Clustering of Gene Expression Time-Series Data

C. S. Möller-Levet
Department of Electrical Engineering and Electronics,
UMIST, Manchester M60 1QD, U.K.

K.-H. Cho
School of Electrical Engineering, University of Ulsan, Ulsan, 680-749, South Korea.

H. Yin
Department of Electrical Engineering and Electronics,
UMIST, Manchester M60 1QD, U.K.

O. Wolkenhauer*
Department of Computer Science, University of Rostock, 18051 Rostock, Germany.
E-mail: wolkenhauer@informatik.uni-rostock.de, Tel./Fax: +49 (0)381 498 33 35 / 33 99.

November 19, 2003

* *To whom correspondence should be addressed.*

Contents

1 Time-series **3**
1.1 Statistical Analysis of Time-series . . . 3
1.2 Gene expression time-series . . . 5

2 Clustering analysis **5**
2.1 Representation and modelling . . . 5
2.2 Similarity measures . . . 6
2.3 Clustering algorithm . . . 7
2.4 Validity measures . . . 7

3 Similarity of gene expression time-series **8**
3.1 Similarity requirements for co-expression . . . 8
3.1.1 Scaling and shifting problems. . . . 9
3.1.2 Unevenly distributed sampling points . . . 9
3.1.3 Shape: internal structure. . . . 10
3.2 Similarity in time-series literature . . . 10
3.2.1 Transformation based: linear transformation. . . . 11
3.2.2 Temporal structure based . . . 12
3.2.3 Shape . . . 15

4 Clustering of gene expression time-series **16**
4.1 Number of clusters . . . 16
4.2 Varying membership . . . 17
4.3 Outliers . . . 17
4.4 Noise . . . 17

5 Literature review **18**
5.1 Literature review on time-series similarity . . . 18
5.2 Literature review on gene expression time-series clustering . . . 20

6 Future work **24**

7 Conclusions **26**

Introduction

Clustering analysis is a multivariate technique for data mining [1] which develops meaningful subgroups of individuals or objects (Everitt 1974, Jain and Dubes 1988). In the gene expression context, the analysis is used to identify subsets of genes that behave similarly along time under the set of test conditions; that is, to cluster gene expression time-series data. In clustering, as in any data analysis, issues ranging from problem definition to a critical diagnosis of the results must be addressed. As a first step, in Section 1 we define the properties of time-series followed by the particular characteristics of gene expression time-series. The statistical analysis of time-series is briefly introduced and its suitability for gene expression time-series is discussed. Section 2 describes general aspects involved in clustering analysis. Next, in Section 3, the similarities of gene expression time-series are discussed and the principal requirements described. In Section 4 the clustering requirements of gene expression time-series are described. In Section 5 an extensive literature review considering related work on time-series similarity and clustering of gene expression time-series is presented. Section 6 discusses future work involving a briefly proposal of the development of a suitable clustering algorithm. Finally, Section 7 concludes the report.

1 Time-series

In this section the general characteristics of time-series are described and their statistical analysis is briefly introduced. Then, the particular characteristics of gene expression time-series are defined and their suitability for conventional statistical analysis is discussed.

A time-series is often defined as a series of values of a variable taken in successive periods of time. The variables come from a variety of different domains, from engineering (e.g. (Notohardjono and Ermer 1986)) to scientific research (e.g. (Tilman and Wedin 1991, Aerts and De Cat 2003)), finance (e.g. (Boschen et al. 2003)) and medicine (e.g.(Guo et al. 2003, Yum and Kim 2003)). The range, noise, scaling and shifting factors of the values that such variables can take depend on the nature of the variables and the instrument utilised to measure them. The instants in time at which the measurements are taken are known as time points. The length between time points can vary or be constant and is called sampling interval. There is a well-established area in statistical analysis of data dedicated to the study of time-series. The statistical analysis of time-series (Anderson 1958, Box and Jenkins 1976) accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. In general most of the analysis is focused towards univariate[2] time-series with a large number of measurements and equally distant time points.

1.1 Statistical Analysis of Time-series

There are two main goals in the statistical analysis of time-series: identifying the nature of the phenomenon represented by the series of observations and predicting

[1] Data mining is "the nontrivial extraction of implicit, previously unknown and potential useful information from data" (Frawley et al. 1992). In this research we focus on temporal data mining (Antunes and Oliveira 2001, Roddick and Spiliopoulou 2002), in particular clustering analysis.

[2] Univariate time-series: one type of measurement made repeatedly on the same object or individual.

future values of the time-series variable. Many of the statistical techniques used in time-series analysis are regression analysis techniques or analogues of them (Anderson 1958, Box and Jenkins 1976, Kendall 1976). Selecting a suitable mathematical model is the first step in the analysis of a time-series. After choosing the model, it is possible to estimate parameters and check for the goodness of fit to the data. The fitted model can then be possibly used to understand the mechanism generating the series or to forecast. There are many methods to model time-series. The selection of the appropriate technique will depend on the application and the user's preference.

An example of a model is the auto regressive (AR) model:

$$x_t = \delta + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + A_t \tag{1}$$

where x_t is the time-series, A_t is noise, and

$$\delta = \left(1 - \sum_{i=1}^{p} \phi_i\right) \mu \tag{2}$$

where μ is the process mean. An autoregressive model is a linear regression of the current value of the series against one or more prior values of the series. The value of p is called the order of the AR model. AR models can be analysed with one of several methods, for example the standard linear least squares technique.

Important properties of the time-series for their statistical analysis are autocorrelation, trend, seasonality and stationarity.

Autocorrelation refers to the correlation of a time-series with its own past and future values. It is the correlation of the time-series with itself but shifted in time k time points, k is usually called the lag. Autocorrelation complicates the application of statistical tests by reducing the effective sample size. There are several tools for assessing the autocorrelation: time-series plot, lagged scatter plot and autocorrelation function.

Trend represents a linear or most often nonlinear component that changes over time and does not repeat.

Seasonality are the periodic fluctuations displayed by many time-series. The analysis of seasonality is formally defined as correlation dependency of order k between each i^{th} element of the series and the $(i\text{-}k)^{th}$ element and measured by autocorrelation (Kendall 1976).

Stationarity in time-series is a common assumption in many analysis techniques. A stationary time-series has the property that the mean, variance and autocorrelation structure do not change over time. If the time-series is not stationary some transformations can be applied to achieve stationarity (Kendall 1976). The data can be differentiated, given a series z_t a new series $y_i = z_1 - z_{i-1}$ can be created. If the data contain a trend, a curve can be fitted to the data and then the residuals from that fit can be modelled. When the variance is non-constant, it might be stabilized by taking the logarithm or square root of the series.

A very popular procedure in time-series analysis is smoothing the data, which removes random variation and shows trends and cyclic components (Box and Jenkins 1976). The most common technique is the moving average smoothing which replaces each element of the series by either the simple or weighted average of n surrounding elements, where n is the width of the smoothing window. Figure 1 shows a continuous sinusoidal function (continuous line) which was sampled and added a random error

(squares), followed by a moving average smoothing process (circles); it can be seen that extreme values are eliminated after the smoothing process.

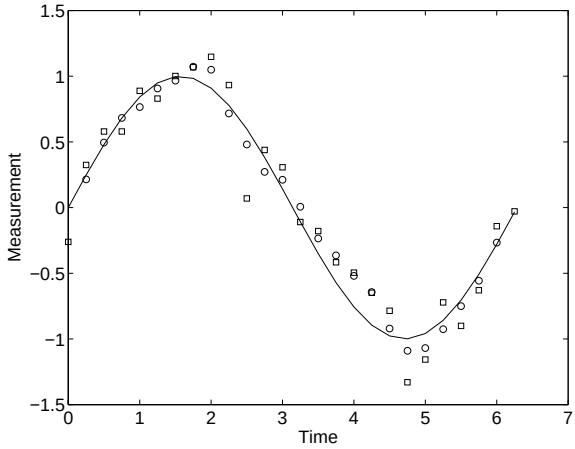


Figure 1: A continuous sinusoidal function (continuous line) is sampled and added a random error (squares), also shown, the smoothed function (circles).

1.2 Gene expression time-series

Gene expression time-series have two main characteristics, they are very short (i.e., four to twenty samples) and are usually unevenly sampled.

The existing literature on short time-series focuses primarily on testing and estimating autocorrelation. Samples under 50 observations are already considered too short for a classical statistical analysis. Calculation of the autocorrelation at different lags is an essential instrument to identify the dependency structure of the series but the estimation of the autocorrelation is biased with small samples, (Bence 1995, Arnau and Bono 2001).

2 Clustering analysis

This section briefly describes several steps involved in clustering analysis. The first step is the representation and modelling of the time-series. The next step is the definition of a similarity measure between time-series which make sense with the selected representation or model utilised. The third step is the definition of a clustering algorithm to group the represented or modelled time-series according to the similarity previously defined. The last step is the validation or scoring of the clustering results.

2.1 Representation and modelling

The model used to represent the time-series has a great impact on the similarity measure adopted, and therefore, in the clustering outcome. Antunes and Oliveira (2001) distinguish four main groups for temporal sequences representation: time-domain continuous, transformation based, discretisation based, and generative models.

In **time-domain continuous representation** the simplest approach is to represent a time-series using the original elements, ordered by their instant of occurrence without any preprocessing. Other alternatives are transformations related to

the length of the series. If the series are too long to be able to manage them, they can be shortened using, for example, piecewise linear functions. There are several approaches to segment the time-series in order to reduce the noise or to reduce dimensionality.

In **transformation based representations** the idea is to transform the initial sequence from time to another domain, and then use a point in the new domain to represent each original series. For example, the use of Discrete Fourier Transformation (DFT) to transform the sequence in a point in the frequency domain choosing the k first frequencies and then representing each sequence as a point in the k-dimensional space (Agrawal et al. 1993). Other examples could be the Discrete Wavelet Transform (DWT) (Popivanov and Miller 2002).

