Clustering of Unevenly Sampled Gene Expression Time-Series Data

C. S. Möller-Levet [a], F. Klawonn [b], K.-H. Cho [c], H. Yin [a], O. Wolkenhauer [d,*]

[a] *Department of Electrical Engineering and Electronics, University of Manchester Institute of Science and Technology, Manchester M60 1QD, U.K.*

[b] *Department of Computer Science, University of Applied Sciences, D-38302 Wolfenbüttel, Germany.*

[c] *School of Electrical Engineering, University of Ulsan, Ulsan 680-749, South Korea.*

[d] *Department of Computer Science, Systems Biology & Bioinformatics Group, University of Rostock, Albert-Einstein Str. 21, 18059 Rostock, Germany.*

Abstract

Time course measurements are becoming a common type of experiment in the use of microrarrays. The temporal order of the data and the varying length of sampling intervals are important and should be considered in clustering time-series. However, the shortness of gene expression time-series data limits the use of conventional statistical models and techniques for time-series analysis. To address this problem, this paper proposes the Fuzzy Short Time-Series (FSTS) clustering algorithm, which clusters profiles based on the similarity of their relative change of expression level and the corresponding temporal information. One of the major advantages of fuzzy clustering is that genes can belong to more than one group, revealing distinctive features of each gene's function and regulation. Several examples are provided to illustrate the performance of the proposed algorithm. In addition, we present the validation of the algorithm by clustering the genes which define the model profiles in Chu *et al* (1998, *Science*, vol. 284, pp. 699-705). The fuzzy c-means, k-means, average linkage hierarchical algorithm and random clustering are compared to the proposed FSTS algorithm. The performance is evaluated with a well-established cluster validity measure proving that the FSTS algorithm has a better performance than the compared algorithms in clustering similar rates of change of expression in successive unevenly distributed time points. Moreover, the FSTS algorithm was able to cluster in a biologically meaningful way the genes defining the model profiles.

Key words: Fuzzy clustering, Unevenly sampled, Short time series, Gene expression

1 Introduction

Microarrays revolutionise the traditional way of one gene per experiment in molecular biology [1]. With microarray experiments it is possible to measure simultaneously the activity levels of thousands of genes. An appropriate clustering of gene expression data can lead to meaningful classification of diseases, identification of co-expressed functionally related genes, logical descriptions of gene regulation, etc.

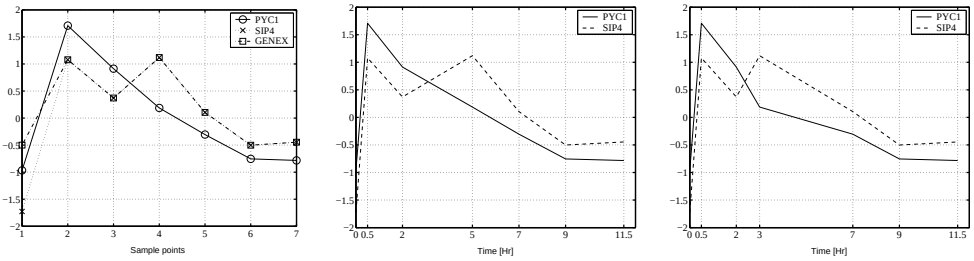
Time course measurements are becoming a common type of experiment in the use of microrarrays. The particularity of time-series, which has to be considered in the clustering analysis, is the temporal information: the measurements ordered in time and sampled at specific intervals. An appropriate similarity measure for gene expression time-series should be able to identify similar shapes which are formed by the relative change of expressions in temporal information.

This paper is organised as follows: The effects of the temporal information in the comparison of shapes are discussed first, followed by the related work. The next section defines the short time-series (STS) distance, develops the fuzzy short time-series (FSTS) algorithm using the standard fuzzy c-means algorithm (FCM) as a template, and provides simple examples to demonstrate its performance. Then, the validation of the algorithm is presented by clustering the genes which define the model profiles in [2]. The fuzzy c-means, k-means, average linkage hierarchical algorithm and random clustering are used for comparison. A well-established validity measure relevant to gene expression clustering is applied to evaluate the quality of the clusters. In addition, the results are discussed using the external biological criteria. Then, the scopes and limitations of the FSTS algorithm are discussed. Finally, conclusions are presented in the final section summarising the presented research.

2 Temporal Information and Clustering

To visualise the effects of the temporal information in the comparison of shapes consider the following example. The microarray analysis of *Saccharomyces cerevisae* by Chu *et al.* (1998) shows that PYC1 and SIP4 are two of the 52 genes that were induced rapidly and transiently after transfer to sporulation medium. PYC1 is involved in the gluconeogenesis pathway as a pyruvate carboxylase and SIP4 is a transcription factor, which interacts with the SNF1

* Corresponding author
Email address: ow@informatik.uni-rostock.de (O. Wolkenhauer).
URL: www.sbi.uni-rostock (O. Wolkenhauer).



(a) PYC1, SIP4 and GENEX

(b) Original time point measurements

(c) Altered length of sampling interval

Fig. 1. (a) When comparing the similarity of SIP4 and GENEX to PYC1, it can be observed that PYC1 and SIP4 have a more similar induction after transfer to sporulation medium. (b) PYC1 and SIP4 at the original sampling points. (c) PYC1 and SIP4 with a shortened length of sampling interval between time points three and four. In all figures the vertical axis correspond to the standardised log_{10}(expression ratio).

protein kinase. These genes were part of the handpicked genes selected for the "metabolic" model profile used in [2]. Consider a synthetic gene (GENEX), whose standardised [1] expression values are identical to those of SIP4 except for the first time point which has a higher expression. Figure 1(a) illustrates the resulting profile along with the standardised values of PYC1 and SIP4. When comparing the similarity of SIP4 and GENEX to PYC1, it can be observed that PYC1 and SIP4 have a more similar induction period after transfer to sporulation medium. That is, the relative change of expression from the first to the second measurement of PYC1 is more similar in SIP4 than in GENEX, while all the other changes are equal in SIP4 and in GENEX. However, when using the Euclidean distance to assess the similarity, PYC1 is more similar to GENEX ($d_E = 1.45$) than to SIP4 ($d_E = 1.57$). The Euclidean distance is invariant with respect to the order of the observations, therefore, the direction of change of expression (e.i. up-down) is not considered. The next element to consider is the length of sampling intervals. By including this not only the direction of the change is considered but also the rate of change. Biological processes are sampled at shorter intervals of time when intense biological activity is taking place. For example, the "metabolic" model profile is characterised by a rapid and transient induction after transfer to sporulation medium. To be able to identify this rapid induction, a short sampling interval between the first and the second measurement is required. Overall, SIP4 follows a similar profile to PYC1, except for the transition from the third to the fourth time point. Interestingly, the relevance of this difference in the comparison of the profiles is related to the length of the sampling interval. Figure 1(b) and (c) show the expression levels of PYC1 and SIP4 at the original sampling intervals and at modified intervals, respectively. Although the expression levels are the

