Assessing Agreement of Clustering Methods with Gene Expression Microarray Data

Xueli Liu[1,*], Sheng-Chien Lee[2], George Casella[1], Gary F. Peter[3,*]

[1] *Department of Statistics and Genetics Institute, University of Florida, Gainesville, FL 32611*
[2] *Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL 32611*
[3] *School of Forest Resources and Conservation, Genetics Institute, University of Florida, Gainesville, FL 32611*

Abstract

In the rapidly evolving field of genomics, many clustering and classification methods have been developed and employed to explore patterns in gene expression data. Biologists face the choice of which clustering algorithm(s) to use and how to interpret different results from the various clustering algorithms. No clear objective criteria have been developed to assess the agreement and compare the results from different clustering methods. We describe two generally applicable objective measures to quantify agreement between different clustering methods. These two measures are referred to as the local agreement measure, which is defined for each gene/subject, and the global agreement measure, which is defined for the whole gene expression experiment. The agreement measures are based on a probabilistic weighting scheme applied to the number of concordant and discordant pairs from two clustering methods. In the comparison and assessment process, newly-developed concepts are implemented under the framework of reliability of a cluster. The algorithms are illustrated by simulations and then applied to a yeast sporulation gene expression microarray data. Analysis of the sporulation data identified $\sim 5\%$ (23 of 477) genes which were not consistently clustered using a neural net algorithm and K-means or

pam. The two agreement measures provide objective criteria to conclude whether or not two clustering methods agree with each other. Using the local agreement measure, genes of unknown function which cluster consistently can more confidently be assigned functions based on co-regulation.

Key words: Clustering algorithms; Agreement measure; Microarray gene expression data.

---

1 Introduction

A central goal of biologists is to elucidate the biochemical functions of genes and their roles in growth, development, and adaptation. Rapid advances in microarray techniques enable simultaneous monitoring of gene expression levels for thousands of genes. To uncover patterns in microarray-based gene expression data, clustering and classification are among the most important analytical methods. Their importance lies in the assumption that genes with similar changes in expression during growth or development are co-regulated and thus are involved in the same or similar biological processes. Hence, co-expressed genes are part of regulatory networks that provides some insight into their biological functions. In the case of genes with no sequence similarity with genes of known function from which their biochemical function cannot be predicted, their co-expression may be some of the only information available about these genes.

Two broad types of clustering and classification methods exist: One is based on heuristic algorithms, for example, K-means clustering [1], self-organizing maps

* Corresponding authors: Xueli Liu, P.O. Box 118545, Gainesville, FL 32611-8545; Phone: 352-392-1941 ext. 212, E-mail: xueli@stat.ufl.edu; Gary F. Peter, P.O. Box 110410, Gainesville, FL 32611-0410, Phone: 352-846-0896, Email: gfpeter@ufl.edu.

[2], hierarchical clustering [3], the recently developed re-sampling based tight clustering [4] and so on, which do not require an underlying statistical model for the mean and the covariance structure; the other methods use model-based algorithms, for example, supervised clustering based on multivariate Gaussian mixtures [5], model-based clustering algorithms [6,7], and so on. Given one particular dataset, different clustering algorithms are very likely to generate different clusters. This is almost always the case when analyzing large-scale gene expression data from microarrays. Therefore, biologists face the problem of choosing the appropriate clustering algorithm for their data and interpreting differences arising from the various clustering algorithms. To address these problems, appropriate quantitative agreement measures of the clustering methods need to be developed.

Guidelines to develop such quantitative measures in genomic data analysis were pioneered by Yeung, Haynor, and Ruzzo [8]. They provided a quantitative, data-driven framework to compare different clustering algorithms. By defining a figure-of-merit (FOM) scale, they rated the predictive power of a clustering arrangement based on a leave-one-out technique. Along this line, very recently, Thalamuthu *et al.* [9] performed a comprehensive comparative study to evaluate the effectiveness of several commonly used clustering methods. They proposed a weighted Rand index to measure similarity of two clustering algorithms and assessed the performance of the methods by a predictive accuracy analysis through verified gene annotations. They found that tight clustering and model-based clustering consistently outperform other clustering methods.

Nearly at the same time as Yeung et al., 2001 [8], Wu *et al.* [10] followed another way to resolve a similar problem. They assigned likely cellular functions

with confidence values to new yeast proteins by making use of a database of clusters produced from different clustering algorithms. Recently, Monti *et al.* [11] developed a method of clustering validation and class discovery called "consensus clustering". Their approach, in conjunction with some re-sampling techniques, provides a way to assess the stability of discovered clusters and to represent the consensus across multiple runs of a clustering algorithm. The consensus algorithm was implemented in a single clustering method with appropriately perturbed data.

More recently, Swift *et al.* [13] proposed a fusion of the approaches described by Wu [10] and Monti [11] to generate both robust and consensus clusters of gene expression data. In this method a consensus matrix is constructed to produce robust clusters which include all full agreement pairs across all clustering methods. Threshold values can be set to relax the full agreement requirement to allow for the inclusion of more genes. A weighted-kappa metric is used to compare the resultant clusters from the different methods. However, the assignment of the strength of the agreement from the value of weighted-kappa is still somewhat arbitrary rather than fully objective. It is therefore ambiguous whether or not the robust or consensus clusters are significantly better than the other clusters.

