Correcting the loss of cell-cycle synchrony in clustering analysis of microarray data using weights

Fenghai Duan and Heping Zhang

Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520-8034, USA

Received on July 11, 2003; revised on February 23, 2004; accepted on February 27, 2004
Advance Access publication May 27, 2004

ABSTRACT

Motivation: Due to the existence of the loss of synchrony in cell-cycle data sets, standard clustering methods (e.g. k-means), which group open reading frames (ORFs) based on similar expression levels, are deficient unless the temporal pattern of the expression levels of the ORFs is taken into account.

Methods: We propose to improve the performance of the k-means method by assigning a decreasing weight on its variable level and evaluating the ‘weighted k-means’ on a yeast cell-cycle data set. Protein complexes from a public website are used as biological benchmarks. To compare the k-means clusters with the structures of the protein complexes, we measure the agreement between these two ways of clustering via the adjusted Rand index.

Results: Our results show the time-decreasing weight function—\( \exp\left(-\frac{1}{2}\left(t^2/C^2\right)\right)\) which we assign to the variable level of k-means, generally increases the agreement between protein complexes and k-means clusters when C is near the length of two cell cycles.

Contact: Heping.Zhang@Yale.edu

INTRODUCTION

Motivation

Analysis of microarray data has attracted a great deal of attention. The number of papers referring to microarray data has increased at an exponential rate during the past 5 years. One of the most commonly analyzed data sets is Spellman et al.’s (1998) yeast cell-cycle data set that includes detailed cell-cycle synchronization information, namely α-factor, elutriation and cdc15 synchronized experiments, as well as Cho et al.’s (1998) cdc 28 synchronized ones.

The transcriptions of many genes oscillate in cell cycles and reach a peak within every cell cycle. From cell-cultured samples, we can only observe the averaged mRNA levels for all cells at various cycles. Thus, it is imperative to examine the consequence of using mixed-cycle data as opposed to synchronized-cell data in statistical analyses. This issue has not been adequately investigated in microarray data analysis (e.g. Spellman et al., 1998).

Cells lose their synchrony property when released from the synchronization point because they move across the cell cycle at different rates. This process has a cumulative effect over time. As a result, the measured mRNA levels deviate increasingly from the true expression level when the measurement time is further away from the starting point. The clustering methods that fail to take into account this phenomenon could produce misleading results. To our knowledge, the extent of how significantly this phenomenon may affect gene clustering has not been studied.

In order to assess this impact, we select protein complexes as biological benchmarks from an existing database constructed by Jansen et al. (2002). They showed that the correlation between two proteins in the same protein complex is significantly higher than the correlation between randomly selected protein pairs. They further demonstrated that the proteins in the permanent complexes reveal higher correlations than the proteins in the transient complexes. From a biological viewpoint, the major difference between permanent and transient complexes is that the permanent complexes could be maintained for a longer time throughout the cell cycle or be maintained through more cellular conditions than transient ones (Jansen et al., 2002).

Furthermore, the interaction between two proteins in the same protein complex tends to be the least affected by the ubiquitous post-translational modification compared to other protein interactions, such as proteins in the same regulatory network (e.g. interaction between transcriptional factor and its target), proteins in the same metabolic pathway (e.g. interaction between two enzymes) and so on. An example of post-translational modifications is phosphorylation—at least 30% of all proteins are thought to contain covalently bound phosphate (Pandey and Mann, 2000).
The above observations are important and form the basis for using the same database as Jansen et al. (2002) to evaluate the performance of clustering methods.

Outline and summary
In the Data and Methods section, we conduct a simulation study to evaluate the synchrony-losing function that measures the extent to which a sampled mRNA level loses its synchrony property over time. Analyses of existing databases are conducted to assess the performance of clusters constructed from k-means. We use protein complexes as the biological clusters and the adjusted Rand index (ARI) (Hubert and Arabie, 1985) as the measure of agreement between two clusterings. We also discuss how to correct for the loss-of-synchrony (LOS) in clustering analysis.

