Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 29 (2008) 232–242
Pattern Recognition Letters
www.elsevier.com/locate/patrec

A modified correlation coefficient based similarity measure for clustering time-course gene expression data ☆

Young Sook Son *, Jangsun Baek

Department of Statistics, Chonnam National University, Gwangju 500-757, Republic of Korea

Received 29 August 2006; received in revised form 11 August 2007
Available online 10 October 2007

Communicated by M. Singh

Abstract

Gene expression levels are often measured consecutively in time through microarray experiments to detect cellular processes underlying regulatory effects observed and to assign functionality to genes whose function is yet unknown. Clustering methods allow us to group genes that show similar time-course expression profiles and that are thus likely to be co-regulated. The correlation coefficient, the most well-liked similarity measure in the context of gene expression data, is not very reliable in representing the association of two temporal profile patterns. Moreover, the clustering methods with the correlation coefficient generate the same clustering result even when the time points are permuted arbitrarily. We propose a new similarity measure for clustering time-course gene expression data. The proposed measure is based on the correlation coefficient and the two indices representing the concordance of temporal profile patterns and that of the time points at which maximum and minimum expression levels are measured between two profiles, respectively. We applied the hierarchical clustering method with the proposed similarity measure to both synthetic and breast cancer cell line data. We observed favorable results compared to the correlation coefficient based method. The proposed similarity measure is simple to implement, and it is much more consistent for clustering than the correlation coefficient based method according to the cross-validation criterion.

Keywords: Pearson's correlation coefficient; Spearmann's correlation coefficient; Modified correlation coefficient; Similarity; Clustering; Time-course gene expression data

1. Introduction

Microarrays have been used to simultaneously measure the gene expression levels of thousands of genes. The procedure of microarray experiments is based on hybridization of a specific RNA-sequence to the target cDNA sequences. The resulting gene expression data allows us to investigate the presence of a specific tumor or differences between healthy and diseased tissues. When gene expression levels are measured through microarray experiments at a limited number of consecutive time points, we call them time-course gene expression profiles. The temporal pattern of gene expression levels can be investigated by analyzing time-course data. Gene expression levels in time-course data are usually measured at a small number of time points. Conventional methods for time series analysis, such as not only trend analysis which includes smoothing, decomposition or regression method, but also ARIMA modelling or Fourier analysis, are not suitable for such a short series of time-ordered data. The data is often analyzed by clustering methods. There have been a number of clustering methods for time-course data. These can be divided into two classes depending on whether the data follows any probabilistic model or not.

Among some results using techniques in the first class are Hoon et al. (2002), Peddada et al. (2003), Schliep et al. (2003) and Luan and Li (2003). Hoon et al. (2002)

☆ Supported by the Korea Research Foundation Grant (KRF-2005-204-C00017).
* Corresponding author. Tel.: +82 62 530 3444; fax: +82 62 530 3449.
E-mail address: ysson@chonnam.ac.kr (Y.S. Son).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.09.015

fitted linear spline functions to a small set of time-ordered data in order to estimate the mean temporal profile. Hoon et al. applied k-means clustering to the fitted linear spline functions. Peddada et al. (2003) proposed a clustering algorithm based on the order-restricted inference methodology. Candidate temporal profiles were defined in terms of inequalities among mean expression levels at the time points. They selected genes when a bootstrap-based criterion was met and assigned each selected gene to the best fitting candidate profile. Schliep et al. (2003) used hidden markov models (HMMs) for clustering time-course data. Given the number of clusters, each was represented by one HMM from a finite collection of data encompassing typical qualitative behavior. They found cluster models iteratively and assigned data points to these models that maximize the joint likelihood of clustering and models. Luan and Li (2003) proposed a clustering method based on the mixed effect model with B-splines which was composed of a term to model the population average gene expression profile of each cluster and a term to model the random effect of each gene.

Whereas the previous methods were based on the inference of sophisticated probabilistic models, there have been much simpler similarity based methods for clustering genes with similar temporal profiles. The association measure such as the correlation coefficient is preferable to mathematical distance based measures such as Euclidean, Mahalanobis, or Minkowski distance. This is because the association measure reflects the tendency of changes for each pair of corresponding expression levels in the two profiles. Chu et al. (1998) pre-identified a few candidate temporal profiles and estimated the mean expression at each time point for each profile. Each gene was then assigned to one of the candidate profiles or not assigned into any depending upon the magnitude of the correlation coefficient between the gene's experimentally determined profile and each of the candidate profiles. Heyer et al. (1999) employed a jack-knifed correlation coefficient for clustering genes from time-course experiments. Balasubramaniyan et al. (2005) proposed a clustering method for the analysis of microarray time-course experiments that used a local shape-based similarity measure based on the Spearman rank correlation. The method was able to detect similar and even time-shifted sub-profiles. Even though the correlation coefficient was very simple to calculate and easy to use, it was not very efficient to detect the characteristics of temporal patterns. As Peddada et al. (2003) noted, a large correlation coefficient does not necessarily indicate two similarly shaped profiles, nor does a small correlation coefficient necessarily indicate differently shaped profiles. Moreover, the clustering methods, using the correlation coefficient as the similarity measure, generated the same clustering result even when the time points were permuted arbitrarily.

In this paper we propose a simple similarity measure as an association index between two time-course profiles. This measure preserves the concordance of temporal profile patterns and thus can be easily used in any dissimilarity based clustering method. The proposed measure is a hybrid of the correlation coefficient and two indices. This method preserves the similarity information of trajectory patterns of the expression levels on the time interval and that of time points where maximum and minimum expression levels are attained, respectively, between two profiles.

We describe the proposed similarity measure for clustering time-course data in Section 2. In Section 3, we illustrate that the proposed similarity measure is preferable to the correlation coefficient with a few synthetic time-course profiles. We show that the proposed measure outperforms the conventional one through a simulation experiment with noisy data. We also apply the clustering method with our similarity measure to the real breast cancer cell line data. We, then, compare the result to that of other methods. Some concluding remarks are in the last section.

2. New correlation coefficient based similarity measures

As there are dependencies between gene expression levels belonging to subsequent time points, it is important to consider clustering techniques that reflect the inherent time dependencies. In most cases of time-course microarray data analysis (Peddada et al., 2003; Schliep et al., 2003; Luan and Li, 2003), genes are considered to be in the same cluster if their trajectory patterns of expression levels are similar, that is, if they have the same shape of profile (which may be either monotone increasing, or monotone decreasing, or up–down, or down–up, or cyclical, etc.), and have the same time points where maximum and minimum levels occur.

The gene expression data usually have noises because of systematic error and/or measurement error. If there are systematic errors, various normalization methods can remove systematic effects. When the noises in the expression profiles are due to random measurement error, we take replicate measurements at the same time points and average them to reduce the random error effect. We use the average expression profiles of replicates for clustering.

Suppose a time-course experiment includes n time points denoted by $t_1, t_2, \ldots, t_n$, and there are p genes for clustering. Let x_{i,t_k} be the expression level of gene i at the time points $t_k, i = 1,2,\ldots,p, k = 1,2,\ldots,n$. Then $\boldsymbol{x}_i = (x_{i,t_1}, x_{i,t_2}, \ldots, x_{i,t_n})$ is the profile of gene i. The Pearson correlation coefficient, between the expression profiles of gene i and gene j, is defined as

$$R_{i,j} = \frac{\sum_{k=1}^{n}(x_{i,t_k} - \bar{x}_i)(x_{j,t_k} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{i,t_k} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{j,t_k} - \bar{x}_j)^2}}, \quad i,j = 1,2,\ldots,p,\ i \neq j,$$

where $\bar{x}_i = \sum_{k=1}^{n} x_{i,t_k}/n$, $\bar{x}_j = \sum_{k=1}^{n} x_{j,t_k}/n$, and $-1 \leqslant R_{i,j} \leqslant 1$. The Pearson correlation coefficient measures the similarity of the changes in the expression levels of two profiles. Specifically it measures the strength of the linear relationship between two profiles.

