Validating Clustering for Gene Expression Data

Yeung K. Y.*
Computer Science and Engineering
Box 352350
University of Washington
Seattle, WA 98195, USA

Haynor D. R.
Radiology
Box 357115
University of Washington
Seattle, WA 98195, USA

Ruzzo W. L.
Computer Science and Engineering
Box 352350
University of Washington
Seattle, WA 98195, USA

August 18, 2000

running head: comparing clustering algorithms on gene expression data

key words: gene expression analysis, clustering, comparison of clustering algorithms, cluster validation

Submitted for publication

*To whom correspondence should be addressed

Abstract

***Motivation**: Many clustering algorithms have been proposed for the analysis of gene expression data, but little guidance is available to help choose among them. We provide a systematic framework for assessing the results of clustering algorithms. Clustering algorithms attempt to partition the genes into groups exhibiting similar patterns of variation in expression level. Our methodology is to apply a clustering algorithm to the data from all but one experimental condition. The remaining condition is used to assess the predictive power of the resulting clusters—meaningful clusters should exhibit less variation in the remaining condition than clusters formed by chance.*

***Results**: We successfully applied our methodology to compare seven clustering algorithms on four gene expression data sets. We found our quantitative measures of cluster quality to be positively correlated with external standards of cluster quality.*

***Availability**: The software is under development.*

***Contact**: kayee@cs.washington.edu*

Supplementary information:

http://www.cs.washington.edu/homes/kayee/cluster.html

1 Introduction

In an attempt to understand complicated biological systems, large amounts of gene expression data have been generated by researchers ([DeRisi *et al.*, 1997] and [Wen *et al.*, 1998]). Because of the large number of genes and the complexity of biological networks, clustering is a useful exploratory technique for analysis of gene expression data. Many clustering algorithms have been proposed for gene expression data. For example, [Eisen *et al.*, 1998] applied the average-link hierarchical clustering algorithm to identify groups of co-regulated yeast genes. [Ben-Dor and Yakhini, 1999] reported success with their CAST algorithm. [Tamayo *et al.*, 1999] used self-organizing maps to identify clusters in the yeast cell cycle and human hematopoietic differentiation data sets.

Assessing the clustering results and interpreting the clusters found are as important as generating the clusters [Jain and Dubes, 1988]. In much of the published clustering work on gene expression, the success of clustering algorithms is assessed by visual inspection using biological knowledge (for example, [Michaels *et al.*, 1998] and [Eisen *et al.*, 1998]). To the best of our knowledge, there is no systematic data-driven method to quantitatively evaluate gene expression clustering results.

Figure 1 shows the importance of assessing clustering results. Figures 1(a) and 1(b) show clusters found in two different data sets having several hundred genes measured under 79 conditions. In both cases, clusters were identified at least in part by visual inspection (using the TREEVIEW software [Eisen *et al.*, 1998]) of the dendrogram produced by an average-link hierarchical clustering algorithm (the CLUSTER software, also by [Eisen *et al.*, 1998]). In the figures, genes are ordered so that similar genes are placed adjacent to each other along the vertical axis. The experimental conditions along the horizontal axis are not clustered. The color intensity of each cell in the figure is proportional to the measured gene expression ratio, with bright red representing the most positive and bright green being most negative. Thus, vertical stripes of similar color and intensity indicate similarities in responses of these genes under the same experimental conditions. Figure 1(a) shows a cluster identified by [Eisen *et al.*, 1998] from a data set with 2467 yeast genes. Known functional roles of many of the genes in this cluster support the suggestion that the cluster represents a biologically related group of genes. Figure 1(b), however, although it shows a pattern nearly as striking as that in Figure 1(a), was obtained from a simulated data set that does *not* contain any intrinsic clusters (see the supplementary web site for the details of the simulated data set). Without validation procedures, results of clustering algorithms may easily be misinterpreted. Moreover, different clustering algorithms may yield markedly different clusters.

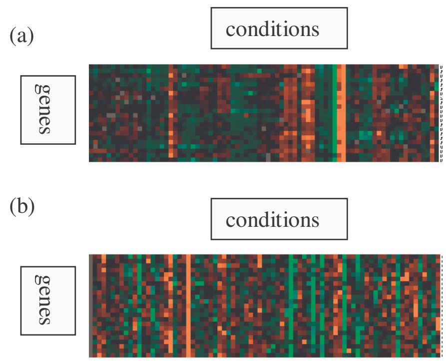

Figure 1: (a) A cluster from [Eisen *et al.*, 1998]. (b) A striking pattern in simulated data lacking intrinsic pattern.

A biologist with a gene expression data set is faced with the problem of choosing an appropriate clustering algorithm for his or her data set. Our paper provides a quantitative data-driven framework to evaluate and compare different clustering algorithms.

2 Our Approach

Our method for assessing the quality of clustering results is motivated by the jackknife approach [Efron, 1982]. We apply a clustering algorithm to all but one experimental condition in a data set, and use the left-out condition to assess the predictive power of the clustering algorithm. We define a scalar quantity called the *figure of merit* (FOM), which is an estimate of the predictive power of a clustering algorithm.