Discretisation based representations translate the initial time-series with real-valued elements to a discretised sequence. An example of this translation is the Transitional State Discrimination (TSD) approach (Möller-Levet, Cho and Wolkenhauer 2003), where the time-series are discretised according to the difference of values between successive time points.

In **generative models representation** the idea is to obtain a model that can be viewed as a generator for the time-series obtained. For example, Hidden Markov Models (HMM) (Ji et al. 2003) or Auto Regressive models (AR) (Ramoni et al. 2002).

2.2 Similarity measures

After representing the time-series appropriately a similarity measure is needed in order to determine if they match.

Depending on the application, the similarity measure should be able to deal with elements such as outlying points, noise and scaling problems and existence of gaps and other time axis distortion. Antunes and Oliveira (2001) distinguishes four main groups for similarity measure based on the type of temporal sequences representation.

Similarity in time-domain continuous representation. The most common distance is the Euclidean distance, where each series of length k is viewed as a point in an k-dimensional space. In order to deal with noise, scaling and translation problems, it can be done by determining if a portion of a sequence fits another, having both a previous linear transformation (Das et al. 1997). In order to deal with small distortions in the time axis the technique of Dynamic Time Warping (DTW) was proposed (Sankoff and Kruskal 1983). Dynamic time warping aligns two sequences so that a predetermined distance measure is minimized.

Similarity in transformation based methods. The simplest approach when dealing with transformed series is to compare the points representing each time-series in each sequence. The comparison of the points is usually given by Euclidean distance.

Similarity in discrete spaces. In this case the simplest approach is to compare each symbol of the series. Also, there exist some specific measures for this representation, such as the Hamming distance[3].

Similarity for generative models. In this case the similarity measure between sequences can be obtained directly by how close the data fits one of the available models. For stochastic generative models (e.g. Markov chains and mixture models)

[3]Hamming distance $d(x, y)$ is the number of coordinates where (the sequences) x and y differ, or: $d(x, y) = |1 \leq i \leq n | x_1 \neq y_i|$.

the probability that a given sequence was generated by a given model is used for the comparison.

2.3 Clustering algorithm

The clustering algorithms can be divided into two main groups:

Hierarchical techniques. Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. The subtypes of this clustering method differ in the rule by which it is decided which two small clusters are merged or which large cluster is split. The final result of the algorithm is a tree of clusters called a dendrogram, which shows how the clusters are related. By cutting the dendrogram at a desired level a clustering of the data items into disjoint groups is obtained.

Partitioning-Optimisation techniques. The partitioning techniques differ from the hierarchical techniques in that they admit relocation of the elements; this allows poor initial partitions to be corrected at a later stage. A centroid or a cluster representative may represent each cluster; this is some sort of summary description of all the objects contained in a cluster. These techniques can be considered as attempts to partition the data set in a way that optimises some predefined criterion. Most of these techniques have three distinctive steps: initiation of clusters, allocation of elements to initialised clusters and reallocation of some or all of the elements to other clusters once the initial segmentation has been completed.

When model-based representations are used for clustering, a generative statistical model is proposed for the data and then a likelihood (or posterior probability) derived form this model is used as the criterion to be optimised (Zhong and Ghosh 2002, Fraley and Raftery 1998). In this approach, the merging of the groups in hierarchical clustering or the reallocation techniques in a partitioning-optimisation clustering, are based on a maximum-likelihood criterion and each cluster is represented by a particular model. The type of model, for example a Gaussian or an HMM, has to be specified according to the objectives of the clustering analysis and the properties of the data set. The structure of the chosen model can usually be selected by model selection techniques, and its parameters estimated using the Expectation Maximisation (EM) algorithm (see Appendix).

2.4 Validity measures

The validity process explores whether the clustering algorithm with the specified parameters (number of clusters, similarity measure, model, etc.) can identify the underlying patterns of the considered data set (Höppner et al. 1999). In order to solve this problem, several cluster quality or validity measures have been proposed in the literature. Cluster validity measures quality of a clustering relative to others created by other clustering algorithms, or by the same algorithms using different parameter values.

The validity measure should reflect the quality of the clusters based on the objectives of the clustering algorithm. For example, fuzzy c-varieties algorithm defines the prototypes as r-dimensional linear subspaces of the data space. If r=2, then the prototypes are lines and it would not make sense to use a validity measure which prefers clusters which are compact (i.e, which have a small radius).

As stated in (Heyer et al. 1999), 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 method to identify the best choice of similarity measure and clustering algorithm.

3 Similarity of gene expression time-series

In this section the similarity requirements of gene expression time-series are described.

Similarity is understood as the resemblance, likeness, or equivalence of two objects. It is a relative term, which only makes sense when comparing more than two elements or when a threshold is utilised. Considering three different objects A, B, and C, there are three possibilities for the similarity of A. A can be more similar to B than to C, or more similar to C than to B, or as similar to C as to B. In order to be able to asses the similarity, a quantitative measure of likeness has to be utilised. It is a common practice to use correlation or distance metrics to quantify such resemblance. The suitability and performance of a similarity measure for a specific comparison depend on the nature of the objects to compare. Therefore, the requirements of similarity of gene expression time-series have to be consider in order to design or select the appropriate similarity measure.

3.1 Similarity requirements for co-expression

The general objective of the clustering of gene expression data is the identification of co-expressed genes. However, there is not a clear definition of co-expression in the literature. In general, it is understood that co-expressed genes have similar patterns of expression. Buy then again, what are similar expression patterns? In (Heyer et al. 1999), similar expression patterns are defined as "patterns that rise and fall concordantly". In (Filkov et al. 2002) the similarity function of time-series is mainly based in the up-down weighted patterns of filtered series. In (Schleip et al. 2003) and in (Ji et al. 2003) HMM are utilised based in qualitative behavior: up- down- no change- regulated. It can be seen that the common idea behind the concept of similar expression patterns lies in the direction of change of the expression level across time points. Other approaches lack of a straightforward biological interpretation, for example (Ramoni et al. 2002), where similar expression patterns are those which are generated by the same stochastic process represented by the proposed AR model. Co-expression has not a common and well defined meaning in the literature, which leaves a wide open door for the suggestion for new and alternative clustering procedures. In this section the basic biological-based requirements of gene expression time-series comparison are summarised.

Three basic similarity requirements have been identified; the similarity measure should be able to handle:

1. Scaling and shifting problems

2. Unevenly distributed sampling points

3. Shape (internal structure)

3.1.1 Scaling and shifting problems.

A promoter is a structural regulatory sequence recognised by the sigma factor of the RNA polymerase holoenzyme (the protein that is used to read the DNA for transcription). Genes that share a common sequence will therefore share their expression, they will be switched on at the same time but not necessarily at the same level. The reason is that the recognition efficiency is not the same for every gene having that promoter. This is one of the situations leading to scaled and shifted expressions. Therefore, scaling and shifting factors in the expression level hide similar expressions and have to be eliminated or not considered when assessing the similarity between expression profiles. Other possible sources for scaling and shifting problems are intrinsic of the microarray experiment (e.g. label efficiency of the dyes) which are usually eliminated in the normalisation procedure. In a mathematical sense scaling and vertical (i.e. of the expression level) shifting refer to the case were linear transformations are present. Considering two time-series x and y, y is a linear transformation of x if it can be expressed as $y = mx + b$. The scaling factor is m and b is the vertical shift. Figure 2(a) shows an example of the effects of scaling and vertical shifting.

Synchronisation of biological processes is not an easy task, common processes may unfold at different times in different experiments or individuals producing horizontal shifts in the resulting time-series. For a few number of time points the identification of horizontal shifts can possibly be made after the clusters are obtained while for longer series temporal aligning techniques can be utilised. Depending on the purposes of the clustering analysis the horizontal shift might or might not be considered. For example, in the DNA microarray analysis by Spellman et al. (1998), samples from yeast cultures are synchronised by three independent methods[4] to create a comprehensive catalogue of yeast genes whose transcript levels vary periodically within the cell cycle. In this case the authors used a Fourier transformation to identify periodicity, when using this transform time shift is ignored. Later, Aach (2001) grouped these time-series based on their time alignment. The author identified that small sample size and high measurement noise decrease alignment stability.

3.1.2 Unevenly distributed sampling points

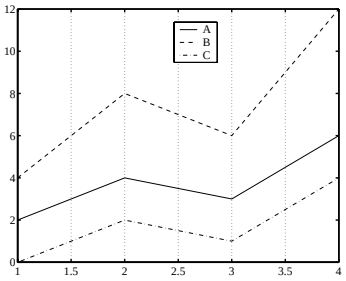
Since gene expression time-series rely on samples of the actual biological process, the higher the sampling frequency the more information one has to recreate the actual process. However, it is not possible to achieve high sampling frequency in microarray experiments due mainly to the time and resources that would require. Therefore, biological processes are sampled at shorter intervals of time when intense biological activity or when the activity of interest is taking place, leading to unevenly distributed sampling points. In consequence, the length of the sampling interval is informative and should be consider in similarity comparisons. Consider the following three time-series: F=[3.5 3 4 4], G=[2 2 3 3] and H=[1 1 2 1.5], as shown in Figure 2(b). The absolute errors of time-series G and F, and G and H at each time point are $e(G, F)$=[1.5 1 1 1] and $e(G, H)$=[1 1 1 1.5], respectively. It can be seen that they have the same overall absolute error. When the sampling interval is considered the rate of change across time (i.e. slope) can be obtained. The rate is obtained by taking the ratio of the difference of expression level in neighboring time points and the length of the corresponding sampling interval. The absolute error of the slopes of time-series G and

[4] These methods are: α factor arrest, elutriation and arrest of a *cdc*15 temperature-sensitive mutant.