To address these concerns of the method described by Swift *et al.*, we developed local and global measures for assessing the agreement of different clustering methods. The global agreement measure is calculated for the whole microarray experiment, and the local agreement measure is calculated for each gene/subject. We then utilize the reliability concept from [12] and apply our proposed agreement measures to bootstrapped samples to obtain a sampling distribution of agreement measures for one specific clustering algorithm. The

sampling distribution provides a reference for assessing whether partitions from a different clustering algorithm vary within an acceptable range. We also determine local agreement measures for each gene. If a gene with unknown function has a reasonably high local agreement measure, its co-regulation with other genes of known function can be more confidently predicted. We illustrate the method with a yeast sporulation gene expression microarray data and draw conclusions on the agreement of two clustering algorithms and possible class assignments for stably clustered unknown genes.

The two general agreement measures are described in Section 2. Section 3 shows the algorithm for assessing agreement and the results of a simulation study. Section 4 shows the results of applying the agreement measures to a yeast sporulation gene expression microarray data, and in Section 5 we discuss our findings.

2 Agreement Measures

2.1 Concepts and Notation

Let V_1, V_2 be two vectors which record the clustering results for a gene expression data set of size n with the numbers of clusters k_1 and k_2, respectively. Without exception, we refer to clustering result one whenever we mention V_1 and refer to clustering result two whenever we mention V_2. Here the clustering results can be cluster membership assignments from different clustering algorithms, or they can be partitions using the same clustering method but applied respectively to the original data and perturbed data (see details in sections 3.1 and 3.2). Throughout this paper, unless otherwise noted, sub-

ject/sample/individual represents one gene in the gene expression data.

For method $k, k = 1, 2$, and each pair of subjects $(i, j), i = 1, \ldots, n; j = 1, \ldots, n$, we define $c_{ij}^{(k)} = 1$ if $V_k(i) = V_k(j)$ and $c_{ij}^{(k)} = 0$ otherwise, i.e., $c_{ij}^{(k)} = 1$ if subject i and subject j fall into the same cluster and 0 otherwise. We denote by $C^{(1)}$ the matrix with elements $c_{ij}^{(1)}$ and by $C^{(2)}$ the matrix with elements $c_{ij}^{(2)}$. We will refer to matrices $C^{(1)}$ and $C^{(2)}$ as symmetric pairwise matrices for clustering methods one and two, respectively. Note that elements of $C^{(k)}, k = 1, 2$ are always independent when taken in pairs (because in one cluster membership assignment, knowing that subject i and j are in the same cluster provides no information on whether subject i and k are in the same cluster or not) but may not be so in triples because of transitivity. For each pair of genes (i, j), the $c_{ij}^{(1)}$ and $c_{ij}^{(2)}$ takes and only takes one of the four combinations:

1. $c_{ij}^{(1)} = 1$ and $c_{ij}^{(2)} = 1$;
2. $c_{ij}^{(1)} = 1$ and $c_{ij}^{(2)} = 0$;
3. $c_{ij}^{(1)} = 0$ and $c_{ij}^{(2)} = 1$;
4. $c_{ij}^{(1)} = 0$ and $c_{ij}^{(2)} = 0$.

We denote by a, b, c, d the number of counts of these four combinations respectively, or equivalently,

$$a = \sum_{i,j} c_{ij}^{(1)} c_{ij}^{(2)}, \quad b = \sum_{i,j} c_{ij}^{(1)} (1 - c_{ij}^{(2)}), \quad c = \sum_{i,j} (1 - c_{ij}^{(1)}) c_{ij}^{(2)}, \quad d = \sum_{i,j} (1 - c_{ij}^{(1)})(1 - c_{ij}^{(2)}).$$

We define a consensus matrix $D = C^{(1)} + C^{(2)}$. By its definition, elements in the consensus matrix D can only take the value of $0, 1$, or 2.

2.2 Two General Agreement Measures

We propose two measures of agreement, a global and a local, to objectively determine the agreement between two clustering methods. The global agreement measure focuses on the overall agreement of two clustering algorithms across the whole gene experiment study. It takes all possible pairs of genes into account. On the other hand, the local agreement is defined for each gene/subject to indicate whether a particular gene/subject is consistently clustered with other genes by different methods.

The global agreement measure for vectors V_1 and V_2 is defined as follows:

$$R_g = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}, \tag{1}$$

where w_1, w_2, w_3, w_4 are weights for the above four types of pairs respectively and satisfy $\sum_{i=1}^{4} w_i = 1$. A straightforward way is to assign equal weights to all four types of pairs and thereby simplifying the deductions and calculations. Unfortunately, equal weights do not account for the fact that the probability that two genes are in the same group is usually not the same as the probability that two genes are not in the same group. Therefore rather than using equal weights, we prefer to assign different weights to these four types of pairs according to a probabilistic weighting scheme.

Intuitively, the weights should be inversely proportional to the probabilities of the events. Hence if the probability of two genes in the same group is higher than the probability that they are in different groups, a lower weight is assigned to the first combination count a and a higher weight is assigned to the fourth combination count d, and vice versa. In practice, we choose weights as follows: We assume there are K clusters in the data (K can be estimated

by, for example, the GAP statistic as proposed in [22]). We further assume the probability that one gene falls into cluster $1, 2, \ldots, K$ with probability of $p_k, k = 1, \ldots, K$ such that $\sum_k p_k = 1$. With this notation, $A = \sum_k p_k^2$ and $B = 1 - A$ respectively are the probabilities that two genes are in the same cluster and in different clusters. According to the relations between weight and probability discussed above, we will use

$$w_1 = \frac{1}{A^2}, w_2 = w_3 = \frac{1}{AB}, w_4 = \frac{1}{B^2}. \tag{2}$$

In a real data analysis, p_k is usually unknown, consequently, we will substitute the empirical estimates of these probabilities.