Figure of merit: Intuitively, a clustering algorithm has possible biological significance if genes in the same cluster tend to have similar expression levels in additional experiments that were not used to form the clusters. We estimate this predictive power by removing one experiment, E, from the data set, clustering genes based on the remaining data, and then measuring the within-cluster similarity of expression values in experiment E. Our figure of merit is the RMS deviation in the left-out condition E of the individual gene expression levels relative to their cluster means. Each condition can be used as the validation condition, so we compare clustering algorithms using the sum of FOM's over all the conditions. The *adjusted figure of merit* is the figure of merit divided by a factor that compensates for a statistical bias with many clusters. The figure of merit and adjusted figure of merit are formally defined in Section 3.

A *small* figure of merit indicates a clustering algorithm having *high* predictive power. We compare clustering algorithms with the same number of clusters, and over a range of number of clusters.

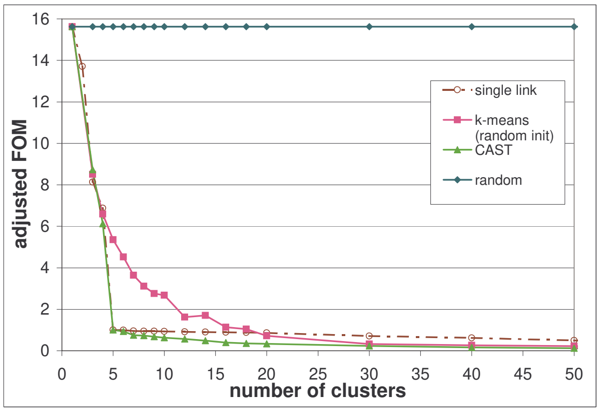

Figure 2: Adjusted FOM's of clustering algorithms on a simulated data set with 5 clusters.

An example with clusters: In Figure 2, three clustering algorithms – hierarchical single-link, k-means (with random initialization), and CAST [Ben-Dor and Yakhini, 1999] – are compared on a simulated data set. As a control, we also include the "random" algorithm, which simply places genes into random clusters. The algorithms are described in Section 5. The simulated data set has 420 genes and 17 conditions, and contains 5 clusters. Refer to the supplementary web site for the details of the simulated data set. Figure 2 shows that all three "real" clustering algorithms exhibit FOM's that are much better than random placement of genes into clusters. Additionally, CAST and single-link show markedly better FOM's than k-means for 5 to 10 clusters. This is no accident. Examination of the cluster results with 5 clusters reveals that the CAST and single-link algorithms have perfectly separated the data into its five underlying clusters, whereas k-means tended to split the largest cluster into two parts, while incorrectly merging two of the smaller clusters. (K-means is a randomized algorithm; 7 of 10 runs exhibited this error.) Overall, we see that our figure of merit is correctly discriminating among these algorithms on this data set.

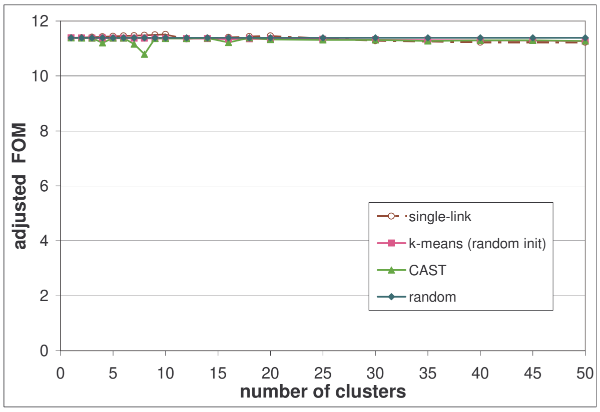

Figure 3: Adjusted FOM's of clustering algorithms on a simulated data set with no clusters.

An example without clusters: Figure 3 shows the adjusted FOM's of the same clustering algorithms on a simulated data set generated according to the same model as in Figure 1(b), which contains *no* intrinsic clusters. All three clustering algorithms have FOM's similar to random clustering on this data set, in sharp contrast to the data set with intrinsic patterns (Figure 2). These simulated data sets show that our methodology can successfully discriminate between data sets with and without intrinsic patterns, and among algorithms.

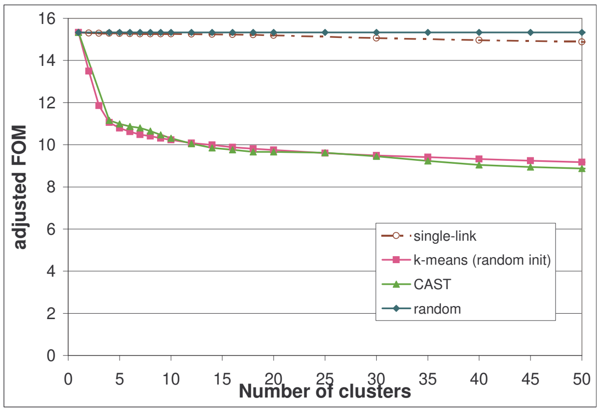

Figure 4: Adjusted FOM's of clustering algorithms on Cho *et al.*'s yeast cell cycle data set.