[1] Standardised values (zero mean and standard deviation of one) are utilised to eliminate shifting and scaling factors.

same for both cases, the similarity between the profiles is different, since the difference of the profiles between time points three and four is emphasised by the short sampling interval in (c). That is, having the same absolute values of expression, the rate of change is increased with smaller sampling intervals. However, the Euclidean distance and the correlation coefficient discard the length of the sampling interval. In addition, given the shortness of the expression time-series, a large correlation coefficient does not necessarily indicate two similar profile shapes, nor does a small correlation coefficient necessarily indicate different profile shapes [3].

Motivated by the need to differentiate these cases and considering the scarce number of time points, we introduce a new similarity measure, which can capture the temporal information of the time-series data to evaluate the similarity of temporal gene expression profiles. Based on the advantages of fuzzy clustering for extracting biological insights [4], we have developed the fuzzy short time-series (FSTS) clustering algorithm. One of the most significant advantages of fuzzy clustering is that genes can belong to more than one group, revealing distinct aspects of their functions and regulations. Other advantages include its intuitive biological interpretation, and its conceptual, computational and algorithmic simplicity, which are usually the main disadvantages of model-based algorithms.

2.1 Related work

The idea of conserving the temporal order in the measurements have been treated recently by several authors using algorithms with different degrees of complexity. In [5] a similarity function for time-series is proposed according to up-down weighted patterns of filtered series where the filtering process has several user-defined thresholds. In [3], the proposed algorithm is based on the statistical theory of ordered-restricted inference that makes explicit use of ordering information. Candidate temporal profiles are defined in terms of inequalities among mean expression levels at the time points. In [6], a first order autoregressive model is used to represent the clusters following an agglomerative clustering. In [7], a two-step approach is introduced to identify groups of points ordered in a line configuration in particular locations and orientations of the data-space. These groups correspond to expressions in the time domain which have the same parameters of linear transformation between successive time points. In [8] a clustering method based on hidden Markov models is presented, the approach assumes that in each group, gene expressions are generated by a Markov chain with certain probability models. All of these methods consider the temporal order but do not include the length of sampling interval in the analysis. In [9], cubic interpolation is used to reduce the non-uniform spacing between measurements, however, the work is not in-

tended to assess the similarity of the profiles. Some researches have already addressed this important issue in model-based clustering. For example, in [10] statistical spline estimation are used to represent time-series gene expression profiles as continuous curves. The method takes account of the actual duration each time point represents and weights time points differently according to the sampling interval. However, the method does require data to be sampled at a sufficiently high rate, hence, the authors use one of the largest data sets available. Although the cubic splines are commonly used, they are not suitable for short time-series [11]. Later, in [12], time-course gene expression data is clustered using a mixed-effect model with B-splines. The authors utilised "long" gene expression time-series (12 and 18 time points) and four equally spaced knots. However, it is not always possible to use equally spaced knots if the series are unevenly sampled and equally spaced knots cannot properly reflect the unevenly distributed time points. In this paper we present an intuitive but systematic approach to include the temporal information in the comparison of shapes taking advantages of already well-established fuzzy c-means algorithm. We illustrate the algorithm with several simulated data sets and validated it using a real experimental data set, i.e. Chu *et al.* (1998) *S. cerevisiae* data set.

3 Algorithm and Implementation

3.1 Short Time-Series Distance

This section presents a measure of similarity for microarray time-series data. The proposed similarity measure is driven by the concept of similarity and the particular characteristics of the time-series generated with microarray experiments. First, there is no clear definition of what "similar" time-series are in a biological context. However, it is generally understood that similar expression profiles correspond to similar shapes of expression. Therefore, it is a common practice to use lines between time points rather than isolated points for aiding a visual comparison. Second, these series have two main properties given by the nature of the experiments generating them: they are short and usually unevenly sampled. When the time-series are short, traditional statistical analyses are not always suitable. For example, in the case of an autoregressive model [6], the order of the model is very restricted by the low number of time points available in gene expression time-series. In [11] the authors identified that conventional techniques for time-series analysis, such as Fourier analysis or autoregressive or moving-average modelling are not suitable for the small number of data points as in most gene expression time-series data. As an alternative, the authors proposed to model the time-series with linear splines. The problem of short time-series has being identified in other fields and has been treated with a particular focus on shape comparisons [13], that is, using

the idea of up-down patterns.

The objective here is to define a similarity measure that can capture the temporal information to evaluate the similarity of temporal gene expression profiles. We approach the problem by considering the time-series as piecewise linear functions and measuring the difference of slopes between them. In [14], the expression level at each time point and the slopes between time points are included in the comparison of profiles. However, the slopes were calculated based on a reduced time interval of one, not taking into account the variable time intervals. By measuring the difference of the true slopes, we are able to include in a meaningful way the length of sampling intervals, while considering the shape (i.e. up-down patterns) of the series. The length of sampling interval can be understood as a weight; the farther apart in time the expressions are, the less weight they have in the comparison. Considering a gene expression profile $x = [x_1, x_2, \ldots, x_{n_t}]$, where n_t is the number of sampling time points, the linear function $x(t)$ between two successive time points t_k and $t_{(k+1)}$ can be defined as $x(t) = \beta_k t + \delta_k$, where $t_k \leq t \leq t_{(k+1)}$, and $\beta_k = (x_{(k+1)} - x_k)/(t_{(k+1)} - t_k)$ and $\delta_k = (t_{(k+1)}x_k - t_k x_{(k+1)})/(t_{(k+1)} - t_k)$. The proposed STS distance corresponds to the square root of the sum of the squared differences of the slopes obtained by considering time-series as linear functions between measurements. The STS distance between two time-series x and v is thus defined as:

$$d^2_{\mathrm{STS}}(x, v) = \sum_{k=1}^{n_t} \left(\frac{v_{(k+1)} - v_k}{t_{(k+1)} - t_k} - \frac{x_{(k+1)} - x_k}{t_{(k+1)} - t_k} \right)^2. \quad (1)$$

The STS distance is able to identify that SIP4 is more similar to PYC1 than GENEX and it considers the length of sampling intervals.