The global agreement measure is connected to well known statistical approaches for specific sets of weights. For example:

1. If we take $w_1 = w_4 = w_2 = w_3 = \frac{1}{4}$, the global agreement measure is equivalent to the Rand index measure [16], which measures the proportion of agreements.
2. Let $w_1 = w_2 = w_3 = \frac{1}{3}, w_4 = 0$, we then obtain the Jaccard's index [17,18].
3. If we know the truth of the cluster membership assignment, say, V_2 is the true cluster membership, the global agreement measure is actually $1-$ misclassification rate.

Besides the global agreement measure, it is also of interest to investigate whether a particular gene is consistently grouped or separated from other genes by different clustering methods. For example, if two clustering methods consistently group gene A with genes B and C and consistently separate gene A from genes D and E, then the two methods are in complete agree-

ment. However, such groupings are not always the same when two clustering methods are compared. To quantify the agreement of the clustering methods for a particular gene, we propose a local agreement measure for each individual gene/subject. The local agreement measure supplies more detailed and comprehensive information on how a particular gene is grouped by different clustering methods. The information on individual genes helps to refine the clustering methods, to assign co-regulation of genes, and to screen out dubious genes/subjects with very low agreement measures. Such information can be valuable to biologists because it alerts them to be more cautious about interpreting these differences arising from different clustering algorithms.

The definition of a local agreement measure between V_1 and V_2 for the gene/subject i is given as follows:

$$R_i = \frac{w_1 a_i + w_4 d_i}{w_1 a_i + w_2 b_i + w_3 c_i + w_4 d_i}, \tag{3}$$

where a_i, b_i, c_i, d_i are the number of counts of the four types of pairs based on $c_{ij}^{(1)}, j = 1, 2, \ldots, n$ and $c_{ij}^{(2)}, j = 1, 2, \ldots, n$ with i being fixed, for example, $a_i = \sum_{j=1}^{n} c_{ij}^{(1)} c_{ij}^{(2)}$. The connection between the local agreement measures $R_1, R_2, \ldots, R_n$ and the global measure can be seen as $R_g = \frac{\sum_{i=1}^{n} R_i}{n}$ when $w_1 = w_2 = w_3 = w_4 = \frac{1}{4}$ are the equal weights. For consistency and ease of interpretation, the weights used in the local agreement measures are equal to those used in the global agreement measure.

3 Algorithm and Simulation

3.1 Assessing the Agreement between Two Partitions

Our goal is not to evaluate the merits of a particular clustering algorithm, rather, our interest is to assess the agreement between the results of two clustering methods applied on the same data set. Therefore, we use two well known and well implemented algorithms, K-means with random initialization and Partitioning Around Medoids (*pam*), to illustrate our agreement measures [1,19]. To facilitate comparisons between different clustering methods, we adapt known parameters wherever appropriate. For example, the number of clusters is assumed to be known.

With a dataset, two clustering methods and the formula given in the previous section, we can easily calculate the agreement measures. However, the value at which the agreement between the two methods is satisfactory needs to be objectively related to the agreement strength. In [13], Swift *et al.* graded values of the weighted-kappa metric from zero to one into five levels of agreement strength from "poor" to "very good". In this grading a weighted-kappa of zero indicates non-agreement for all genes; whereas, a weighted-kappa of one indicates full agreement. However, the grading between 0 and 1 is somewhat arbitrary. Consequently, the conclusion on whether the agreement is satisfactory is still subjective although such subjectivity is usually enough to give researchers a rough idea on how two clustering methods agree with each other.

In order to construct a fully objective measure, we first tested the reliability of one clustering algorithm, referred to as the reference algorithm, in light of appropriately perturbed data [11,12]. When we perturb the data using

some random noise with mean 0 and variance σ^2, the same method often produces different clustering results. If we denote by V_1 the clustering result of the original data and V_2 the clustering result of the perturbed data, we can compute the agreement measures for V_1 and V_2. The sampling distribution of these agreement measures can be obtained by repeating the perturbation many times. Under this framework, because the two clustering results are the same when the observed noise (perturbation) σ^2 is below a critical level σ_0^2, we can not distinguish the two clustering results. A special case is when there is little perturbation and the clustering algorithm performs stably; the perturbed clustering results will be very similar to the original clustering membership assignment and the agreement measures will be close to 1. A natural yet conservative choice of σ_0^2 would be based on the predicted variation, which can be computed by differencing the adjacent two measurements within the same subject (see equation (5) in section 3.2).

That said, the spread of the sampling distribution of the global agreement measures between clustering results of the perturbed data and the original data indicates reasonable departure from the null hypothesis. Consequently, in order to produce a reasonable level of concordance or agreement, the global agreement measures should be at least the corresponding lower quantiles of the sampling distribution (obtained by perturbing data with Normal(0, σ_0^2)), which becomes our objective merit to determine whether the algorithms agree. Similarly, the sampling distribution of the local agreement measures will provide objective merit to determine whether one gene/subject is consistently clustered or not by the two clustering algorithms.

3.2 Algorithm

Both the K-means and *pam* clustering algorithms are powerful and popular data mining techniques with which to group data having similar characteristics or features [20]. As an illustration, we compare K-means with *pam* clustering, where K-means is the reference algorithm. Results are similar if we change the reference algorithm to *pam* (results not shown). The detailed procedure is as follows.