A real data set: Figure 4 shows the adjusted FOM's of the same clustering algorithms on a yeast cell cycle data set [Cho *et al.*, 1998]. Two facts are evident. First, although the decline in FOM is not as dramatic as that shown on the (low noise) synthetic data in Figure 2, there is none the less a significant decline, strongly suggesting that this data does contain intrinsic patterns. Second, as in Figure 2, the graph reveals distinct differences among the algorithms, although different ones than in Figure 2. Namely, k-means (with random initialization) and CAST are nearly indistinguishable, but both are decisively better than single-link, which performs no better than random placement into clusters. (See Section 6 for further discussions.)

How many clusters are really present? Ideally, we would like to be able to compare proposed clusterings having different numbers of clusters. Unfortunately, determining the "right" number of clusters in real data is a long-standing and very difficult problem [Jain and Moreau, 1987]. To the best of our knowledge, this has been rigorously approached only when there is a well-grounded statistical model of the data [Banfield and Raftery, 1993], which is not yet available for gene expression data. Although our method may give some hints as to how many clusters the data support, and although our adjusted figure of merit will be neutral to number of clusters under appropriate assumptions (see Theorem 1 in Section 3), it still seems safest to compare the adjusted figures of merit of two algorithms only when they are generating the same number of clusters.

Clustering algorithms typically have parameters that directly (for example, k-means) or indirectly (for example, CAST's similarity threshold) determine the number of clusters. To compare the figures of merit of clusters produced by two different algorithms, we adjust the parameters so that the number of clusters is the same in both cases. Plots of adjusted FOM versus number of clusters, as shown above, then give an overall picture of the behavior of the clustering algorithm. Figure 4 is fairly typical. The relatively steep decline in the adjusted FOM's when the number of clusters is small probably reflects genuine progress in producing more homogeneous clusters, whereas, the more gradual declines for larger number of clusters probably reflect purely statistical effects of increasing the number of clusters, for example, isolating outliers in singleton sets.

Related work: Our approach has some similarity to *leave-one-out cross validation* in machine learning. In leave-one-out cross validation, the objective is to estimate the accuracy of a *classifier*, an algorithm that maps an unlabeled instance to a label, by *supervised learning* [Kohavi, 1995]. The labels of the objects to be classified are assumed to be known. The idea is to hide the label of each object in turn, and to estimate the label of the object using a classifier. This is in contrast to our approach in which we do *not* assume any prior information of the genes to evaluate the quality of clustering results. Instead, we define figures of merit, which are estimators of the *predictive* power of clustering algorithms, to assess the quality of clustering results.

Validating clustering results is a well-studied problem in statistics. [Jain and Dubes, 1988] divides cluster validation procedures into two main categories: external and internal criterion analysis. External criterion analysis validates a clustering result by comparing it to a given "gold" standard which is another partition of the objects. The "gold" standard can be obtained by an independent process based on information other than the given data set. There are many statistical measures that assess the agreement between an external criterion and a clustering result. For example, [Milligan *et al.*, 1983] and [Milligan and Cooper, 1986] evaluated the performance of different clustering algorithms and different statistical measures of agreement on both synthetic and real data. We use external criteria to validate our FOM methodology in Section 7, but reliable external criteria are rarely available when analyzing gene expression data. Internal criterion analysis uses information from within the given data set to represent the goodness of fit between the input data set and the clustering results. For example, compactness and isolation of clusters are possible measures of goodness of fit.

For validation of clustering results, external criterion analysis has the strong benefit of providing an independent, hopefully unbiased assessment of cluster quality. On the other hand, external criterion analysis has the strong disadvantage that an external "gold standard" is rarely available. Internal criterion analysis avoids the need for such a standard, but has the alternative problem that clusters are validated using the same information from which clusters are derived. Different clustering algorithms optimize different objective functions or criteria. Assessing the goodness of fit between the input data set and the resulting clusters is equivalent to evaluating the clusters under a different objective function. Our approach compromises these two extremes: *no* external standard is required, and the clustering results are evaluated based on homogeneity of the *hidden* data that are not available to the clustering algorithms. A fictitious example, which illustrates possible merits of our approach, is shown in the Appendix.

3 Figure of Merit

A *figure of merit* is an estimate of the predictive power of a clustering algorithm. A typical gene expression data set contains measurements of expression levels of n genes measured under m experimental conditions. Suppose a clustering algorithm is applied to the data from conditions $1, \ldots, (e-1), (e+1), \ldots, m$, and condition e is used to estimate the predictive power of the algorithm. Suppose there are k clusters, $C_1, C_2, \ldots, C_k$. Let $R(g, e)$ be the expression level of gene g under condition e in the raw data matrix. Let $\mu_{C_i}(e)$ be the average expression level in condition e of genes in cluster C_i. The *2-norm* figure of merit, $FOM(e, k)$, for k clusters using condition e as validation provides an estimate of the mean error of predicting the expression levels from the average expression levels of the clusters in condition e. The 2-norm FOM is essentially the RMS deviation in the left-out condition e of the individual gene expression levels relative to their cluster means:

$$FOM(e,k) = \sqrt{\frac{1}{n} * \sum_{i=1}^{k} \sum_{x \in C_i} (R(x,e) - \mu_{C_i}(e))^2}$$

Each of the m conditions can be used as the left-out experimental condition. The *aggregate figure of merit*, $FOM(k) = \sum_{e=1}^{m} FOM(e,k)$, is an estimate of the total predictive power of the algorithm over all the conditions for k clusters in a data set.