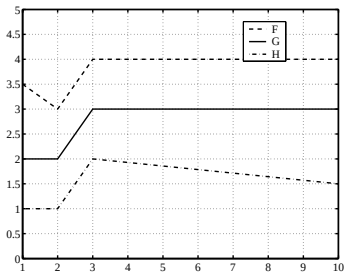
F and G and H are $e(s_G, s_F)$=[0.5 0 0] and $e(s_G, s_H)$=[0 0 0.0714], respectively. Now it is evident that G is more similar to H than to F when considering the rate of change of expression, given by the length of sampling intervals.

3.1.3 Shape: internal structure.

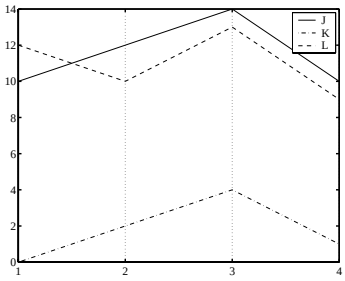
The main difference between a set of measurements and a time-series is the internal structure, therefore a time-series can not be treated as independent identically distributed data. The internal structure can be described by different models and in general it is reflected in the shape of the series. In microarray experiments, the intensity of gene expression is not relevant, instead, the relative change of intensity characterised by the shape of the expression profile is regarded as characteristic and informative. Figure 2(c) shows three time-series, J,K and L, where J is more similar to K, when the intensity of the gene expression is considered and J is more similar to L, when the relative change of intensity is considered. A necessary condition for the existence of internal structure or characteristic shape is the temporal order of measurements. Therefore, the similarity function should not allow a change in the order. The internal structure can be represented by a statistical model, by deterministic functions or by symbols describing the series.



(a) Scaling and shifting: three time-series, A, B and C, where B is A scaled by 2, and C is A shifted by 2.



(b) Sampling interval: three unevenly sampled time-series with different rate of change of expression across time.



(c) Shape: three time-series, J,K and L, where J is more similar to K, when the intensity of the gene expression is considered and J is more similar to L, when the relative change of intensity is considered.

Figure 2: Elements involved in gene expression time-series similarity.

3.2 Similarity in time-series literature

We have identified several approaches to compare times-series in the literature:

1. Transformation based: linear transformation
2. Temporal structure based
3. Shape based

3.2.1 Transformation based: linear transformation.

A transformation $T : R^n \rightarrow R^m$ is a linear transformation if it satisfies:

$$\begin{aligned} T(u+v) &= T(u)+T(v) \\ T(cu) &= cT(u) \end{aligned}$$

When Euclidean distance (of standardised time-series) or correlation are used as similarity measures, similarity between two time-series can be understood as the strength of their linear relationship. See the appendix for a short review of some common distances, the Euclidean distance and correlation coefficient. Möller-Levet, Cho and Wolkenhauer (2003) present the function relating the Euclidean distance and correlation coefficient of standardised time-series, showing that the more are time-series linearly related the smaller is the Euclidean distance between them after standardisation. In the temporal database field, Das et al. (1997) considers that two time-series are similar if there is a linear function f such that a long subsequence of X can be approximately mapped to a long subsequence of Y using f. This is illustrated in Figure 3 on complete sequences. However, two linearly related time-series can present opposite shapes as discussed and illustrated in section 5, showing that this concept of similarity can fail in identifying similar shapes.

In gene expression time-series data this concept of similarity is too naive and superficial since the internal structure of the data set is of most interest. However, as stated before, linearly transformed time-series are regarded as scaled and shifted similar expressions. This similarity can be uncovered simply by normalising[5] or standardising[6] the data set as illustrated in Figure 4.

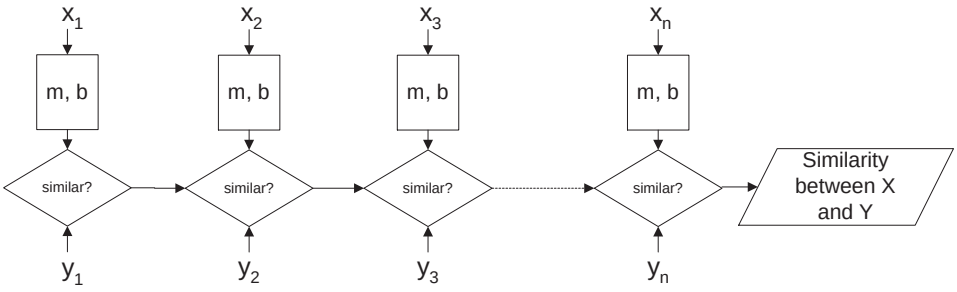


Figure 3: Considering two time-series $X = [x_1, x_2, x_3, \ldots, x_n]$ and $Y = [y_1, y_2, y_3, \ldots, y_n]$, the figure shows similarity based on the degree to which one time-series can be expressed as the linear transformation of another, such that $Y = mX + b$.

Dynamic Time Warping (DTW) Algorithms align two time-series against each other. This tool emerged originally for speech recognition in the 1970's. In speech recognition the alignment is necessary due to different parts of the words being spoken at different rates. Therefore, the best alignment of the word to be recognised and a reference pattern, identifies the word to be the same as the one represented by the reference pattern. In contrast, in unsupervised gene expression clustering there are not reference patterns to match the expression profiles with. Therefore, the similarity

[5]The time-series can be normalised by subtracting the mean and dividing by the highest value.

[6]The time-series can be standardised by subtracting the mean and dividing by the standard deviation.

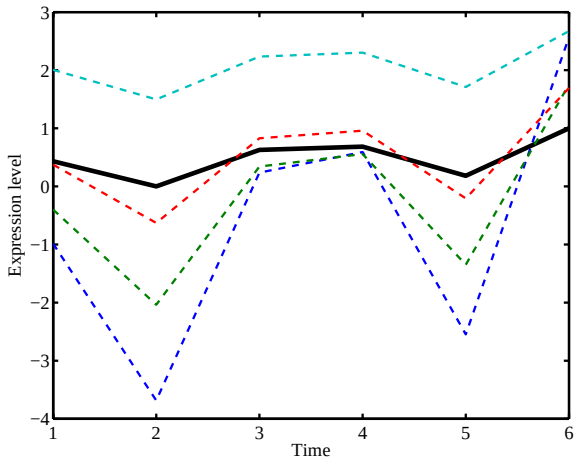


Figure 4: The four dashed time-series are different linear transformations of a given time-series. The continuous line represents the four identical patterns obtained after the normalisation of the dashed time-series.

problem goes beyond the alignment. Alignment scores are based on Euclidean distance, implying that the similarity between already aligned sequences becomes the conventional Euclidean distance between the two series. However, in some cases the alignment of the series is essential when comparing series from different experiments and DTW had shown to have a good performance (Aach 2001).

In conclusion, the similarity based in linear transformation is very basic and does not consider the internal structure and the sampling interval. However, linearly related gene expression time-series are can be considered as shifted and scaled similar patterns. By normalising or standardising the data set, linearly related expressions are uncovered.

3.2.2 Temporal structure based

The underlying assumption of the statistical model is that the series can be well characterised as a parametric random process, and that the parameters of the stochastic process can be estimated in a precise, well-defined manner. Several models have been utilised for gene expression time-series including: normal mixture, autoregressive, hidden Markov and splines models.

In a **normal mixture model-based** approach each gene is assumed to have come from a mixture of multivariate normal densities with different means and certain parameterizations of the covariance matrix (Yeung, Fraley, Murua, Raftery and Ruzzo 2001, Fraley and Raftery 1998). In this approach there is not consideration towards the temporal structure of the data and the length of sampling intervals.

A more appropriate model for this application is the **autoregressive model**. In its simplest form, an autoregressive (AR) model is a linear regression equation which links the current value of some variable to its value in the previous period and a constant term. The order of the AR is the number of past values that are considered to generate the actual value. Given the size of the time-series in gene expression, AR of order one is the most appropriate (Ramoni et al. 2002). When using AR, the internal

structure considered is that the current value of the series is a linear combination of the p most recent past values of itself plus an error term , which incorporates everything new in the series at time t that is not explained by the past p values. In this approach time-series are similar when they are generated by the same stochastic process represented by the proposed model as shown in Figure 5. There are two main questions involved in the selection of the AR model: the order of the model and the method for parameter estimation. The parameters are calculated from the time-series using optimisation methods such as maximum likelihood and least squares (Box and Jenkins 1976). The autoregressive model is limited by the requirement of stationarity, that is, the system generating the time-series should be time invariant. Although several techniques can be used for converting non-stationary time-series into stationary ones, it is not always possible to meet the requirement. In this approach the temporal structure of the data is established by an AR model, but the length of sampling intervals is not considered.

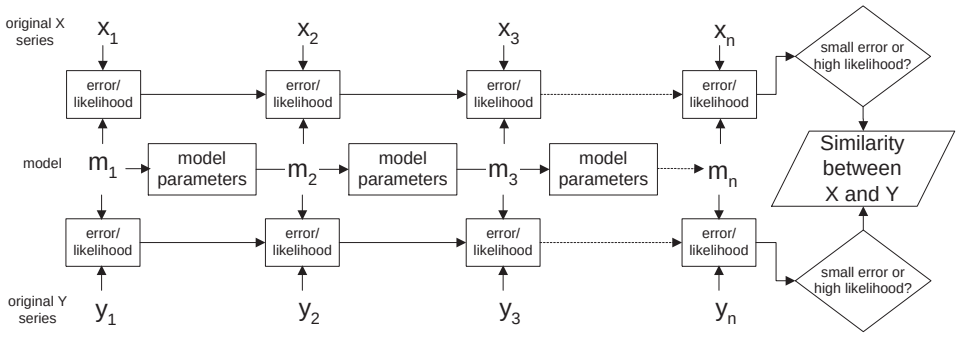


Figure 5: Considering two time-series $X = [x_1, x_2, x_3, \ldots, x_n]$ and $Y = [y_1, y_2, y_3, \ldots, y_n]$, and an autoregressive model with parameters (δ,$\phi_1 ... \phi_p$ and ϵ), such that $x_t = \delta + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \epsilon_t$. The similarity of X and Y is given by how well they fit the proposed model in terms of a small error or a high likelihood to be generated by the system represented by the selected model.