DATA AND METHODS
Yeast cell-cycle data
We use the yeast cell-cycle data set collected and analyzed by Spellman et al. (1998), which also includes Cho et al.’s (1998) yeast cell-cycle data. This cDNA microarray data set is publicly available at http://genome-www.stanford.edu/cellcycle/. It contains 18 α-factor synchronized, 24 cdc15 synchronized and 17 cdc28 synchronized as well as 14 elutriation synchronized time-series data points. It also contains two time points for the clb2 and two time points for cln3 induction experiments. So, the total number of time points is 77. All time points were normalized so that the average log-transformed intensity over the course of the experiments equals zero (Spellman et al., 1998). The database includes 6178 open reading frames (ORFs) (Cho et al., 1998) and 97% of them are known and predicted genes.

Protein complex data
The protein complexes in our analysis were obtained from the website at http://bioinfo.mbb.yale.edu/expression/integrates/. We extracted 199 protein complexes of yeast Saccharomyces cerevisiae involving 874 ORFs that have 80% or more available data points. All the protein complexes were originally from the well-known yeast S. cerevisiae database (http://www.mips.biochem.mpg.de).

Simulated data
Before applying our approach to real data, we first evaluated the performance of our weighted k-means on simulated data sets. Let \( y_{it} \) denote the expression of the \( i \)-th gene at time \( t \) in the simulated data sets, then we can describe it as the sum of a fixed component and a random component. Namely,

\[
y_{it} = f_{it} + \epsilon_{it},
\]

where the fixed component, \( f_{it} \), corresponds to its gene expression temporal pattern while the random one, \( \epsilon_{it} \), corresponds to the noise introduced during the measurement. For the fixed component we refer to Quackenbush (2001) and use nine distinct temporal patterns over 10 time points (time \( t \) ranges from unit 1 to unit 10). We use Gaussian noise with mean zero and time-increasing variance \( \sigma^2(t) = 0.16 \exp(t^2/\xi^2) \), where \( \xi \) is a constant that controls the variance increasing rate over time. In our simulation, it is set at 5. We consider 50 genes in the experiment and replicate the simulation 100 times, resulting in 100 simulated data sets.

k-Means clustering algorithm
Given the pre-specified number of clusters, the \( k \)-means algorithm allocates the observations into different groups in order to minimize the within-group sum of squares. Suppose we have \( n \) objects and \( m \) variables: \( y_{it} \) for \( i = 1, 2, \ldots, n \) and \( t = 1, 2, \ldots, m \). Then, the within-group sum of squares is

\[
\min_{S_k} \left( \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{t=1}^{m} (y_{it} - \bar{y}_{kt})^2 \right),
\]

where \( K \) is the pre-specified cluster number, \( S_k \) the set of objects in the \( k \)-th cluster and \( \bar{y}_{kt} \) the mean for the variable \( t \) over cluster \( k \).

Weights can be incorporated into the within-group sum of squares at two levels. One is at an object level that is illustrated in the NAG C library developed by Numerical Algorithms Group Ltd (2001). The other is at the variable level, as will be introduced and used shortly. Our modified clustering is based on the within-group weighted sum of squares defined below; we refer to this as weighted \( k \)-means clustering.

The weighted within-group sum of squares is defined as

\[
\min_{S_k} \left( \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{t=1}^{m} w_i(y_{it} - \bar{y}_{kt})^2 \right)
\]

at the object level, and

\[
\min_{S_k} \left( \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{t=1}^{m} w_t(y_{it} - \bar{y}_{kt})^2 \right)
\]

at the variable level, where \( w_i \) and \( w_t \) denote the corresponding weights in the formulae. When we perform the weighted \( k \)-means clustering at the variable level, we can simply modify our working data set by multiplying each variable by our weight function (i.e., \( \sqrt{w_i} \)) and then using the standard \( k \)-means software. Our analysis is performed using Eisen’s CLUSTER program (http://rana.lbl.gov/) (Eisen et al., 1998).