We have also evaluated other definitions of figure of merit; see [Yeung *et al.*, 2000] for details.

Adjusted Figure of Merit

With isolated exceptions, all the data sets (both real and synthetic) we have considered exhibit declining figures of merit under all algorithms, including the random algorithm, as the number of clusters increases. Two factors contribute to this. First, the algorithms may be finding higher quality clusterings, as they subdivide large, coarse clusters into smaller, more homogeneous ones. Second, simply increasing the number of clusters will tend to decrease the FOM.

The analysis in this section partially quantifies the behavior of the figure of merit, and formally defines the *adjusted figure of merit*, which corrects for the second effect.

We assume the following idealized model. Suppose that the n genes fall into c true classes, with the ith class containing $\alpha_i n$ genes, where $0 < \alpha_i < 1$ and $\sum_{i=1}^{c} \alpha_i = 1$. Further assume that the expression levels of genes in class i under condition e are independent normally distributed random variables with mean $\mu_{i,e}$ and variance $\sigma_{i,e}^2$.

Suppose we apply a clustering algorithm to the n genes to obtain k clusters, where $k \geq c$. We assume that the clustering algorithm is perfect, in the sense that each cluster contains genes from only one class. Assume there are $\alpha_i k$ clusters containing class i genes. (This assumption is valid if the clustering algorithm favors equal-sized clusters. However, the analysis is otherwise independent of the sizes of the clusters within each class.)

Theorem 1 *With the above assumptions, the expected aggregate 2-norm FOM, $FOM(k)$, is $\sqrt{\frac{n-k}{n}}\bar{\sigma}$, where $\bar{\sigma}$ is a weighted average of the $\sigma_{i,e}$, independent of k. Specifically, $\bar{\sigma} = \sum_{e=1}^{m} \sqrt{\sum_{i=1}^{c} \alpha_i \sigma_{i,e}^2}$.*

Proof Outline: Suppose the measured expression levels of the $\alpha_i n$ genes in true class i, condition e are $x_{1,e}, \ldots, x_{\alpha_i n, e}$. Let $\bar{x}_e = \sum_{i=1}^{\alpha_i n} x_{i,e} / \alpha_i n$. Then the expected value of $\sum_{i=1}^{\alpha_i n} (x_{i,e} - \bar{x}_e)^2$ is $(\alpha_i n - 1)\sigma_{i,e}^2$. Subdividing this cluster into $\alpha_i k$ smaller nonempty sub-clusters would reduce these genes' expected contribution to the $FOM(e,k)$ to $(\alpha_i n - \alpha_i k)\sigma_{i,e}^2$. Hence, $FOM(k) = \sum_{e=1}^{m} \sqrt{\frac{1}{n} * \sum_{i=1}^{c} (\alpha_i n - \alpha_i k)\sigma_{i,e}^2} = \sqrt{\frac{n-k}{n}} \sum_{e=1}^{m} \sqrt{\sum_{i=1}^{c} \alpha_i \sigma_{i,e}^2}$. $\square$

If the assumptions in Theorem 1 are satisfied and round-off errors ($\alpha_i k$ is assumed to be an integer) are ignored, the rate of decline of 2-norm figures of merit as k, the number of clusters, increases should be $\sqrt{\frac{n-k}{n}}$. Define the *adjusted figure of merit* of k clusters to be $FOM(e,k)/\sqrt{\frac{n-k}{n}}$. The figures in Section 2 were shown using the adjusted figure of merit. In our results on real data (for example, Figure 4), the adjusted FOM of the random algorithm is very close to constant with respect to the number of clusters, despite the fact that the yeast cell cycle data set and the clustering algorithms probably violate key assumptions in the foregoing analysis.

4 Data Sets

The Barrett's Esophagus Data Set

Barrett's esophagus (BE) is a condition in which the normal squamous esophageal mucosa is replaced by a metaplastic columnar epithelium. It develops as a complication of chronic gastroesophageal reflux disease (GERD) and predisposes to the development of adenocarcinomas of the esophagus and cardia. Patients with Barrett's esophagus frequently have symptoms of GERD, such as heartburn and indigestion, and frequently seek medical attention before the development of cancer. The standard care for many patients is periodic endoscopic surveillance with surgery reserved for the subset of patients who develop esophageal adenocarcinoma. Thus endoscopic biopsies from patients with BE can be acquired from all stages of disease and provide highly favorable material for studying human neoplasia in vivo.

The Barrett's data set [Barrett *et al.*, 2000] consists of 7306 genes and 10 conditions. Fresh biopsies of each tissue from 2 to 3 patients were pooled prior to RNA extraction. The 10 conditions consist of 4 pools of esophageal squamous biopsies, 4 pools of Barrett's esophagus biopsies, one pool of duodenal biopsies, and one pool of gastric biopsies. The data set was filtered using the GENECLUSTER software [Tamayo *et al.*, 1999], and 795 genes passed the filter of absolute change 300 and relative change 4. The data set was then normalized to have mean 0 and variance 1.

The Rat CNS Data Set

The rat CNS data set was obtained by reverse transcription-coupled PCR to study the expression levels of 112 genes during rat central nervous system development [Wen *et al.*, 1998] over 9 time points. As suggested in [Wen *et al.*, 1998], the raw data was normalized by the maximum expression level for each gene. The data set was then augmented with slopes (differences between consecutive time points) to capture parallel trajectories of the time course data. This results in a data set with 112 genes and 17 conditions.

