Journal of Biosciences and Medicines, 2013, 1, 16-21 **JBM**
http://dx.doi.org/10.4236/jbm.2013.13004 Published Online December 2013 (http://www.scirp.org/journal/jbm/)

Fuzzy clustering of time series gene expression data with cubic-spline

Yu Wang, Maia Angelova, Akhtar Ali

Mathematical and Intelligent Modeling Lab, Northumbria University, Newcastle upon Tyne, UK
Email: yu.wang@northumbria.ac.uk, maia.angelova@northumbria.ac.uk, akhtar.ali@northumbria.ac.uk

Received September 2013

ABSTRACT

Data clustering techniques have been applied to extract information from gene expression data for two decades. A large volume of novel clustering algorithms have been developed and achieved great success. However, due to the diverse structures and intensive noise, there is no reliable clustering approach can be applied to all gene expression data. In this paper, we aim to the feature of high noise and propose a cubic smoothing spline fitted for the time course expression profile, by which noise can be filtered and then groups genes into clusters by applying fuzzy c-means clustering on the resulting splines (FCMS). The discrete values of radius of curvature are used to compute the similarity between spline curves. Results on gene expression data show that the FCMS has better performance than the original fuzzy c-means on reliability and noise robustness.

Keywords: Fuzzy c-Means; Cubic Spline; Noise; Radius of Curvature

1. INTRODUCTION

With the development of DNA sequencing techniques and microarray technology, genomic research has achieved a great success. A wealth of biological data has been extracted from microarrays. Analysis of these data on the molecular level is revolutionary in medicine because they are highly informative. Innovative models are needed instead of straightforward adaptations of existing methodologies. Clustering of gene expression data provides an efficient way to extract information from these large-scale datasets [1]. The underlying assumption in clustering gene expression data is that co-expression indicates co-regulation, thus clustering should identify genes that share similar functions.

Generally, there are two categories of gene expression data: static and time series [2,3]. In static expression experiments, a snapshot of the expression of genes in different samples is measured, while in time series expression experiments, a temporal process is measured. Another important difference between these two types of data is that while static data from a sample population are assumed to be independent identically distributed, while time series data exhibit a strong autocorrelation between successive points. Most previous works analyzing time series expression used methods developed originally for static data by neglecting the time series characteristics [1]. More recently several new algorithms specifically targeting time series expression data were presented. A very popular procedure in time-series analysis is smoothing the data, which removes random variation and shows trends and cyclic components [4]. Bar-Joseph *et al*. [5] used statistical spline estimation to represent time-series gene expression profiles, however, the method requires data that has been sampled at a sufficiently high rate. In addition, cubic splines are used for gene expression time-series, however no appropriate similarity metric is adopted [6]. Later, Luan and Li [7] proposed a mixed-effects model using cubic B-splines for gene expression time-series. However, it is not always possible to define equally spaced knots if the series are unevenly sampled.

In this paper, we focus on the time series characteristics and proposed an integrate approach for clustering time series microarray data. The approach was composed of two steps: the first one is modeling gene expression profiles by cubic spline curves. By tuning the smoothing parameter, gene expression data can be smoothed with statistical consideration. The second step is fuzzy c-means on the radius of curvature of the smoothed curves after discretization.

2. BACKGROUND

2.1. Fuzzy c-Means Algorithm

Large volumes of clustering algorithms have been applied to the analysis of gene expression data, such as *k*-means [8], hierarchical clustering [9] and SOM [10]. However, most of these algorithms are restricted to a one

to one mapping strategy: one gene belongs to exactly one cluster. In biology, genes can participate in more than one genetic network and are frequently coordinated by a variety of regulatory mechanisms. To address this feature, fuzzy clustering is applied for gene expression data analysis [11].

The fuzzy *c*-means clustering algorithm (FCM) is actually a variation of the *k*-means clustering algorithm, which allows one object belong to more than one cluster [12]. The FCM assigns a membership degree to each object in the data [13]. The centroids are computed based on the degree of memberships of data points. The algorithm is an iterative optimization that minimizes the cost function defined as follows:

$$J = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m} d_{ij}^{2}(x_j; v_i) \tag{1}$$

where C and N denote the number of clusters and data points respectively, u_{ij} represent the membership of point j in the i th cluster, v_i is the i th cluster center, which satisfy: $0 \le u_{ij} \le 1$ and $\sum_{i=1}^{C} u_{ij} = 1$. d_{ij}^{2} is the distance between feature vector x_j and prototypes v_i. The original formulation of FCM uses prototypes and inner-product induced norm metric for d_{ij}^{2} given by:

$$d_{ij}^{2}(x_j; v_i) = \left\| x_j - v_i \right\|_A^2 = (x_j - v_i)^T A (x_j - v_i) \tag{2}$$

The parameter m controls the fuzziness of the resulting partition. If $m \to 1$, the fuzzy clustering will turn into hard clustering. The prototypes are simply the means of the clusters. If $m \to \infty$, the partition approaches maximal fuzziness, and a gene will be assigned to all clusters equally.