In a **Hidden Markov Model** (HMM) the observations are a probabilistic function of the state and the transition from one state to another is also given by a probability function. A HMM (Rabiner 1989) is characterised by the following:

1. N, the number of states in the model.

2. M, the number of distinct observation symbols $V = \{v_1, v_2, \cdots, v_M\}$ per state, that is, the discrete alphabet size.

3. $A = a_{ij}$, the state transition probability, where $a_{ij} = P[q_{t+1} = S_j | q_t = S_i]$, $1 \leq i, j \leq N$. That is, given the current state q_t been S_i, a_{ij} is the probability that the next state q_t is S_i. A is illustrated in Table 1.

4. $B = b_j(k)$, the observation symbol probability distribution in state j, where $1 \leq j \leq N$ and $1 \leq k \leq M$. B is illustrated in Table 2.

5. $\pi = \pi_i$ the initial state distribution, where $\pi_i = P[q_1 = S_i]$, $1 \leq i \leq N$.

Given appropriate values of N, M, A, B and π, the HMM can be used as a generator to give an observation sequence $O = O_1 O_2 \cdots O_T$, where T is the number of observations in the sequence. The complete parameter set of the model has a compact notation $\lambda = (A, B, \pi)$.

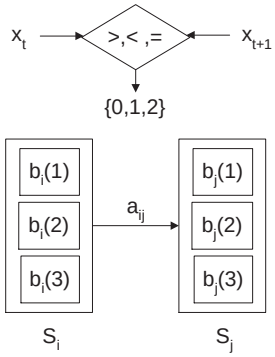


Figure 6: The time-series are transformed to a three letter-alphabet sequences describing the shape of the time-series, down-pattern ($x_t > x_{t+1}$) $v_1 = 1$, up-pattern ($x_t < x_{t+1}$) $v_2 = 2$, and no change-pattern ($x_t = x_{t+1}$) $v_3 = 3$. In the model, the transition from state S_i to S_j has a probability of a_{ij}, and at state S_i, the probability of having an observation of 'down' is $b_i(1)$, of 'up' is $b_i(2)$ and 'no change'is $b_i(3)$.

Table 1: The state transition probability $A = a_{ij}$

	S_0	S_1	...	S_N
S_0	a_{00}	a_{01}	...	a_{0N}
S_1	a_{10}	a_{11}	...	a_{1N}
⋮	⋮	⋮	...	⋮
S_N	a_{N0}	a_{N1}	...	a_{NN}

Table 2: Observation symbol probability $B = b_j(k)$, with $M = 3$ and $V = [1\,2\,3\,]$.

	v_1	v_2	v_3
S_0	$b_{0(1)}$	$b_{0(2)}$	$b_{0(3)}$
S_1	$b_{1(1)}$	$b_{1(2)}$	$b_{1(3)}$
⋮	⋮	...	⋮
S_N	$b_{N(1)}$	$b_{N(2)}$	$b_{N(3)}$

In this approach it is assumed that each gene expression profile has been generated by a Markov chain with certain probability. Therefore, the temporal structure of the time-series is considered in the HMM, given that the next state is a probabilistic function of the current state. However, as in the AR model, the length of the sampling interval is not incorporated in the similarity assessment.

In a **Spline Model** the time-series are represented by a model which defines a curve

in time. Splines are piecewise polynomials with pieces that are smoothly connected together.

The main elements of the B-splines are:

1. The number n of B-splines connected together. That is, the number of functions in which the interval $0 \leq t \leq T$ is divided, where T is the time of the last measurement.

2. The degree k of the n B-splines.

3. The number p of joining points of the polynomials which are called knots.

For a spline of degree k, each segment is a polynomial of degree k, which should suggest that k+1 coefficients are needed to describe each piece. However, there is an additional smoothness constraint that imposes the continuity of the spline and its derivatives up to order (k-1) at the knots, so that, effectively, there is only one degree of freedom per segment. A cubic spline is a piecewise cubic polynomial such that the function, its derivative and its second derivative are continuous at the knots.

This approach considers the shape of the profiles and could consider the length of sampling interval if the knots are properly defined. Bar-Joseph et al. (2002) used statistical spline estimation to represent time-series gene expression profiles, however, the method require data that has been sampled at a sufficiently high rate (Bar-Joseph et al. 2002). In addition, whereas cubic splines are used more commonly, for the usual shortness of gene expression time-series, they are not suitable (de Hoon et al. 2002). Later, Luan and Li (2003) proposed a mixed-effects model using cubic B-splines utilising "long" gene expression time-series (12 and 18 time points), and four equally spaced knots. However, it is not always possible to define equally spaced knots if the series are unevenly sampled. In addition, equally spaced knots can not properly reflect the unevenly distributed time points.

In general this is a good approach for time-series similarity which considers a temporal structure by considering the shape of the profiles and could consider the length of sampling intervals. In this case the actual value of the series is not related with the previous one by a specific model or function, but the values are a function of the time.

A different approach not based in a probabilistic model but which considers a temporal structure is the one presented in (Möller-Levet, Cho and Wolkenhauer 2003). In this case the proposed structure is: $x_{t+1} = m_{t+1}x_t + b_{t+1}$ and $1 \leq t \leq$ number of time points. The similarity is based on the resemblance of the parameters of linear transformation between time points as illustrated in Figure 7. Once again, this approach considers a temporal structure but fails to consider the length of sampling interval.

3.2.3 Shape

This approach has a straight forward biological interpretation. In this case, the up and down patterns of the series are considered to calculate the similarity between two series. Although it seems to be a simpler approach, special attention has to be given to the selection of the elements used to describe the shape. One possible approach is the use of slopes, in (Wen et al. 1998) 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

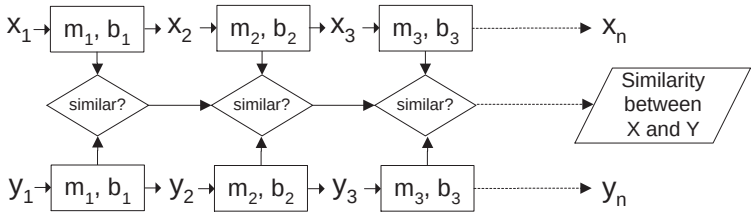


Figure 7: Considering two time-series $X = [x_1, x_2, x_3, \ldots, x_n]$ and $Y = [y_1, y_2, y_3, \ldots, y_n]$, and the model $x_{t+1} = m_{t+1}x_t + b_{t+1}$, the similarity of the series is given by the resemblance of the parameters m and b of linear transformation between time points.

the variable time intervals. In (Möller-Levet, Klawonn, Cho and Wolkenhauer 2003) time-series are considered as piecewise linear functions and the slopes calculated with the real length of sampling intervals are compared. However, as the measurements are weighted proportionally inverse to the length of sampling intervals, very long sampling intervals could have too low impact in the comparison, while short sampling intervals could have too much impact. In addition, the temporal order of the slopes is not considered. Other approaches are based on the discretisation of the series according to the direction of the change of expression. In these cases one or several thresholds are used to define an event (i.e. up, down or no change) and have to be defined. This last approach is somewhat restrictive since biological information (i.e. the amount and rate of change of expression level) is lost with the discretisation.

4 Clustering of gene expression time-series

After having described the requirements of similarity for gene expression time-series and the approaches found in the literature. This section specifies the requirements for clustering the aforementioned series.

Several requirements can be identified for the clustering algorithm, in specific it should be able to handle:

1. Unknown number of clusters
2. Varying membership
3. Outliers
4. Noise

4.1 Number of clusters

Unsupervised clustering is the most common approach for clustering gene expression data. This means that there is no previous knowledge of the number and characteristics of the clusters forming the data. Different clustering algorithms have special techniques for identifying the number of clusters. A common approach to identify the number of clusters is the use of validity measures; the data is clustered defining different number of clusters and the best value of the validity measure will identify the most convenient number. For example, in (Ji et al. 2003) a validity measure based in the FOM presented in (Yeung, Haynor and Ruzzo 2001) is utilised successfully to identify the number of

clusters. In fuzzy clustering the fuzziness of the partition can be used to evaluate the goodness of the results for different number for clusters (Höppner et al. 1999). Other approaches do not have to define a number of clusters, because it is obtained as a result of the partition, but additional parameters have to be defined. For example, in the CAST algorithm (Ben-Dor and Yakhini 1999) the number of clusters do not has to be determined. Instead, a parameter called *affinity threshold* is used to determine what is the minimum similarity required between an object and a cluster for that object to be a member, and not all the genes are assigned to a cluster. In this approach each cluster is formed by alternating between adding and removing genes from the current cluster until such time that changes no longer occur or a maximum of iterations has been executed. Therefore, the number of clusters obtained from the partition of the data set depends on the selection of the affinity threshold. In general, a clustering algorithm for gene expression data should be able to identify the hidden number of clusters by appropriate means relevant to the algorithm.

4.2 Varying membership

Several researches have identified and emphasize the importance of overlapping clusters in gene expression clustering analysis, (Ji et al. 2003, Gasch and Eisen 2002). The partition of genes into classical sets implies that each gene has been associated with a single biological function or process which, may be an oversimplification of the biological system. Therefore, genes should be allowed to have varying probability or membership degree to different clusters to allow connection of genes to more than one clusters, revealing distinct aspects of their function and regulation. The EM algorithm utilised for model-based clustering allows for partial credit to different clusters, that is, genes have varying probability to belong to different clusters. Other approach is fuzzy clustering, which allows genes to belong to more than one group by assigning different degrees of membership to each cluster.

4.3 Outliers

Given the restricted number of time points in gene expression time-series, an outlier has a high influence on the similarity measure. The clustering algorithm should be able to minimise this impact. A good example is (Heyer et al. 1999), where the authors use the jack-knife correlation. This measure corresponds to the minimum of all the possible calculations of correlation between two series, where each calculation is done with the omission of a different time point. In this way, the effects of outliers in the clustering procedure are reduced. Another approach is the identification of outliers as a previous step to the clustering procedure as suggested in a section 6.