3.2 Fuzzy Short Time-Series Clustering Algorithm

The wide variety of clustering algorithms available from various disciplines are distinguished by the way in which they measure distances between objects and the way they group the objects based upon the measured distances [15]. In the previous section we already discussed the way in which we desire the "distance" between objects to be measured; hence, in this section, we focus on grouping the objects based upon the measured distance. For this purpose we select the fuzzy clustering scheme as a template for our development, since fuzzy sets are a more realistic approach to address the concept of similarity than classical sets. A classical set has a crisp or hard boundary where the constituting elements have only two possible values of membership. In contrast, a fuzzy set has fuzzy boundaries where each element is given a degree of membership providing information about the influence of a given gene for the overall characteristics of the cluster. In addition, a fuzzy approach inherently

accounts for noise in the data because it extracts trends, not precise values.

Fuzzy clustering is a partitioning-optimisation technique [16–18]. The objective function that measures the desirability of partitions is described by,

$$J(x, v, u) = \sum_{i=1}^{n_c} \sum_{j=1}^{n_g} u_{ij}^{w} \, d^2(x_j, v_i) \tag{2}$$

where n_c is the number of clusters, n_g is the number of vectors to cluster, u_{ij} is the value of the membership degree of the vector x_j to the cluster i, $d^2(x_j, v_i)$ is the squared distance between vector x_j and prototype v_i, and w is a parameter (usually set between 1.25 and 2), which determines the degree of overlap of fuzzy clusters.

The minimisation of the fuzzy objective function is a nonlinear optimisation problem that can be solved using various methods. The most common method is the Picard Iteration through the first-order conditions for stationary points of the function. Figure 2 illustrates the iterative procedure of the fuzzy c-means algorithm, the most well known fuzzy clustering algorithm. Considering other fuzzy extensions, the convergence is independent of the change in the distance function if the distances are all positive and the prototypes are calculated accordingly to the minimisation of the objective function. A full review of the minimisation and convergence of the FCM objective function can be found in [19].

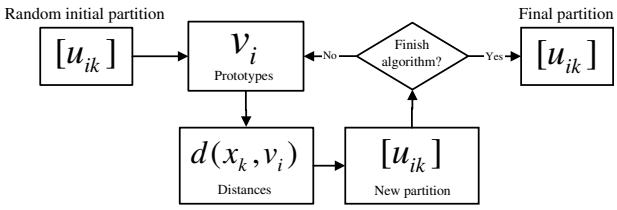


Fig. 2. Diagram of the iterative procedures for the FCM clustering algorithm. Considering the partition of a set $X = [x_1, x_2, \ldots, x_{n_g}]$, into $2 \leq n_c < n_g$ clusters, the fuzzy clustering partition is represented by a matrix $U = [u_{ik}]$, whose elements are the values of the membership degree of the object x_k to the cluster i, $u_i(x_k) = u_{ik}$.

In order to integrate the STS distance into the conventional fuzzy clustering scheme, it is necessary to obtain the value of the prototype v_i that minimises (2), when (1) is used as the distance. Substituting (1) into (2) we obtain

$$J(x, v, u) = \sum_{i=1}^{n_c} \sum_{j=1}^{n_g} u_{ij}^{w} \sum_{k=1}^{n_t} \left(\frac{v_{i(k+1)} - v_{ik}}{t_{(k+1)} - t_k} - \frac{x_{j(k+1)} - x_{jk}}{t_{(k+1)} - t_k} \right)^2 . \tag{3}$$

The partial derivative of (3) with respect to v_{ik} is:

$$\begin{aligned}\frac{\partial J(x,v,u)}{\partial v_{ik}} = &\sum_{j=1}^{g} 2u_{ij}^{w} \frac{\left(a_k v_{i(k-1)} + b_k v_{ik} + c_k v_{i(k+1)}\right)}{\left(t_k - t_{(k+1)}\right)^2 \left(t_k - t_{(k-1)}\right)^2} + \\ &\sum_{j=1}^{g} 2u_{ij}^{w} \frac{\left(d_k x_{j(k-1)} + e_k x_{jk} + f_k x_{j(k+1)}\right)}{\left(t_k - t_{(k+1)}\right)^2 \left(t_k - t_{(k-1)}\right)^2}\end{aligned} \tag{4}$$

where

$$\begin{aligned} a_k &= -(t_{(k+1)} - t_k)^2 \quad b_k = -(a_k + c_k) \quad c_k = -(t_k - t_{(k-1)})^2 \\ d_k &= (t_{(k+1)} - t_k)^2 \quad e_k = -(d_k + f_k) \quad f_k = (t_k - t_{(k-1)})^2 \,. \end{aligned}$$

Setting (4) equal to zero and solving for v_{ik} we have

$$a_k v_{i(k-1)} + b_k v_{ik} + c_k v_{i(k+1)} = m_{ik} \tag{5}$$

where

$$m_{ik} = -\frac{\sum_{j=1}^{n_g} u_{ij}^{w} \left(d_k x_{j(k-1)} + e_k x_{jk} + f_k x_{j(k+1)}\right)}{\sum_{j=1}^{n_g} u_{ij}^{w}} \,.$$