Weight function
As illustrated in Simulation data, we describe the temporal expression pattern of a gene by a two-component function:

\[
g(t) = f(t) + \epsilon(t),
\]

where \( f(t) \) is an oscillating function for the fixed effect and \( \epsilon(t) \) the random error. To simulate the cumulative effect of LOS over time \( t \), we consider a weight function \( \phi(t) \) with an
increasing variance over time, e.g. \( N(0, \sigma_t^2) \), where \( \sigma_t \) is an increasing function of time \( t \). Then, after LOS, the value of the gene expression level at time \( t \) is

\[
h(t) = \int g(t + x)\phi(x)\,dx
\]

\[
= \int f(t + x)\phi(x)\,dx + \int \varepsilon(t + x)\phi(x)\,dx.
\]

Considering a Fourier transformation on \( f(t) = \sum_{k=1}^{\infty} c_k e^{ikr} \), and after some algebraic manipulations, we have

\[
\int f(t + x)\phi(x)\,dx = \int \sum_{k=1}^{\infty} c_k e^{ik(t+x)} \frac{1}{\sqrt{2\pi\sigma_t}} e^{-(1/2)(x^2/\sigma_t^2)}\,dx
\]

\[
= \sum_{k=1}^{\infty} \exp\left(-\frac{k^2\sigma_t^2}{2}\right) c_k e^{ikx}\,dx.
\]

Therefore, after LOS, the fixed component becomes \( \sum_{k=1}^{\infty} e^{-k^2\sigma_t^2/c_k} e^{ikr} \). A simple choice for \( \sigma_t^2 \) is \( t^2/C^2 \), where \( C \) controls how fast the variance increases over time. In other words, the effect of LOS on the fixed component can be accounted for by assigning a decreasing weight function of \( t \) to each addend in its Fourier transformation.

The gene expression from the random variable \( \varepsilon \) is

\[
\int \varepsilon(t + x)\phi(x)\,dx.
\]

Instead of using continuous weights, we discretize the integration interval and use the weighted sums. For example, for a normal density \( N(0, \tau^2) \), we discretize the range \([-3\tau, 3\tau]\) into \( m \) equally spaced intervals and then calculate the summation of the weighted normal random noise. The sum is \( \sum_{i=1}^{m} w_i/\varepsilon_i / \sum_{j} w_j \), where \( \varepsilon_i \)'s are independent and identically distributed Gaussian random variables with variance \( \tau^2 \). The \( w_i \)'s are the weights derived from \( w_i = \exp(-(-3 + (6i/m))^2)/2) \), and \( i \) is from 0 to \( m \). Let \( x_i \) denote \([[-3 + (6i/m)]\), then the sum (variance) will be

\[
\sum_{i} \exp(-x_i^2/2) / \left[ \sum_{i} \exp(-x_i^2/2) \right]^2,
\]

which does not depend on time \( t \). Therefore, after LOS, the 'new' noise follows a normal distribution with mean 0 and a constant variance.

Combining the LOS effects to the fixed and the random components, we get the 'smoothed' function by LOS,

\[
h(t) \approx \sum_{k=1}^{\infty} e^{-k^2\tau^2/c_k} e^{ikr} + \sum_{i} \exp(-x_i^2) / \left[ \sum_{i} \exp(-x_i^2/2) \right]^2\varepsilon.
\]

Let

\[
\varepsilon^* = \frac{\sum_{i} \exp(-x_i^2)}{\left( \sum_{i} \exp(-x_i^2/2) \right)^2} \varepsilon.
\]

When \( k = 1 \),

\[
h(t) = e^{-(1/2)(t^2/C^2)} c_1 e^{it} + \varepsilon^*
\]

\[
= e^{-(1/2)(t^2/C^2)} \left( c_1 e^{it} + \exp(1/2)(t^2/C^2) \times \varepsilon^* \right).
\]

This equation reveals that the 'net effect' of LOS is to increase the random noise of the temporal expression pattern over time. In other words, the accuracy of the measurement becomes worse over time. Consequently, we should assign a time-decreasing weight to the temporal expression pattern when we do clustering. If we include only one Fourier-transformed term, the weight function in the weighted \( k \)-means can be chosen as the inverse of the standard deviation function, \( \exp[-(1/2)(t^2/C^2)] \).