4.4 Noise

It is well known that microarray experiments are subject to a large experimental error producing very noisy measurements. The clustering algorithm should be able to handle these common levels of noise. There are several approaches to achieve this. For example, in fuzzy clustering a noise cluster can be added to the partition (Dave 1991). This cluster attracts all those genes which do not show a minimum level of similarity to the rest of the clusters and reduces the influence of this group in the whole partition. Some algorithms based in stochastic models have remarked

the possibility of incorporating noise in the model (Fraley and Raftery 1998, Yeung, Fraley, Murua, Raftery and Ruzzo 2001).

5 Literature review

This section presents the related literature divided in two sections, the first section concentrates in time-series similarity and the second on clustering gene expression time-series.

5.1 Literature review on time-series similarity

The statistical literature on time-series is vast, however, it has not studied similarity notions that would be appropriate for data mining applications such as clustering analysis. Most of the work done so far in similarity of time-series come from the problem of similarity queries in the field of temporal databases. Similarity queries can be classified into two categories: whole matching or subsequence matching. Whole matching refers to the comparison of two complete sequences, while subsequence matching, as its name implies, is the comparison of a small sequence to small sequences in a complete sequence. The tendency in the literature is to focus on the subsequence matching.

Agrawal et al. (1993) was the first to examine similarity matching of time-series data. The authors present an indexing structure for fast similarity searches over time-series databases. They use a form of dimensionality reduction, (i.e., feature extraction), where the time-series are represented as points in a low dimensional feature space. They use the Euclidean distance to measure the similarity of the time-series represented by the first few coefficients of their Fourier transformation since the Euclidean distance is conserved after the transformation. They showed that a few coefficients (1-3) are adequate to provide good performance, which is increased with the number and length of sequences. They worked with time-series of 400 time points in a whole matching scheme. Later, Faloutsos et al. (1994) examine the problem of subsequence matching extending the idea of (Agrawal et al. 1993).

A different approach was introduced by (Agrawal et al. 1995), proposing a new model for time-series similarity where two sequences are considered similar if they have enough non-overlapping time-ordered pairs of subsequences that are similar. Then, two subsequences are considered similar if after a given transformation one can be enclosed within an envelope of a specified width drawn around the other. They used windows of size w of 5 to 20 elements to form the subsequence for further matching, dealing with scaling and shifting by normalising each window to a range (-1,+1). They use L_∞ norm as the distance measure between subsequences, which are considered as points in a w-dimensional space. They assume equally sampled time-series, therefore, they do not consider different lengths of sampling interval. Also, by using the L_∞ norm the shape is not consider, although is somewhat delimited by the threshold of similarity used.

The concept of longest common subsequence was then used by (Das et al. 1997), defining similarity as follows: two sequences $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ are F-similar, if there is a function $f \in F$ such that a long subsequence $X' = (x_{i_1}, \ldots, x_{i_{\gamma n}})$ of X can be approximately mapped to a long subsequence $Y' = (y_{j_1}, \ldots, y_{j_{\gamma n}})$ of Y using f. They propose f to be a linear transformation.

$$y_{jk}/(1+\varepsilon) \leq ax_{ik} + b \leq y_{jk}(1+\varepsilon) \tag{3}$$

where $\varepsilon \leq 1$, $0 \leq \gamma$, $1 \leq k \leq \gamma n$. The matched subsequences allow for a number of "holes" in the original sequences, conserving the same relative order in X and Y. As observed by the authors, the main differences of this approach compared with (Agrawal et al. 1995) are that the later model does not allow outliers within windows of a specified length w, and the linear function can vary slightly in the length of a matched common subsequence. (Das et al. 1997) present some experimental results using equally sampled time-series of equal length. The distance between two sequences is obtained by subtracting from the total length of the series the length of the longest common subsequence. A distance matrix is created which is the input for a clustering software package. This idea can handle scaling and shifting very efficiently but it does not consider shape and length of sampling interval. Figure 8 shows two linearly related time-series ($y = -1.8428x + 0.9968$) presenting opposite shapes to illustrate how this concept of similarity fails to identify similar shapes.

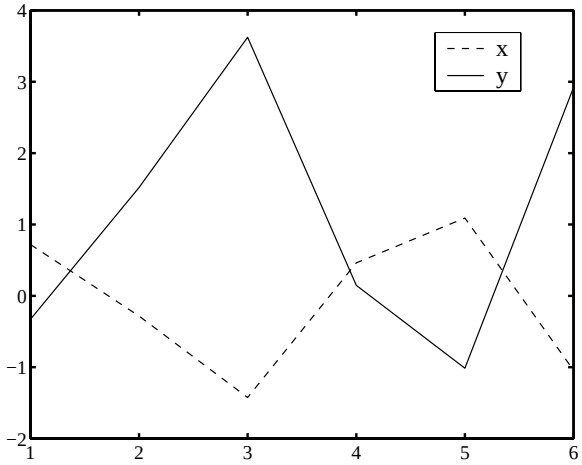


Figure 8: Two linearly related time-series $y = -1.8428x + 0.9968$ presenting opposite shapes.

Keogh and Pazzani (1998) introduce a time-series representation consisting of piecewise linear segments to represent shape and a weight vector that contains the relative importance of each individual linear segment. The total weight associated with a sequence of a given length is constant, regardless of how many segments are used to represent it. Considering two time-series, x and y, the metric measure they present is given by:

$$d(x,y) = \sum_{k=1}^{n} x_{wk} y_{wk} |(x(t_k) - y(t_k)) - (x(t_{(k+1)}) - y(t_{(k+1)}))| \tag{4}$$

where x_{wk} and y_{wk} are the weights for the segment k, $x(t_k)$ and $x(t_{(k+1)})$ are the values of the series at time point t_k and $t_{(k+1)}$, respectively, and as for y. This metric measures how close corresponding segments from x and y are to being parallel.

Todorovski et al. (2002) proposed a new qualitative similarity measure for short time-series and use it in a hierarchical clustering scheme. The distance distinguishes the up-down patterns. It correspond to the normalised sum of scores given to up, down or equal change. It is a simple and intuitive measure suitable for short time-series and a statistical test of the significance of change is suggested.

5.2 Literature review on gene expression time-series clustering

A very good paper for gene expression clustering is the one of Heyer et al. (1999). This presents an analysis procedure for the clustering analysis of the yeast *Saccharomyces cerevisiae* data set by Cho et al. (1998). The authors present as a first step their proposed definition of similarity. Haven studied the disadvantages of the correlation coefficient and Euclidean distance as a similarity measurement, they proposed what they called, the jack-knife correlation. For a pair of genes, i and j, let ρ_{ij} denote the correlation of the pair i and j; also, let $\rho_{ij}^{(l)}$ denote the correlation of the pair i and j computed with the lth observation deleted. For a data set with t observations, they define the jackknife correlation J_{ij} as $J_{ij} = min\rho_{ij}^{(l)} \cdots \rho_{ij}^{(2)}, \cdots \rho_{ij}^{(t)}, \cdots \rho_{ij}$. Then they proceed to the grouping of genes quantifying the quality of the clusters by their diameter, defined as $1 - min_{i,j \epsilon C} S_{ij}$, in which s is the similarity measure being used, and i, j are genes in cluster C. The focus of the algorithm is to find large clusters that have a quality guarantee. Transitivity is ensured by finding clusters whose diameter does not exceed a given threshold value d, so any two genes in a cluster have a jackknife correlation value that is at least $1 - d$. The jacknife correlation is insensitive to the outlier effect and it captures the shape of an expression pattern, although it does not consider the sampling interval.

(Lukashin and Fuchs 2001) presents a strategy which organizes the search of the optimal number of clusters simultaneously with the optimisation of the clustering. They normalise the profiles such that the expression level varies between 0 and 1. The similarity metric they use is the Euclidean distance and they use a function to optimise the clustering by minimising the sum of distances within clusters. To minimise the function they apply the simulated annealing algorithm (Kirkpatrick et al. 1983). They present a very well structured strategy for obtaining the optimal number of clusters. Considering that the optimal number of clusters depends primarily on the variation between profiles within a given data set, they treat the problem introducing a cutoff distance D and postulate that the assumption that vectors i and j belong to the same cluster is incorrect if the distance between two vectors is larger than D.

(Yeung, Fraley, Murua, Raftery and Ruzzo 2001) present a model-based approach. In general, the model-based approach assumes that the data is generated by a finite mixture of underlying probability distributions, and in this case multivariate normal distributions. In the Gaussian mixture model, each component is modelled by the multivariate normal distribution with parameters μ_k (mean vector) and Σ_k (covariance matrix). Geometric features such as shape volume and orientation of each component are determined by the covariance matric Σ_k. In this approach the problem of determining the number of clusters and of choosing an appropriate clustering method become statistical model choice problems (Fraley and Raftery 1998). For a complex model a small number of clusters may suffice, whereas for simple models, a large number of clusters to fit the data adequately may be needed. The authors state that the advantages over CAST are the selection of number of clusters and an appropriate model. However, in CAST it is not necessary to define the number of clusters. So, it can not be considered as an advantage.

In (de Hoon et al. 2002) the authors identified that the use of 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 present in most of the gene expression time-series data. They propose to model the time-series with

linear splines[7], and cluster the resulting models using k-means clustering. The authors remark that cubic splines are used more commonly, however, the linear spline functions are more suitable for this application given the restricted number of time points. They outline a strategy based on fitting linear spline functions to time-series using the maximum likelihood method and Akaike's Information Criterion, (Akaike 1974). The significance of the gene expression measurements is assessed by applying Student's t-test and only the genes considered to be significantly affected by the experiment are considered. The percentage variance explained for each gene is used as a measure of the goodness of fit of the linear spline function. The authors are able to determine how many measurements are needed at each time point (replicates) in order to estimate the linear spline function reliably. The accuracy of the selection of knots is related to the number of replicates available. The conventional approach when dealing with replicates would use the average at each time point while this approach uses their linear spline estimates.