The Yeast Cell Cycle Data Set

The yeast cell cycle data set [Cho *et al.*, 1998] shows the fluctuation of expression levels of approximately 6000 genes over approximately two cell cycles (17 time points). By visual inspection of the raw data, [Cho *et al.*, 1998] identified 420 genes showing significant variation over the course of the experiment. The data set was normalized as in [Tamayo *et al.*, 1999]: the 17 conditions were divided into two panels (which correspond to two cell cycles) and were normalized to have mean 0 and variance 1 within each panel.

The Ovary Data Set

A subset of the ovary data set ([Schummer *et al.*, 1999], [Schummer, 2000]) is used. The ovary data set was generated by hybridizing randomly selected cDNAs from normal and neoplastic ovarian tissues to membrane arrays. The subset of ovary data set we used contains 233 clones and 24 samples, 7 of which are derived from normal tissues, 4 from blood samples, and the remaining 13 from ovarian cancers in various stages of malignancy. The 233 clones were sequenced, and they correspond to 4 different genes.

5 Clustering Algorithms

We implemented three partitional clustering algorithms: the *Cluster Affinity Search Technique* (CAST) [Ben-Dor and Yakhini, 1999], an *iterative* partition algorithm [Yeung *et al.*, 2000] and the *k-means* algorithm [Jain and Dubes, 1988]. Three hierarchical clustering algorithms were also implemented: single-link, average-link and complete-link [Jain and Dubes, 1988].

We also investigated the effect of initialization methods for the k-means algorithm. One implementation initializes the algorithm with randomly selected genes as centroids. Another implementation uses the results from average-link as initial centroids.

We also implemented random clustering as a benchmark for evaluating the performance of a clustering algorithm. A random clustering with k clusters is obtained by placing k randomly selected genes into separate bins, then placing the remaining genes into the same bins uniformly at random. An algorithm whose figure of merit is little better than that of a random clustering is probably producing poor clusters.

6 Results and Discussion

In this section, we compare the performance of various clustering algorithms on the Barrett's esophagus data set [Barrett *et al.*, 2000], the rat CNS data set [Wen *et al.*, 1998], the yeast cell cycle data set [Cho *et al.*, 1998], and the ovary data set [Schummer, 2000]. The results using another data set [Tamayo *et al.*, 1999] (not shown here) can be found in our supplementary web site or [Yeung *et al.*, 2000]. The correlation coefficient was used to compute pairwise similarities of genes. In our experiments, the random clustering algorithm was repeated 1000 times, the iterative algorithm was run 10 times and the k-means algorithm with random initialization was run 30 times to obtain reliable FOM's. We also studied the variation in FOM reported by the randomized algorithms. The standard deviations of the FOM of k-means (with random initialization) and the random clustering algorithms are under 10% of the average FOM.

The Barrett's Esophagus Data Set

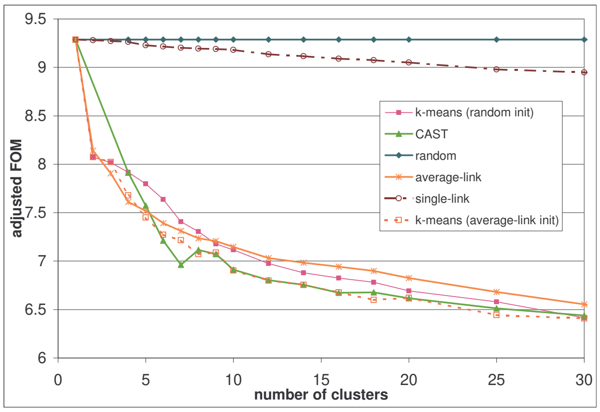

Figure 5: Adjusted FOM's of clustering algorithms on the Barrett's esophagus data set (795 genes).

Figure 5 shows the adjusted FOM's on the Barrett's esophagus data set. For clarity of the figure, the results of complete-link and iterative algorithms are not shown. For the detailed results, please refer to our supplementary web site. When the number of clusters is above 30 (data not shown), the decline in the adjusted FOM's is very gentle, so only 1 to 30 clusters are shown. Figure 5 shows a drastic decline in FOM up to 10 clusters, thus suggesting the number of clusters of interest is around 10. Moreover, the FOM's of CAST and k-means with average-link initialization are comparable and are the best for 10 clusters.

With our FOM analysis in mind, we produced clustering results on the full data set (all 10 conditions) with 10 clusters using CAST, average-link and k-means with average-link initialization. Then, we compared the clusters with prior biological knowledge. Cytokeratins have a tissue specific profile and can be used to distinguish and purify tissues of interest in flow cytometry assays. Probe sets for twenty members of the cytokeratin genes passed the initial filtering criteria. The ten clusters generated by CAST and k-means with average-link initialization correctly placed each of the cytokeratins in their respective clusters. In contrast, average-link with 10 clusters had a higher FOM, generated two small clusters that include only two and three members and did not correctly assign the cytokeratins to tissue specific clusters. Therefore, clusterings of cytokeratins suggest that CAST and k-means with average-link initialization produce more robust clusters than average-link, which confirm our FOM analysis.