When outliers or measurement errors exist in data, the Spearman rank correlation coefficient, utilizing the ranks of the data, is preferred. The Spearman rank correlation coefficient, between two profiles of gene i and gene j, is given by

$$S_{i,j} = 1 - \frac{6}{n(n^2-1)} \sum_{k=1}^{n} \{r_{\boldsymbol{x}_i}(x_{i,t_k}) - r_{\boldsymbol{x}_j}(x_{j,t_k})\}^2,$$

where $r_{\boldsymbol{x}_i}(x_{i,t_k})$ is the rank of x_{i,t_k} in the profile $\boldsymbol{x}_i = (x_{i,t_1}, x_{i,t_2}, \ldots, x_{i,t_n})$, $i = 1, 2, \ldots, p$. The Spearman correlation coefficient measures the strength of the curvilinear monotonic relationship between two profiles. That is, it can measure the monotonic association between the profiles even when they do not show a linear relationship.

Suppose the sequences $\boldsymbol{x}_1 = (2, 3, 6, 4, 7)$ and $\boldsymbol{x}_2 = (1, 2, 3, 5, 3)$ are two expression profiles measured at five consecutive time points. By permuting the first and fourth measurements for each profile, we get $\boldsymbol{y}_1 = (4, 3, 6, 2, 7)$ and $\boldsymbol{y}_2 = (5, 2, 3, 1, 3)$. Fig. 1a and b show the pairs of profiles $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{y}_1, \boldsymbol{y}_2$, respectively. The two profiles in Fig. 1a display different patterns; in that $\boldsymbol{x}_1$ has an up–down–up shape, whereas $\boldsymbol{x}_2$ is an up–down profile. On the other hand, the two profiles in Fig. 1b arguably display similar patterns. Both change down–up–down–up over time. The Pearson correlation coefficient for each pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ is equally 0.439. The Spearman correlation coefficient for each pair is 0.667, also the same as each other. They produce the same measure of association not only for the profiles with different relationships, but also for those permuted at different time points arbitrarily. The values of both correlation coefficients seem to be too large for the different pair, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and too small for similar pair, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, as a similarity measure.

The same correlation coefficients, as well, can be obtained for similar and different pairs of profiles. A large coefficient does not necessarily indicate two similarly shaped profiles, nor does a small coefficient necessarily confirm differently shaped profiles. Therefore the correlation coefficient may not be a reliable measure of association when the experimental time points are few. In order to remedy the shortcomings of the correlation coefficient, we propose a new association measure by adding a device for keeping track of the profile shape over time to the correlation coefficient.

Let $\text{slope}(i, t_k, t_{k+1})$ be the slope of the straight line going through (t_k, x_{i,t_k}) and $(t_{k+1}, x_{i,t_{k+1}})$ of gene i. That is

$$\text{slope}(i, t_k, t_{k+1}) = \frac{x_{i,t_{k+1}} - x_{i,t_k}}{t_{k+1} - t_k}.$$

This measures the rate of expression level change of gene i from t_k to t_{k+1}. With this slope information, we define the following function L being used to describe the profile shape which may be a combination of up, down, and no change:

$$L_{i,t_k,t_{k+1}} = \begin{cases} 1, & \text{slope}(i, t_k, t_{k+1}) > 0, \\ -1, & \text{slope}(i, t_k, t_{k+1}) < 0, \\ 0, & \text{slope}(i, t_k, t_{k+1}) = 0. \end{cases}$$

Now we define a concordance index between the profile shapes of gene i and gene j as

$$A_{i,j} = \sum_{k=1}^{n-1} I(L_{i,t_k,t_{k+1}} = L_{j,t_k,t_{k+1}})/(n-1),$$

where $I(D)$ is 1 if D is true, and 0 otherwise. Also, note that $0 \leqslant A_{i,j} \leqslant 1$. Since this index indicates the proportion of times that the two profiles change in the same direction over time intervals, we can see how similarly two distinct profiles behave over time.

Where the profile attains a minimum or a maximum is another issue with respect to the shape of profiles. Let $T_i^{\min}$ and $T_i^{\max}$ be the time points at which the expression level of the gene i attains a minimum and maximum respectively. Let $M_{i,j}$ be defined as follows:

$$M_{i,j} = \begin{cases} 1 & \text{if } T_i^{\min} = T_j^{\min} \text{ and } T_i^{\max} = T_j^{\max}, \\ 0.5 & \text{if either } T_i^{\min} = T_j^{\min} \text{ or } T_i^{\max} = T_j^{\max}, \\ 0 & \text{if } T_i^{\min} \neq T_j^{\min} \text{ and } T_i^{\max} \neq T_j^{\max}. \end{cases}$$

Then $M_{i,j}$ tells whether the minimum and/or maximum level time points between gene i and gene j are matched or not.

Since the concordance index $A_{i,j}$ counts only the number of time intervals in which there is an agreement in the sign of the change between two profiles, it may lose the

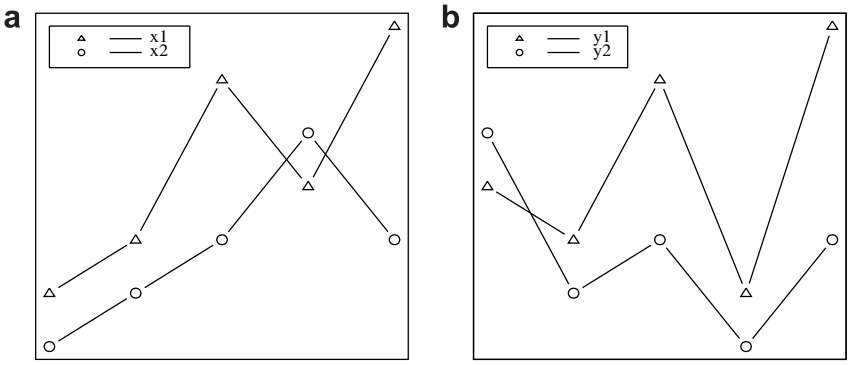

Fig. 1. The pairs of artificial profiles (a) two genes with different profiles (b) two genes with similar profiles.

information of the size of the change. Therefore instead of simply looking at the sign of the change, we could take into account its size and compute the correlation coefficient between the differences of the profiles in the time intervals. Let $\boldsymbol{d}_i = (d_{i1}, d_{i2}, \ldots, d_{i(n-1)})$ where $d_{ik} = x_{i,t_{k+1}} - x_{i,t_k}$, then a new concordance index $A^*_{i,j}$ is defined as

$$A^*_{i,j} = (\text{Pearson correlation}(\boldsymbol{d}_i, \boldsymbol{d}_j) + 1)/2.$$

As an alternative index to $M_{i,j}$, one may use $M^*_{i,j}$ which utilizes the distances between two profiles' time points where the max/min is attained. That is,

$$M^*_{i,j} = 1 - \frac{|T_i^{\min} - T_j^{\min}| + |T_i^{\max} - T_j^{\max}|}{2(n-1)}.$$

Now we consider new association measures between two time-course profiles, $Y^{R1}_{i,j}, Y^{R2}_{i,j}, Y^{S1}_{i,j}, Y^{S2}_{i,j}$ being composed of different component indices as follows:

$$\begin{aligned}
Y^{R1}_{i,j} &= \omega_1 R^*_{i,j} + \omega_2 A_{i,j} + \omega_3 M_{i,j},\\
Y^{R2}_{i,j} &= \omega_1 R^*_{i,j} + \omega_2 A^*_{i,j} + \omega_3 M^*_{i,j},\\
Y^{S1}_{i,j} &= \omega_1 S^*_{i,j} + \omega_2 A_{i,j} + \omega_3 M_{i,j},\\
Y^{S2}_{i,j} &= \omega_1 S^*_{i,j} + \omega_2 A^*_{i,j} + \omega_3 M^*_{i,j},
\end{aligned}$$

where $R^*_{i,j} = (R_{i,j}+1)/2$, and $S^*_{i,j} = (S_{i,j}+1)/2$, $i,j = 1,2,\ldots,p$, $i \neq j$, and $\omega_k \in [0,1]$ is the weighting coefficient of the kth factor composing new measure with $\sum_{k=1}^{3}\omega_k = 1$. Note that the value of new measures is between 0 and 1. The two profiles are very similar in both shape and time points for the peak and valley if the value of new measure is close to 1. The profiles are distinct to each other if the value is close to 0. It will be shown in the next section that $Y^{R1}_{i,j}$ and $Y^{S1}_{i,j}$ are preferable to $Y^{R2}_{i,j}, Y^{S2}_{i,j}$ and the conventional correlation coefficients. Therefore we propose $Y^{R1}_{i,j}$ and $Y^{S1}_{i,j}$ as new similarity measures.

