in the following [19]:

$$\begin{cases} c_k^{j-1} &= \sum_{l\in Z} h_{l-2k} c_l^j, \\ d_k^{j-1} &= \sum_{l\in Z} g_{l-2k} c_l^j . j = N, N-1, \ldots, 1, k \in Z. \end{cases} \quad (1)$$

And if $\{h_l\}_l$ and $\{g_l\}_l$ are orthogonal, there is $g_l = (-1)^l h_{1-l}$. While in the bio-orthogonal condition there are four filters (two group filters): decomposition filters $\{h_l\}_l$, $\{g_l\}_l$, reconstruction filters $\{\widetilde{h}_l\}_l$, $\{\widetilde{g}_l\}_l$, where $g_l = (-1)^l \widetilde{h}_{1-l}$, and $\widetilde{g}_l = (-1)^l h_{1-l}$.

1.3. Computation of global fractal scale

The fractal scales are expressed in wavelet transform [20]. If the signal $f \in L^2(R)$ is bounded and phase continuous and there exists certain α to make the wavelet transformation of f satisfy

$$|\langle f, \psi_{j,k}\rangle| \le c2^{-j(\alpha+\frac{1}{2})}, j = N-1, \ldots, N-M, k \in Z,$$

where $c > 0$ is a constant, M is the level number of decomposition, N is the initial level number. Then the fractal scale of f is α.

Considering the high frequency part $\{d_k^j\}_{k,j}$, to get the global fractal scale is to solve the maximum α and minimum c, which satisfy the below inequations:

$$|d_k^j| \le c2^{-j\alpha}, j = N-1, N-2, \ldots, N-M, k \in Z.$$

For the discrete initial signal $\{c_l^N, l = 0, \ldots, 2^N - 1\}$, we need to get the maximal high frequency signals in each level: $d_j^* = \max_k |d_k^j|$, where $d_j^* > 0$.

The problem becomes to solve c and α to satisfy

$$d_j^* \le c2^{-j\alpha}, j = N-1, N-2, \ldots, N-M.$$

Suppose $b_j^* = \log_2 d_j^*, b = \log_2 c$, and $\beta_j = b - j\alpha - b_j^*$, then using the least square estimation we can get α and b to minimize $\sum_j \beta_j^2$. To have enough data for least square estimation and for the stability of algorithm, only $M > 2$ is utilized in our implementation.

The fractal scale in the Mth level is $\alpha - \frac{1}{2}$. The global fractal scale in the Mth level is expressed as $\alpha_{M,\,global}$.

1.4. Computation of local fractal scale

After original signals are decomposed completely, we get high frequency signals of the former M levels and compute the local fractal scale in the Mth level.

The high frequency signals of the former M levels are

$$\begin{cases} d_k^{N-1}, & k = 0, 1, \ldots, 2^{N-1} - 1, \\ d_k^{N-2}, & k = 0, 1, \ldots, 2^{N-2} - 1, \\ \ldots, & \ldots, \\ d_k^j, & k = 0, 1, \ldots, 2^j - 1, \\ \ldots, & \ldots, \\ d_k^{N-M}, & k = 0, 1, \ldots, 2^{N-M} - 1. \end{cases} \quad (2)$$

In the Mth level, the number of the local fractal scale is 2^{N-M}. And the sth local fractal scale can be computed by using the following wavelet coefficients in Fig. 1.

$$\begin{cases} d_k^{N-M}, & k = s, \\ d_k^{N-M+1}, & k = 2s, 2s+1, \\ \ldots, & \ldots \\ d_k^j, & k = 2^{j-N+M}s, 2^{j-N+M}s + 1, \\ & \ldots, 2^{j-N+M}s + 2^{j-N+M} - 1, \\ \ldots, & \ldots \\ d_k^{N-1}, & k = 2^{M-1}s, 2^{M-1}s + 1, \\ & \ldots, 2^{M-1}s + 2^{M-1} - 1. \end{cases} \quad (3)$$

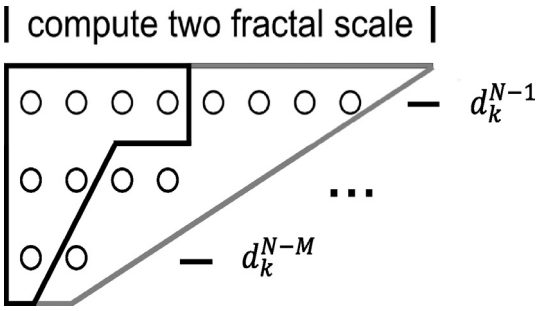


Fig. 1. Wavelet coefficients: the bottom row is d_k^{N-M} , the top row is d_k^{N-1}. From this partition, we can get two groups of high frequency signals corresponding to two blocks and so obtain two local fractal scale features.

$$\begin{aligned} &\left|d_k^j\right| \le c2^{-j\alpha}, \\ &j = N-1, \ldots, N-M, \\ &k = 2^{j-N+M}s, \ldots, 2^{j-N+M}s + 2^{j-N+M} - 1. \end{aligned} \tag{4}$$

The computation method is similar to the global fractal scale. The difference is the field of k becomes smaller. The sth local fractal scale in Mth is expressed as $\alpha^s_{M,local}$ $(M > 2)$.