Gasch and Eisen (2002) proposed the use of fuzzy clustering for extracting biological insights for gene-expression data. They utilised a modified fuzzy c-means clustering to identify overlapping clusters of yeast genes based on published gene-expression data following the response of yeast cells to environmental changes. The algorithm was modified in two ways: fist, they performed three successive cycles of fuzzy c-means clustering, with the second and third rounds of clustering performed on subsets of the data. The second modification is the initialisation of the algorithm by seeding prototype centroids with the eigen vectors identified by PCA of the respective data set. The authors used the Pearson correlation as a distance measure. The number of clusters has to be selected, however, it is shown that fuzzy c-means appears to be less sensitive to over-fitting, because the genes are not forced to belong to only a single cluster. One of the most significant advantages of fuzzy c-means clustering is that genes can belong to more than one group, revealing distinct aspects of their function and regulation. The fuzzy k-means was chosen for its conceptual and algorithmic simplicity. One of the limitations identified by the authors is the selection of meaningful cutoffs for the membership degree, which was alleviated by the use of visualization software for the clustering results.

Ramoni et al. (2002) present a Bayesian method for model-based clustering of gene expression time-series. The method represents gene expression time-series as autoregressive equations and uses an agglomerative procedure to search for the most probable set of clusters given the available data. The authors consider that two time-series are similar when they are generated by the same stochastic process. To reduce the effort of selecting which time-series are merged in the agglomerative process, they use a heuristic strategy based on a measure of similarity between the time-series (e.g. Euclidean distance, correlation coefficient). The method identifies the number of clusters and partitions the gene expression time-series in different groups on the basis of the principled measure of the posterior probability of the clustering model. The method has two components: a stochastic description of a set of clusters, from which they derive a probabilistic scoring metric, and a heuristic search procedure. The derivation of the scoring metric assumes that the processes generating the data can be approximated by autoregressive models. The model assume the time-series are stationary, which can be observed with a series of plots, nevertheless the authors claimed that in

[7] A spline function is a continuous function formed by piecewise liner functions, which are connected to each other at knots.

their experience, the clustering process seems to be largely unaffected by the presence of nonstationary time-series. After some tests they found the order one for the autoregressive model had the best performance. The order of the autoregressive model is somewhat restricted by the number of time-points which is very small for gene expression data. The sampling interval is not considered by this method. A possible drawback of this method is that the model-free distances are calculated on the raw data, therefore shifting and scaling factors are not considered. However it does cluster based in the temporal properties of the data set.

(Filkov et al. 2002) is a very comprehensive paper which discusses several issues of the analysis for microarray time-series data. The similarity function of time-series proposed by these authors, is mainly based in the up-down weighted patterns of filtered series. The filtering removes changes of expression which are less than a predefined threshold by labelling time points as as local minima, maxima and in-between allowing a given expression error level. The points which do not pass the error threshold are eliminated. Local minima and maxima are connected and the normalised expression changes smaller than a threshold are erased. The remaining points may or may not be merged. Then, narrow picks (i.e., points that vary significantly in relative expression from their adjacent points, typically a factor of 2) are eliminated. Finally, the similarity measure between two time-series x and y is given by scoring each slope e_x of x against each slope e_y of y after the filtering, as shown in the following equation.

$$S_g = \sum_{\text{all } e} d\left(1 - \frac{\delta}{\delta_{\max}}\right) \Big/ \sqrt{n_a n_b} \tag{5}$$

where e are all the slopes remaining after filtering, d is the comparison of up-down shape ($d = 1$ if the signs of the slopes agree, otherwise, $d = -1$), $\delta_{\max}$ is the maximum allowable time difference between the middle of e_x and e_y (interactions between pairs of slopes with time difference $> \delta_{\max}$ are considered biologically meaningless and are simply ignored), δ is the observed time difference between the middle of e_x and e_y, n_x and n_y are the number of slopes in x and y respectively. It can be seen from equation (5) that $(1 - \delta/\delta_{\max})$ is the weight of d, corresponding to the closeness in time of the measurements. Another contribution of this paper related to the similarity problem is the comparison of the Hamming distance. Concluding that the correlation coefficient is inadequate as a similarity measure of two-letter-alphabet sequences.

(Ji et al. 2003) presents a model-based clustering method based on hidden Markov models which produce clusters of quality comparable to two prevalent clustering algorithms (k-means and SOM). This model-based approach assumes that each gene expression profile has been generated by a Markov chain with certain probability. They determined the number of clusters using a validity measure and testing for different number of clusters. The algorithm is more sensible to the number of clusters that the conventional algorithms which the authors consider an advantage for using their validity measure for selecting the number of clusters. The original data set of N time points is standardised followed by a transformation to a three-letter-alphabet sequence (0=no change, 1=up or 2=down) aided by a tolerance factor. A simple HMM was constructed for the transformed sequences with $N - 1$ states, where each state could generate a character (0, 1, or 2) according to a distribution representing the regulation trend at this state. For convenience they added a dummy "Begin" state and a dummy "End" state. So, the gene transformed sequence could be generated commencing at state "Begin", choose a transition to another state and generate the character (0, 1,

or 2) based on the distribution at this state. Then choose a transition to the next state and generate the next character, and so on until the "End" state is reached. The model was then trained with the Baum-Welch method (Rabiner 1989) and the probability of a sequence given the HMM was calculated with a forward-backward algorithm (Rabiner 1989). Each cluster is represented by a HMM. The parameters of the model are initialised randomly and optimised iteratively, such that sequences that had a higher probability of a particular HMM had a greater influence on re-estimating the parameters of that HMM. Since the HMMs were initialised randomly assigning values to the elements of the model, each calculation could potentially generate very different clusters. Therefore, every two genes that cluster more than 40 times over the 100 clustering calculations are preserved and those that cluster for fewer than 40 times are excluded. The authors remark that partitioning genes into disjoint set may be an oversimplification of the biological system. In this approach, genes have varying probability to belong to the different clusters, allowing connection of genes to more than one cluster. As the authors remark, the method is not very fast, because it has to train the parameters of the model with the gene expression data set.

Schleip et al. (2003) used the HMM within a model-based clustering framework. The authors regard the model-based methods as the only ones which assume the different experiments to be dependent, which can be further questioned. The authors do not really capture the qualitative behavior of time-series since they have to be further separated in order for them to be qualitatively similar. Starting from an initial collection of HMMs encompassing typical qualitative behavior (up- down- regulated), an iterative procedure finds cluster models and an assignment of data points to these models that maximises the joint likelihood of the clustering. In Ji et al. (2003) the starting value of the parameters of the models are initialised randomly and they are adjusted iteratively such that sequences that had a higher probability for a particular HMM had a greater influence on re-estimating the parameters of that HMM. In contrast, in this paper the starting point HMMs encompass typical qualitative behavior and then followed an adapted k-mean algorithm for the partition (the prototypes are the models, and the parameters of the models are estimated in every iteration). The authors use a *noise cluster*, which in this context is a simple model that can generate all possible expression profiles with less probability and that is excluded from training. A re-assignment of a profile to the model that maximises its likelihood only occurs when the likelihood exceeds that of the profile under the noise model. In order to select the number of clusters they merge small clusters and separate big clusters. The clustering method produce clusters that may contain many different forms of prototype appearances and further analysis within one cluster is needed. In this paper special emphasis is made to partially supervised learning.

In (Luan and Li 2003) the authors propose a mixed-effects model using cubic splines. In this modelling framework, the observed time-series are treated as samples taken from underlying continuous smooth process. Treating the clustering problem as a mixture model problem, they assume that the observed gene time-series come from a mixture of C probability distributions with the cth probability defined by the mixed-effects model with cubic splines proposed. After fitting the mixture model in the framework of the mixed-effects model using an EM algorithm, the authors obtained the smooth mean gene expression curve for each cluster. Then, for each gene, they obtained the best linear unbiased smooth estimate of the gene expression trajectory over time, combining data from that gene and other genes in the same cluster. The

number and locations of the knots for the B-splines corresponding to the mean function and the random effects have to be specified. Cubic B-splines with four equally space knots are used for both mean and random effects, assuming the same spline basis for all the clusters. The number of clusters is determined via BIC scores. They compare their approach with the AR model presented in (Ramoni et al. 2002) concluding that a simple low order AR model is not appropriate for modelling possible non-linear relationship between the gene expression levels at different time points. Also, that the simple AR model as used in (Ramoni et al. 2002) cannot be used for modelling time-trend, which can sometimes be precisely what differentiates among different gene clusters.

6 Future work

Microarray data can be analysed using a wide variety of clustering algorithms with which different information can be obtained. There is not a omnipotent algorithm which can extract every single biological information hidden in the data. Some algorithms will identify different kinds of relations among genes useful for different analysis of the data. For example, for some experiments the objective is to identify genes which picked at a particular time (Cho et al. 1998), in others it is to find genes following similar cycles of expression (Spellman et al. 1998), in others the pattern of gene induction under particular biological conditions (Chu et al. 1998), in others genes that are switch on and off concordantly (Heyer et al. 1999).

In the next two months we will develop a clustering algorithms which meets the required characteristics specified in earlier sections which are not completely followed by any of the existing algorithms proposed for gene expression time-series clustering. The objective of the algorithm will be the identification of genes which truly behave similar across time, meaning that we will not concentrate in the complete partition of the data set, but in the identification of very strongly similar groups. The partition of the whole set forces genes to be grouped together even when they have a small probability or a small membership degree to a given cluster, usually corrupting the model or the prototype, and therefore, the clusters become less suitable for further analysis. Clusters formed using this approach might be good starting points for further analysis of expression data since they have a high quality and the the selection of a number of clusters is not required since it is implicit in the data set.