Equation (5) yields an underdetermined system of equations. We know the relations of the prototype values among the time points, but not the absolute value at each time point. That is, we know the slope but not the absolute level. By adding two known fixed time points we can solve the underdetermined system of equations for any n_t. If we add the same two time points to all the time-series the similarity is not altered. If the fixed values are zero, a general solution is easier to obtain. The length of the sampling interval between the first real time point and the last fixed time point acts as a weight to the first real time point. Additional time points should be $t_1 = -1$ and $t_2 = 0$, and the original time points should be scaled down by substraction to start as $t_3 = 1$. This avoids altering v with the added fixed time points, since with this configuration, the values of a, c and f for the extra time points equal one and do not affect the products. The prototypes can be calculated as shown in the following equation,

$$\begin{aligned} v(i,n) = \sum_{r=2}^{n-3} [m_{ir} \prod_{q=1}^{r-1} c_q \left(\prod_{q=r+1}^{n-1} a_q + \prod_{q=r+1}^{n-1} c_q + \sum_{p=r+3}^{n} \prod_{j=p-1}^{n-1} c_j \prod_{j=r+1}^{p-2} a_j \right) / \prod_{q=2}^{n-1} c_q] + \\ [m_{i(n-1)} \prod_{q=1}^{n-2} c_q + m_{i(n-2)} \prod_{q=1}^{n-3} c_q (a_{(n-1)} + c_{(n-1)})] / \prod_{q=2}^{n-1} c_q \end{aligned} \tag{6}$$

where $1 \leq i \leq n_c$, $3 \leq n \leq n_t + 2$ (since $v(i,1) = 0$ and $v(i,2) = 0$), $m_{i1} = 0$ and $c_1 = 1$.

STEP 1: Initialisation

n_g : number of genes
n_t : number of time points
X : gene expression matrix (GEM) $[n_g \times n_t]$
n_c : number of clusters
w : fuzziness parameter
α : threshold for membership
ϵ : termination tolerance
t : time points

STEP 2: Addition of two fixed time points

$t_1 = -1$,
$t_2 = 0$,
$t_{k+2} = t_k - t_1 + 1 \quad \text{for} \quad 1 \leq k \leq n_t$,
$X = [0_{[n_g \times 2]} \; X_{[n_g \times n_t]}]$.

STEP 3: Initialisation of the partition matrix

Initialise the partition matrix randomly, $U^{(0)}$ $[n_c \times n_g]$.

STEP 4: Repeat for $l = 1, 2, ...$

3.1 Compute the cluster prototypes:
$v(i,1)^{(l)} = 0$,
$v(i,2)^{(l)} = 0$,
For $v(i,n)^{(l)}$ use Equation (6) $1 \leq i \leq n_c$, $3 \leq n \leq n_t + 2$.
3.2 Compute the distances using Equation (1), $1 \leq i \leq n_c$ and $1 \leq j \leq n_g$.
3.3 Update the partition matrix:
if $d_{\mathrm{STS}ij} > 0$ for $1 \leq i \leq n_c$, $1 \leq j \leq n_g$, use Equation (7);
otherwise $u_{ij}^{(l)} = 0$ if $d_{\mathrm{STS}ij} > 0$, and $u_{ij}^{(l)} \in [0,1]$ with $\sum_{i=1}^{n_c} u_{ij}^{(l)} = 1$.

Until $\|U^{(l)} - U^{(l-1)}\| < \epsilon$.

Fig. 3. Pseudocode of the FSTS clustering algorithm.

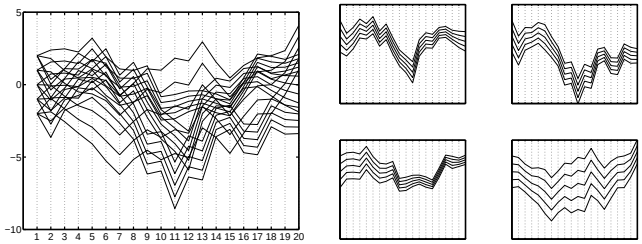
The change of the distance function has no effect in the optimisation of (2) with respect to the membership degree, therefore, u_{ij} can be calculated as in the FCM algorithm,

$$u_{ij} = \frac{1}{\sum_{q=1}^{n_c} \left(d_{\mathrm{STS}}(x_i, v_j)/d_{\mathrm{STS}}(x_i, v_q)\right)^{1/(w-2)}}. \tag{7}$$

The same scheme of the iterative procedure as for the FCM, described in Figure 2 is followed, but the distance and the prototypes are calculated using (1) and (6), respectively. Figure 3 provides the pseudocode of the FSTS clustering algorithm. The three user-defined parameters in the FSTS algorithm, i.e., the number of clusters n_c, the threshold of membership to form the final crisp clusters α, and the fuzziness parameter w are the same as in the FCM. The selection of w for an optimal performance of fuzzy clustering for microarray data is addressed in [20].

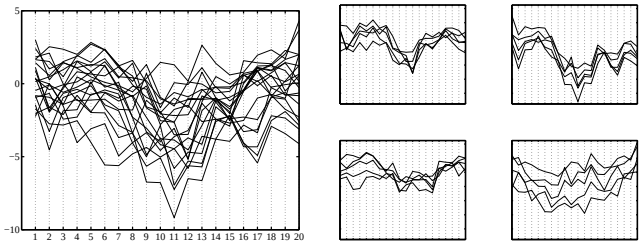
3.3 Illustrative examples

Simulated data sets are used to illustrate the FSTS clustering algorithm. In the figures displaying the simulated data sets, Figures 4, 5, and 7, the horizontal



(a) Simulated Data (b) Constituting clusters

Fig. 4. All the algorithms tested (FCM, FSTS, KM and HC) are able to identify the constituting clusters.



(a) Simulated Data with added noise (b) Constituting clusters

Fig. 5. Only the fuzzy clustering algorithms (FCM and FSTS) are able to identify successfully the constituting clusters.

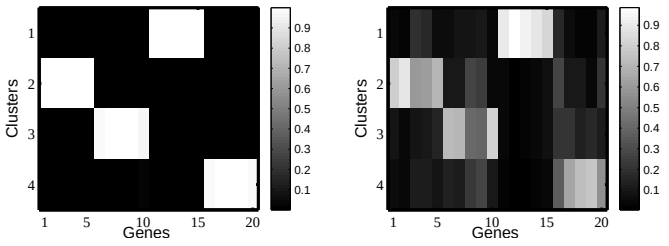
axis denotes time and the vertical axis denotes the expression level.

For the first test the data set is created as described in the following. Four groups are created where each group of five time-series has the same parameters of linear transformation between time points, as shown in Table 1. That is, for the group i, $1 \leq i \leq 4$, $x_{j(k+1)} = a_{ik}x_{jk} + b_{ik}$ with $1 \leq k < n_t$ and $1 \leq j \leq 5$. The values of a and b were obtained randomly from normal distributions ($\mu_a = 1$, $s.d._a = 0.15$ and $\mu_b = 0$, $s.d._b = 1$.) for each group.