1.5. Multi-resolution shape mixture model clustering

In this section, using the multi-resolution fractal features, we group the gene expression profiles based on the mixture model method [17,21–23]. In mixture models, each component probability distribution corresponds to a cluster. The number of clusters is usually treated as an external parameter determined by the Bayesian information criterion (BIC) [18].

Given data y with independent multivariate observations $y_1, y_2, \ldots, y_n$, the likelihood for a Gaussian mixture model with G components is defined as:

$$L(\theta_1, \ldots, \theta_n, \tau_1, \ldots, \tau_n) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(y_i|\theta_k),$$

where f_k and θ_k are the density and parameters of the kth component in the mixture and τ_k is the probability that an observation belongs to the kth component. τ_k satisfies

$$\sum_{k=1}^{G} \tau_k = 1, \tau_k \ge 0.$$

At most cases, f_k is the multivariate normal (Gaussian) density ϕ_k with parameterized by its mean μ_k and covariance matrix Σ_k as follows:

$$\phi_k(y_i|\mu_k, \Sigma_k) = \frac{\exp\{\frac{1}{2}(y_i - \mu_k)^T \Sigma_k^{-1}(y_i - \mu_k)\}}{\sqrt{\det(2\Pi\Sigma_k)}}.$$

Data regarded as generated by mixtures of multivariate normal densities are characterized by groups or clusters centered at the means μ_k, with increased density for points nearer the mean.

The components are ellipsoidal, centered at the means μ_k. The covariance matrix Σ_k determines their other geometric features, which include shape, volume and orientation. Each covariance matrix is parameterized by eigenvalue decomposition as follows:

$$\Sigma_k = \lambda_k D_k A_k D_k^T,$$

where D_k is the orthogonal matrix of eigenvectors, A_k is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_k, and λ_k is a scalar. The orientation of the principal components of Σ_k is determined by D_k, while A_k determines the shape of the density contours; λ_k specifies the volume of the corresponding ellipsoid. The detailed description of Σ_k can be seen in the Table 1.

Table 1
Description of covariance matrix Σ_k.

	Method	Distribution	Volume	Shape	Orientation
E		Univariate	Equal		
V		Univariate	Variate		
EII	λI	Spherical	Equal	Equal	
VII	$\lambda_k I$	Spherical	Variate	Equal	
EEI	λA	Diagonal	Equal	Equal	
VEI	$\lambda_k A$	Diagonal	Variate	Equal	
EVI	λA_k	Diagonal	Equal	Variate	
VVI	$\lambda_k A_k$	Diagonal	Variate	Variate	
EEE	λDAD^T	Ellipsoidal	Equal	Equal	Equal
EEV	$\lambda D_k A D_k^T$	Ellipsoidal	Equal	Equal	Variate
VEV	$\lambda_k D_k A D_k^T$	Ellipsoidal	Variate	Equal	Variate
VVV	$\lambda_k D_k A_k D_k^T$	Ellipsoidal	Variate	Variate	Variate

The EM algorithm [24,25] is a general approach to maximum likelihood estimation. The Bayesian information criterion (BIC) is performed well for model selection, which adds a penalty to the log likelihood based on the number of parameters. The BIC is defined as follows:

$$\mathrm{BIC} = 2L_M(y, \hat{\theta}) - m_M \log(n),$$

where $L_M(y, \hat{\theta})$ is the maximized log likelihood for the model M and data y, m_M is the number of independent parameters to be estimated in the model M, and n is the number of observations in the data.

For different fractal features, the different mixture model and parameters are selected. For global fractal features, the univariate mixture model cluster are applied, which only include two possible models – equal variance (denoted E) or varying variance (denoted V). For local fractal features, the multivariate mixture model clusters are applied, which include considerable more parameters selection.

2. Results

2.1. Data collection and preprocessing

The gene expression data used in this paper is an expression profile of 6178 genes for 77 time sample points of the yeast cell cycle data [26].

The preprocessing is also needed, including a natural logarithm transformation, removing the genes with more than 15% missing values, imputing the missing values using KNN algorithm [17] and removing the genes with small variance. The number of the selected gene is 533.

2.2. Multi-resolution fractal feature of gene expression

After the normalization, such time series gene expression profiles are decomposed by orthogonal wavelets and the corresponding multi-resolution fractal feature are extracted.

To show excellent ability to capture shape characteristics of the multi-resolution fractal features, the original expression profiles and multi-resolution fractal features at different scales are plotted and compared in Fig. 2 for genes with different similarities.

