Clustering Analysis of Gene Expression Time Series Data

Huang-Cheng Kuo¹*, Cheng-Che Wu¹, Tsung-Lung Lee¹, Jen-Peng Huang²

¹Department of Computer Science and Information Engineering
National Chiayi University, Chiayi 600, Taiwan
²Department of Information Management
Southern Taiwan University of Technology, Tainan 710, Taiwan
*E-mail: hckuo@mail.nctu.edu.tw

Abstract

Microarray is used to generate large amount of gene expression data and observing the differences among gene expression levels. Gene expression time series data represents the trend of gene behaviors. Clustering is a popular analysis for gene expression time series data. Genes in the same cluster have similar behavior. Cluster analysis helps people investigate the relativity among genes.

We propose a similarity measurement, named LCSS (Longest Common Subseries Similarity), to overcome the influence of “shift-effect,” which is not well handled in commonly used similarity measurements. Also, we use the sequential pattern mining technique to decide the number of clusters for partitioning-based clustering algorithm. A mechanism based on nearest neighbor is used as the criteria for objects relocation in the clustering algorithm.

Keyword: Microarray, Gene Expression, Time series, Clustering, Similarity Measurement

1. Introduction

Following the accomplishment of sequence-tagged site, biotechnology has been paid much attention to and has become one of the most popular researches. Genes, the inheritance factors of organisms, dominate the physiology of growth, disease, behavior, and so on. Gene expressions can be examined to discover the functionalities of organisms. Biologists are interested in what genes are expressed at what part of an organism’s life cycle. Several materials have been developed for gene functionalities discovery, and microarray is an extensively used material. Microarray experiments are frequently used for carrying out pair-wise comparisons between samples and to discover the co-expression genes according to monitor the expression levels for thousands of genes. Clustering is currently the most frequently used technique to analyze gene expression time series [1].

Tendencies of gene expression and the intensity of gene expression are concerned from the viewpoint of biology. Genes with similar expression tendency are considered that they may have similar regulative mechanism in physiological functions. Gene expression time series clustering is to discover the co-expressions of genes and to retrieve the patterns of gene expressions. Determining the similarities between objects is one of the major factors for clustering and may influence the quality of clustering. There are many different similarity definitions for time series. Most of them are based on distance measurements and correlation coefficients, such as Euclidean distance measurement and Pearson correlation coefficient. But, these similarity measurements do not deal with important characteristics of microarray time series data. Those properties are scale, baseline shift, outliers, and shift effect [2]. We devise a similarity definition, called Longest Common Subseries Similarity (LCSS), to
take these properties into consideration.

The major issue of partition-based clustering method is that users have to specify the target number of clusters before clustering. This is usually done by means of try-and-error to pick up a number from a large range. In our study, the target number of clusters is identified by translating time series into sequences of symbols and applying frequent pattern mining technique on the sequences. For the partition-based clustering, we use nearest neighbor concept to set the criteria for re-locating objects among clusters.

In the second session, a similarity measurement, LCSS, is described. A partition-based clustering algorithm which uses nearest neighbors as the criteria of objects relocation is discussed in the third session. In the forth session, the experiment results are presented, and the fifth session is the conclusions and future works.

2. Similarity Measurement for Gene Expression Time Series

There are four characteristics when calculating similarities between time series: (1) scale, (2) baseline Shift, (3) outliers, and (4) shift effect.

2.1 Characteristics of time series similarity

A. Scale and baseline shift

Two time series with scale effect is that the general shape of the time series may be identical, but the scale may be somewhat different. Two time series with baseline shift effect indicates that the actual baseline values may differ. Both challenges will reduce the accuracy of time series similarities when using distance based similarity measurements. One possible solution is to normalize time series, and another possible solution is to use correlation based similarity measurements for time series similarities.

B. Outliers

In gene expression time series, the number of time points is usually restricted. So, an outlier will have a high influence on the similarity measure. A possible solution is to skip outliers and to reduce the impact on time series similarities. Jackknife correlation measurement is an adaptive similarity metric to calculate similarity by skipping one outlier.