Because \( C \) is an important factor in the weight function, it is important to discuss how it should be chosen. In Spellman et al.'s (1998) yeast cell-cycle data, \( C \) could be understood as a factor controlling how fast the synchronized cells are losing their synchrony property over time. However, the implication of this constant varies according to the specific microarray experiment. To choose a useful weight function, it is necessary to determine \( C \) first. Although we do not have a method to find exactly the best \( C \), we will provide a heuristic method that is reasonable from a biological viewpoint.

As an initial step, it is reasonable to let \( C \) be in the same range as \( t \). For example, in the simulation study, \( t \) varies from 1 to 10, so \( C \) will be between 1 and 10. In the real data set, \( t \) is in the range of 0–390; so the initial range for \( C \) can be between 1 and 400. While we analyze specific data sets, we will also discuss how to determine the range of \( C \) to improve the performance of clustering.

**SIMULATION STUDY**

We have described the design of our simulation. For each of the nine gene expression patterns, we simulate 50 genes and order them as \((50, 50, 50, 50, 50, 50, 50, 50, 50)\) from the first pattern to the last one.

**Standard \( \kappa \)-means**

By pre-specifying the number of clusters at nine for standard \( \kappa \)-means clustering, e.g. the last of the 100 simulated data sets lead to the clusters 78, 30, 50, 56, 63, 70, 42, 31, 30. This example shows there are 78 genes in cluster one, 30 genes in cluster two and so on. Therefore, it is obvious that the clusters from the standard \( \kappa \)-means deviate substantially from the original membership.

We use ARI to assess the degree of agreement between \( \kappa \)-means clustering and the 'true clustering' (the original pattern) (Hubert and Arabie, 1985; Yeung et al., 2001a). This index has mean 0 for random partitions and reaches its maximum value of 1 for the two identical partitions.

The average ARI between the original pattern and the standard \( \kappa \)-means clusters from the 100 simulated data sets is
The averaged ARI as a function of $C$ (solid curve) from the weighted $k$-means clustering for the 100 simulated data sets. Here, $C$ is the constant in the weight function, $\exp[-(1/2)(t^2/C^2)]$. The dashed horizontal line is at the ARI value from the standard $k$-means clustering.

0.4, which is far below 1. This indicates that there is inadequate performance by the standard $k$-means method in this simulation experiment.

**Weighted $k$-means at the variable level**

Next, we perform the weighted $k$-means clustering at the variable level for the same 100 simulated data sets. We use the inverse of the standard deviation, $\exp[-(1/2)(t^2/C^2)]$, as the weight function as discussed above. It is important to note that, in choosing our weight function (i.e., $C$), we do not assume any knowledge of $\xi (=5)$ (see simulation data section for the meaning of $\xi$). Instead, we consider a range of $C$ as displayed in Figure 1. It is clear from this figure that the improvement becomes evident when $C$ is greater than 2. In other words, we do not have to use the ideal $C$ to benefit from the weighted $k$-means. Furthermore, when $C$ is equal to 5, the average ARI from the 100 simulated data sets is 0.79, which doubles the average ARI from the standard $k$-means. More specifically, the clusters constructed from the last of the 100 simulated data sets are 49, 46, 51, 53, 50, 50, 50, 51, 50—with 49 genes in cluster one, 46 genes in cluster two and so on. This is very close to the original grouping, and much better than the clusters formed by the standard $k$-means as presented above. This example demonstrates the benefit of using our weight function in discovering the underlying clusters.

**PROTEIN COMPLEXES AND YEAST CELL-CYCLE DATA**

In this section, we first use all the protein complexes as the biologic reference. Then, we use the large protein complexes (more than 10 proteins) followed by the permanent and transient complexes as defined by Jansen et al. (2002) for large protein complexes.

**All protein complexes**

The ARI between the standard $k$-means and the 199 protein complexes is 0.039. Besides the inherent noise in the microarray data, the main reason behind such a small ARI is the relatively large number of small complexes, which makes a match of two clusters less likely. This will become more apparent when we restrict our attention to large complexes.