We get 0.555 and 0.805 as the value of $Y^{R1}_{i,j}$ defined with the Pearson correlation coefficient for the pairs $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ and $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ in Fig. 1 respectively. The calculated values of $Y^{S1}_{i,j}$, defined with Spearman correlation coefficient, are 0.583 and 0.833 for the same pairs respectively. The weights, used in the calculation, are $\omega_1 = 1/4$, $\omega_2 = 1/2$, and $\omega_3 = 1/4$. A large similarity index (above 0.8) is obtained for similar profiles in Fig. 1b, whereas a relatively small index (below 0.6) is obtained for different profiles in Fig. 1a. Therefore the proposed measures can distinguish the pair of profiles with similar pattern from those with different patterns by producing a large value for the former.

The performance of the proposed similarity measure depends on the proper choice of the weights of the three components, $\omega_1, \omega_2, \omega_3$. When we cluster microarray profiles, the true memberships of the profiles are not known. Therefore it is impossible to select the optimal weights of the proposed index producing the best clustering results for any given clustering algorithm. We can, however, take an indirect step to select the optimal weights by utilizing clustering consistency measure. The clustering algorithm given to group genes based on their time-course profiles is considered to be consistent if a similar clustering result is obtained when the data is used with some of them lost. Datta and Datta (2003) suggested three cross-validation measures for clustering consistency. We picked the average proportion of the non-overlap measure among them. For each $k = 1, 2, \ldots, n$ repeat the clustering algorithm for each of the n data sets obtained by deleting the observations at time t_k. For each gene, $1 \leqslant i \leqslant p$, let C^{i,t_k} denote the cluster containing gene i in the clustering based on the data set with time t_k observations deleted. Let $C^{i,0}$ be the cluster in the original data containing gene i. Then the average proportion of non-overlap measure is

$$V = \frac{1}{pn}\sum_{i=1}^{p}\sum_{k=1}^{n}\left(1 - \frac{n(C^{i,t_k} \cap C^{i,0})}{n(C^{i,0})}\right),$$

where $n(C)$ is the number of elements in cluster C. This measure computes the average proportion of genes that are not put in the same cluster by the clustering method under consideration on the basis of the full data and the data obtained by deleting the expression levels at one time point at a time. We expect V to be small for a good clustering method.

We use the average proportion of non-overlap measure V to select the optimal values for the weighting coefficients ω_1, ω_2, and ω_3. That is, we calculate Vs for various choices of weights satisfying $\omega_k \in [0,1]$ and $\sum_{k=1}^{3}\omega_i = 1$. We, then, pick the values of weights that generate the minimum V.

3. Numerical analysis

3.1. Synthetic data

We first illustrate that the proposed similarity measure is preferable to the correlation coefficient through a similar example in Peddada et al. (2003). Fig. 2 shows synthetic expression levels for four genes at six time points. The correlation coefficients, $R_{i,j}$ and $S_{i,j}$, the proposed similarity

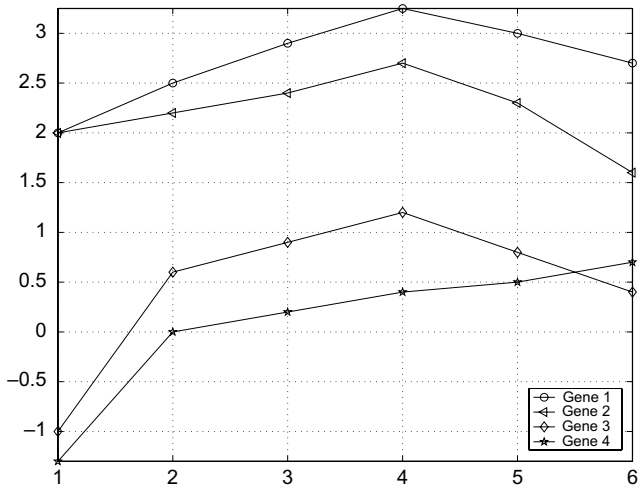

Fig. 2. Four synthetic time-course profiles with six time points.

Table 1
Similarity indices

(i,j)	$R_{i,j}$	$S_{i,j}$	$A_{i,j}$	$M_{i,j}$	$A^*_{i,j}$	$M^*_{i,j}$	New similarity index			
							$R_{i,j}$ based		$S_{i,j}$ based	
							$Y^{R1}_{i,j}$	$Y^{R2}_{i,j}$	$Y^{S1}_{i,j}$	$Y^{S2}_{i,j}$
(1,2)	0.5920	0.7714	1.0	0.5	0.9786	0.5	0.7730	0.7677	0.8179	0.8125
(1,3)	0.9348	0.8857	1.0	1.0	0.9191	1.0	0.9837	0.9635	0.9714	0.9512
(1,4)	0.8300	0.6000	0.6	0.5	0.7811	0.8	0.7325	0.8528	0.6750	0.7953
(2,3)	0.5651	0.9429	1.0	0.5	0.8524	0.5	0.7663	0.7294	0.8607	0.8238
(2,4)	0.0966	0.0286	0.6	0.0	0.6896	0.3	0.4241	0.5216	0.3929	0.4903
(3,4)	0.8595	0.2571	0.6	0.5	0.9607	0.8	0.7399	0.9050	0.5893	0.7545

$Y^{R1}_{i,j} = 0.5R^*_{i,j} + 0.25A_{i,j} + 0.25M_{i,j}$, $Y^{R2}_{i,j} = 0.5R^*_{i,j} + 0.25A^*_{i,j} + 0.25M^*_{i,j}$.
$Y^{S1}_{i,j} = 0.5S^*_{i,j} + 0.25A_{i,j} + 0.25M_{i,j}$, $Y^{S2}_{i,j} = 0.5S^*_{i,j} + 0.25A^*_{i,j} + 0.25M^*_{i,j}$.