3.2.1 Data Standardization

Given one dataset $y_i(t_j)$ where $i = 1, 2, \ldots, N$ denotes the indices of genes/subjects, and $j = 1, 2, \ldots, J$ denotes the indices of time points. Subjects similar to each other as measured by some dissimilarity measures will be grouped together.

For quantitative variables, there are two main types of dissimilarity measures Euclidean distance and correlation. Euclidean distance is a measure of the magnitude/absolute differences whereas correlation is a measure of the trends or relative differences. For instance, the Euclidean distance between two J dimensional vectors (x, y) would be much different from that between $(ax + b, y)$ where a, b are both not zero. This is because $ax + b$ may take large values depending on the choice of a, b, which, in turn, will dominate the distance between $(ax + b, y)$. On the other hand, the correlation between (x, y) and that between $(ax + b, y)$ stays the same for $a > 0$, which is often a desirable property.

The choice of which dissimilarity measure to use should be based on what types of similarities in the data researchers are trying to identify. In microarray

experiments [14,21] suggested that a correlation measure is preferable when clustering time course data. Since K-means uses the Euclidean distance as the dissimilarity measure, we propose to standardize the data $y_i(t_j)$ first by subtracting its mean $\bar{y}_i = \frac{1}{J}\sum_{j=1}^{J} y_i(t_j)$ and dividing by its standard deviation $s_i = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}(y_i(t_i) - \bar{y}_i)^2}$. It is self-evident that the two dissimilarity measures are equivalent after standardization. In this paper, unless otherwise noted, the clustering algorithm is applied to the standardized data, which we will still denote by $y_i(t_j)$.

3.2.2 Perturbation Model

We first apply the K-means algorithm with a total of K clusters to the standardized data $y_i(t_j)$. We denote by $\hat{C}_0$ the clustering results. To assess the reliability of K-means, we perturb the data according to the model as follows:

$$y_i^*(t_j) = y_i(t_j) + \epsilon_j, \tag{4}$$

where y_i^* is the perturbed data for the ith gene/subject. The random errors $\epsilon_j, j = 1, \ldots, J$ are a variant (because of the standardization step) of potential measurement errors introduced from experimental procedures including sample acquisition, scanning, and/or cross-hybridization. These small random errors are from a normal distribution with mean 0 and variance σ_0^2, where σ_0^2 are based on the predicted variation. Specifically, we first remove the trend from each curve by subtracting its corresponding cluster center, then estimate the predicted variation of gene/subject i by differencing the adjacent two measurements. It can be computed from the following equation:

$$\sigma_i^2 = \sum_{j=2}^{J} \frac{(e_i(t_j) - e_i(t_{j-1}))^2}{2(J-1)}, \quad i = 1, 2, \ldots, N, \tag{5}$$

where $e_i(t_j) = y_i(t_j) - c(i)(t_j)$ with $c(i)$ being the cluster center to which subject i belongs. Once again, please note that $y_i(t_j)$ is standardized with mean of 0 and variance of 1 before computing σ_i^2, because each gene has its own expression intensity and it is unreasonable to combine the raw variances (the interest is in the shape of the gene expression profile, but not the magnitude). Based on this definition of σ_i^2, σ_0 is taken as the arithmetic average of σ_i, i.e., $\sigma_0 = (\sigma_1 + \sigma_2 + \ldots + \sigma_N)/N$.

The perturbation model described above can be used to evaluate the reliability of any clustering method against potential experimental errors.

3.2.3 Agreement Between K-means and Pam

As stated before, we apply the K-means clustering procedure on each perturbed data set. Let us denote by $\hat{C}_b$ the clustering results for the b-th dataset, $b = 1, 2, \ldots, B$. The global and the local agreement measures are computed between each $\hat{C}_b$ and $\hat{C}_0$, where $\hat{C}_0$ serves as the common reference. The sampling distribution of these agreement measures will be shown in the form of a histogram. The *pam* algorithm is then applied to the original dataset $y_i(t_j)$. We denote by $\hat{C}_{pam}$ the clustering results. Agreement measures are also computed between $\hat{C}_{pam}$ and $\hat{C}_0$.

With these computations and sampling distributions, we still need some notation in order to objectively assess the agreement between the two clustering algorithms, K-means and *pam*. Let us denote by $R_{g\alpha}, R_{l\alpha}$ the lower α quantile value of the global and local agreement measures from the sampling distribution respectively, where we take $\alpha = 0.05$. We also denote by $R_{g,pam}, R_{l,pam}$ the corresponding global and local agreement measures between *pam* and the

reference algorithm K-means (omitted in the index). For the global agreement measure, if $R_{g,pam} \leq R_{g\alpha}$, then there is little concordance between these two methods. On the other hand, if $R_{g,pam} \geq R_{g\alpha}$, then we have no evidence to say that the methods are disconcordant. Thus, when two clustering methods are in better agreement, the global agreement measure is higher than the lower α quantile values of the corresponding sampling distribution. Similar arguments hold for the local agreement measures. By checking local agreement measures, we can single out genes for further investigation which are clustered inconsistently by the two clustering algorithms, for example, we may look at a third clustering algorithm to enhance the confidence level of assigning membership of these genes. Please refer to section 4 for more details on this.