C. Shift effect

Two time series are considered to have shift effect indicates temporal gaps between two time series. In other words, two time series might look very distant, but they might be similar if a few items are dropped off from the two time series. Figure 1 illustrates an example of two time series with shift effect. If the 10th point of time series X and the 2nd point of time series Y are dropped off, the two time series are quite match.

Dynamic programming technique can be used to discover the best alignment between time series when shift effect exists [3]. In our research, the dynamic programming technique is used to discover the longest common segments between time series for similarity.

2.2 LCSS similarity measurement

According to the background of time series similarity
measurements, we propose a similarity measurement, LCSS (Longest Common Subseries Similarity), to calculate similarities between gene expression time series in our research. The concept of LCSS is to discover the longest common segments between two time series, and the longer common segments indicates the higher similarity. Given two time series \( X = \{x_1, x_2, \ldots, x_n\} \) and \( Y = \{y_1, y_2, \ldots, y_n\} \), both \( X \) and \( Y \) are normalized first to avoid the influences of scale and baseline shift. Z-score normalization is used to normalize time series. For discovering the longest common segments between time series, unlike symbol, expression levels are numeric values and one problem is how to determine two numeric values are “alike”. The definition of \( \varepsilon \)-match is used to determine whether two time points are alike.

**Definition 1:** Given two time series \( X = \{x_1, x_2, \ldots, x_n\} \) and \( Y = \{y_1, y_2, \ldots, y_n\} \) with length \( n \), and a parameter \( 0 < \varepsilon \leq 1 \), the time points \( x_i \) and \( y_j \) are considered equivalent if and only if

\[
0 \leq |x_i - y_j| \leq \text{MAX}\left(\varepsilon |x_i|, |y_j| \times \varepsilon\right),
\]

where \( 1 \leq i, j \leq n \), and the parameter \( \varepsilon \) is for controlling the toleration between two time points. If the inequality is satisfied with \( x_i, y_j \), then \( x_i, y_j \) are called \( \varepsilon \)-match.

By the definition of \( \varepsilon \)-match, two time points which are \( \varepsilon \)-match are considered to be alike, and the largest number of \( \varepsilon \)-match between two time series can be discovered.

Let LCSL (Longest Common Subseries Length) be the largest number of \( \varepsilon \)-match time points between two time series. LCSL can be computed by dynamic programming. Then we can obtain the longest common subseries in both time series. For example, given two time series \( X = \{0.32, 0.47, 0.89, 0.66, 1.02, 0.79, 0.54, 0.21\} \) and \( Y = \{0.21, 0.12, 0.38, 0.69, 0.98, 1.11, 0.84, 0.52\} \), and \( \varepsilon = 0.5 \). The longest common subseries of \( X \) is \{0.32, 0.47, 0.89, 0.66, 1.02, 0.79, 0.54\} and the longest common subseries of \( Y \) is \{0.21, 0.38, 0.69, 0.98, 1.11, 0.84, 0.52\}. Therefore, we have a simple similarity definition of time series \( X \) and \( Y \) which is the ratio of LCSL, with respect to a certain \( \varepsilon \), to the length \( n \). In this example, the LCSL of time series \( X \) and \( Y \) is 7, and the similarity between \( X \) and \( Y \) is \( 7/8 = 0.875 \).

Moreover, in order to distinguish the similarities of time series which the LCSL are equal, weights are applied to the similarities. Pearson correlation coefficient, denoted as \( r(X', Y') \) where \( X' \) and \( Y' \) are subseries is used as weight. Therefore, the longest common subseries similarity of time series \( X \) and \( Y \) with length \( n \) is given by:

\[
LCSS(X, Y) = \left(\frac{\text{LCSL}(X, Y)}{n}\right)^{1/(1+\varepsilon)}.
\]

3. Nearest Neighbors Based Clustering Algorithm

In this session, a target number of clusters prediction mechanism is proposed first. Then, the nearest neighbors based clustering algorithm is discussed.