Several analysis had follow this approach, for example the CAST algorithm. As mention in an earlier section, this technique groups genes based in a measure of their "affinity" to the clusters. Not all the genes meet the affinity threshold, and the resulting clusters are not affected by these genes. Another example is the algorithm presented in (Heyer et al. 1999), where the focus of the algorithm is to find large clusters that have a quality guarantee by finding clusters whose diameter does not exceed a given threshold value d. So any two genes in a cluster have a similarity value that is at least $1 - d$. As in the CAST algorithm, the number of clusters is inherent from the data set. Other similar approach is the adaptive quality-based clustering algorithm from De Smet et al. (2002). This method is an iterative two-step algorithm, the first step is to find a sphere in the data space where the density of expression profiles is locally maximal, based on a preliminary estimate of the radius of the clusters. In the second step, an optimal radius of the cluster is derived so that only significantly co-expressed genes are included in the cluster. CAST and the adaptive quality-based

clustering are two of the leading algorithms for clustering of gene expression, however, they are not designed to meet the specific requirements of temporal data.

Table 3: Main requirements for gene expression time-series clustering and proposed solution.

Requirement	Solution
Number of clusters	Inherent from the data set
Varying membership	Fuzzy clustering
Outliers	Previous identification
Noise	Noise clustering

Table 3 presents the main requirements for gene expression time-series clustering and the proposed solutions. As commented in the previous paragraph, given the particular objectives of the proposed algorithm, the number of clusters is inherent in the data set. Following to the next point of the table, we will use a fuzzy clustering approach which will allow varying membership degrees. Gasch and Eisen (2002) successfully proposed the use of fuzzy clustering for extracting biological insights for gene-expression data, showing the biological relevance of varying membership degrees. Next, the outliers will be identified before hand to avoid their effect in the clustering procedure. Then, in order to appropriately deal with noise and not so obvious outliers, a noise cluster will be implemented.

Identification of outliers. As seen in a previous section, in (Heyer et al. 1999) the proposed similarity measure reduces the effect of outlier values, however, with this method false outliers could be identified since it is just based in the comparison of two genes. Outliers often have a relatively large distance to all of the data groups and are equally shared among the groups. This can be used to identify possible outliers. Figure 9 shows the Euclidean distance between pairs of series with one time point omitted at a time. It can be seen that the pairs which include the series number one, have a similar behavior. They have smaller distance when time point two is omitted, identifying successfully the outlier value of the series number one in the second time point. For large data sets this plot is not informative, but the main idea can still be used. This method identifies an outlier value in context with the similarity measure being used. When the Euclidean distance is used on un-standardised data, outliers will be those values with a high absolute value. There are other methods to identify outliers based in the statistical properties of the data set (Benett and Lewis 1984), which will be properly reviewed in the following months.

The similarity measure. We will implement a hand tailored similarity measure based in the similarity requirements defined in an earlier section. In (Möller-Levet, Klawonn, Cho and Wolkenhauer 2003) a new similarity measure was proposed, but its robustness was limited by the optimisation problems of the fuzzy objective function. We will study the possibility of optimising the objective function using evolutionary strategies, allowing the implementation of a more sophisticated similarity function.

Artificial data set. In order to find out whether an algorithm can meet the specified objectives, we have to define test data with which the performance of the technique can be evaluated for characteristic cases.

Validation. The algorithm will be validated with real biological data set as well with quantitative measures of quality.

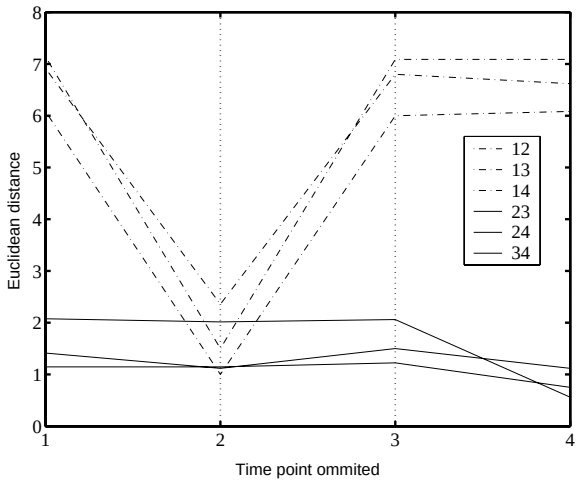


Figure 9: Euclidean distance between pairs of series with one time point omitted at a time.

7 Conclusions

An extensive literature review of similarity of time-series and clustering of gene expression time-series was perform, which help to the elaboration of an outline of basic requirements for the clustering of the aforementioned series. In the following months a new approach will be presented which will use the knowledge acquired in the development of the previous methods and will meet the specific requirements identified in this report.

Appendix

Distance functions

A scalar function, d(x,y), of the ordered pair of vectors x, y, is a distance function if it satisfies the following axioms of a distance measure on $\mathbb{R}^n$:

$$\begin{aligned} d(x,y) &\geq 0 \text{ and } d(x,y) = 0 \text{ if } x = y, \\ d(x,y) &= d(y,x), \\ d(x,y) &\leq d(x,z) + d(z,y) \text{ for any } z. \end{aligned} \tag{6}$$

The most common distance function is the Euclidean distance, which is defined as the distance measured along a straight line from one point to another in the data space. It is the square root of the sum of the squared differences between all the dimensions of two elements:

$$d_E(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{7}$$

Figure 10 shows three expression profiles, A, B and C, where $d_E(A,B) = d_E(C,B) = 2$. However, it can be seen that B is more similar to C than to A. In fact, C is just a linear transformation of B, $C = B - 0.7559$, while A is a completely different pattern.

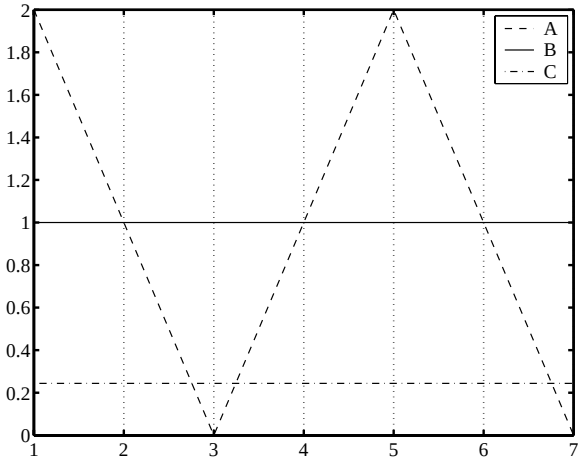


Figure 10: Euclidean distance: three expression profiles, A, B and C, where $d_E(A, B) = d_E(C, B) = 2$.

Although it is frequently used, it is very weak for time-series. It is affected by scaling and shifting. Also, shape, order in the measurements and length of sampling interval are not considered since the distance is formed by sums of intensity differences at each time point.

Another common distance function is the City block, which is defined as the rectilinear route measured parallel to the axes, it corresponds to the sum of the distances on each dimension.

$$d_{CB}(x, y) = \sum_{i=1}^{n} |x_i - y_i|. \tag{8}$$

Both Euclidean distance and City block distance are examples of the more general Minkowski measure, where Euclidean corresponds to m=2 and City block to m=1, [Everitt, 97].

$$d_{Mm}(x, y) = (\sum_{i=1}^{n} (x_i - y_i)^m)^{1/m}. \tag{9}$$

A valuable feature of the Euclidean distance, is that it is preserved under orthonormal transforms. Other distance functions, where $m \neq 2$, to not have this property (Agrawal et al. 1993).

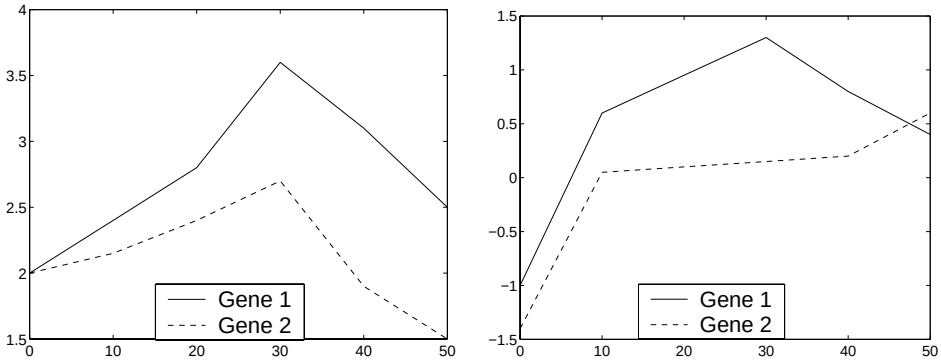
Correlation: a statistical relationship

The Pearson product-moment correlation coefficient, ρ, is a statistical term which measures the linear relationship between two variables. The statistical significance of a correlation coefficient depends on the sample size, defined as the number of independent observations. If time-series are autocorrelated, the sample size must be adjusted downward to account for dependence of successive observations when evaluating significance. The statistical significance of a calculated ρ can be computed if populations from which samples were drawn are normally distributed; otherwise, when the assumption of normality is not satisfied, the procedure can be justified for large samples. Most of the gene expression time-series come from an unknown distribution (Kruglyak and Tang 2001) and are usually very short. Therefore, in this case, a large

correlation coefficient does not necessarily indicate two similarly shaped profiles, nor does a small correlation coefficient necessarily indicate differently shaped profiles. As in Peddada et al. (2003), Figure 11 presents an example to illustrate this point.

$$\rho = \frac{cov(x,y)}{\sqrt{var(x)\ var(y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{10}$$

Considering the time-series as vectors, the correlation coefficient measures the cosine