In the FCM, we solve the optimization problem using Lagrange multipliers [12]:

$$J = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m} d_{ij}^{2}(x_j; v_i) + \lambda\left(\sum_{i=1}^{C} u_{ij} - 1\right) \tag{3}$$

The problem is solved by iteratively updating degrees of membership withfixed centers and updating centers with fixed degrees of membership. The closed-form formulas for updates are derived by taking the partial derivatives with respect to both and setting them to zero. The partition matrix and the cluster center of FCM are estimated by (4) and (5).

$$u_{ij} = \frac{(1/d_{ij}^{2}(x_j, v_i))^{1/(m-1)}}{\sum_{i=1}^{c}(1/d_{ij}^{2}(x_j, v_i))^{1/(m-1)}} \tag{4}$$

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{j=1}^{N} u_{ij}^{m}} \tag{5}$$

2.2. Cubic Spline

In numerical analysis, Cubic-spine has been widely used in image processing and computer graphics. A cubic spline is a piecewise third-order polynomial which is smooth in the first derivative and continuous in the second derivative. It is further called natural if the second derivatives at its boundaries are enforced to be both zero. Given a set of coordinates $(x_0, y_0), \ldots, (x_n, y_n)$, we seek a natural cubic spline function $f(x)$ that.

$$y_i = f(x_i) \qquad 0 \le i \le n \tag{6}$$

Let us define (x_i, x_{i+1}) is one interval, the second derivatives of the function $f(x)$ at x_i and x_{i+1} are $f''(x_i)$ and $f''(x_{i+1})$ respectively. The cubic spline on the interval can be rewrite as a cubic polynomial,

$$f(x) = Af(x_i) + B(x_{i+1}) + Cf''(x_i) + Df'''(x_{i+1}) \tag{7}$$

$$x \in [x_i, x_{i+1}]$$

where

$$A = \frac{x_{i+1} - x}{x_{i+1} - x_i}$$

$$B = \frac{x - x_i}{x_{i+1} - x_i}$$

$$C = \frac{1}{6}(A^3 - A)(x_{i+1} - x_i)^2$$

$$D = \frac{1}{6}(B^3 - B)(x_{i+1} - x_i)^2$$

3. FUZZY c-MEANS WITH CUBIC SPLINE (FCMS)

This integrated clustering algorithm includes two steps: the first one is modeling gene expression data by cubic spline, by which noise and variation can be filtered and the clustering algorithm can be more effective. The second step is to partition the data based on the spline, due to the continuity of the curve, we introduce a geometry term, radius of curvature, to extract the feature of the curve.

3.1. Modeling Gene Expression with Cubic Spline

Due to biological and experimental factors, noises are intensive in gene expression measurements [14,15]. However, the FCM is sensitive to the noise, which causes improper positioning of the prototypes as the noise "attract" the prototypes. Therefore, the effective way is denoising and recovering the original values of the data. Many denoising methods are borrowed from signal system and image processing to deal with noise, such as low pass filtering [16], wavelet denoising [17], etc. these me-

thods only focus on the signal characteristic of the noise without consideration of time series attributes, therefore no significant results are reported so far. According to Dejean *et al.* [18], cubic spline can represent time series gene expression appropriately. However, interpolating cubic splines to time series expression data may inadvertently attribute significance to measurements dominated by noise due to over-fitting.