The global fractal features of the gene YAL003W and gene YAL064C are computed as 1.01148 and 0.21087 respectively. From (a) and (b) in Fig. 2, the original expression profiles and the local fractal features of the gene YAL003W and the gene YAL064C are totally different.

The gene expression profiles of YAL040C and YAL005C have the similar shape changes. The global fractal features of YAL040C and YAL005C are 0.2378 and 0.23142. From (c) and (d) in Fig. 2,

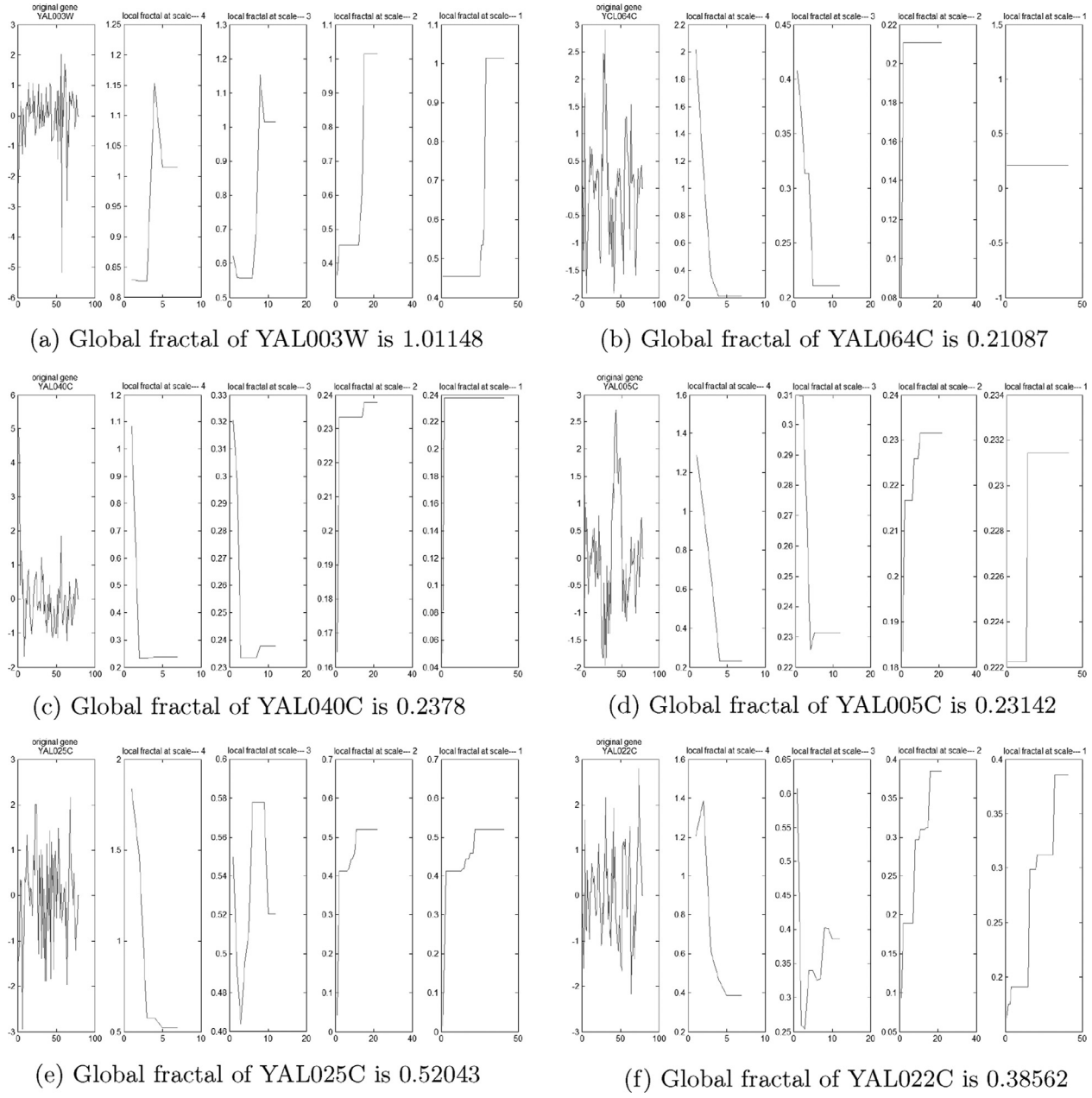


Fig. 2. Original gene expression profiles and multi-resolution fractal features.

they have the similar global and local fractal features. Furthermore, the multi-resolution local fractal features have subtal difference.

The global fractal features of YAL025C and YAL022C are computed as 0.52043 and 0.38562, respectively. From Fig. 2, the multi-resolution local fractal features between (e) and (f) have more similar than (a) and (b), but have more distinct than (c) and (d).

The gene expression profile of YAL003W with largely different global fractal feature from others genes in Fig. 2 can easily be discriminative from its original gene expression. These figures show that multi-resolution fractal feature is a very powerful tool to explore the shape change at different scales.