indices $Y^{R1}_{i,j}, Y^{S1}_{i,j}$ defined with $A_{i,j}, M_{i,j}$, and $Y^{R2}_{i,j}, Y^{S2}_{i,j}$ defined with $A^*_{i,j}, M^*_{i,j}$ between two genes are listed in Table 1. The weights for the proposed indices are $\omega_1 = 1/2$ and $\omega_2 = \omega_3 = 1/4$. Gene 1 and gene 2 display similar patterns, in that both have an up–down shape and attain a peak value at the fourth time point. Their Pearson correlation coefficient $R_{1,2}$ is, however, only 0.5920. On the other hand, gene 3 and gene 4 show apparently different patterns over time. One gene pattern increases monotonically whereas the other has a peak at the fourth time point and decreases after that. They have a high correlation coefficient of $R_{3,4} = 0.8595$, which is much higher than that between gene 1 and 2. Gene 3 and 4 may be grouped together by a correlation-based approach while gene 1 and 2 may not. When we use the proposed similarity index defined with $R_{i,j}, A_{i,j}$, and $M_{i,j}$ to measure the association between two genes, we get $Y^{R1}_{1,2} = 0.7730$ being slightly higher than $Y^{R1}_{3,4} = 0.7399$. We applied the complete-linkage hierarchical clustering method to the data with both the conventional correlation coefficients and the proposed similarity measures. Given the number of clusters of 2, genes 1, 3, 4 are grouped together excluding gene 2 when the Pearson correlation coefficient is used as a measure of association. Gene 2 is excluded because the three lowest correlation coefficients are $R_{2,4} = 0.0966$, $R_{2,3} = 0.5651$, and $R_{1,2} = 0.5920$. On the other hand, when the proposed measure is used genes 1, 2, 3 are grouped together, with gene 4 excluded, since the three lowest similarity measures are all with gene 4; $Y^{R1}_{2,4} = 0.4241$, $Y^{R1}_{1,4} = 0.7325$, and $Y^{R1}_{3,4} = 0.7399$. When we use either the Spearman correlation coefficient $S_{i,j}$ or the proposed measure $Y^{S1}_{i,j}$ defined with $S_{i,j}, A_{i,j}$ and $M_{i,j}$ to cluster the data, genes 1, 2 and 3 are determined to be in the same class excluding gene 4 since the three lowest indices are all with gene 4 as well.

Even if the clustering results are the same for the methods using $S_{i,j}$ and $Y^{S1}_{i,j}$, the quality of association between the two profiles may be different. Note that the Spearman correlation coefficient between gene 2 and 3 is 0.9429. It is larger than the coefficient 0.8857 between genes 1 and 3. That is, gene 2 is determined to be more similar to gene 3 than gene 1 is similar to gene 3 when the Spearman correlation coefficient is used as a measure of association. Gene 2, however, attains its minimum at the sixth time point while genes 1 and 3 do at the first time point.

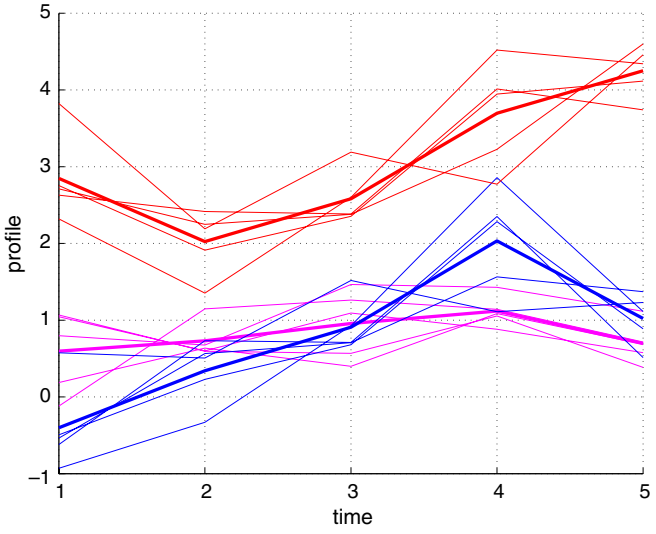

Fig. 3. Five gene expression profiles with their sample mean profile from each of three clusters.

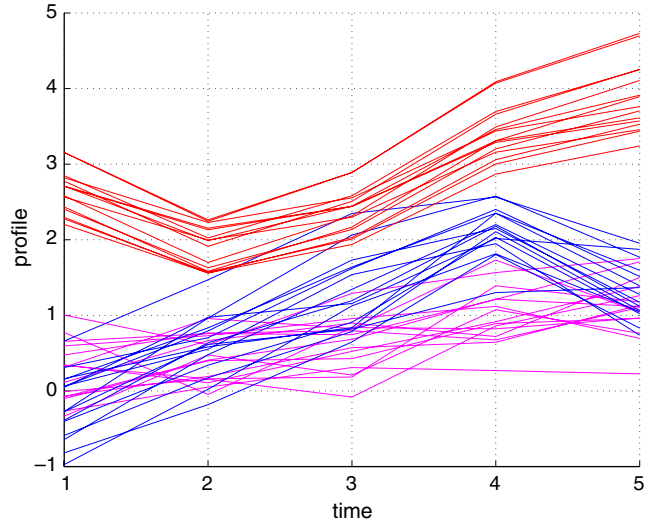

Fig. 4. Sample mean profiles from each of three clusters.

Therefore gene 3 is more closely related to gene 1 than to gene 2. The values of the proposed measure defined with $S_{i,j}$, $Y^{S1}_{1,3} = 0.9714$ and $Y^{S1}_{2,3} = 0.8607$ account for the correct strength of association among genes 1, 2, and 3 since $Y^{S1}_{1,3} > Y^{S1}_{2,3}$. It is shown that $Y^{R1}_{i,j}$, $Y^{S1}_{i,j}$, $Y^{S2}_{i,j}$, and $S_{i,j}$ are useful for the correct clustering while $R_{i,j}$ and $Y^{R2}_{i,j}$ are not. The three lowest values for $M^*_{i,j}$ are all with gene 2, and $Y^{R2}_{i,j}$ fails to cluster the genes correctly. Thus $R_{i,j}$-based clustering methods may classify similar profiles to different clusters or cluster genes with different profiles. The proposed similarity index is one of the possible improvements of the correlation coefficient as the measure of association for clustering time-course profiles.

Table 2
Agreement comparison of clustering results by the conventional correlation coefficients and the new similarity indices with true classification based on Rand index

Sample size	σ	Similarity measure	Rand index		p-Value (paired t-test)
			Mean	Std. dev.	
45	0.1	$R_{i,j}$	1.000	0.000	
		$Y^{R1}_{i,j}$	0.999	0.007	0.9929
		$Y^{R2}_{i,j}$	0.834	0.066	1.0000
		$S_{i,j}$	0.780	0.083	
		$Y^{S1}_{i,j}$	0.999	0.007	$<10^{-4}$
		$Y^{S2}_{i,j}$	0.852	0.072	$<10^{-4}$
	0.5	$R_{i,j}$	0.796	0.084	
		$Y^{R1}_{i,j}$	0.820	0.078	$<10^{-4}$
		$Y^{R2}_{i,j}$	0.794	0.064	0.8574
		$S_{i,j}$	0.741	0.118	
		$Y^{S1}_{i,j}$	0.813	0.079	$<10^{-4}$
		$Y^{S2}_{i,j}$	0.795	0.071	$<10^{-4}$
	1.0	$R_{i,j}$	0.687	0.110	
		$Y^{R1}_{i,j}$	0.709	0.085	$<10^{-4}$
		$Y^{R2}_{i,j}$	0.709	0.095	$<10^{-4}$
		$S_{i,j}$	0.630	0.127	
		$Y^{S1}_{i,j}$	0.701	0.090	$<10^{-4}$
		$Y^{S2}_{i,j}$	0.704	0.095	$<10^{-4}$
90	0.1	$R_{i,j}$	1.000	0.002	
		$Y^{R1}_{i,j}$	0.999	0.003	1.0000
		$Y^{R2}_{i,j}$	0.820	0.048	1.0000
		$S_{i,j}$	0.784	0.067	
		$Y^{S1}_{i,j}$	0.999	0.003	$<10^{-4}$
		$Y^{S2}_{i,j}$	0.851	0.061	$<10^{-4}$
	0.5	$R_{i,j}$	0.729	0.074	
		$Y^{R1}_{i,j}$	0.780	0.074	$<10^{-4}$
		$Y^{R2}_{i,j}$	0.749	0.082	0.8574
		$S_{i,j}$	0.573	0.082	
		$Y^{S1}_{i,j}$	0.770	0.082	$<10^{-4}$
		$Y^{S2}_{i,j}$	0.731	0.106	$<10^{-4}$
	1.0	$R_{i,j}$	0.628	0.129	
		$Y^{R1}_{i,j}$	0.659	0.089	$<10^{-4}$
		$Y^{R2}_{i,j}$	0.644	0.112	0.0003
		$S_{i,j}$	0.566	0.115	
		$Y^{S1}_{i,j}$	0.668	0.077	$<10^{-4}$
		$Y^{S2}_{i,j}$	0.642	0.109	$<10^{-4}$

$Y^{R1}_{i,j} = 0.5R^*_{i,j} + 0.25A_{i,j} + 0.25M_{i,j}$, $Y^{R2}_{i,j} = 0.5R^*_{i,j} + 0.25A^*_{i,j} + 0.25M^*_{i,j}$.
$Y^{S1}_{i,j} = 0.5S^*_{i,j} + 0.25A_{i,j} + 0.25M_{i,j}$, $Y^{S2}_{i,j} = 0.5S^*_{i,j} + 0.25A^*_{i,j} + 0.25M^*_{i,j}$.