The Rat CNS Data Set

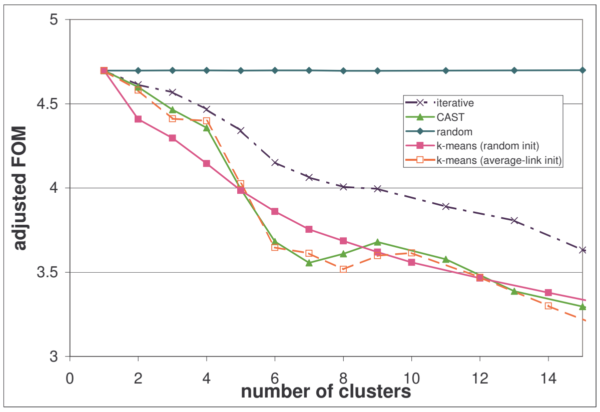

Figure 6: Adjusted FOM's of partitional clustering algorithms on the rat CNS data (112 genes).

Figure 6 and Figure 7 show the adjusted FOM's between 1 to 15 clusters, which is expected to be the range of interest. ([Wen *et al.*, 1998] categorized genes in the rat CNS data set into four families using biological knowledge.)

In Figure 7, hierarchical complete-link and average-link achieve lower adjusted FOM's than single-link. Single-link tends to produce "chained" clusters in which genes at opposite ends of a cluster can be very dissimilar [Anderberg, 1973]. Our results are consistent with the general belief that single-link is less desirable than complete- and average-link. Average-link clustering may be more tolerant of outliers than complete-link clustering, which may explain the reduced FOM's obtained by complete link on this data.

Figure 6 shows that both initialization methods (random and average-link) of k-means achieve comparable FOM's to CAST. Figure 7 shows that k-means initialized with average-link results achieve lower FOM's than average-link alone. Therefore, the iterative k-means step after average-link improves the quality of clusters. Note that k-means with average-link initialization is a deterministic algorithm, which is more suited for further analysis.

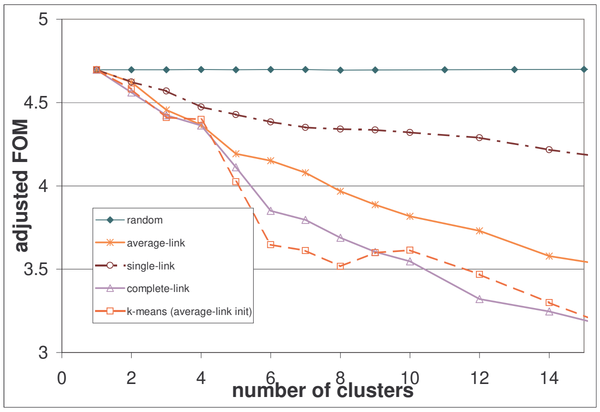

Figure 7: Adjusted FOM's of hierarchical clustering algorithms on the rat CNS data set (112 genes).

The Yeast Cell Cycle Data Set

Figure 4 on page 2 shows the adjusted FOM's of four clustering algorithms for 1 to 50 clusters on the 420 gene yeast cell cycle data set. We computed the adjusted FOM's over all possible number of clusters (*i.e.*, from 1 to 420 clusters). We find that the rate of decline of the adjusted FOM's with respect to the number of clusters is relatively gentle above 50. So, we believe that the range of interest is between 1 to 50 clusters. Below 25 clusters, k-means with random initialization and CAST have comparable FOM's and lower FOM's than the iterative algorithm. Above 35 clusters, CAST achieves a slightly lower FOM, but all three algorithms are close.

We also compared other clustering algorithms on the yeast cell cycle data set (data not shown). Again, the average-link and complete-link outperform the single-link. The average-link and complete-link algorithms have slightly higher FOM's than CAST and k-means. The k-means algorithm with average-link initialization achieves lower FOM's than average-link (data not shown).

The Ovary Data Set

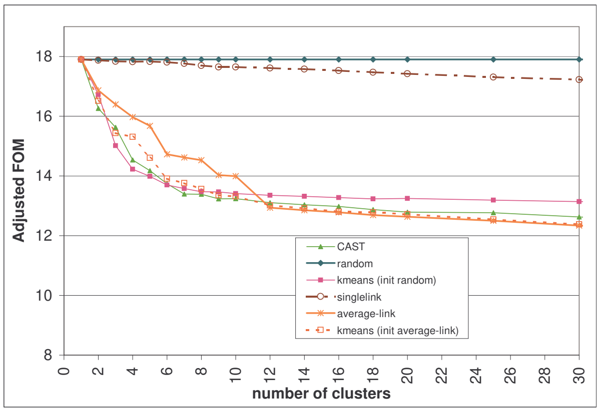

Figure 8: Adjusted FOM's of clustering algorithms on the ovary data set (233 clones).

Figure 8 shows the adjusted FOM's on the ovary data set from 1 to 30 clusters. The CAST and k-means algorithms (with both random and average-link initializations) achieve the lowest FOM's on this data. Both algorithms show a drastic decline of FOM's up to around 4 to 6 clusters. The 233 clones in this data set correspond to 4 genes. Therefore, our result suggests the "correct" number of clusters.