Despite a small value (0.039), the ARI is significantly above 0. This implies that the match is beyond chance. This significance level is assessed by a non-parametric permutation test as follows:

1. We randomly permute and assign the proteins into the 199 complexes.
2. An ARI is computed after each permutation.
3. After 200 random permutations, the average ARI is $-4.1 \times 10^{-4}$ and the standard deviation is 0.0015, giving the one-sided $P$-value less than $1 \times 10^{-6}$ for 0.039.

Now, we examine the performance of weighted $k$-means. Table 1 reveals the percentage increases as a result of using the weighted $k$-means versus the standard $k$-means. For the weighted $k$-means, the increases are quite large (>40%) when $C$ is between 100 and 200. Overall, the largest increase is near 200 for $C$. The improvement reduces as $C$ moves away from 200. This underscores the importance of a well-balanced weight function because information can be lost by over-adjustment of ‘loss-of-synchrony’ correction.

**Large protein complexes**

Here, as in Jansen et al. (2002), we examine 25 large protein complexes that contain more than 10 proteins in each complex. These complexes include cytoplasmic ribosomal, SAGA

<table>
<thead>
<tr>
<th>Protein Complexes and Yeast Cell-Cycle Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>In this section, we first use all the protein complexes as the biologic reference. Then, we use the large protein complexes (more than 10 proteins) followed by the permanent and transient complexes as defined by Jansen et al. (2002) for large protein complexes.</td>
</tr>
</tbody>
</table>

### Table 1. ARI between $k$-means clustering and protein complex clustering

<table>
<thead>
<tr>
<th>Protein Complexes</th>
<th>ARI Standard</th>
<th>ARI C = 100</th>
<th>ARI C = 200</th>
<th>ARI C = 300</th>
<th>ARI C = 400</th>
</tr>
</thead>
<tbody>
<tr>
<td>All $^a$</td>
<td>0.039</td>
<td>0.056 (44%)</td>
<td>0.058 (49%)</td>
<td>0.044 (13%)</td>
<td>0.042 (8%)</td>
</tr>
<tr>
<td>Large $^b$</td>
<td>0.12</td>
<td>0.13 (8%)</td>
<td>0.14 (17%)</td>
<td>0.12 (0%)</td>
<td>0.12 (0%)</td>
</tr>
<tr>
<td>Permanent $^c$</td>
<td>0.24</td>
<td>0.30 (25%)</td>
<td>0.31 (29%)</td>
<td>0.30 (25%)</td>
<td>0.26 (8%)</td>
</tr>
<tr>
<td>Transient $^d$</td>
<td>0.062</td>
<td>0.071 (15%)</td>
<td>0.074 (19%)</td>
<td>0.068 (10%)</td>
<td>0.067 (8%)</td>
</tr>
</tbody>
</table>

The rows correspond to different protein complexes with which clusters from the $k$-means methods are compared. The columns correspond to different $C$ values used in the weight function $\exp[-(1/2)(t^2/C^2)]$. The percentage values in parenthesis indicate the relative increase of ARI using the weighted $k$-means relative to the standard $k$-means. $^a$All protein complexes. $^b$Large protein complexes. $^c$Permanent versus transient (including other) protein complexes.
complex and nuclear pore complex among others. Jansen et al. (2002) further classify them into 13 permanent complexes (e.g. cytoplasmic ribosome), and 12 transient (e.g. SAGA complexes) or other complexes (e.g. nuclear pore complex).

From Table 1, we see an overall improvement in terms of the ARIs for clusters from both the standard and weighted $k$-means methods, when compared to all protein complexes. Nonetheless, there are still notable improvements in ARI for $C$ between 100 and 200 from the weighted $k$-means over the standard one, and particularly so (17%) when $C$ equals 200.