3.3 Simulation Study

To evaluate the agreement measures, we used simulated gene expression patterns of nine clusters over 10 time points similar to the one described in [15]. Briefly, only one array from one sample was obtained at each time point during the course of the study. A total of 10 samples were analyzed. This is an independent sampling scheme with an identity correlation matrix because a sample at one time points represents an independently sampled unit. There were $49, 42, 51, 53, 57, 46, 38, 51, 56$ genes respectively in the nine clusters (empirical genomic evidence [9] suggests that the cluster sizes are often different).

With these data we added random noise (Normal($\mu = 0, \sigma^2 = 0.16$)) to each element of the $log2$ expression values. We estimate the perturbation noise level in the data based on (5) and $\sigma_0 = 0.56$. Figure 1 (a)-(b) shows heatmaps of the simulated examples with and without perturbations whereas (c) (no

apparent pattern at all) shows differences between (a) and (b). Although the cluster patterns are apparently similar in (a) and (b), it can be seen that the experimental variability weakens the signal intensity as expected (c).

As mentioned in section 2.2, the weights in computing both agreement measures should be inversely proportional to the probabilities of two genes in the same group. However, since the probabilities $p_k, k = 1, \ldots, K$, where p_k is the probability that one gene falls into cluster $1, 2, \ldots, K$, are often unknown, we will use the empirical estimates for these probabilities, i.e., the proportion of genes in each cluster obtained from clustering the standardized raw data. The empirical weight estimates in this simulation study are $w_1 = 0.787$, $w_2 = 0.10$, $w_3 = 0.10$, $w_4 = 0.012$ based on equation 2. Since these weights are data-driven estimators, they may be different values for different data sets.

In the simulation, we conducted $B = 10,000$ runs according to the procedure described in section 3.2. Figure 2 (a) presents histograms of the global agreement measures obtained from perturbing the simulated data, which will serve as the reference distributions. It shows an example where $\alpha = 0.05$, $R_{g\alpha}$ is 0.84, and the global agreement measure of K-means and *pam* is $R_{g,pam} = 0.91$. This global agreement measure clearly exceeds the threshold value 0.84 and we draw the conclusion that with this simulated data set K-means and *pam* are in agreement. We also conducted 100 replicate runs with data sets of similar kind (nine clusters, ten time points, $\sigma = 0.4$). With σ_0 being estimated from (5) and $B = 100$ for each data set, we observe that K-means and *pam* agree100% of the time. This result is fully expected because K-means and *pam* have high global agreement measures because their working principles are very similar and the noise level in the simulated data is in a reasonable range.

Figure 3 (a) presents variation of mean local agreement measures for all genes.

We link every gene's strength of co-expression with other genes within the same cluster to its mean local agreement measure. As observed, genes with very high mean local agreement measures in the same cluster are always/strongly co-expressed despite the influence of perturbation. In contrast, genes with very low mean local agreement measures are rarely/weakly co-expressed with other genes and perturbation often displaces them to other clusters. Genes with medium mean local agreement measures are those that are sometimes co-expressed with genes in the same cluster and sometimes not. A closer examination of the range of mean local agreement measures of each cluster clearly reveals that genes in the fourth and fifth clusters are most strongly co-expressed (with the range of $[0.92, 0.96]$ and $[0.91, 0.96]$, respectively).

Figure 3 (b)-(d) show histograms of local agreement measures of three exemplary genes. Figure 3 (b) shows the histogram of a gene that clustered consistently in most of the perturbations because of it is strongly co-expressed with other genes. Its mean is 0.96 with lower and upper 5% quantiles being $0.95, 0.98$ respectively. Figure 3 (c) shows the histogram of a gene that clustered consistently in many perturbations; whereas, in some perturbations it did not cluster consistently. Its mean is 0.77 with lower and upper 5% quantiles being $0.41, 0.93$ respectively. Figure 3 (d) shows the histogram of a gene that did not cluster consistently in most of the perturbations because it is weakly co-expressed with other genes. Its mean is 0.62 with lower and upper 5% quantiles being $0.33, 0.90$, respectively. Comments on the distributions of these agreement measures are given in section 5.

Beyond the results presented above, we investigated the effects of sample size (or the number of chips, here corresponding to the number of time points), and increasing levels of noise in the generation of data (σ^2) on the perfor-

mance of the agreement measures. Experimental results with six noise levels $(0.1, 0.2, 0.3, 0.4, 0.6, 0.8)$ are similar with slight variations: The global agreement measures correspond to lower quantiles $(0.87, 0.85, 0.82, 0.78, 0.74, 0.71)$ of sampling distribution as the noise levels increases. Four sample sizes $(5, 7, 8, 10)$ are tested. The corresponding lower 5% quantiles of the global agreement measures ($B = 200$, $\sigma = 0.4$, $\sigma_0 = 0.47$) are $0.77, 0.81, 0.82, 0.83$, respectively, which clearly indicates that the global agreement measures correspond to higher quantiles of sampling distribution as the sample size increases. These results indicate that the proposed measures are suitable for objectively assessing the agreement of clustering algorithms.

4 Application to Yeast Gene Expression Microarray Data

In addition to testing the measures with the simulated gene expression data, we illustrate the methods with actual budding yeast gene expression data from a sporulation experiment reported by Chu *et al.* [14]. In this experiment, yeast cells were shifted to sporulation media and mRNA samples were taken at the time intervals of 0, 30 minutes, and $2, 5, 7, 9$, and 11 hours representing different stages during sporulation. Chu *et al.* identified over 1000 of the 6200 genes on the cDNA spotted array as expressed at different levels relative to control diploid vegetative stage [14]. During sporulation, about half were down-regulated and half were up-regulated. The 477 up-regulated genes were assigned to seven temporal patterns with a neural net algorithm. For our analysis, we focused on these 477 up-regulated genes and their corresponding cluster assignments (see additional file 1) from Chu *et al.* [14].