(a) Similar shaped profiles (b) Different shaped profiles

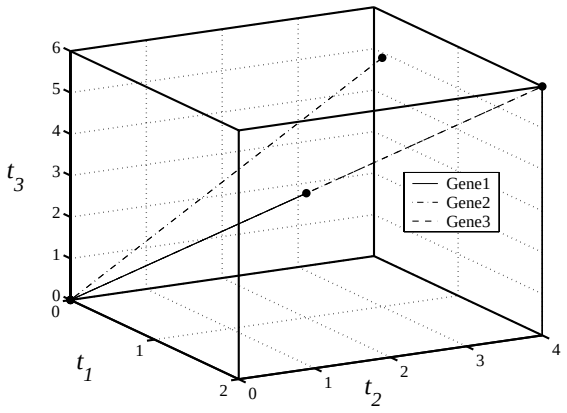
Figure 11: Correlation coefficient and profile shape. (a) Similar shaped profiles with $\rho = 0.56$, (b) Different shaped profiles with $\rho = 0.82$.

of the angle between the expression vectors minus their means. Therefore, vectors with a correlation coefficient of zero are said to be orthogonal (uncorrelated), with a correlation coefficient of one are parallel in the same direction (positively correlated), and with a correlation coefficient of -1 are parallel but in opposite direction (negatively correlated). All vectors parallel on the same direction have similar expression profiles in the time domain. These profiles have the same shape and could be scaled as illustrated in figure 12. Shifted vectors are un-shifted by the subtraction of the mean, as can be seen from equation (10)

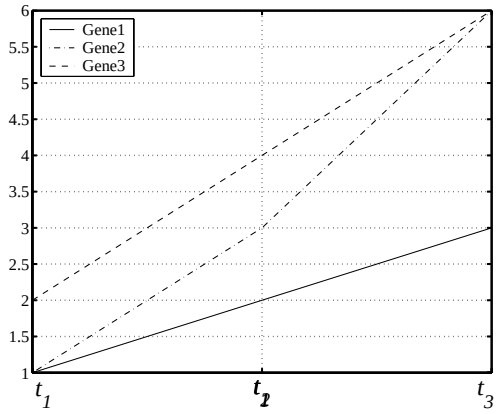
The correlation-based rest on the assumption that the set of observations for each gene are independent and identically distributed. The correlation or Euclidean distance are invariant with respect to the order of observations: if the temporal order of a pair of series is permuted their correlation or Euclidean distance will no change. Additionally, the correlation coefficient is very problematic in cases when we are dealing with very short time-series, measured at ten or less time points.

Expectation-Maximisation algorithm

In the EM algorithm, the Expectation (E) steps and Maximisation (M) steps alternate. In the E-steps, the probability of each observation belonging to each cluster is estimated conditionally on the current parameter estimates. In the M-step, the model parameters are estimated given the current group membership probabilities. When the EM algorithm converges, each observation is assigned to the group with the maximum conditional probability.



(a) Vectors in the data space



(b) Equivalent time-series in the time domain

Figure 12: Data space and equivalent time domain. (a) Three vectors in the data space where Gene1 and Gene3 are parallel, thus, $\rho = 1$ (Gene3 is a linear transformation of Gene1, Gene3=2(Gene1)). (b) The corresponding time-series of vectors in (a), the three profiles show an increase of the expression level at each time point, although the rate of Gene1 and Gene3 is constant (although different), while the rate of G2 is not constant.

References

Aach, J.: 2001, Aligning gene expression time series with time warping algorithms, *Bioinformatics* **17**(6), 459–508.

Aerts, C. and De Cat, P.: 2003, Beta cep stars from a spectroscopic point of view, *Space Science Reviews* **105**(1-2), 453–459.

Agrawal, R., Faloutsos, C. and Swami, A.: 1993, Efficient similarity search in sequence databases, *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms (FODO'92)*, Chicago, USA.

Agrawal, R., Lin, K. I., Sawhney, H. S. and Shim, K.: 1995, Fast similarity search in the presence of noise, scaling , and translation in time-series databases., *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB'95)*, Zurich, Switzerland, pp. 490–501.

Akaike, H.: 1974, A new look at the statistical model identification, *IEEE TRans. Automat. Contr.* **AC-19**, 716–723.

Anderson, T.: 1958, *The Statistical Analysis of Time Series*, John Wiley & Sons, Inc., London, England.

Antunes, C. M. and Oliveira, A. L.: 2001, Temporal data mining: an overview, *KDD 2001 Workshop on Temporal Data Mining*, San Francisco, EUA.

Arnau, J. and Bono, R.: 2001, Autocorrelation and bias in short time series: An alternative estimator, *Quality & Quantity* **35**, 365–387.

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

Ben-Dor, A. and Yakhini, Z.: 1999, Clustering gene expression patterns, *RECOMB.*, Lyon France, pp. 33–42.

Bence, J. R.: 1995, Analysis of short time series: correcting for autocorrelation, *Ecology* **76**(2), 628–639.

Benett, V. and Lewis, T.: 1984, *Outliers in statistical data*, John Wiley & Sons, Norwich, Great Britain.

Boschen, F., J. and Weise, C.: 2003, What starts inflation: Evidence from the OECD countries, *Journal of Money Credit and Banking* **35**(3), 323–349.

Box, G. E. and Jenkins, G. M.: 1976, *Time Series Analysis forcasting and control*, Holden-Day, Inc., Oakland, California.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. and Davis, R. W.: 1998, A genome-wide transcriptional analysis of the mitotic cell cycle, *Mol. Cell* **2**, 65–73.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. and Herskowitz, I.: 1998, The transcriptional program of sporulation in budding yeast, *Science* **282**, 699–705.

Das, G., Gunopulos, D. and Mannila, H.: 1997, Finding similar time series, *Principles of Data Mining and Knowldege Discovery*, pp. 88–100.

Dave, R. N.: 1991, Characterization and detection of noise in clustering, *Pattern Recognition Letters* **12**, 657–664.

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

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B. and Moreau, Y.: 2002, Adaptive quaqlity-based clustering of gene expresion profiles, *Bioinformatics* **18**(5), 735–746.

Everitt, B.: 1974, *Cluster Analysis*, Heinemann Educational Books, London, England.

Faloutsos, C., Fanganathan, M. and Monolopoulos, Y.: 1994, Fast subsequence matching in time-series, *Proceedings 1994 ACM SIGMOD Conference*, Mineapolis, MN, pp. 419–429.

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

Fraley, C. and Raftery, A. E.: 1998, How many clusters? Which clustering method? Answers via model-based cluster analysis, *The Computer Journal* **41**(8), 578–588.

Frawley, W., Piatetsky-Shapiro, G. and Matheus, C.: 1992, Knowledge discovery in databases: An overview, *AI Magazine* **13**(3), 57–70.

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

Guo, J.J., Curkendall, S., Jones, J., Fife, D., Goehring, E. and She, D. W.: 2003, Impact of cisapride label changes on codispensing of contraindicated medications, *Pharmaceopidemiology and Drug Safety* **12**(4), 295–301.

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

Höppner, F., Klawonn, F., Kruse, R. and Runkler, T.: 1999, *Fuzzy Cluster Analysis*, John Wiley & Sons, Chichester, England.

Jain, A. K. and Dubes, R. C.: 1988, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ.

Ji, X., Li-Ling, J. and Sun, Z.: 2003, Mining gene expression data using a novel approach based on hidden Markov models, *FEBS* **542**, 125–131.

Kendall, S. M.: 1976, *Time-Series*, Charles Griffin & Company Ltd., London.

Keogh, E. J. and Pazzani, M. J.: 1998, An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback, *in* R. Agrawal, P. Stolorz and G. Piatetsky-Shapiro (eds), *Fourth International Conference on Knowldege Discovery and Data Mining (KDD'98)*, pp. 239–241.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P.: 1983, Optimization by simulated annealing, *Science* **220**, 827–846.

Kruglyak, S. and Tang, H.: 2001, A new estimator of significance of correlation in time series data, *Journal of Computational Biology* **8**(5), 463–70.

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

Lukashin, A. and Fuchs, R.: 2001, Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters, *Bioinformatics* **17**(5), 405–414.

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

Möller-Levet, C. S., Klawonn, F., Cho, K.-H. and Wolkenhauer, O.: 2003, Fuzzy clustering of short time-series and unevenly distributed sampling points, *LNCS, Proceedings of the IDA2003.*

Notohardjono, B. D. and Ermer, D. S.: 1986, Time-series control charts for correlated and contaminated data, *J. Eng. Ind-T. ASME* **108**(3), 219–226.

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

Popivanov, I. and Miller, R. J.: 2002, Similarity search over time-series data using wavelets, *Proceeding of the 18th International Conference on Data Engineering (ICDE'02).*

Rabiner, L. R.: 1989, A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE*, Vol. 77, pp. 257–286.

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

Roddick, J. and Spiliopoulou, M.: 2002, A survey of temporal knowledge discovery paradigms and methods, *IEEE Transactions on Knowledge and Data Engineering* **14**(4), 750–767.

Sankoff, D. and Kruskal, J.: 1983, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison Wesley.

Schleip, A., Schönhuth, A. and Steinhoff, C.: 2003, Using hidden markov models to analyze gene expression time course data, *Bioinformatics* **19**(1), i255–i263.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B.: 1998, Comprehensive identification of cell cycle-regulated genes of yeast *Saccharamyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell* **9**, 3273–3297.

Tilman, D. and Wedin, D.: 1991, Oscillations and chaos in the dynamics of a perennial grass, *Nature* **353**(6345), 653–655.

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

Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. and Somogyi, R.: 1998, Large-scale temporal gene expression mapping of central nervous system development, *Proc. Natl Acad. Sci. USA* **95**, 334–339.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L.: 2001, Model-based clustering and data transformations for gene expression data, *Bioinformatics* **17**(10), 977–987.

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

Yum, M. K. and Kim, J. H.: 2003, A very-short-term intermittency of fetal heart rates and developmental milestone, *Pediatric Research* **53**(6), 915–919.

Zhong, S. and Ghosh, J.: 2002, A unified framework for model-based clustering, *Smart Engineering System Design: Neural Networks, Fuzzy Logic, Evolutionary Programming, Data Mining and Complex Systems. ANNIE 2000*, Vol. 10 of *Intelligent Engingeering Through Artificial Neural Networks.*, ASME Press.