Table 1
Simulated profile $x = [x_1, x_2, \ldots, x_{n_t}]$. A group of time-series with similar shapes can be obtained by changing the initial value.

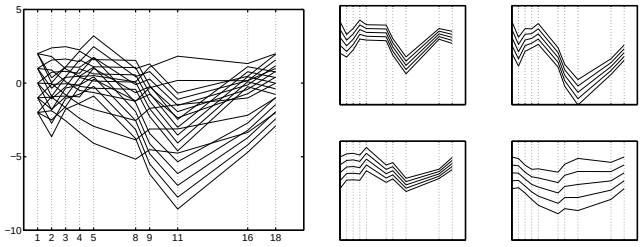
Time points	Value
x_1	initial value
x_2	$a_2x_1 + b_2$
x_3	$a_3x_2 + b_3$
$\vdots$	$\vdots$
x_{n_t}	$a_{n_t}x_{(n_t-1)} + b_{n_t}$

The resulting simulated time-series, shown in Figure 4, is clustered using FCM, FSTS, k-means (KM) and average linkage hierarchical clustering (HC) algorithms. The number of clusters was set to four for all the algorithms. The Euclidean distance was used for KM and the correlation coefficient for HC.



(a) Simulated data set (b) Simulated data set with added noise

Fig. 6. FSTS clustering results, the membership degree of each gene to each cluster is mapped into a grayscale. The genes are ordered in the horizontal axis according to the original clusters.



(a) Unevenly resampled simulated data (b) Constituting clusters

Fig. 7. Only the FSTS algorithm is able to identify the constituting clusters successfully.

All the algorithms are able to identify the four clusters successfully (FCM 41 out of 50 runs, FSTS 50 out of 50 runs, KM 23 out of 50 runs and HC 50 out of 50 runs). The clustering parameters are $w = 1.6$ and $\alpha = 0.4$ for the two fuzzy algorithms.

In the second test we added (independent) random noises at each measurement of the data set shown in Figure 4. Normal variations with mean zero and standard deviation 0.65 were used. The resulting data set is displayed in Figure 5. The number of clusters was set to four for all the algorithms. In this case only the fuzzy clustering algorithms (FCM and FSTS, $w = 1.65$ and $\alpha = 0.35$) were able to identify successfully the constituting clusters. Figure 6 presents the FSTS clustering results of the first and the second test using membership plots. These plots map the membership degrees into grayscales. Here, the genes are ordered in the horizontal axis according to the original clusters. It can be seen that for the original data set the memberships are well defined being all above 0.9, while for the noisy data set the memberships are more spread but the clusters can still be identified with an α value of 0.35. Fuzzy clustering provides a natural approach to dealing with noise. In contrast, hard clustering assumes that the constituting clusters are well defined, that is, the elements can only belong to one cluster.

The third test considers a subset of the simulated data set shown in Figure 4. The original data set was "resampled" selecting 10 time points randomly out of the 20 original time points. The resulting data set, simulating unevenly sampled time-series, is shown in Figure 7. The number of clusters was set to four for all the algorithms. In this case, only the FSTS algorithm ($w = 1.2$ and $\alpha = 0.6$) can identify the four clusters successfully, while FCM and HC can only identify two clusters correctly and KM does not produce consistent results. Choosing different parameters for the FCM did not lead to any improvement. The original information of each series is kept in the closest time points, and lost as time points are farther apart. The FSTS can extract this information by weighting the comparisons with the length of sampling intervals.

4 Clustering of Temporal Profiles

We clustered a subset of the microarray data on the transcriptional program of sporulation in budding yeast collected and analysed by Chu *et al.*, (1998) [2]. DNA microarrays containing 97% of the known or predicted genes of *Saccharomyces cerevisiae* were used to explore the temporal program of gene expression during meiosis and spore formation. Changes in the concentrations of mRNA transcripts from each gene were measured at seven uneven time intervals. The authors distinguished seven temporal patterns of induced transcription. They chose a set of representative genes from each of the seven expression patterns, and the average for each set was calculated to create the seven model profiles. We clustered the genes used to produce these models, these genes are listed in Table 2. The available external validation and unevenly sampling intervals of this small set of genes make it ideal for our comparative study. As in [2], the ratio of each gene's mRNA level to its mRNA level in vegetative cells just before transfer to sporulation medium was calculated, followed by a log-transformation.

4.1 Cluster Validation

As observed in [21], not all related genes are similarly expressed, and some unrelated genes have similar expression patterns. Therefore, external biological validation cannot be used as the only means to identify the best choice of similarity measure and clustering algorithm. In this paper we utilised a validity index in addition to the external biological validation.

[2] Available from http://cmgm.stanford.edu/pbrown/sporulation.

Table 2
Experimental data set: Temporal model profiles.

Metabolic	Early I	Early II	Early-Mid	Middle	Middle-Late	Late
ACS1	ZIP1	KGD2	YBL078C	YSW1	CDC27	SPS100
PYC1	YDR374C	AGA2	QRI1	SPR28	DIT2	YKL050C
SIP4	DMC1	YPT32	PDS1	SPS2	DIT1	YMR322C
CAT2	HOP1	MRD1	APC4	YLR227C		YOR391C
YOR100C	IME2	SPO16	KNR4	ORC3		
CAR1		NAB4	STU2	YLL005C		
		YPR192W	YNL013C	YLL012W		
			EXO1			

Cluster validity measures the goodness of a clustering relative to others created by other clustering algorithms, or by the same algorithms using different parameter values. In this study the correctness of the clustering results is quantitatively evaluated by one of the well-established validity methods that has been used for gene expression data. In addition, random clustering [3] is utilised as a control in the comparison [22].

The Dunn's validity index [23] identifies clusters that are "compact and well separated". The index for a specific number of clusters n_c is defined as

$$\mathrm{D}_{n_c} = \min_{1\le i\le n_c}\left(\min_{1\le j\le n_c, j\neq i}\left(\frac{\delta(c_i,c_j)}{\max\limits_{1\le k\le n_c}\Delta(c_k)}\right)\right) \tag{8}$$

where $\delta(c_i, c_j)$ defines the distance between clusters c_i and c_j (intercluster distance); $\Delta(c_k)$ represents the intracluster distance of cluster c_k. The distance function used in this study is the proposed STS distance, as it is relevant to the clustering of unevenly sampled time-series. This validity index has been used as a meaningful validation method for gene expression data by Bolshakova and Azuaje (2003). Large values of the index indicate the presence of compact and well-separated clusters, in this case, the compactness and separation refer to the slopes formed between time points.