Our FOM methodology suggests that CAST and k-means algorithms achieve higher quality clusters than other algorithms. We applied CAST, average-link and k-means (with average-link initializations) to the full ovary data set (with all 24 conditions) to obtain four clusters. Each of the clusters from CAST and k-means contains mostly clones from one gene only, while average-link combined clones from two genes into one cluster. This is evidence that clustering algorithms with lower FOM's produce higher quality clusters.

7 External Validation of FOM

In addition to the comments from Section 6, we will describe experiments, based on real data, to validate the use of our figure of merit as a measure of the predictive power of clustering algorithms. Two possible concerns about our FOM methodology are (a) that the particular FOM chosen, while intuitively appealing, may be a poor indicator of cluster quality, and (b) that holding out data from one condition will result in significant degradation in the quality of clusters found by any algorithm. Success of the method on artificial data, as illustrated in Figures 2 and 3, is useful evidence to the contrary, but such data may only weakly reflect the complexities of real data sets.

For the rat CNS and yeast cell cycle data sets, functional categorizations of a subset of the genes in the given data set are known. These functional categorizations were derived from information other than gene expression data, but the expression data sets are expected to reflect the functional categories to a certain degree. We compare clustering results to these known functional categories and evaluate the degree to which FOM reflects this external standard of cluster quality. Note that our cluster validation methodology is applicable whether or not an external standard is available.

Agreement Between Two Partitions

The adjusted Rand index [Hubert and Arabie, 1985] assesses the degree of agreement between two partitions of the same set of objects. [Milligan and Cooper, 1986] recommended the adjusted Rand index as the measure of agreement even when clusters across different hierarchy levels (*i.e.,* different numbers of clusters) are compared. The adjusted Rand index has the maximum value of 1, which means perfect agreement between the external criterion and the clustering result. The expected value of the adjusted Rand index in the case of random clusters is 0. A higher adjusted Rand index means a higher correspondence to the "gold standard". In this paper we will focus on results using the adjusted Rand index. Similar results were also obtained for other external indices; see [Yeung *et al.*, 2000] for details.

External Validation Methodology

Given a functional categorization of genes into k classes, we apply a clustering algorithm to all conditions except condition e to produce k clusters. We then compute $FOM(e, k)$ and the adjusted Rand index of the clustering result compared against the given functional categorization. This process is repeated for different clustering algorithms, and (for randomized algorithms) for different random trials of the same algorithm. Finally, we plot the adjusted Rand index against $FOM(e, k)$ for these experiments. Recall that a high adjusted Rand index indicates high similarity of a clustering result to the given functional categorization, and that a low figure of merit indicates high predictive power. Thus, we take a downward trend in the data as evidence for the predictive power of FOM for comparing clustering algorithms—rising FOM tends to indicate declining correlation of the clustering to the external standard. This is exactly the behavior observed in the experiments described below. In the following sections, results of leaving out one particular experiment are shown. We investigated the effect of leaving out other experimental conditions, and found that the figures shown are typical.

External validation results on rat data

[Wen *et al.*, 1998] categorized genes in the rat CNS data set into four families using biological knowledge. The k-means (with random and average-link initializations), CAST and random algorithms were applied to 16 conditions (condition 0 omitted) of the 112 genes in the rat CNS data set to produce four clusters. Figure 9 shows the result of plotting the adjusted Rand index against $FOM(0, 4)$. Three trials of k-means with random initializations are shown in Figure 9. The intermediate results of successive iterations of CAST and k-means are also shown. The random algorithm was repeated 30 times. Successive iterations of k-means and CAST tend to show lower FOM's and higher adjusted Rand indices.

As shown in Figure 9, the adjusted Rand index for random partitions reveals the expected lack of correlation to [Wen *et al.*, 1998]'s categorization, whereas the most of the points resulting from runs of the other clustering algorithms show significantly stronger correlations. Furthermore, there is an obvious trend in the data, with lower (better) FOM associated with higher (better) adjusted Rand index—i.e., clusterings having better FOM also tend to have higher correlation to [Wen *et al.*, 1998]'s biologically informed categorization. Although the adjusted Rand indices are not high in absolute terms, the points plotting the adjusted Rand index versus FOM scores for random partitions are tightly clustered;

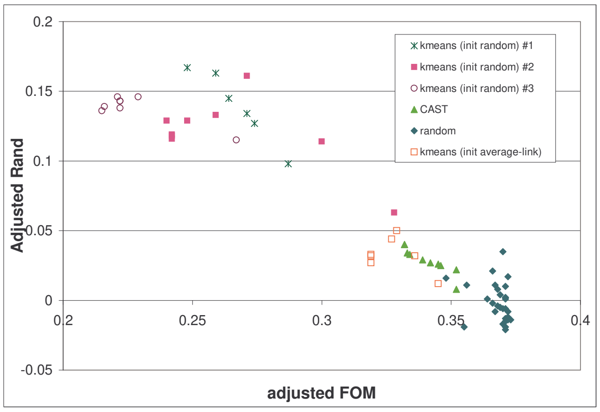

Figure 9: The adjusted Rand index (based on Wen *et al.*'s four categories) against $FOM(0, 4)$ on the rat CNS data set.

the majority of the other points lie several standard deviations away from the mean of this control group, showing that clusterings of this quality in either metric are very unlikely to have arisen by chance. Hence, we conclude that clustering results with low FOM's tend to have a relatively high correspondence with the functional categorization in Wen *et al.* on the rat CNS data set.