To infer meaningful gene expression trends over time, we wish to fit natural cubic splines to expression data in a smooth fashion. Define each gene expression.

$$y_i^j = f(t_j) + \varepsilon_{ij} \tag{8}$$

where y_i^j denotes the observation for the i th gene at time t_j, f is a continuous and differentiable function, and ε_{ij} are independent and identically distributed random variables satisfying classical assumptions

$$E(\varepsilon_{ij}) = 0, \quad Var(\varepsilon_{ij}) = \sigma^2 \tag{9}$$

A standard curve fitting process is to minimize the residual sum of squares (RSS) in **Figure 1**:

$$\text{RSS} = \sum_{i=0}^{n} (y_i - f(t))^2 \tag{10}$$

Simultaneously, in order to avoid over-fitting (the curve passes through all the data points), parameterizing *f*(*t*) by a set of pre-specified basis functions. A more flexible strategy is to impose a smoothness condition, which is also scientifically desirable here. Here, we adopt a standard constraint used in the statistics literature, *i.e.*

$$\int \left| f''(t) \right|^2 \mathrm{d}t < \eta \tag{11}$$

where η is a specific constant. We seek a cubic smoothing spline $f(t_j)$ for each gene (18), which shall be both reasonably smooth and also reasonably close to its observation value y_i^j. As a standard practice for spline smoothing, a cubic smoothing spline (**Figure 1**) can be found by minimizing the following combined function

$$L = \lambda \sum_{i=0}^{n} (y_i - f(t))^2 + (1-\lambda) \int \left| f''(t) \right|^2 \mathrm{d}t \tag{12}$$

The first term of **Eq.**12 is residual sum of squares, which quantifies the closeness to gene expression data points, and the second term, is the integrated squared second derivative, which quantifies the smoothness of the fitted spline. The smoothing parameter, $\lambda \in [0,1]$, is used to control the trade-off between the above two contradictory criteria. If setting λ = 0, a straight line is generated from an ordinary linear least-squares regression. If setting λ = 1, it leads to a cubic interpolating spline by passing through all values. The selection of λ can be found in [18].

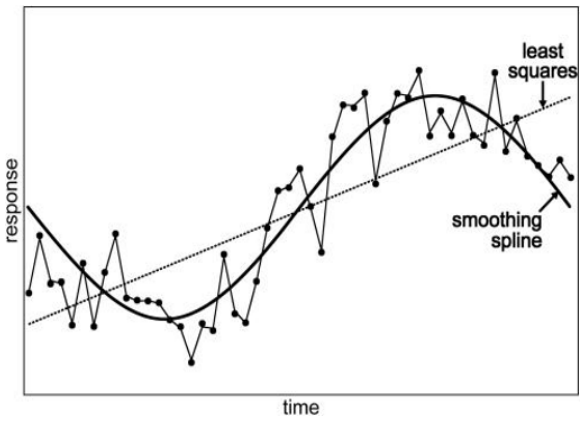


Figure 1. Cubic spline modeling gene expression profiles.

3.2. Similarity Metric

The similarity metric is generally required in every clustering method, which has a crucial influence on the clustering result. Conventional choices are Euclidean distance, Pearson correlation [19]. However, both of the two metrics cannot be applied in continuous vectors. Moreover, Euclidean computed the magnitude change by neglecting the meaningful shape information. Pearson correlation coefficient is not capable of uncover nonlinear pattern. In this paper, we propose a new similarity metric to calculate the similarity between gene profiles. After cubic spline smoothing, gene expression profiles are transformed from discrete values to continuous curves. Following [15], the discrete values of radius of curvature can be considered as one important feature vector for the curve. Similarity between two curves is calculated by normalizing the dot product of the vectors.

For curves, the radius of curvature at a given point is the radius of a circle that mathematically best fits the curve at that point. It can be seen from **Figure 2** that the radiuses of the two cycles represent the radius of curvature at the two different points in the curve. However, a cubic spline curve sometimes contains inflection points. At these points, curvature value becomes zero and the radius of curvature value becomes infinity. Therefore, curvature is used for computation instead of the radius of curvature

In geometry, the radius of curvature is the inverse of the curvature. In the case of a plane curve, the radius of curvature can be computed by [20],

$$R = \left| \frac{(1+y'^2)^{3/2}}{y''} \right| \tag{13}$$

where

$$y' = \frac{\mathrm{d}y}{\mathrm{d}x} \quad y'' = \frac{\mathrm{d}^2 y}{\mathrm{d}x^2}$$

To adjust the various sizes of the curves, the total

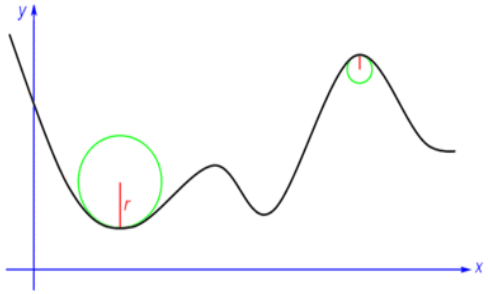


Figure 2. Radius of curvature of a curve.