To assess the performance of the proposed similarity measure in practice, we carried out the analysis on a large number of data sets with some noises. Suppose the time-course of a gene in a specific cluster follows the shape of the mean profile of the cluster, but with additional random gene-specific location and scale shifts. The actual observations at discrete time points, however, also contain normally distributed measurement errors and are assumed to be measured repeatedly l times. Let μ_{ik} be the mean profile of the ith gene's time-course expression replicates in the kth cluster, and let K denote the total number of clusters. Now we model the lth time-course expression measurement of gene i in cluster k at time point t_j considering both measurement error, random gene-specific location, and scale shifts:

$$x_{lijk} = \mu_{ik}(t_j; \alpha_{ik}, \beta_{ik}) + \epsilon_{lijk},$$

where μ_{ik} is the mean profile of the ith gene; α_{ik} and β_{ik} explain the random deviation of μ_{ik} from the cluster mean profile in location and scale, respectively. This random deviation is not due to measurement error, but an inherent, gene-specific shift from the cluster mean profile. The measurement error ϵ_{lijk} follows Normal$(0, \sigma^2)$. We assume there are $K = 3$ clusters and the time points are $t_1 = 1$, $t_2 = 2$, $t_3 = 3$, $t_4 = 4$, $t_5 = 5$. Suppose we have $\mu_{i1}(t_j) = \beta_{i1}t_j/(2\pi) + \alpha_{i1}$, $\mu_{i2}(t_j) = 2 - \beta_{i2}|t_j - 4|/2 + \alpha_{i2}$, and $\mu_{i3}(t_j) = 2 + \beta_{i3}|t_j - 2|/2 + \alpha_{i3}$, then each gene expression profile in different clusters is generated 5 times $(l = 1, 2, \ldots, 5)$ randomly from the following three functions:

$$x_{lij1} = \beta_{i1}t_j/(2\pi) + \alpha_{i1} + \epsilon_{lij1},$$
$$x_{lij2} = 2 - \beta_{i2}|t_j - 4|/2 + \alpha_{i2} + \epsilon_{lij2},$$
$$x_{lij3} = 2 + \beta_{i3}|t_j - 2|/2 + \alpha_{i3} + \epsilon_{lij3},$$

where $\alpha_{ik} \sim$ Uniform$(-0.5, 0.5)$, $\beta_{ik} \sim$ Uniform$(1, 2)$ and $\epsilon_{lijk} \sim$ Normal$(0, \sigma^2)$ with $\sigma = 0.1, 0.5, 1.0$, $i = 1, 2, \ldots, p$, $l = 1, 2, \ldots, 5$. The mean profile of the kth cluster is $E(\mu_{ik})$. For example, the mean profile of the third cluster is $2 + (3/4)|t - 2|$, $1 \leqslant t \leqslant 5$. Let $\bar{x}_{ijk}$ denote the sample

Table 3
The average proportion of non-overlap V for various weights

ω_1		ω_2		ω_3		V
$R^*_{i,j}$	$S^*_{i,j}$	$A_{i,j}$	$A^*_{i,j}$	$M_{i,j}$	$M^*_{i,j}$	
1						0.3075
1/2		1/4		1/4		0.2232
1/2			1/4		1/4	0.2639
1/3		1/3		1/3		0.2210
1/3			1/3		1/3	0.2719
1/2				1/2		0.1627
	1					0.2544
	1/2	1/4		1/4		0.1497
	1/2		1/4		1/4	0.3199
	1/3	1/3		1/3		0.1682
	1/3		1/3		1/3	0.2797
	1/2	3/10		1/5		0.1248

mean of gene i in cluster k at time point t_j. We use these gene sample mean profiles $\bar{\boldsymbol{x}}_{i\cdot k} = (\bar{x}_{i1k}, \bar{x}_{i2k}, \ldots, \bar{x}_{i5k})$ as a data for clustering, $i = 1, 2, \ldots, p$, $k = 1, 2, 3$. Fig. 3 shows five gene expression profiles (thin lines) with their sample mean profile (thick line) in each cluster, respectively, where $\sigma = 0.5$. The pink profiles are from the first cluster; the blue profiles, from the second cluster; and, the red profiles from the third cluster.

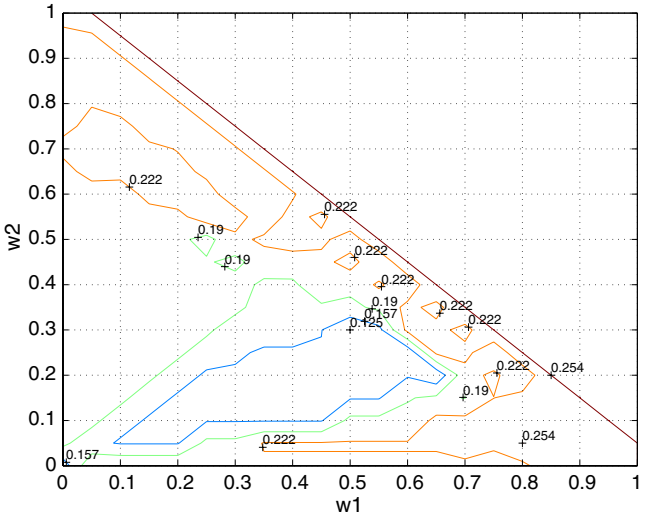

Fig. 5. The contour plot of V for various choices of weights of $Y^{S1}_{i,j}$.

A 1000 data sets consisting of a total of 45 sample mean profiles ($p = 15$ gene sample mean profiles of five replicates across five time points from each of three clusters, respectively,) were generated. A data set with 45 sample mean profiles, generated from the model with $\sigma = 0.5$, is shown in Fig. 4. We applied the complete-linkage hierarchical clustering method with the Pearson and Spearman correlation coefficient and the proposed similarity measures $Y^{R1}_{i,j}, Y^{R2}_{i,j}, Y^{S1}_{i,j}, Y^{S2}_{i,j}, (\omega_1 = 1/2, \omega_2 = \omega_3 = 1/4)$ to each of the 1000 simulated data sets to determine which clustering result agreed better with the true classification of expression profiles. We repeated the same experiment for another 1000 data sets consisting of a total of 90 sample mean profiles ($p = 30$ gene sample mean profiles of five replicates from each of three clusters, respectively).