3.1 Number of clusters

A serious challenge of partition-based clustering is to specify the target number of clusters and this is usually done by try-and-error.

In gene expression time series data, genes have the great part of similar expression segments may be considered to group in the same class. In other words, genes have the same expression patterns may be grouped into one cluster. According to the idea, the number of clusters can be generated by discovering the expression patterns during gene expression time series data.

We use frequent pattern mining technique to discover expressions patterns by translating time series into sequences. Given a time series \( T = \{t_1, t_2, \ldots, t_n\} \) with length \( n \), a sequence \( S = \{i_1, i_2, \ldots, i_n\} \)
is translated from time series $T$ by:

$$i_k = \left\lfloor \frac{t_{k+1} - t_k}{0.5 \times \sigma_T} \right\rfloor, 1 \leq k \leq n-1,$$

where $\sigma_T$ is the standard deviation of time series $T$. The gene expression sequence dataset is generated from the gene expression time series dataset, and is used to discover expression patterns.

Several frequent pattern mining techniques can be used such as Apriori-based algorithms [4]. In frequent pattern mining, the min_support parameter is the threshold that a pattern should be supported by at least min_support sequences.

By mining frequent patterns, gene expression time series data is separated into several groups according to the frequent patterns, and the groups are regarded as the initial clusters for gene expression time series clustering.

### 3.2 Re-location by nearest neighbors

Many clustering methods have been used to cluster gene expression time series data. In this paper, a kind of partition-based clustering algorithm which used nearest neighbors as the relocation criteria is proposed for gene expression time series.

Nearest neighbors based clustering algorithm uses nearest neighbors to replace the centroid-based technique in K-means algorithm. The general conception is that one object may exist in the cluster which contains the most nearest neighbors. A parameter $k$ is used to determine the number of nearest neighbors for one object in nearest neighbors based clustering.

**Definition 2:** For an object $g$, a candidate cluster of $g$ is a cluster which contains at least one nearest neighbor of $g$.

**Definition 3:** The cohesion of one cluster $C$ indicates the average similarity between objects in cluster $C$.

$$\text{cohesion}(C) = \frac{\sum_{i \neq j, i \in C} \text{sim}(i, j)}{|C|}$$

For clustering gene expression time series data by nearest neighbors clustering algorithm, all objects (genes) are separated into several clusters by mining frequent patterns which has been discussed in section 4.1. And the $k$ nearest neighbors of each object $g_i$ must be discovered and stored. Subsequently, for each object $g_i$, all the candidate clusters are collected and kept in one list structure $\text{cand\_list}$. The candidate clusters kept in $\text{cand\_list}$ are ranked according to the number of nearest neighbors they contain, and the rank of each candidate cluster can be regarded as the priority of relocation. Afterward, an iterative algorithm starts to reassign all objects into suitable clusters.

When one object $g_i$ is reassigned from one cluster $C_i$ to another cluster $C_j$, the relocated criterion which the cohesion of $C_j$ does not decrease when $g$ joins must be followed. For each object, the major candidate cluster which has the largest rank and satisfies the relocated criterion will be found. Finally, all objects are reassigned to the When one object $g$ is reassigned from one cluster $C_i$ to another cluster $C_j$, the relocated criterion which the cohesion of $C_j$ does not decrease when $g$ joins must be followed. For each object, the major candidate cluster which has the largest rank and satisfies the relocated criterion will be found. Finally, all objects are reassigned to the where $t$ is the number of re-location iterations; $n$ is the number of objects; $c$ is the number of clusters; and $k$ is the number of nearest neighbors.

### 4. Experiment Results

In this paper, the accuracy of similarity measurements and clustering algorithms will be compared.