The global fractal feature can capture the global property of original gene expression profile well and truly. In Fig. 3, the histogram of global fractal feature for the 533 gene expression profiles is plotted, from which the statistical property with similar global fractal feature can be inferred.

2.3. Multi-resolution shape mixture model clustering

For global fractal features, the univariate mixture model cluster is applied, which only includes two possible models – equal variance (denoted E) or varying variance (denoted V). The better numbers of clusters are 2, 3 and 4, where the number clusters 2 has the optimal BIC. In (a) of Fig. 4, the BIC values are plotted against different Σ_k and the number of clusters.

For local fractal features, the multivariate mixture model cluster is applied, which includes considerable more parameters selection. In (b) of Fig. 4, the BIC values for local fractal features mixture model clustering are plotted against different Σ_k and the number of clusters. It is clear that the optimal BIC is the mixture model with 4 components, ellipsoidal distribution, variable shape, variable volume and variable orientation. Each cluster includes the number of genes are 110, 158, 175 and 90, respectively. In order to show the clustering result, we recur to the PCA of local fractal

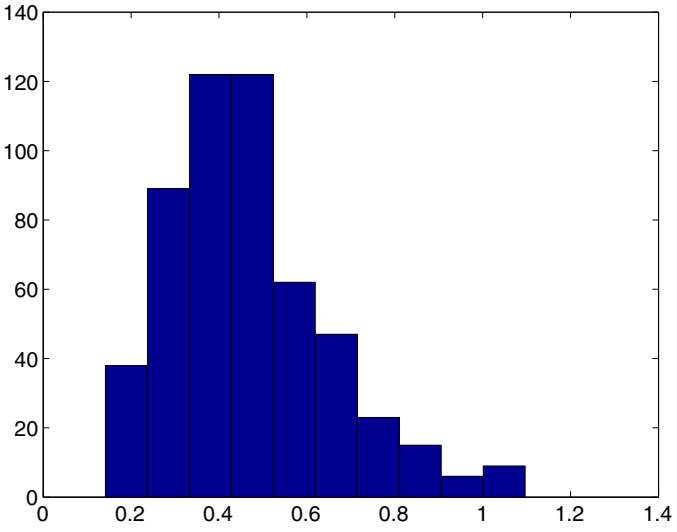


Fig. 3. Histogram of global fractal features of the 533 gene expression profiles.

features. In (a) of Fig. 5, the two largest principle components occupy the majority variance. In (b) of Fig. 5, the clustering result is provided in terms of the two largest principle components.

The genes in the same cluster tend to have the similar functions and more interaction. Therefore we use the known yeast biological knowledge to further evaluate our clustering results from the following two aspects. At first, the pathway enrichment analysis [27] is applied to the clustered genes to identify whether the genes in the same cluster have the significant functions. The significant enriched KEGG pathways for each cluster are listed with false discovery rate smaller than 0.05 in the Table 2. Secondly, the yeast protein–protein interactions are collected from the Harvard Yeast Interactome project [28]. The number of the interactions and the *p*-value of the interactions for each cluster are calculated and listed in Table 3. The protein–protein interaction graphs for each cluster are provided in Fig. 6. From the Table 3 and Fig. 6, the clustered genes are highly interacted.

2.4. Comparison with other clustering methods

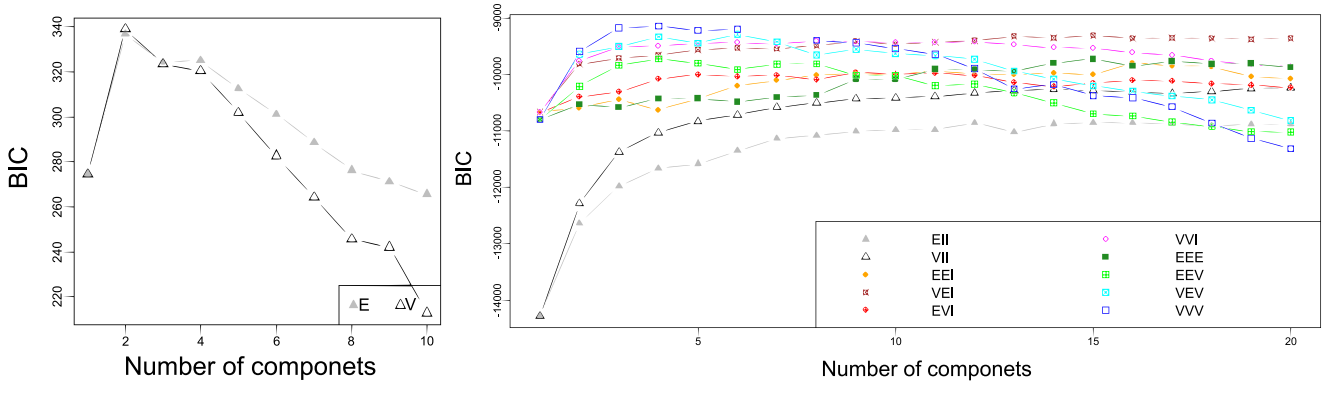
In this paper, several widely used methods for clustering gene expression including APcluster [5,6], Mfuzz [11,12], mixture model [4], *k*-means [1] and QUBIC bicluser algorithm [7] are selected to compare with our proposed algorithm. These methods are carried on yeast time-course gene expression profiles as above. The number of clusters are chosen as 4 in order to be more convenient to compare the results. For Mfuzz, mixture model and *k*-means, we can directly take the number of clusters as 4. For APcluster algorithm, the compute the similarity matrix of gene expression beforehand are computed using Pearson correlation coefficients. Then an exemplar-based agglomerative clustering is used to obtain a hierarchy of clusters from a set of clusters previously computed by affinity propagation. The right number of clusters can be extracted from the dendrogram of hierarchy of clusters. The right number of clusters of APcluster is also 4. For QUBIC bicluster algorithm, the biggest four biclusters with relative continuous time points are extracted. For clustering of gene expression profiles, the most important assessment criteria is the biological significance of the grouped genes. Through pathway enrichment analysis, the significant enriched KEGG pathways of the grouped genes identified by different clustering methods are computed. In Table 4, the enriched KEGG pathways with FDR<0.05 for four clusters identified by APcluster are listed, where the cluster 1 has not significant biological pathway. The enriched KEGG pathways with FDR<0.05 for four clusters identified by Mfuzz are shown in Table 5, where the cluster 2 has not significant biological pathway. In Table 6,