We first applied both the K-means and *pam* clustering algorithms to the log-

ratio data (re-scaled so that the mean of each gene is 0 and standard deviation is 1). The empirical weight estimates in this study are $w_1 = 0.64$, $w_2 = 0.16$, $w_3 = 0.16$, $w_4 = 0.04$ based on equation 2. In the analysis that follows, the first goal was to determine whether, globally, *pam* and the published results agree with those obtained from K-means. The second goal was to use the local agreement measures to identify genes from the published results that did not cluster consistently with K-means and *pam* and to provide biologically meaningful interpretations.

The stability of the K-means clusters was assessed by perturbing the standardized raw data. The perturbation is done according to the model given in equation (4). For ϵ, we use normal random noise with mean of 0 and standard deviation of 0.3. According to equation (5), $\sigma_0 = 0.31$ was calculated from the standardized data. We repeat the-perturbation-then-cluster with K-means 9999 times. The result is $10,000$ K-means clustering results, one for the standardized raw data and 9999 for perturbed data.

Figure 2 (b) gives the sampling distribution of the global agreement measure. At $\alpha = 0.05$, $R_{g\alpha}$ is 0.711. The global agreement measure between K-means and *pam*, is $R_{g,pam} = 0.836$. The global agreement measure of K-means and the published result from Chu *et al.* [14] is 0.777. These two global agreement measures both exceed the threshold value $R_{g\alpha} = 0.711$. Therefore, K-means and Chu *et al.* cluster results agree globally with each other. The high global agreement indicates that the assigned clusters in general were consistent between the three methods. This is of significance, because each clustering method needs to be parameterized for optimal results.

We propose that the criterion for choosing the objective cut-off value of the global agreement measure is given in terms of percentiles of its sampling dis-

tribution. Particularly, to determine whether the global agreement measure 0.77 is good or not, the objective cut-off value in this example is 0.711 and the answer is yes. One should note that, however, as discussed in the simulation studies, percentiles of the sampling distribution are affected by the sample size and noise level. In order to get a correct interpretation of the global agreement measure and the subsequent clustering results, we need to relate the value to its position in the corresponding sampling distribution.

Local agreement measures were calculated for the 477 genes up-regulated during sporulation. Similar to the simulation study, we present the variation of mean local agreement measures for all genes in Figure 4 (a). A quick examination of the range of mean local agreement measures shows that the genes in the Metabolic class are most strongly co-expressed (with the range $[0.83, 0.96]$). Figure 4 (b)-(d) show histograms of local agreement measures of three exemplary genes, with high, medium, and low mean local agreement measures, respectively. The histograms look similar to those from the simulated data.

To further explore the differing results of these clustering algorithms, we selected some specific genes based on their local agreement measures for investigation. The selection criteria of genes is as follows: If the local agreement measure of a gene satisfies $R_{l,pam} > R_{l\alpha}$ and $R_{l,\mathrm{Chu}} < R_{l\alpha}$, then the gene is consistently clustered by K-means and *pam*, but not so by K-means and the neural net algorithm. By this criteria, $\sim 5\%$ of the genes (23) do not cluster consistently with the different algorithms. Table 1 summarizes these genes. The lack of agreement using the local measure suggests that their classification needs to be reevaluated.

Seven of these 23 genes show no similarity with other sequences of known function in the database. Of these seven genes with unknown function three

(*YDL050C, YNL171C, YPL114W*) had dubious open reading frames that are unlikely to code for functional proteins. For these genes with dubious open reading frames, one possibility is that they represent pseudo-genes, whose expression is less tightly regulated due to their divergence and loss of function. Interestingly, three of the five genes that were categorized as metabolic are involved in DNA replication and cell division, suggesting that these genes may have been incorrectly clustered.

5 Discussion

Many clustering algorithms are available for grouping genes with similar expression profiles. However, due to different data structures and characteristics it is unrealistic to assume that one clustering algorithm will perform the best with all the datasets. Consequently, it is necessary to consider appropriate quantitative measures of comparing and assessing agreement between clustering methods.

In this paper, we introduce two generally applicable approaches to objectively quantify the agreement between different clustering methods: (1) a global agreement measure for an experiment and (2) a local agreement measure for each gene/subject. The global agreement measure, defined for the whole gene expression experiment, measures the overall strength of agreement for all of the genes between different clustering methods; whereas, the local agreement measure, defined for each gene/subject, measures whether a particular gene/subject is consistently clustered with other genes by the different methods. Both measures provide us with objective guidelines to draw conclusions about the agreement between two clustering algorithms. Application of these

agreement measures with both simulated and real gene expression microarray data demonstrate the strength of the proposed quantitative measures.

By their nature, the agreement measures are designed to evaluate the extent that two clustering algorithms agree with each other and further enhance the confidence of assigning membership to novel genes. For example, if a gene with unknown function has a high local agreement measure, its co-expression with other genes with known functions can be more confidently predicted. On the other hand, by choosing the appropriate cutoff, dubious genes with relatively low local agreement measures can be screened out for further investigation into why the agreement was low. A low local agreement score does not necessarily mean that this gene is problematic (an outlier), because the basis for this low score may lie in the analytical methods as well as in its behavior on microarrays.