length of radius of curvature is rescaled to 1. The radius of curvature must be calculated according to the knot sequence of the knot vector. Similarity between curve A and curve B is evaluated by normalizing the dot product of the vectors.

$$S = \frac{a \bullet b}{|a||b|} \tag{14}$$

By definition, the curvature of a curve is nonnegative, which limits its application to gene expression similarity metric. **Figure 3(a)** shows an example of two curves having the same radius of curvature and the same counter-clockwise direction while **Figure 3(b)** shows that two curves have the same radius of curvature and the same clockwise direction. However, the two curves have totally different expression trend. In many cases it is useful to ascribe a sign to the curve. The choice of the sign is usually connected with the tangent rotation the curvature of the curve is positive when its tangent rotates counter-clockwise. The curvature of the curve is negative when its tangent rotates clockwise. However, this simple use of the direction cannot capture the gene profiles similarity. We modify the radius of curvature by adding a sign function of the first derivative of the curve.

$$R = \operatorname{sgn}(y') \bullet \frac{(1+y'^2)^{3/2}}{y''} \tag{15}$$

The Algorithm for the simulations is given below

Step 1: construct the cubic spline to modeling the time series gene expression data according to **Eq.12**.
Step 2: select N points of the cubic spline and calculate the radius of curvature in these points by **Eq.13**.
Step 3: Compute the similarity according to equation **Eq.14**.
Step 4: Run Fuzzy c-means Clustering algorithms based on the new similarity metric.
Step 5: Evaluate the result by validity measures.

4. DATA, EXPERIMENTS AND RESULT

4.1. Data

The complete yeast gene expression profiles include 6200 genes measured every 10 min during two cell cycles in 17 hybridization experiments. Cho *et al*. [14]

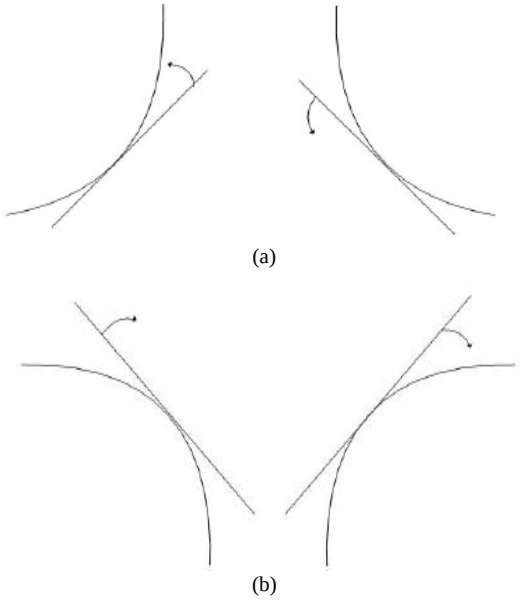


Figure 3. Comparison of the directions of curves.

selected 384 genes whose expression levels peak at different time points corresponding to the five phases (G1, S, G2/M, M/G1 and S/G2) of cell cycle. Further, Yeung *et al*. [21] extract 237 genes from the yeast cell cycle data which correspond to four categories: DNA synthesis and replication, organization of centrosome, nitrogen and sulphur metabolism, and ribosomal proteins

4.2. Results

Adjusted Rand index (ARI) is a measure of agreement between two partitions: one is the clustering result and the other is the standard partition. The value of ARI varies from 0 to 1 and higher value means that the clustering result is more similar to the standard partitions. Suppose T is the true clustering of a gene expression data set based on domain knowledge and C a clustering result given by some clustering algorithm. Let a denote the number of gene pairs belonging to the same cluster in both T and C, b is the number of pairs belonging to the same cluster in T but to different clusters in C, c is the number of pairs belonging to different clusters in T but to the same cluster in C and d is the number of pairs belonging to different clusters in both T and C.

$$\mathrm{ARI}(T,C) = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \tag{16}$$

Silhouette width index (SWI) is a measure of tightness and separation of clusters, which is used to assess the level of statistical significance of clusters. The Silhouette width for the i^{th} sample in cluster X_j is defined as,

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{17}$$

where $a(i)$ is the average distance between the i^{th} sample and all of the samples included in X_j, and $b(i)$ is the minimum average distance between the i^{th} sample and all of the sample and all of the samples clustered in X_k ($k=1,....,c; k \neq j$). When $s(i)$ is close to 1, one may infer that the i^{th} sample has been well clustered. Thus, for a given cluster X_j ($j=1,....,c$), it is possible to calculate a cluster Silhouette S_j, which characterizes the heterogeneity and isolation properties of such a cluster,

$$S_j = \frac{1}{m}\sum_{i=1}^{m} s(i) \tag{18}$$

Before experiments, the data was log2 transformed to make symmetry between negative and positive fold change and normalized to obtain a mean expression value of one for each gene. This ensures that genes which share the same expression pattern have similar gene expression vectors. λ = 0.8. Each algorithm run 10 times with randomly initialization, results are obtained by the averages value.