Table 2
Enriched KEGG pathways for each cluster identified by multi-resolution shape mixture model (FDR<0.05).

Clusters	Enriched KEGG pathways (FDR < 0.05)
Cluster 1	Ribosome
Cluster 2	Meiosis – yeast
	Carbon metabolism
	Citrate cycle (TCA cycle)
	Microbial metabolism in diverse environments
	Glyoxylate and dicarboxylate metabolism
Cluster 3	Cell cycle – yeast
	Meiosis
	DNA replication
	Metabolic pathways
Cluster 4	Cell cycle
	DNA replication

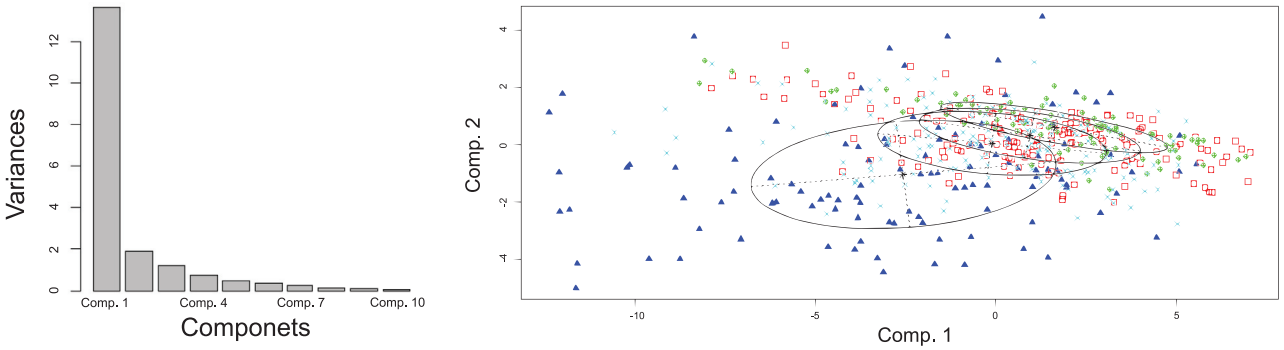
Table 3
The significance of protein–protein interactions for each cluster identified by our proposed algorithm.

Clusters	Num. interaction	Num. genes	PPI enrichment *p*-value
Cluster 1	243	107	4.09e−6
Cluster 2	355	175	5.89e−5
Cluster 3	224	90	0
Cluster 4	280	158	7.50e−5