In order to compare clustering results obtained by different measures with the true classification, we use the Rand index (Rand, 1971) as a measure of agreement. Suppose that A and B are two different clustering results. Let SS be the number of pairs of objects that are placed in the same class in A and in the same cluster in B. Let SD be the number of pairs of objects in the same class in A but not in the same cluster in B. Let DS be the number of pairs of objects in the same cluster in B but not in the same class in A; and, let DD be the number of pairs of objects in different classes and different clusters in both partitions. The quantities, SS and DD, can be interpreted as agreements

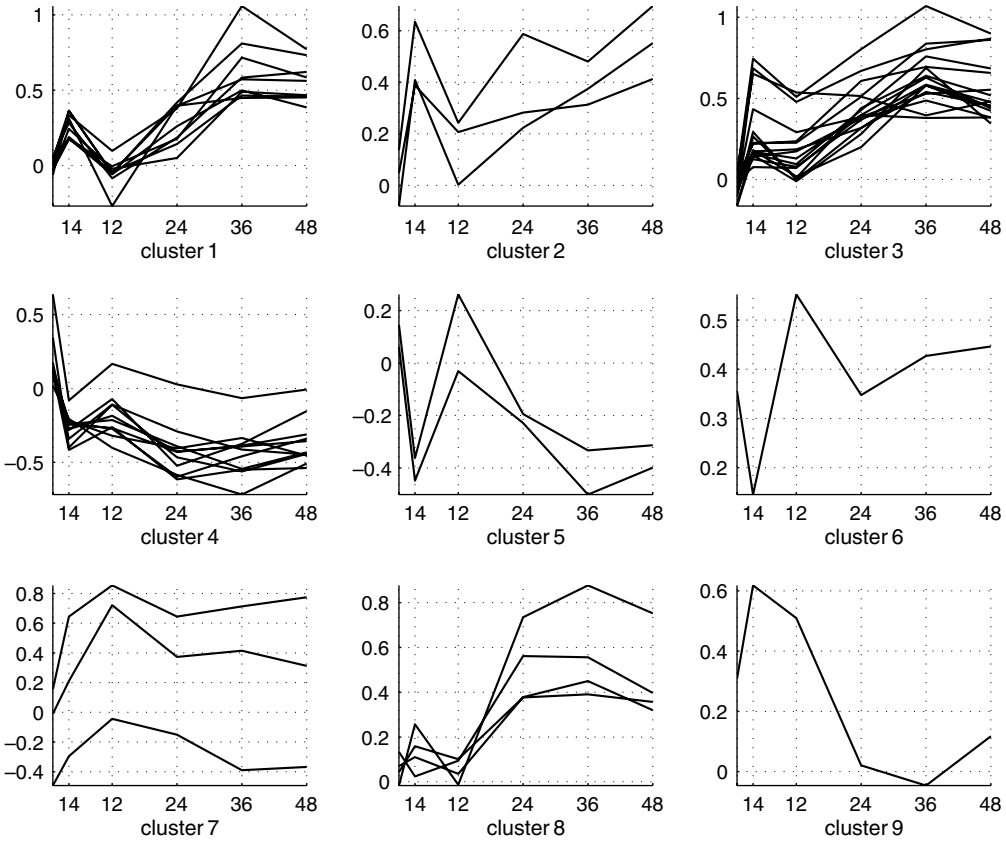

Fig. 6. The clusters classified by $R_{i,j}$.

and SD and DS as disagreements. Then the Rand index (RI) is $(SS+DD)/(SS+SD+DS+DD)$. The Rand index lies between 0 and 1. The two clustering results agree perfectly if the index is 1. For an illustration of the Rand index calculation, suppose there are four different gene profiles x_1, x_2, x_3, x_4 Let the clustering result A be $A_1 \cup A_2$, where $A_1 = \{x_1, x_2\}$ and $A_2 = \{x_3, x_4\}$, and let another clustering result B be $B_1 \cup B_2 \cup B_3$, where $B_1 = \{x_1\}$, $B_2 = \{x_2\}$, and $B_3 = \{x_3, x_4\}$. $SS = 1$ because (x_3, x_4) is the only pair of gene profiles that are placed in the same class in A and in the same cluster in B. $SD = 1$ since (x_1, x_2) is also the only pair of gene profiles in the same cluster in A but not in the same cluster in B. Having no pair of gene profiles in the same cluster in B but not in the same cluster in A leads to $DS = 0$. There are 4 pairs of gene profiles, (x_1, x_3), (x_1, x_4), (x_2, x_3), (x_2, x_4), in different classes in A and different clusters in B, so $DD = 4$. Therefore the Rand index in this example is $(1+4)/(1+1+0+4) = 5/6$.

Table 2 contains the agreement comparison of clustering results by the Pearson/Spearman correlation coefficient and their modified similarity indices for 1000 data sets of different sample sizes with $\sigma = 0.1, 0.5, 1.0$. In each case the mean and standard deviation of 1000 Rand indices were obtained. The agreement of the clustering results by all measures with the true classification got higher as σ gets small. The paired t-test with a significance level of 0.05 showed that the true mean Rand index by each proposed similarity measure (except for $Y^{R2}_{i,j}$) was significantly higher than the true mean Rand index by its conventional counterpart $R_{i,j}$ and $S_{i,j}$, respectively, for the data with medium

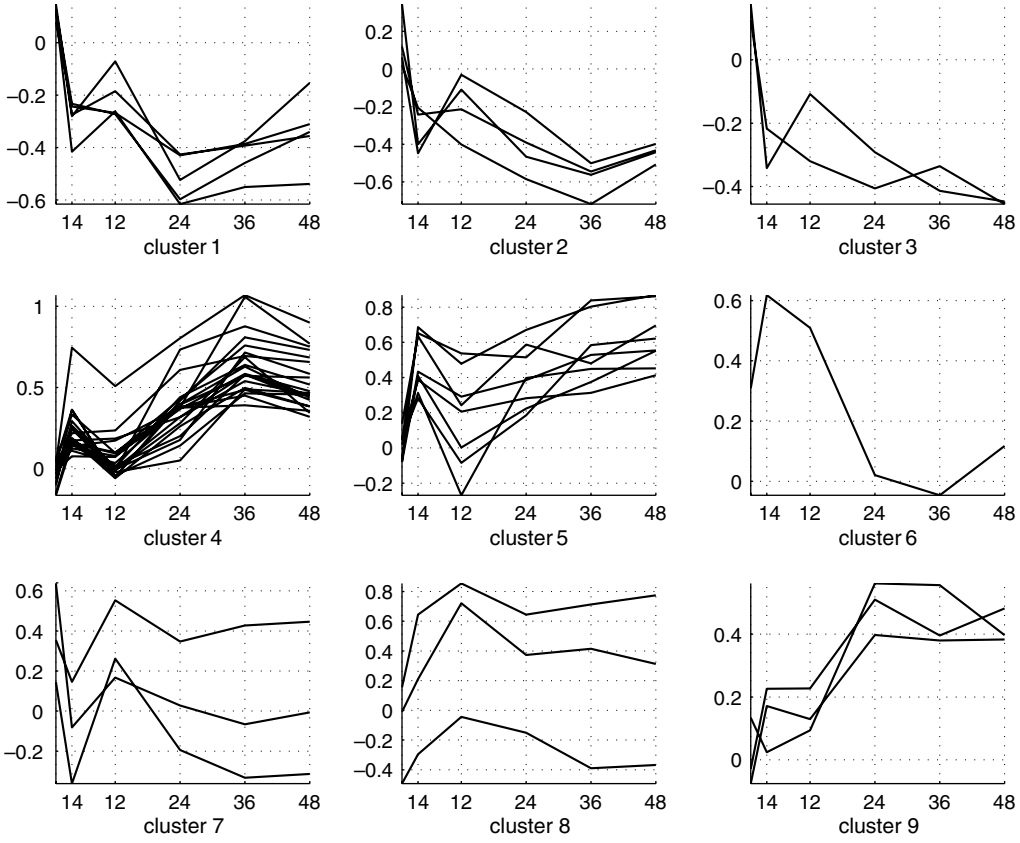

Fig. 7. The clusters classified by $Y^{S1}_{i,j}$.