The proposed concept and method can be easily applied to any biological dataset or other datasets. As an example, we applied it to budding yeast gene expression data from the sporulation experiment reported by Chu *et al.* [14]. *Chu et al.* used a neural net algorithm, to assign 477 up-regulated genes to seven temporal patterns [14]. The global agreement and all but 5% of the local agreement measures were consistent between the neural net algorithm and K-means or *pam* clusters. This is somewhat surprising given that neural net algorithms work like blackboxes, and it is not straightforward to figure out their relation with the other two clustering algorithms.

To objectively compare the agreement between two clustering algorithms, one clustering algorithm needs to be chosen first as the reference. The reliability of the reference algorithm is then evaluated by perturbing the data according to a model. A straightforward extension can be used to compare multiple

clustering methods. To address this problem, two options are available: either compare them pairwise or choose one clustering method as the reference and then compare all the other methods with this reference. In the latter case the reference clustering method can be chosen randomly (usually one that is widely used and easy to implement). If the clustering methods form an equivalent class under the agreement measures, then we can draw pairwise comparison conclusions even when one is randomly chosen.

Missing values for genes in microarray data may affect downstream analysis, such as clustering and network analysis [23–27]. In particular, the stability of clusters is reduced as the proportion of missing values increases when some concordance measures, for instance, the Conserved Pairs Proportion (CPP) in [23], which is a variant of the Rand index or the Jaccard score in [27] are used. This work also provides an alternative way to perturb the data and consequently, assess the stability of clusters accordingly using the proposed agreement measures.

It remains open as to the best way to find the threshold values for the agreement measures. Ideally it is more convenient if we can identify the exact distribution of the proposed agreement measures in light of the perturbed data. Unfortunately, this is not an easy task, as the sampling distributions are not necessarily continuous (because they are defined based on count data); they can be bimodal or even multimodal, depending on the experimental variation σ_0^2 used to perturb the data. As σ_0 increases, the deviation from the true cluster membership becomes bigger. As noted by one referee, the availability of biological replicates may improve the estimation of σ_0, and thus the performance of the clustering algorithms and subsequent analysis. For the time being, the empirical lower quantiles of the reference distribution serve

well as the threshold values for the agreement measures. Thus, an advance in the methods described here would be to derive the exact or approximate distributions of the agreement measures.

The R-code implementing the agreement measures algorithm is included in the additional files (see additional file 2 for simulation study and additional file 3 for real data analysis). It provides a readily useable tool for biologists whose interest is to quantify the agreement of different clustering algorithms and thus arrive at meaningful biological interpretation of the clustering results.

Acknowledgements

The authors want to thank Drs. Mark Yang and Josè Mira for their very helpful discussions and comments to the manuscript. Casella was partially supported by NSF grant Number DEB 0540745. They would also like to thank the referees and the associate editor for their helpful and clever insights, which have improved the paper significantly.

References

[1] Hartigan, JA, Wong, MA, 1979. A K-means clustering algorithm. Appl. Statist. v28. 100–108.

[2] Tamayo, P, Slonim, D, Mesirov, J, Zhu, Q, Kitareewan, S, Dmitrovsky, E, Lander, ES, Golub, TR, 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. Natl Acad. Sci. USA, v96. 2907–2912.

[3] Eisen, MB, Spellman, PT, Brown, PO, Botsein, D, 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl Acad. Sci. USA, v95. 14863–14868.

[4] Tseng, GC, Wong, WH, 2005. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. Biometrics. v61. 10–16.

[5] Qu, Y, Xu, S, 2004. Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. Bioinformatics. v20. 1905–1913.

[6] Yeung, KY, Fraley, C, Murua, A, Raftery, AE, and Ruzzo, WL, 2001. Model-based clustering and data transformations for gene expression data. Bioinformatics. v17. 977–987.

[7] Luan, Y, Li, H, 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics. v19. 474–482.

[8] Yeung, KY, Haynor, DR, Ruzzo, WL, 2001. Validating clustering for gene expression data. Bioinformatics. v17. 309–318.

[9] Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC, 2006. Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics. v 22. 2405–2412.

[10] Wu, LF, Hughes, TR, Davierwala, AP, Robinson, MD, Stoughton, R, Altschuler, SJ, 2002. Large-scale prediction of saccharomyces cerevisiae gene function using overlapping transcriptional clusters. Nat. Genetics. v32. 255–265.

[11] Monti, S, Tamayo, P, Mesirov, P, Golub, T, 2003. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning. v52. 101–118.

[12] Kerr, K, Churchill, G, 2001. Bootstrapping clustering analysis: Assessing the reliability of conclusions from microarray experiments. Proc. Natl. Acad. Sci. USA, v98. 8961–8965.

[13] Swift, S, Tucker, A, Vinciotti, V, Martin, N, Orengo, C, Liu, X, and Kellam, P, 2004. Consensus clustering and functional interpretation of gene-expression data. Genome Biol. v5: R94.

[14] Chu, S, DeRisi, J, Eisen, M, Mulholland, J, Botstein, D, Brown, P, Herskowitz, I, 1998. The transcriptional program of sporulation in budding yeast. Sci. v282. 699-705.

[15] Quackenbush, J, 2001. Computational analysis of microarray. Nat. Rev. Genet. v2. 418-427.

[16] Rand, WM, 1971. Objective criteria for the evaluation of clustering methods. J. Am. Statist. Assoc. v66. 846–850.

[17] Sneath, PHA, Sokal, RR, 1973. Numerical taxonomy: The principles and practice of numerical classification. W. H. Freeman, San Francisco, California.