The index was used to evaluate the results of the simulated data sets of the previous section. Tables 3 to 5 show the results.

4.2 Methods and Results

First, the data set is standardised to remove shifting and scaling factors. Next, it is clustered using FSTS, FCM, KM, the average linkage HC algorithm and

[3] Random clustering is a random grouping of the data into a predefined number of clusters.

Table 3
Summary of the Dunn's Index for FSTS, FCM, KM, HC and random clustering of first simulated data.

	Mean	Median	Std. Dev.	Coef. of Var.
FSTS	1.4660	1.4660	0	0
FCM	1.1890	1.4660	0.5598	0.4708
KM	0.9071	1.4660	0.6864	0.7566
HC	1.4660	1.4660	0	0
Random	0.0388	0.0384	0.002	0.0515

Table 4
Summary of the Dunn's Index for FSTS, FCM, KM, HC and random clustering of second simulated data.

	Mean	Median	Std. Dev.	Coef. of Var.
FSTS	0.6614	0.6614	0	0
FCM	0.6614	0.6614	0	0
KM	0.4794	0.4898	0.131	0.2732
HC	0.4843	0.4843	0	0
Random	0.2445	0.2388	0.0190	0.0777

Table 5
Summary of the Dunn's Index for FSTS, FCM, KM, HC and random clustering of third simulated data.

	Mean	Median	Std. Dev.	Coef. of Var.
FSTS	0.8093	0.8093	0	0
FCM	0.1833	0.1581	0.0551	0.3006
KM	0.3384	0.1581	0.3050	0.9013
HC	0.1520	0.1520	0	0
Random	0.0282	0.0275	0.0012	0.0425

random clustering. The clustering parameters are $w = 1.6$ and $\alpha = 0.4$ for the two fuzzy algorithms. As for the simulated data set, the Euclidean distance was used for KM and the correlation coefficient for HC. The number of clusters is set to seven, since that is the number of model profiles constituting the data set. The data set is clustered by all the above algorithms for 50 times and the Dunn's index value of each method is calculated as described in (8). Table 6 summarises the results showing that the Dunn's index for the FSTS results have the highest value.

Table 6
Summary of the Dunn's Index for FSTS, FCM, KM, HC and random clustering of experimental data.

	Mean	Median	Std. Dev.	Coef. of Var.
FSTS	0.132	0.132	0.004	0.028
FCM	0.111	0.106	0.007	0.063
KM	0.090	0.080	0.026	0.285
HC	0.113	0.113	0	0
Random	0.024	0.024	0.001	0.067

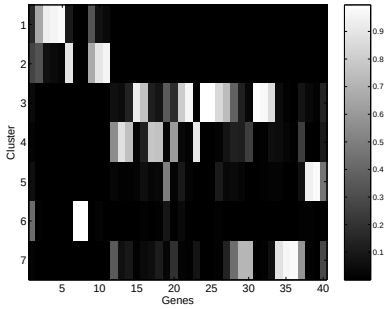
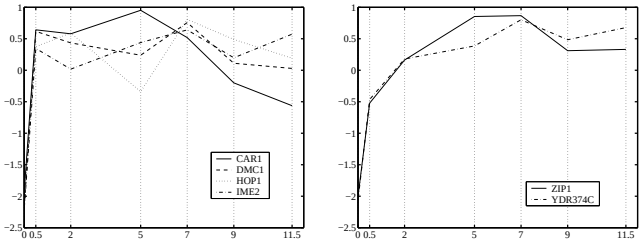


Fig. 8. FSTS clustering results, the membership degree of each gene to each cluster is mapped into a grayscale. The genes are ordered based on Table 2.



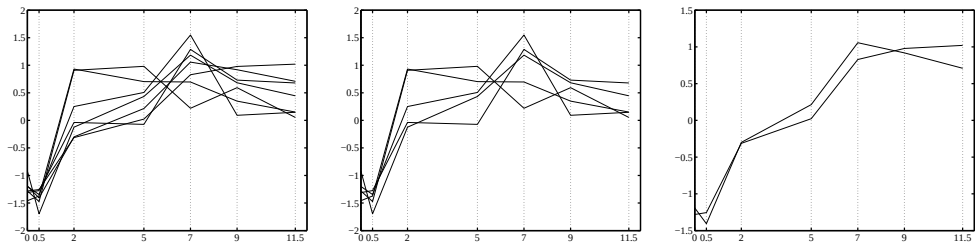
(a) Cluster 2 (b) ZIP1 and YDR374 in cluster 6

Fig. 9. Early I model profiles.

Although the validity measures are a useful tool for assessing clustering results, the results are discussed in the following using the external biological criteria. Figure 8 presents the FSTS clustering results using a membership plot. The genes are ordered based on Table 2. In Figures 9, 10, 11, and 12, the horizontal axis denotes time and the vertical axis denotes the normalised log_{10}(expression ratio).

Metabolic model profile. Four (PYC1, SIP4, CAT2, YOR100C) of the six genes defined for this profile are clustered together forming cluster 1. CAR1 has a higher membership degree to cluster 2, which is formed by Early I genes. Cluster 2 is illustrated in Figure 9(a). CAR1 is not induced rapidly and transiently after transfer to sporulation medium, instead it has a sustained induction after 0.5 hr., peaking at time 5 hr. ACS1 does not show a high membership for a single cluster as illustrated in Figure 8 (gene 1), its membership is mainly distributed between cluster 2 and cluster 6, both formed by Early I genes.

Early I model profile. The genes forming this model profile are characterised by an early induction, detectable 0.5 hour after transfer to sporulation medium, and sustained expression throughout the rest of the time course. Three (DMC1, HOP1, and IME2) of the five genes defined for this model profile were clustered together in cluster 2, illustrated in figure 9(a). The other two genes which form this model, ZIP1 and YDR374C shown in Figure 9(b), are grouped together in cluster 6. These two genes are very similar to each other and well



(a) Cluster 4 (b) Early II profiles in (a) (c) Early-Mid profiles in (a)

Fig. 10. Cluster 4 is formed by genes belonging to the Early II and Early-Mid model profiles.