It can be seen from **Table 1** that the clustering results by FCMS outperforms FCM on ARI and SWI. This not only indicates that clusters generated by FCMS are better in intra compactness and inter separateness, but also illustrates that clusters include more biological significance.

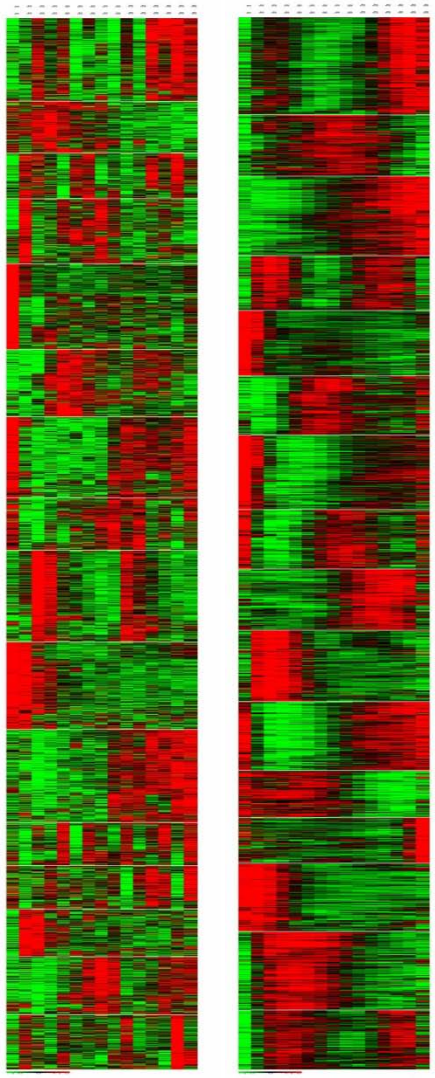
Heatmap [22] is used to graphically represent multidimensional gene expression data which have been subjected to clustering algorithms. **Figure 4** shows the quality of clusters of *Yeast* 2945 [23]. It can be seen in **Figure 4** that the FCMS shows better separated and homogeneous clusters than FCM.

5. CONCLUSION

Conventional partition clustering methods are frequently used for gene expression analysis without consideration of the noise and variations in expression that do not fit into any global pattern. In this paper, we present an integrated fuzzy clustering approach, FCMS, uses spline estimation to represent gene time-series expression profiles as continuous curves, by which noise can be filtered and meaningful data will be preserved. Similarity is an crucial factor for clustering, we introduce a new geometry term of radius of curvature, which can capture the similarity between curves. Results demonstrate that our clustering method has substantial advantages over FCM for time-series gene expression data.

Table 1. Comparison of FCM and FCMS.

		Clusters number	ARI	SWI
Yeast 384	FCM	5	0.5563	0.3545
	FCMS	5	**0.5681**	**0.4012**
Yeast 237	FCM	4	0.4658	0.3242
	FCMS	4	**0.4821**	**0.3869**



(a) (b)

Figure 4. Cluster structure plot generated by GEDAS [22]. (a) Cluster structure of FCM; (b) Cluster structure of FCMS.

6. ACKNOWLEDGEMENTS

This work was partly supported by the European project FP7-Marie Curie Actions-IRSES-MARSIQEL.

REFERENCES

[1] Belacel, N., Wang, Q. and Cuperlovic-Culf, M. (2006) Clustering methods for microarray gene expression data. *Omics*: *A Journal of Integrative Biology*, **10**, 507-531.

[2] Tsai, T.H., Milhorn, D.M. and Huang, S.K. (2006) Microarray and gene-clustering analysis. *Methods in Molecular Biology*, **315**, 165-174.