Table 4
Rand indices between the result of Peddada et al. (2003) and those of various linkage methods using new similarity indices with different weights

Similarity measure	Hierarchical clustering methods			
	Complete	Average	Centroid	Ward
$R_{i,j}$	0.7665	0.8563	0.8563	0.8057
$Y^{R1}_{i,j} = 0.5R^*_{i,j} + 0.25A_{i,j} + 0.25M_{i,j}$	0.9755	0.9600	0.9600	0.9053
$Y^{R2}_{i,j} = 0.5R^*_{i,j} + 0.25A^*_{i,j} + 0.25M^*_{i,j}$	0.8286	0.8963	0.8963	0.8286
$Y^{R1}_{i,j} = 0.5R^*_{i,j} + 0.5M_{i,j}$	0.9020	0.9473	0.9437	0.9053
$S_{i,j}$	0.9265	0.9200	0.9200	0.9543
$Y^{S1}_{i,j} = 0.5S^*_{i,j} + 0.25A_{i,j} + 0.25M_{i,j}$	0.9771	0.9657	0.9657	0.9249
$Y^{S2}_{i,j} = 0.5S^*_{i,j} + 0.25A^*_{i,j} + 0.25M^*_{i,j}$	0.9298	0.9249	0.9167	0.9020
$Y^{S1}_{i,j} = 0.5S^*_{i,j} + 0.3A_{i,j} + 0.2M_{i,j}$	0.9771	0.9657	0.9657	0.9771

($\sigma = 0.5$) and large ($\sigma = 1.0$) measurement error since the p-value was less than 10^{-4}. There was no difference between the Rand indices by $R_{i,j}$ and $Y_{i,j}^{R1}(Y_{i,j}^{R2})$ for the data with small ($\sigma = 0.1$) measurement error. For the data with $\sigma = 0.1$, the Rand index by $S_{i,j}$ is significantly lower than that by $Y_{i,j}^{S1}$ and $Y_{i,j}^{S2}$. Therefore the proposed similarity measures outperformed the conventional correlation coefficient in the more realistic cases where the profiles contained both gene-specific deviations and significant measurement errors.

3.2. Application to a breast cancer cell line data

Lobenhofer et al. (2002) treated the MCF-7 breast cancer cell line with 17β-estradiol or ethanol, and got samples at 1, 4, 12, 24, 36 and 48 h after treatment. At each time point, eight hybridizations were performed. Peddada et al. (2003) used the estimated profiles that were obtained using each gene's eight replications using the method of Hwang and Peddada (1994). They identified the most biologically meaningful 50 genes and grouped them into nine

Table 5
Genes classified according to different measures

CloneID	Gene name	Functional category	P	$R_{i,j}$	$Y_{i,j}^{S1}$
417226	v-myc viral oncogene homolog	Transcription/chromatin structure	1	4	8
110022	Cyclin D1	Cell cycle	2	6	8
428733	Protein kinase C, delta	Cellular signaling	2	5	8
362059	Laminin, alpha 3, kalinin, epilegrin	Extracellular matrix/cell structure	2	7	7
417503	EST	Unknown	2	7	7
248613	v-myb viral oncogene homolog	Transcription/chromatin structure	2	7	7
563187	CDC6	Cell Cycle	3	3	9
321207	Polymerase (DNA directed), epsilon	DNA replication/repair	3	8	9
196676	Replication factor C (activator 1)4	DNA replication/repair	3	3	9
129140	MAD2L1	Cell cycle	4	3	4
248008	Deoxythymidylate kinase	Cell cycle	4	3	4
489092	Deoxythymidylate kinase	Cell cycle	4	3	4
285427	CSE1L	Cell cycle	4	3	4
359119	CDC28 protein kinase 2	Cell cycle	4	3	4
415639	Serine/threonine kinase 15	Cell cycle	4	1	4
488059	Tubulin, gamma 1	Cell cycle	4	3	4
563809	CDC20	Cell cycle	4	3	4
293274	Cyclin-dependent kinase inhibitor 3	Cell cycle	4	1	4
49950	Flap structure-specific endonuclease 1	DNA replication/repair	4	3	4
346838	Minichromosome maintenance deficien 3	DNA replication/repair	4	3	4
359465	Dihydrofolate reductase	DNA replication/repair	4	1	4
487757	Ligase I, DNA, ATP-dependent	DNA replication/repair	4	8	4
49940	Replication factor C (activator 1) 5	DNA replication/repair	4	8	4
52713	Vitronectin	Extracellular matrix/cell structure	4	1	4
339075	Karyopherin alpha 2	Protein degradation/synthesis/targeting	4	1	4
136609	v-myb homolog-like 1	Transcription/chromatin structure	4	3	4
198205	v-myb homolog-like 2	Transcription/chromatin structure	4	8	4
229509	Coagulation factor V	Miscellaneous	4	1	4
200573	EST	Unknown	4	3	4
366842	EST	Unknown	4	1	4
264117	Cathepsin D	Cell cycle	5	2	5
150163	Neuropeptide Y receptor Y1	Cellular signaling	5	3	5
238545	ADP-ribosylation factor-like 3	Cellular signaling	5	3	5
242182	Protein kinase inhibitor beta	Cellular signaling	5	3	5
509614	High-mobility group protein 1	Transcription/chromatin structure	5	1	5
510595	Lactate dehydrogenase A	Miscellaneous	5	2	5
470480	Autocrine motility factor receptor	Miscellaneous	5	2	5
487407	Insulin induced gene 1	Miscellaneous	6	1	5
361381	Myeloid cell leukemia sequence 1	Apoptosis	7	4	1
145093	Myeloid cell leukemia sequence 1	Apoptosis	7	4	1
485875	EFEMP 1	Extracellular matrix/cell structure	7	4	1
34821	CHRNA 4	Miscellaneous	7	4	1
359191	Protein kinase H11	Cellular signaling	8	9	6
180789	Low density lipoprotein-related protein 1	Protein degradation/synthesis/targeting	8	5	2
162479	E74-like factor 3	Transcription/chromatin structure	8	4	2
430235	H2B histone family, member Q	Transcription/chromatin structure	8	4	2
545242	STAT 1	Transcription/chromatin structure	8	4	1
268652	p21/CIP 1	Cell cycle	8	4	2
29682	Protein kinase C binding protein 1	Cellular signaling	9	4	3
365147	v-erb-b2 homolog 2	Cellular signaling	9	4	3

(P: Peddada et al., $R_{i,j}$: Pearson correlation coefficient, $Y_{i,j}^{S1}$: proposed measure).

clusters. We used the sample mean of eight replicated profiles as each gene's time-course profile. We clustered these 50 genes' time-course profiles by the complete-linkage hierarchical clustering method with the conventional correlation coefficient (Pearson, Spearman) and the new similarity measures $Y_{i,j}^{R1}, Y_{i,j}^{R2}, Y_{i,j}^{S1}, Y_{i,j}^{S2}$, respectively, and compared their results. Since the proper number of classes for this data was 9 (Peddada et al., 2003), we clustered the profiles into nine groups.

We used the average proportion of non-overlap measure V for clustering consistency. The values of V for $Y_{i,j}^{R1}, Y_{i,j}^{R2}, Y_{i,j}^{S1}, Y_{i,j}^{S2}$ with various choices of ω_1, ω_2 and ω_3 are shown in Table 3. When the number of classes is 9, we obtained $V = 0.1627$ for the method with $Y_{i,j}^{R1}$ where $\omega_1 = 1/2, \omega_2 = 0$ and $\omega_3 = 1/2$; and $V = 0.3075$ for the method with $R_{i,j}$. The value of V for $Y_{i,j}^{S1}$ where $\omega_1 = 1/2, \omega_2 = 3/10$ and $\omega_3 = 1/5$ was 0.1248 while that for $S_{i,j}$ was 0.2544. The new similarity indices (except for $Y_{i,j}^{S2}$) produced more consistent clustering results than the conventional correlation coefficients $(R_{i,j}, S_{i,j})$ did.