#### 4.1 Similarity measurements experiments

In the experiments of similarity measurements, the
Cho/Spellman dataset which contains 6178 ORFs in yeast is used. The Cho/Spellman data is comprised of four time series datasets which have been synchronized with distinct methods. It is available at http://genome-www.stanford.edu/cellcycle. And the information about Cho/Spellman dataset is represented in Table 1 which the Data set indicates the methods used to synchronize the yeast cells; Start indicates the start time point of experiments; Terminate indicates the terminate time point of experiments; Period indicates the intervals between continuous time points; # time points indicates the number of time points; and # orfs indicates the number of time series that have no missing values in them.

For comparing the accuracy of different measurements, the alpha dataset which has fewer missing values and constant samples intervals in the Cho/Spellman dataset is most suitable. In alpha dataset, 343 pairs of genes have been discovered to have activation by Filkov etc. [5], and the similarity distribution of the 343 activation is shown in Table 2.

In order to recognize the effectiveness of similarity measurements, additional 657 gene pairs which are distinct from the 343 activation in the alpha dataset will be joined with the 343 activation. The 1000 pairs are used to calculate the similarities and analyze the similarity distribution. Because the similarity domain of Pearson correlation is -1 to 1, we adjusted the domain as 0 to 1 by treating the correlation between -1 and -0.8 as the similarity between 0 and 0.1 and treating the correlation between 0.8 and 1.0 as the similarity between 0.9 and 1.0 identically. The similarity distribution of Pearson correlation coefficient and LCSS is illustrated in Figures 2 and 3.

Table 1: Cho/Spellman gene expression time series dataset

<table>
<thead>
<tr>
<th>Period</th>
<th># time points</th>
<th># orfs</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 min</td>
<td>18</td>
<td>3361</td>
</tr>
<tr>
<td>20 min (10 ~ 70, 250 ~ 290)</td>
<td>24</td>
<td>3453</td>
</tr>
<tr>
<td>10 min (70 ~ 250)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 min</td>
<td>17</td>
<td>1188</td>
</tr>
<tr>
<td>30 min</td>
<td>24</td>
<td>4753</td>
</tr>
</tbody>
</table>

Table 2: Similarity distribution of the 343 activation in the alpha dataset

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>≥-0.8</th>
<th>≥-0.6</th>
<th>≥-0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activation pairs</td>
<td>343</td>
<td>338</td>
<td>316</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>≥-0.2</th>
<th>≥0</th>
<th>≥0.2</th>
<th>≥0.4</th>
<th>≥0.6</th>
<th>≥0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activation pairs</td>
<td>250</td>
<td>179</td>
<td>102</td>
<td>54</td>
<td>23</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 2: The similarity distribution between Pearson correlation coefficient and LCSS (ε = 0.3)
**4.2 Clustering algorithms experiments**

In the experiments, we use the Iyer’s data which contains gene expression levels of 517 human genes in response to serum stimulation over 12 time points. The Iyer’s data and documents is available at the website [http://genome-www.stanford.edu/serum/](http://genome-www.stanford.edu/serum/).

We first compare the number of clusters obtained by giving different minimal supports to the frequent pattern mining algorithm, and the result is shown in Figure 4.

The numbers of final clusters between 10 and 13 are generated when minimal supports are 5% and 10%, respectively, for any number of nearest neighbors. And the numbers of final clusters are both 7 clusters when minimal supports are 15% and 20% for any number of nearest neighbors. From the results, more frequent patterns are generated when using lower minimal supports, so more initial clusters are generated for clustering. When clustering process finished, more clusters will be reserved. Higher minimal supports lead to generate less number of frequent patterns, and fewer clusters will be reserved finally. Subsequently, the quality of clustering algorithm is experimented.