(a) Global fractal features (b) Local fractal features

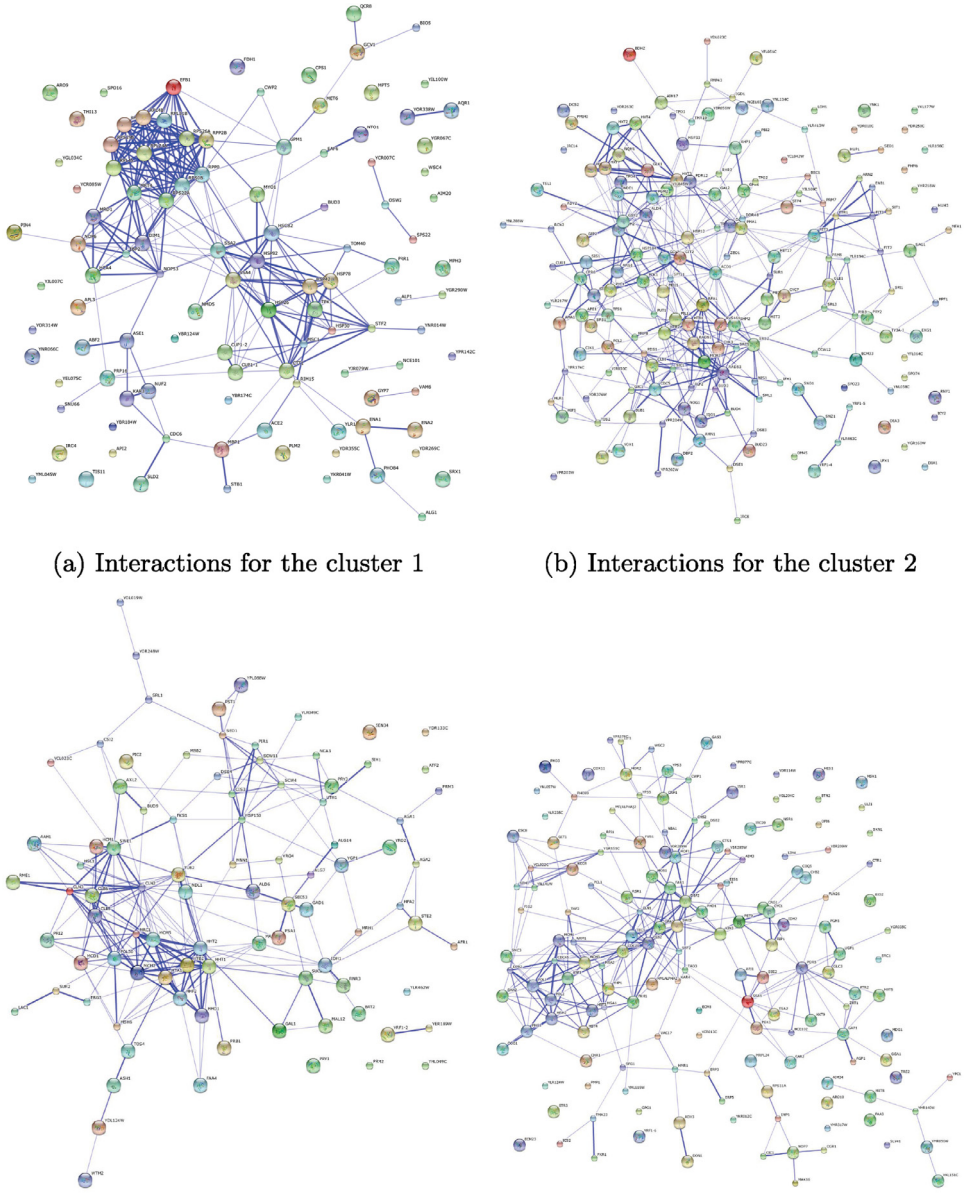
Fig. 4. BIC values for global and local fractal features.



(a) Variance of principal components for local fractal features

(b) Clustering result is provided in term of the two largest principle components

Fig. 5. PCA analysis of local fractal features.



(a) Interactions for the cluster 1

(b) Interactions for the cluster 2

(c) Interactions for the cluster 3

(d) Interactions for the cluster 4

Fig. 6. Protein–protein interactions for each cluster.

Table 4
Enriched KEGG pathways for each cluster identified by APcluster (FDR<0.05).

Clusters	Enriched KEGG pathways (FDR<0.05)
Cluster 1	–
Cluster 2	Ribosome
	Meiosis – yeast
Cluster 3	Cell cycle – yeast
	Biosynthesis of secondary metabolites
	Glyoxylate and dicarboxylate metabolism
	Carbon metabolism
	Meiosis – yeast
	Starch and sucrose metabolism
	Amino sugar and nucleotide sugar metabolism
	DNA replication
	Citrate cycle (TCA cycle)
Cluster 4	Cell cycle – yeast
	Meiosis – yeast
	Mismatch repair
	DNA replication

Table 5
Enriched KEGG pathways for each cluster identified by Mfuzz (FDR<0.05).

Clusters	Enriched KEGG pathways (FDR<0.05)
Cluster 1	Cell cycle – yeast
	Mismatch repair
	DNA replication
	Meiosis – yeast
	Base excision repair
	Purine metabolism
	Nucleotide excision repair
Cluster 2	–
Cluster 3	Starch and sucrose metabolism
	MAPK signaling pathway – yeast
	Protein processing in endoplasmic reticulum
Cluster 4	Cell cycle – yeast
	Meiosis – yeast
	Mismatch repair
	DNA replication

Table 6
Enriched KEGG pathways for each cluster identified by mixture model (FDR<0.05).