The optimal values for the weighting coefficients ω_1, ω_2 and ω_3 were selected by V. We calculated Vs for various choices of weights satisfying $\omega_k \in [0,1]$ and $\sum_{k=1}^{3}\omega_k = 1$; each ω_k increases from 0 to 1 by 0.05 increment. We picked the values of weights that generated the minimum V. $Y_{i,j}^{R1}$ with $\omega_1 = 1/2, \omega_2 = 0, \omega_3 = 1/2$ produced the minimum $V = 0.1627$ among all $Y_{i,j}^{R1}$s and $Y_{i,j}^{R2}$s which are based on $R_{i,j}$. $Y_{i,j}^{S1}$ with $\omega_1 = 1/2, \omega_2 = 3/10, \omega_3 = 1/5$ produced the minimum $V = 0.1248$ among all $Y_{i,j}^{S1}$s and $Y_{i,j}^{S2}$s which are based on $S_{i,j}$, and this is the minimum among all $Y_{i,j}^{R1}, Y_{i,j}^{R2}, Y_{i,j}^{S1}, Y_{i,j}^{S2}$ with possible weights. Fig. 5 is the contour plot of the values of V for various choices of weights of $Y_{i,j}^{S1}$. Therefore $Y_{i,j}^{S1}$ with $\omega_1 = 1/2, \omega_2 = 3/10, \omega_3 = 1/5$ is more consistent for clustering than any other proposed indices including the conventional correlation coefficients $(R_{i,j}, S_{i,j})$ according to the cross-validation criterion.

The clusters classified by $R_{i,j}$ are shown in Fig. 6, and the clusters classified by $Y_{i,j}^{S1}$ with $\omega_1 = 1/2, \omega_2 = 3/10$ and $\omega_3 = 1/5$ are shown in Fig. 7. The highest profile in cluster 4 of Fig. 6 was classified into cluster 7 of Fig. 7. The rest of the profiles grouped in cluster 4 of Fig. 6, which were determined by $R_{i,j}$, were divided into the first three clusters determined by $Y_{i,j}^{S1}$ (clusters 1–3 in Fig. 7). The profiles in cluster 1 of Fig. 7 increased from their minimum level time point 24 to the time point 48 whereas those in cluster 2 of Fig. 7 decreased from time point 24 to their minimum level time point 36. Two profiles in cluster 3 shown on Fig. 7 decreased on the final interval while clusters 1 and 2 had their profiles increasing in the same interval. Therefore the $R_{i,j}$-based method could not distinguish three different groups and combine them into one cluster.

For the comparison of clustering results obtained by these two methods with that of Peddada et al. (2003), one of the biologically significant analysis results for this particular data, we used the Rand index again. The Rand indices between the result of the order-restricted inference-based clustering method in Peddada et al. (2003) and those of various hierarchical linkage clustering methods using the new indices with different weights are shown in Table 4. The clustering results from the method using the new similarity measures (except for $Y_{i,j}^{S2}$) agreed better with that of Peddada et al. (2003) on the partitions of profiles than the result from the one using the conventional correlation coefficient did. The Rand indices between the result of Peddada et al. (2003) and those of complete-linkage hierarchical clustering method using $Y_{i,j}^{S1}$ with the optimal weights and $R_{i,j}$ were 0.9771 and 0.7665, respectively. The clustering results according to the method of Peddada et al. (2003), $R_{i,j}$ and $Y_{i,j}^{S1} = 0.5S_{i,j}^{*} + 0.3A_{i,j} + 0.2M_{i,j}$ are shown in Table 5 along with support for the proposed measure $Y_{i,j}^{S1}$'s better agreement with the result of Peddada et al. (2003).

4. Conclusions

Two time-course gene expression profiles are considered to be in the same group if they have similar shape and the max/min expression levels are measured at similar time points. The correlation coefficient is a measure of association commonly used in a distance based method for clustering profiles. As far as the two temporal profiles have similar expression levels at most of the time points, even though they have different directions for the slopes, the correlation coefficient could be high and the profiles might be put into the same class.

We considered four similarity measures $Y_{i,j}^{R1}, Y_{i,j}^{R2}, Y_{i,j}^{S1}, Y_{i,j}^{S2}$ that can preserve the information of the profile pattern to make up for the weakness of the correlation coefficient as a measure of association for clustering. They are based on either Pearson or Spearman correlation coefficient and the two indices representing the concordance of temporal profile patterns (shapes) and that of the time points at which maximum and minimum expression levels (extreme time points) are measured between two profiles respectively. We considered two types of profile shape concordance indices, $A_{i,j}$ and $A_{i,j}^{*}$. $A_{i,j}$ counts the number of time intervals in which there is an agreement in the sign of the change in each time interval between two profiles, and $A_{i,j}^{*}$ is the correlation coefficient between the changes of each profile in the time intervals. We also considered two types of concordance indices, $M_{i,j}$ and $M_{i,j}^{*}$ for the extreme time points. $M_{i,j}$ is an all-or-nothing measure which checks whether the maximum and minimum time points of two profiles are matched or not. $M_{i,j}^{*}$ on the other hand utilizes the actual distance between two profiles' time points where the max/min is attained.

The applications of the hierarchical clustering method to both synthetic noisy data and the real breast cancer cell line data showed that the proposed similarity measures, $Y_{i,j}^{R1}$ and $Y_{i,j}^{S1}$ are preferable to the conventional correlation coefficients. Applied to the small synthetic experiment and the real data in this research, $Y_{i,j}^{R1}$ and $Y_{i,j}^{S1}$ are better than $Y_{i,j}^{R2}$ and $Y_{i,j}^{S2}$ in clustering consistency, respectively. Especially $Y_{i,j}^{S1}$ outperforms the others including the conventional

correlation coefficients. The quantitative measures, $A^*_{i,j}$ and $M^*_{i,j}$ do not help much to improve the clustering ability when compared with their qualitative counterparts $A_{i,j}$ and $M_{i,j}$, respectively. As $A_{i,j}$ and $M_{i,j}$ are qualitative measures, it seems to be more reasonable to construct a new association measure ($Y^{S1}_{i,j}$) by combining them with the qualitative correlation coefficient ($S_{i,j}$) rather than the quantitative correlation coefficient ($R_{i,j}$).

$Y^{S1}_{i,j}$ was simple to implement, yet obtained very similar clustering result to that of the sophisticated statistical inference-based method (Peddada et al., 2003) for the breast cancer cell line data. Moreover it is much more consistent for clustering than the correlation coefficient according to the cross-validation criterion.

Acknowledgement

We are grateful for helpful comments from two referees that greatly improved the paper.

References

Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., Kämper, J., 2005. Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics 21, 1069–1077.

Chu, Y., DeRisi, J., Eisen, M., Mulholland, J., Bostein, D., Brown, P.O., Herskowtz, I., 1998. The transcriptional program of sporulation in budding yeast. Science 282, 699–705.

Datta, S., Datta, S., 2003. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 19, 459–466.

Hwang, J., Peddada, S.D., 1994. Confidence interval estimation subject to order restrictions. Annals of Statistics 22, 67–93.

Heyer, L.J., Kruglyak, S., Yooseph, S., 1999. Exploring expression data: Identification and analysis of coexpressed genes. Genome Research 9, 1106–1115.

Hoon, M.J.L., Imoto, S., Miyano, S., 2002. Statistical analysis of a small set of time-ordered gene expression data using linear splines. Bioinformatics 18, 1477–1485.

Lobenhofer, E.K., Bennett, L., Cable, P., Li, L., Bushel, P., Afshari, C.A., 2002. Regulation of DNA replication fork genes by 17 beta-estradiol. Molecular Endocrinology 16, 1215–1229.

Luan, Y., Li, H., 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics 19, 474–482.

Peddada, S.D., Lobenhofer, E.K., Li, L., Afshari, C.A., Weinberg, C.R., Umbach, D.M., 2003. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics 19, 834–841.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66, 846–850.

Schliep, A., Schonhuth, A., Steinhoff, C., 2003. Using hidden Markov models to analyze gene expression time course data. Bioinformatics 19, i255–i263.