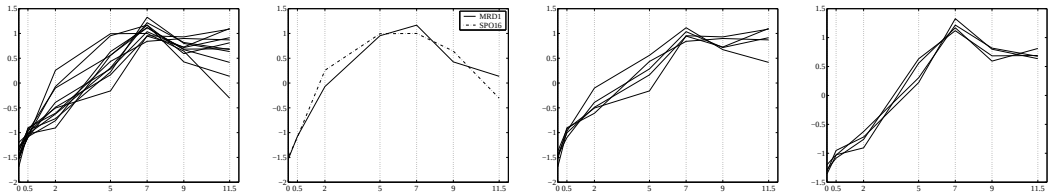
separated from most of the other genes, as illustrated in Figure 8 (genes 7 and 8). The difference between the Early I genes in cluster 6 and cluster 2 is that the formers present an increase in expression until the 7 hour rather than an early sustained one.

Early II model profile. This group is defined by a slightly delayed increase in transcript levels compared to the Early I. These genes were clustered together in cluster 4 illustrated in Figure 10, except for two genes. MRD1 and SPO16 are contained in cluster 3 which is formed by Early-Mid and Middle genes and is shown in Figure 11. These two genes follow a continuously increasing induction until the 7th hr. followed by a decrease of expression. In contrast, the Early II genes belonging to cluster 4 show an fast induction at time 2 hr. and a sustained expression throughout the rest of the time course.

Early-Mid model profile. This group has an early and a posterior middle induction around times 2 hr. and 5-7 hr. These genes were cluster together in cluster 4 (QRI1 and KNR4) and cluster 3 (PDS1, APC4, STU2, YNL013C and EXO1), illustrated in Figures 10 and 11, respectively. QRI1 and KNR4 have a repressed or low expression after transfer to sporulation medium while the Early-Mid genes in cluster 3 are induced at this time. YBL078C does not show a high membership for a single cluster, its membership is mainly distributed between cluster 3 and cluster 5 as illustrated in Figure 8 (gene 19).

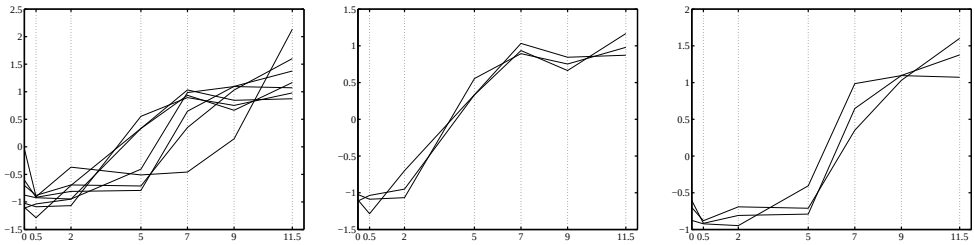
Middle model profile. The genes were strongly induced between 2 and 5 hours. These genes are found in clusters 3 (YSW1, ORC3, YLL005C and YLL012W) and 7 (SPR28, SPS2 and YLR227C), illustrated in Figure 11 and 12, respectively. The difference between the Middle genes in cluster 3 and cluster 7 is in the first three time points; the genes in cluster 3 are induced earlier that those in cluster 7.

Mid-Late model profile. The genes were induced between 5 to 7 hours. The three genes which form this model profile are clustered together in cluster 7 as shown in figure 12.



(a) Cluster 3 (b) MRD1 and SPO16 in (a) (c) Early-Mid profiles in (a) (d) Middle profiles in (a)

Fig. 11. Clusters 3 is formed by genes belonging to the Early II (MRD1 and SPO16), Early-Mid and Middle model profiles.



(a) Cluster 7 (b) Middle profiles in (a) (c) Mid-Late profiles in (a)

Fig. 12. Cluster 7 contains the genes belonging to the Middle and Mid-Late model profiles.

Late model profile. These genes were induced between 7 and 11.5 hours. SPS100 does not show a high membership for a single cluster, its membership is mainly distributed between cluster 4 and cluster 7, as illustrated in Figure 8 (gene 37). YKL050, YMR322C and YOR391C are clustered together in cluster 5.

The FSTS clustering algorithm was able to cluster the 40 genes which form the above seven temporal profiles into seven groups of similar expressions. Several clusters overlap and others are well differentiated. The Chu *et al.* data set has being previously analysed by several authors. In [24], the clustering method of [2] is modified to incorporate bootstrapping and assess the reliability of the results in terms of the stability of the genes. The authors observed high correlation between profiles, specially between Early-Middle and Middle profiles and conclude that these two clusters are too similar to be readily distinguished. These high correlations can be observed in the FSTS clustering results. Early II and Mid-Early profiles are combined in cluster 4, Mid-Early and Middle profiles are combined in cluster 3 and Middle and Mid-Late profiles are combined in cluster 7. It can be observed in Figure 8 that the elements belonging to these profiles show a spread membership among clusters 3, 4 and 7, while Metabolic and Early I profiles are well distinguished in clusters one and two.

5 Discussion

The STS distance is not sensitive to shifting but it is sensitive to scaling, which can be solved by standardising the data set. However, the length of the sampling interval between the last fixed time point and the first real time point acts as a weight to the first time point. If two series are identical but have an offset, this offset will be measured indirectly by the slope between the last fixed time point and the first real time point. So, future work could involve the study of the meaning of "parallel" series after standardisation. A simple solution to avoid weighting the first time point could be the normalisation of the data set by the first value.

In fuzzy clustering, the crisp clustering results are dependent on the cutoff point α. One possible solution for the impact of the selection of a global α could be the choice of local α values, that is, different thresholds for different clusters. With the membership plot presented in this paper it is possible to globally visualise the clustering results, so helping the selection of α. Figure 8 shows the genes ordered by classes in the horizontal axis because the classes are already known. When the classes are unknown, a hierarchical clustering of the membership degrees can be performed to order the genes and obtain a better visualisation of the results. In this way, the membership plot can be utilised to identify genes with similar distribution of membership degree across all the clusters.