Clusters	Enriched KEGG pathways (FDR<0.05)
Cluster 1	–
Cluster 2	–
Cluster 3	Carbon metabolism
	Citrate cycle (TCA cycle)
	Glyoxylate and dicarboxylate metabolism
	Metabolic pathways
Cluster 4	Cell cycle – yeast
	DNA replication
	Meiosis – yeast
	Mismatch repair
	Base excision repair

the enriched KEGG pathways with FDR<0.05 for four clusters identified by mixture model are listed, where the clusters 1 and 2 all have not enriched biological pathway. The enriched KEGG pathways with FDR<0.05 for four clusters identified by *k*-means are shown in Table 7, where the cluster 1 has not significant biological pathway. In Table 8, the results for QUBIC bicluster algorithm are shown, in which the clusters 1 and 2 all have not significant pathways. Compared with the results in Table 2 obtained by our proposed method, the multi-resolution shape mixture model can explore strong biological significance. Each cluster identified by our proposed method has enriched KEGG pathways.

Table 7
Enriched KEGG pathways for each cluster identified by *k*-means algorithm (FDR<0.05).

Clusters	Enriched KEGG pathways (FDR<0.05)
Cluster 1	–
Cluster 2	Carbon metabolism
	Cell cycle – yeast
	Glyoxylate and dicarboxylate metabolism
	Microbial metabolism in diverse environments
	Citrate cycle (TCA cycle)
	Biosynthesis of secondary metabolites
	Metabolic pathways
	Meiosis – yeast
	Biosynthesis of amino acids
	Glycine, serine and threonine metabolism
	2-Oxocarboxylic acid metabolism
Cluster 3	Cell cycle – yeast
	Mismatch repair
	DNA replication
	Meiosis – yeast
	Base excision repair
	Purine metabolism
	Nucleotide excision repair
Cluster 4	Starch and sucrose metabolism
	MAPK signaling pathway – yeast
	Protein processing in endoplasmic reticulum

Table 8
Enriched KEGG pathways for each cluster identified by bicluster algorithm (FDR<0.05).

Clusters	Enriched KEGG pathways (FDR<0.05)
Cluster 1	–
Cluster 2	Starch and sucrose metabolism
Cluster 3	–
Cluster 4	Protein processing in endoplasmic reticulum
	Endocytosis

In addition, in order to further systematically compare coherence between different methods, the co-membership matrix for grouped genes are computed as follows

$$M(i,j) = \begin{cases} 1, & \textit{if gene } i \textit{ and } j \textit{ are in the same cluster} \\ 0, & \textit{otherwise} \end{cases}$$

$i = 1, \ldots, N, j = 1, \ldots, N$
where N is the total number of the clustered genes.

The co-membership matrices are computed for each method except for QUBIC bicluster algorithm due to the specificity of bicluster algorithm. Hubert's statistic [29] is used to validate the coherence between different methods. Let X and Y be co-membership matrix from the grouped genes obtained by different clustering methods, Huberts statistic is defined as follows:

$$H = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(\frac{X(i,j) - \bar{X}}{\sigma_X} \right) \left(\frac{Y(i,j) - \bar{Y}}{\sigma_Y} \right)$$

where $\bar{X}$ and $\bar{Y}$ denote the means of the co-membership matrices of X and Y. The larger absolute value of H shows the better coherence between the co-membership matrices. The Hubert's statistics of co-membership matrices of different clustering methods are listed in Table 9. From Table 9, the coherence of the co-membership matrices among *k*-means, Mfuzz and mixture model are significant higher. APcluster has relatively low consistency with other clustering methods. Our proposed algorithm has relatively higher consistent with other clustering methods, which help to validate the performance of our algorithm and also provide a comprehensive understanding for different clustering methods on gene expression analysis.

Table 9
Hubert's statistics of co-membership matrices of different clustering methods.

Methods	APcluster	Mfuzz	Mixture model	*k*-means
Multi-resolution shape mixture model	0.6089	0.6019	0.6046	0.6264
APcluster		0.4467	0.3829	0.4027
Mfuzz			0.6476	0.7594
Mixture model				0.8497

3. Conclusion

In this paper, a novel multi-resolution shape mixture model algorithm based on multi-resolution fractal features is proposed. Firstly, a multi-resolution fractal features are constructed to capture the shape change of the gene expression time-course profile at different resolutions from global to local view. Then combined with the mixture model, a novel multi-resolution shape mixture model algorithm based on multi-resolution fractal features is proposed. Multi-resolution fractal features include global fractal features and local fractal features at different scales. Five mostly used clustering algorithms for time-course gene expression including including *k*-means, mixture clustering model, APcluster, bicluster and Mfuzz are selected to compare with our proposed algorithm. As we know, it is hard to give a relatively fair and reasonable criteria to estimate the clustering results when the actual cluster structure is unknown. The grouped genes identified by different methods are validated by enrichment analysis of biological pathways and known protein–protein interactions from experiment evidence. The results shows that our proposed algorithm is effective and can explore the gene set with strong biological significance. The experiment on the yeast time-course gene expression profiles shows that our proposed method can explore the clusters with significant biological function. In addition, we compare the co-membership matrices for different clustering methods, which provide a comprehensive understanding and comparison for different clustering methods on gene expression analysis. Therefore our proposed multi-resolution shape mixture model for visualization and analysis of the gene expression time-course profile provides a novel horizons and an effective alternative tool.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant numbers 61272207, 61472158 and 61572228); and the Science-Technology Development Project from Jilin Province (Grant number 20130522118JH).

