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.

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