One of the advantages of fuzzy clustering is that it provides useful tools for evaluating the results by the use of the membership matrix. In this paper, the number of clusters was known in advance in both simulated and real cases, nevertheless, there are several well established methods to estimate the number of clusters, in particular for fuzzy clustering, see for example [25].

6 Conclusions

Clustering algorithms have been developed for various applications and within a range of disciplines. In order to choose the most suitable algorithm for a particular application, the type of experiment and the specific purposes of the research have to be considered. The concept of similarity is at the core of any clustering algorithm and terms such as co-expression and "similar profiles" have to be well defined within the biological context. In this paper we have introduced a metric in which the similarity is based on the rate of change of expression levels across time, which is an intuitive biological idea of similar behavior across time.

Conventional clustering algorithms, based in the Euclidian distance or correlation coefficient are not able to properly reflect the temporal information embedded in the distance metric. We tackled the problem by considering the time-series as piecewise linear functions and measuring the difference of slopes between the functions. Fitting a higher order function will be restricted by the few number of time points. The proposed criterion is intended to provide a simple metric which can be easily interpreted and addresses the particular characteristics of these time-series. Based on the advantages of fuzzy clustering for identifying distinct features of genetic function and regulation [4], we have developed the FSTS algorithm incorporating the new distance measure in the fuzzy-c-means clustering scheme. The FSTS clustering algorithm presents a new approach to cluster time-series. It is particularly suited for short and unequally sampled time-series, a situation that commonly occurs in many practical situations, particularly in biological experiments.

We illustrated the algorithm with simulated data sets and validated it using a subset of the microarray data on the transcriptional program of sporulation in budding yeast. The FSTS algorithm was able to cluster in a biologically meaningful way the genes which define the model profiles defined in [2]. In addition, the FSTS clustering algorithm showed better performance than the conventional algorithms in the clustering of similar rates of change of expression across unevenly distributed time points.

Acknowledgements

The authors would like to thank anonymous reviewers for their helpful comments. C. Möller-Levet was supported by grants from ABB Ltd. U.K., an Overseas Research Studentship (ORS) award by Universities U.K. and Consejo Nacional de Ciencia y Tecnologia (CONACYT); K.-H. Cho acknowledges the support received by a grant from the Korea Ministry of Science and Technology (Korean Systems Biology Research Grant, M10309000006-03B5000-00211) and also by grant No. (R05-2004-000-10549-0) from the Korea Ministry of Science and Technology; and O.Wolkenhauer was supported by a grant from the UK Department for the Environment, Food and Rural Affairs (DEFRA). The work was conducted in collaboration with the Veterinary Laboratories Agency (VLA), Weybridge.

References

[1] P. Brown, D. Botstein, Exploring the new world of the genome with DNA microarrays, Nature Genetics supplement 21 (1999) 33–37.

[2] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. Brown, I. Herskowitz, The transcriptional program of sporulation in budding yeast, Science 282 (1998) 699–705.

[3] S. D. Peddada, E. K. Lobenhofer, L. Li, C. A. Afshari, C. R. Weinberg, D. M. Umbach, Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference, Bioinformatics 19 (7) (2003) 834–841.

[4] A. P. Gasch, M. B. Eisen, Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering, Genome Biology 3 (11) (2002) 0059.1–0059.22.

[5] V. Filkov, S. Skiena, J. Zhi, Analysis techniques for microarray time-series data, Journal of Computational Biology 9 (2) (2002) 317–330.

[6] M. F. Ramoni, P. Sebastiani, I. S. Kohane, Cluster analysis of gene expression dynamics, PNAS 99 (14) (2002) 9121–9126.

[7] C. S. Möller-Levet, K.-H. Cho, O. Wolkenhauer, Microarray data clustering based on temporal variation: FCV with TSD preclustering, Applied Bioinformatics 2 (1) (2003) 35–45.

[8] X. Ji, J. Li-Ling, Z. Sun, Mining gene expression data using a novel approach based on hidden markov models, FEBS 542 (2003) 125–131.

[9] P. D'Haeseleer, X. Wen, S. Fuhrman, R. Somogyi, Linear modeling of mRNA expression levels during CNS development and injury, in: Pacific Symposium on Biocomputing, Hawaii, 1999, pp. 41–52.

[10] Z. Bar-Joseph, G. Gerber, D. K. Gifford, T. S. Jaakkola, I. Simon, A new approach to analyzing gene expression time series data, in: Proceedings of RECOMB, Washington DC, USA, 2002, pp. 39–48.

[11] M. J. L. de Hoon, S. Imoto, S. Miyano, Statistical analysis of a small set of time-orderd gene expression data using linear splines, Bioinformatics 18 (11) (2002) 1477–1485.

[12] Y. Luan, H. Li, Clustering of time-course gene expression data using a mixed-effects model with B-splines, Bioinformatics 19 (4) (2003) 474–482.

[13] L. Todorovski, B. Cestnik, M. Kline, Qualitative clustering of short time-series: A case study of firms reputation data, in: Conference on Data Mining and Warehouses (SIKDD 2002), Ljubljana, Slovenia, 2002.

[14] 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. USA 95 (1998) 334–339.

[15] B. Everitt, Cluster Analysis, Heinemann Educational Books, London, England., 1974.

[16] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[17] F. Höppner, F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis, John Wiley & Sons, Chichester, England., 1999.

[18] O. Wolkenhauer, Data Engineering, John Wiley & Sons, New York, 2001.

[19] J. Bezdek, A convergence theorem for the fuzzy ISODATA clustering algorithms, IEEE Trans. Pattern Anal. Machine Intell. 2 (1) (1980) 1–8.

[20] D. Dembélé, P. Kastner, Fuzzy C-means method for clustering micoarray data, Bioinformatics 19 (8) (2003) 973–980.

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

[22] K. Y. Yeung, D. R. Haynor, W. L. Ruzzo, Validating clustering for gene expression data, Bioinformatics 17 (4) (2001) 309–318.

[23] J. Dunn, Well separated clusters and optimal fuzzy partitions, J. of Cybernetics 4 (1974) 95–104.

[24] M. K. Kerr, G. A. Churchill, Bootstraping cluster analysis: Assessing the reliability of conclussions from microarray experiments, PNAS 98 (16) (2001) 8961–8965.

[25] X. L. Xie, G. Beni, A validity measure for fuzzy clustering, TPAMI 13 (8) (1991) 841–847.