[3] Tang, R. and Muller, H.G. (2009) Time-synchronized clustering of gene expression trajectories. *Biostatistics*, **10**, 32-45. http://dx.doi.org/10.1093/biostatistics/kxn011

[4] Song, J.J., Lee, H.J., Morris, J.S. and Kang, S. (2007) Clustering of time-course gene expression data using functional data analysis. *Computational Biology and Chemistry*, **31**, 265-274. http://dx.doi.org/10.1016/j.compbiolchem.2007.05.006

[5] Bar-Joseph, Z. (2004) Analyzing time series gene expression data. *Bioinformatics* **20**, 2493-2503. http://dx.doi.org/10.1093/bioinformatics/bth283

[6] Bar-Joseph, Z., Gerber, G.K., Gifford, D.K., Jaakkola, T.S. and Simon, I. (2003) Continuous representations of time-series gene expression data. *Journal of Computational Biology*, **10**, 341-356. http://dx.doi.org/10.1089/10665270360688057

[7] Luan, Y.H. and Li, H.Z. (2003) Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, **19**, 474-482. http://dx.doi.org/10.1093/bioinformatics/btg014

[8] Gasch, A.P. and Eisen, M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, **3**, RESEARCH0059.

[9] Boratyn, G.M., Datta, S. and Datta, S. (2006) Biologically supervised hierarchical clustering algorithms for gene expression data. 28*th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBS*'06), New York, 30 August-3 September 2006, 5515-5518.

[10] Wang, J., Delabie, J., Aasheim, H., Smeland, E. and Myklebost, O. (2002) Clustering of the SOM easily reveals distinct gene expression patterns: Results of a reanalysis of lymphoma study. *BMC Bioinformatics*, **3**, 36. http://dx.doi.org/10.1093/bioinformatics/btg014

[11] Zhang, M., Adamu, B., Lin, C.C. and Yang, P. (2011) Gene expression analysis with integrated fuzzy C-means and pathway analysis. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC 2011), Boston, 30 August-3 September 2011, 936-939.

[12] Dembele, D. and Kastner, P. (2003) Fuzzy C-means method for clustering microarray data. *Bioinformatics*, **19**, 973-980. http://dx.doi.org/10.1093/bioinformatics/btg119

[13] Du, P., Gong, J., Syrkin Wurtele, E. and Dickerson, J.A. (2005) Modeling gene expression networks using fuzzy logic. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*: *Cybernetics*, **35**, 1351-1359. http://dx.doi.org/10.1109/TSMCB.2005.855590

[14] Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977-987. http://dx.doi.org/10.1093/bioinformatics/17.10.977

[15] Yang, C.M., Wan, B.K. and Gao, X.F. (2003) Data preprocessing in cluster analysis of gene expression. *Chinese Physics Letters*, **20**, 774-777. http://dx.doi.org/10.1093/bioinformatics/17.10.977

[16] Futschik, M.E. and Carlisle, B. (2005) Noise-robust soft clustering of gene expression time-course data. *Journal of Bioinformatics and Computational Biology*, **3**, 965-988. http://dx.doi.org/10.1093/bioinformatics/17.10.977

[17] Hu, X., Yoo, I., Zhang, X., Nanavati, P. and Debjit, D. (2005) Wavelet transformation and cluster ensemble for gene expression analysis. *International Journal of Bioinformatics Research and Applications*, **1**, 447-460. http://dx.doi.org/10.1504/IJBRA.2005.008447

[18] Dejean, S., Martin, P.G., Baccini, A. and Besse, P. (2007) Clustering time-series gene expression data using smoothing spline derivatives. *EURASIP Journal on Bioinformatics & Systems Biology*, **2007**, Article ID: 70561.

[19] Yin, L., Huang, C.H. and Ni, J. (2006) Clustering of gene expression data: Performance and similarity analysis. *BMC Bioinformatics*, **7**, S19. http://dx.doi.org/10.1186/1471-2105-7-S4-S19

[20] Kuragano, T. and Kasono, K. (2008) Curve generation and modification based on radius of curvature smoothing. *Proceedings of the* 10*th WSEAS International Conference on Mathematical and Computational Methods in Science and Engineering* (MACMESE'08), Bucharest, 7-9 November 2008, 80-87.

[21] Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E. and Ruzzo, W.L. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977-987. http://dx.doi.org/10.1093/bioinformatics/17.10.977

[22] Fu, L. and Medico, E. (2007) FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics*, **8**, 3. http://dx.doi.org/10.1038/10343

[23] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Systematic determination of genetic network architecture. *Nature genetics*, **22**, 281-285. http://dx.doi.org/10.1038/10343