For verifying the accuracy of clustering, the cluster validity assessment, Dunn’s validity index [6], is used. The formula of Dunn’s validity index is given by:

$$D = \min_{1 \leq i, j \leq n} \left\{ \frac{\min_{1 \leq k \leq n} \left\{ d(c_i, c_k) \right\}} {\max_{1 \leq k \leq n} \left\{ d'(c_k) \right\}} \right\},$$

where $d(c_i, c_j)$ is the distance between cluster $i$ and cluster $j$, $d'(c_k)$ is the individual distance in the cluster $k$, and $n$ is the number of clusters. The major implication of Dunn’s validity index is to find the ratio of minimal inter-cluster distance to maximal intra-cluster distance. So, if one clustering algorithm got the large minimal inter-cluster distance and the small maximal intra-cluster distance, and it will get a larger Dunn’s validity index value what it means the algorithm is more effective.

In our experiments, the Dunn’s validity index is modified to be suitable for similarity based measurements, and the formula is given by:

$$D = \max_{1 \leq i \leq n} \left\{ \frac{\min_{1 \leq j \leq n} \left\{ s'(c_k) \right\}} {\max_{1 \leq j \leq n} \left\{ s(c_i, c_j) \right\}} \right\},$$

where $s(c_i, c_j)$ is the similarity between cluster $i$ and cluster $j$, $s'(c_k)$ is the individual similarity in the cluster $k$, and $n$ is the number of clusters. Identically, the higher Dunn’s validity index value indicates the better clustering.

Figure 5 shows the Dunn’s validity index at different minimal supports and different number of nearest neighbors. In Figure 12, we can find that the quality of clustering results is the same nearly and the best clustering results appear when the number of nearest neighbors is 13.

Finally, different clustering algorithms are used to compare the qualities. In our experiments, both K-medoids clustering algorithm and agglomerative hierarchical clustering algorithm are used to compare with nearest neighbors clustering algorithm. The
LCSS similarity measurement which $\varepsilon$ is 0.7 is used by the three clustering algorithm and the experiment results are represented in Figure 6.

![Figure 5](image1.png)

Figure 5: The Dunn’s validity index at different minimal supports and different number of nearest neighbors

![Figure 6](image2.png)

Figure 6: The Dunn’s validity index of different clustering algorithms

5. Conclusions and Future Works

Biological informatics becomes more and more important in recent years. Powerful analytic techniques are essential to retrieve the information from the biological data which grows day by day. In this paper, a partition-based clustering algorithm is proposed to cluster gene expression time series.

For measuring time series similarities, the LCSS measurement is proposed to deal with both similar expression segments and intensity of gene expression simultaneously. LCSS discovers the longest common segments, and therefore scale, baseline shift, outliers, and shift effect problems can be resolved effectively.

For clustering, the challenge to predict the target number of clusters is resolved by mining frequent patterns. Time series which support identical pattern will be grouped in the same cluster, and the clusters are generated more confident.

Finally, a partition-based clustering algorithm used nearest neighbors for object relocation is proposed. From the experiments, the results showed that the nearest neighbors based clustering algorithm has better cluster quality than K-medoids clustering algorithm and hierarchical clustering algorithm.

For the future work, in similarity measurement, a parameter $\varepsilon$ is used to control the toleration of expression level for LCSS, and similarity between two time series may get slight difference with different $\varepsilon$. How to choose a suitable value for $\varepsilon$ can be studied. If the value of $\varepsilon$ is chosen too small, just the pair of time points which are very close will be taken into account and the great majority time points are considered to be dissimilar. Oppositely, if the value of $\varepsilon$ is chosen too large, the great majority time points are considered to be similar and are used to calculate the similarity between time series. The pair of time points which have greater difference may not be ignored when the value of $\varepsilon$ is chosen too large. Sometimes, the value of $\varepsilon$ is chosen too small or too large may lose some information of biology. Therefore, a suitable value for $\varepsilon$ may be necessary to be discovered in the future.

In initial cluster generation, the number of clusters may be influenced by setting different minimal supports. How to discover a suitable minimal support towards data with different scales can be conferred further.

Finally, to visualize gene expression data is useful for biologists. So, how to visualize the clustered data by reducing dimensionalities into lower dimensions is an interested research that can be implemented in the future.

Acknowledgement

This work is partially supported by National Science
Council, Taiwan under grant NSC 93-2213-E-415-005.

References