**Permanent versus transient protein complexes**

Jansen et al. (2002) suggested that relationships are much weaker in transient protein complexes than those in permanent complexes. Therefore, it is interesting to use these two sets of complexes separately and examine the changes in ARI. Two comments are warranted from Table 1. First, the ARIs based on the permanent complexes are four to five times higher than those based on the transient complexes. This confirms the suggestion of Jansen et al. (2002). Second, and most interestingly, the benefit of using the weighted $k$-means becomes evident when the permanent and transient complexes are used separately. Improvement is most dramatic when the permanent complexes are used as the reference. Furthermore, the improvement is seen from a wide range of $C$ (100–400), and again, it reaches its peak near 200.

**Biological justification for $C$**

It is clear from Table 1 that the optimal $C$ is near the length of two cell cycles (200) for the yeast *S. cerevisiae* data. Are there any biological justifications to this finding? We know that there exists synchrony loss in the synchronous cell-culturing process. It has been suggested that the cell synchronization can only be maintained by two to three cell cycles while each cell cycle of yeast *S. cerevisiae* is about 80–110 min in length (http://www.ibvt.uni-stuttgart.de/Staff/Mueller/Projdetails.html). Thus, the synchronization usually lasts till 160–330 min. In other words, after 160–330 min the cell culture synchronization phenomenon becomes ambiguous. Therefore, we should choose $C$ such that the weight $\exp[-(1/2)(t^2/C^2)]$ will drop to (1/2) or 1/10 after 160–330 min. For the yeast *S. cerevisiae* data set, the range of $C$ is around 200 min, that is, the length of two cell cycles.

Also important is that the choice of $C$ provides a general idea as to how the synchronous cells lose their synchrony property over time as seen in the yeast *S. cerevisiae* system used by Spellman et al. (1998). Thus, the per-minute-synchrony-losing rate is approximately $\exp \left[- \left( \frac{t}{2C} \right)^2 \right] = 1.25 \times 10^{-5}$.

**DISCUSSION**

In this work, we demonstrated the benefit of using a weighted $k$-means clustering on both simulated and real data. Although we focused on the $k$-means method, our general idea can be useful in other clustering methods such as hierarchical clustering and self-organizing maps. In addition, our approach for deriving the weight function can be applied to the clustering analyses of other microarray data in different experiments and organisms.

Our finding supports the notation that future synchronized cell-cycle experiments should monitor the degree of synchrony during the preparation of mRNA samples by including ‘oxygen consumption, flow cytometry, Coulter volume and digital microscopy measures of cell-cycle events’ (http://www.talandic.com/main/wave/cell_cycle.html). If such information is gathered, it can be useful for determining the weight functions. In fact, Lähdesmäki et al. (2002) discuss the usage of flow cytometry to correct the smoothing effect caused by cell population asynchrony.

Our analyses reveal the benefit of classifying protein complexes into permanent and transient complexes. Such information is available for a limited group of protein complexes. Further generation and clarification of these complexes will be useful for gene clusterings.

Finally, we must point out that we consider only one approach to deal with the ‘time-increasing noise’ data—i.e. to use the weighted least squares by assigning an inverse-standard-deviation weight to the observations. Other methods of variance stabilization transformations can also be exploited. The comparisons will be addressed in our future work.

**CONCLUSIONS**

Due to the extreme complexity of biological systems and the noise in producing microarrays, it is challenging to draw exact and deterministic conclusions from biological microarray analyses. Our analysis indicates that there exists a certain level of agreement between statistical clustering ($k$-means) and biological clustering (protein complex). Since the $k$-means clustering assigns ORFs into a cluster based on the co-expression of the ORFs, we believe the individual subunits in a protein complex should also display co-expression patterns. We demonstrate that the loss of synchrony affects the formations of clusters and a proper correction reduces the difference between statistically based clusters and biologically based ones.

For Spellman et al.’s (1998) yeast cell-cycle data, we conclude that a weight function $\exp[-(1/2)(t^2/C^2)]$ with $C$ around the length of two cell cycles leads to the greatest improvement in the performance of the $k$-means clustering, although the improvement is quite robust for a broad range of $C$ between one cell-cycle length to three cell-cycles length.

**ACKNOWLEDGEMENT**

This work is supported in part by grants: P01NS35476, R01DA12468, R01DA016750, and K02DA017713 from the National Institutes of Health.
REFERENCES