[18] Magurran AE, 1988. Ecological diversity and its measurement. Chapman and Hall, London, England.

[19] Kaufman, L, Rousseeuw, PJ, 1990. Finding groups in data: An introduction to cluster analysis. Wiley, New York.

[20] Hastie, T, Tibshirani, R, and Friedman, J, 2001. The elements of statistical learning. Springer, New York.

[21] Spellman et al., (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell vol 9. 3273–3297.

[22] Tibshirani, R, Walther, G, and Hastie, T, 1999. Estimating the number of clusters in a dataset Via the Gap statistic. J. of Roy. Statist. Soc. Ser. B v63. 411–423.

[23] de Brevern, A., Hazout, S, and Malpertuy, A, 2004. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. BMC Bioinformatics 5:114.

[24] Bϕ, TH, Dysvik, B, and Jonassen, I, 2004. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Res. v32. e34.

[25] Tuikkala, J, Elo, L, and Aittokallio, T, 2006. Improving missing value estimation in microarray data with gene ontology. Bioinformatics v22. 566–572.

[26] Wong, DS, Wong, FK, and Wood, GR, 2007. A multi-stage approach to clustering and imputation of gene expression profiles. Bioinformatics v23. 998–1005.

[27] Varshavsky, R, Gottlieb, A, Horn, D, and Linial, M, 2007. Unsupervised feature selection under perturbabtions: meeting the challenges of biological data. Bioinformatics v23. 3343–3349.

Table 1
Summary of genes with low local agreement measures

ORF	SGD Name	SGD Function/Gene Product	Temporal Class
YBL015W	ACH1	acetyl-CoA hydrolase	Metabolic
YAR007C	RFA1	replication factor A, heterotrimeric ssDNA binding	Metabolic
YBR088C	POL30	PCNA -DNA polymerase processivity factor	Metabolic
YDL050C		unknown-ORF dubious	Mid-Late
YDL193W	NUS1	Nuclear Undecaprenyl pyrophosphate Synthase	Early II
YDR089W		unknown	Middle
YDR148C	KGD2	alpha-KetoGlutarate Dehydrogenase	Early II
YDR219C	MFB1	Mitochondria-associated F-Box protein	Early-Mid
YDR380W	ARO10	Phenylpyruvate decarboxylase	Mid-late
YDR438W	THI74	THI regulon	Middle
YGR110W		unknown	Early-Mid
YGR224W	AZR1	Plasma membrane transporter -major facilitator superfamily	Early II
YIL132C	CSM2	Protein required for accurate chromosome segregation during meiosis	Metabolic
YJR137C	ECM17	Sulfite reductase beta subunit	Early I
YKL042W	SPC42	spindle pole body component	Early-Mid
YLR054C	OSW2	unknown function proposed to be involved in the assembly of the spore wall	Mid-Late
YKR099W	BAS1	Myb-related transcription factor	Early-Mid
YNL171C		unknown-ORF dubious	Middle
YNL270C	ALP1	basic amino acid transporter	Early-Mid
YNR026C	SEC12	guanine nucleotide exchange factor for Sar1p	Mid-Late
YOR214C		unknown	Early-Mid
YPL111W	CAR1	arginase	Metabolic
YPL114W		unknown-ORF dubious	Middle

Additional Files

Additional File 1: Supplemental Data. Containing budding yeast gene expression data from the sporulation experiment reported by Chu *et al.* [14]. Each row represents one gene. Columns are the gene ORF, ratios at the time intervals of 0, 30 minutes, and $2, 5, 7, 9, 11$ hours, and the temporal classes reported by [14].

Format: CSV, **size:** 22KB.

Additional File 2: R code for Simulation Model. Containing the R code for the simulation study where K-means and *pam* are used to illustrate the algorithm.

Format: rtf, **size:** 20KB.

Additional File 3: R code for Real Data. Containing the R code for implementing the algorithm to the budding yeast gene expression data.

Format: rtf, **size:** 18KB.

Figures

Figure 1- Heatmaps of simulated data

(a) A simulated example of 443 curves in nine clusters. (b) A simulated example obtained by perturbing data in (a) using random normal distribution with mean 0 and $\sigma_0 = 0.4$. (c) Data for differences between (b) and (a). Rows correspond to genes and columns correspond to time points. Color bars on the row side indicate the nine clusters (with the first cluster starting at the bottom then moving to the top).

Figure 2 - Histograms of the global agreement measures

Agreement measure values are plotted on the x-axis, and the y-axis is the frequency that agreement measures fall into each group/bin of the original and 9999 sets of perturbed data. (a) Histogram of the global agreement measures for K-means algorithm for simulation data. (b) Histogram of the global agreement measures for K-means algorithm for the yeast sporulation gene expression data.

Figure 3 - Histograms of the local agreement measures for simulation data

Agreement measure values are plotted on the x-axis, and the y-axis is the frequency that agreement measures fall into each group/bin of the original and 9999 sets of perturbed data. (a) Histogram of mean local agreement measures for all 443 genes. (b)-(d) are histograms for three exemplary genes with high, medium and low mean local agreement measures.

Figure 4 - Histograms of the local agreement measures for the yeast sporulation gene expression data

Agreement measure values are plotted on the x-axis, and the y-axis is the frequency that agreement measures fall into each group/bin of the original and 9999 sets of perturbed data. (a) Histogram of mean local agreement measures for all 477 genes. (b)-(d) are histograms for three exemplary genes with high, medium and low mean local agreement measures.