References

[1] J.A. Hartigan, M.A. Wong, Algorithm AS 136: A *k*-means clustering algorithm, J. R. Stat. Soc. Ser. C (Appl. Stat.) 28 (1) (1979) 100–108.
[2] T. Kohonen, P. Somervuo, Self-organizing maps of symbol strings, Neurocomputing 21 (1) (1998) 19–30.
[3] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (25) (1998) 14863–14868.
[4] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J. Am. Stat. Assoc. 97 (458) (2002) 611–631.
[5] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.
[6] U. Bodenhofer, A. Kothmeier, S. Hochreiter, APcluster: An R package for affinity propagation clustering, Bioinformatics 27 (17) (2011) 2463–2464.
[7] G. Li, Q. Ma, H. Tang, A.H. Paterson, Y. Xu, QUBIC: A qualitative biclustering algorithm for analyses of gene expression data, Nucleic Acids Res. 37 (15) (2009) e101.
[8] X. Wen, S. Fuhrman, G.S. Michaels, D.B. Carr, S. Smith, J.L. Barker, R. Somogyi, Large-scale temporal gene expression mapping of central nervous system development, Proc. Natl. Acad. Sci. 95 (1) (1998) 334–339.
[9] A.T. Kwon, H.H. Hoos, R. Ng, Inference of transcriptional regulation relationships from gene expression data, Bioinformatics 19 (8) (2003) 905–912.
[10] V. Filkov, S. Skiena, J. Zhi, Analysis techniques for microarray time-series data, J. Comput. Biol. 9 (2) (2002) 317–330.
[11] M.E. Futschik, B. Carlisle, Noise-robust soft clustering of gene expression time–course data, J. Bioinf. Comput. Biol. 3 (4) (2005) 965–988.
[12] L. Kumar, M.E. Futschik, Mfuzz: A software package for soft clustering of microarray data, Bioinformation 2 (1) (2007) 5–7.
[13] K. Willbrand, F. Radvanyi, J.-P. Nadal, J.-P. Thiery, T.M. Fink, Identifying genes from up–down properties of microarray expression series, Bioinformatics 21 (20) (2005) 3859–3864.
[14] R. Balasubramaniyan, E. Hüllermeier, N. Weskamp, J. Kämper, Clustering of gene expression data using a local shape-based similarity measure, Bioinformatics 21 (7) (2005) 1069–1077.
[15] S. Robinson, G. Glonek, I. Koch, M. Thomas, C. Davies, Alignment of time course gene expression data and the classification of developmentally driven genes with hidden Markov models, BMC Bioinform. 16 (1) (2015) 196.
[16] T.-Y. Chiu, T.-C. Hsu, C.-C. Yen, J.-S. Wang, Interpolation based consensus clustering for gene expression time series, BMC Bioinform. 16 (1) (2015) 1–17.
[17] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for dna microarrays, Bioinformatics 17 (6) (2001) 520–525.
[18] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464.
[19] S.G. Mallat, Multiresolution approximations and wavelet orthonormal bases of $l^2(r)$, Trans. Am. Math. Soc. 315 (1) (1989) 69–87.
[20] S. Mallat, W.L. Hwang, Singularity detection and processing with wavelets, IEEE Trans. Inf. Theory 38 (2) (1992) 617–643.
[21] J.D. Banfield, A.E. Raftery, Model-based Gaussian and non-Gaussian clustering, Biometrics 49 (3) (1993) 803–821.
[22] C. Fraley, A.E. Raftery, How many clusters? Which clustering method? Answers via model-based cluster analysis, Comput. J. 41 (8) (1998) 578–588.
[23] C. Fraley, A.E. Raftery, Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust, J. Classif. 20 (2) (2003) 263–286.
[24] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B (methodological) 39 (1) (1977) 1–38.
[25] G. McLachlan, T. Krishnan, The EM Algorithm and Extensions, vol. 382, John Wiley & Sons, 2007.
[26] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, Mol. Biol. Cell 9 (12) (1998) 3273–3297.
[27] D.W. Huang, B.T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M.W. Baseler, H.C. Lane, R.A. Lempicki, DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists, Nucleic Acids Res. 35 (2007) W169–W175. suppl 2
[28] H. Yu, P. Braun, M.A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, et al., High-quality binary protein interaction map of the yeast interactome network, Science 322 (5898) (2008) 104–110.
[29] V.S. Tseng, C.-P. Kao, Efficiently mining gene expression data via a novel parameterless clustering method, IEEE/ACM Trans. Comput. Biol. Bioinform. 2 (4) (2005) 355–365.