External validation results on yeast data

Cho *et al.* categorized approximately 380 genes into five phases of cell cycle. Since the 420 genes were identified by visual inspection of gene expression data according to the peak times of genes, we expect clustering results to correspond to the five known categories of genes. The results (with intermediate results of successive iterations of CAST and k-means) for leaving out the first time point ($e = 0$) are shown in Figure 10.

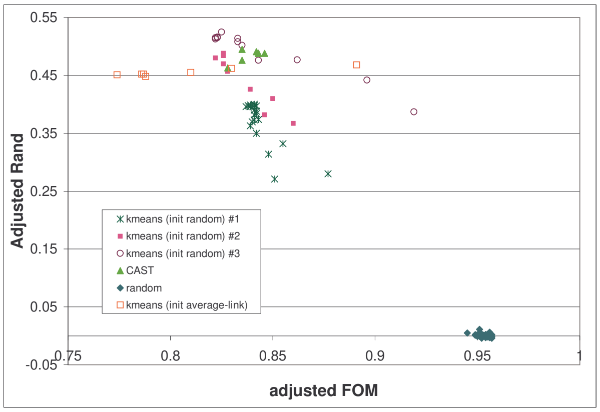

Figure 10: The adjusted Rand index (based on five stages of cell cycle) against $FOM(0, 5)$ on the yeast cell cycle data set.

The general picture is the same as seen in the rat CNS data. There is a downward trend in Figure 10. The k-means (with random and average-link initializations) and CAST algorithms show significantly lower FOM's and higher adjusted Rand indices than random. Again, successive iterations of k-means and CAST tend to show lower FOM's and higher adjusted Rand indices.

Effect of Omitting One Condition

On both the rat CNS and yeast cell cycle data sets, the adjusted Rand indices of clustering results using all 17 conditions are comparable to those when one condition was left out. This shows that leaving out one condition does not have a significant effect on the quality of clustering results.

8 Conclusions and Future Work

Summary of Comparisons: Our results in Section 6 confirm the general belief that average-link and complete-link algorithms tend to be more desirable than single-link. We also show that CAST tends to have relatively high predictive power. We also investigated the effect of different initialization methods for the k-means algorithm, and found that k-means with average-link initialization achieve comparable FOM's to k-means with random initialization and CAST. Furthermore, we show that the iterative k-means step after average-link improves cluster quality. Since CAST and k-means with average-link initialization are deterministic algorithms, the clustering results are reproducible in every run. Based on our results, we would recommend using CAST or k-means with average-link initialization for analysis of gene expression data, and would recommend against using single-link.

Our Contributions: Our main contribution is not the comparison of specific algorithms on specific data sets, but rather the development of a simple, quantitative data-driven methodology allowing such comparisons to be made between any clustering algorithms on any data set. We present experimental evidence that our methodology produces results that are well correlated with biologically relevant external standards on real data sets. Additionally, we present preliminary but interesting comparisons of several important clustering algorithms. Our experience with the Barrett's esophagus and ovary data sets shows that clustering algorithms recommended by our FOM methodology actually produces relatively high quality clusters.

Our methodology can also be used to verify the existence of clusters in a given data set. The graph of adjusted FOM against the number of clusters for the simulated data set without any intrinsic clusters (Figure 3) shows a very different trend than data sets with intrinsic patterns (Figure 2).

Future Work: An interesting direction of further research would be definitions of figures of merits that depend on the similarity metrics used in clustering algorithms. For example, if the goal is to capture anti-correlated genes and the absolute value of the correlation coefficient is used to compute pairwise similarities between genes, our current definition of FOM to measure within-cluster variation is not appropriate. Another interesting direction is to design FOM's that depend on the normalization methods of the data.

To summarize, clustering is a difficult problem. We believe that the methodology introduced in this paper for quantitative comparison of the predictive power of clustering algorithms will prove to be a valuable ingredient in future clustering studies.

Acknowledgement

We would like to thank Mike Barrett, Jeff Delrow and the Reid Lab at the Fred Hutchinson Cancer Research Center for the Barrett's esophagus data set, and for the feedback for our analysis on the Barrett's data set. We would also like to thank Michel Schummer from the Department of Molecular Biotechnology at University of Washington for the ovary data set. We would like to thank Richard M. Karp at University of California, Berkeley for suggesting the iterative algorithm. We would like to thank Amir Ben-Dor for sharing the additional heuristics implemented in their software with us. In addition, we would like to thank Lue Ping Zhao at the Fred Hutchinson Cancer Research Center for his suggestions on modeling simulation data sets. We would like to thank the Whitehead Institute for use of their GENECLUSTER software. We would also like to thank Michael Eisen and the Stanford University for use of their CLUSTER and TREEVIEW softwares. This work is partially supported by NSF grant DBI-9974498.

Appendix

Figure 11 illustrates a fictitious data set in 2 dimensional space in which the data points are on the perimeters of three circles. The average pairwise Euclidean distance is a possible measure of compactness of clusters. In this example, there are two reasonable 2-cluster results: either circles 1 and 2 or circles 1 and 3 can be combined into one cluster. Intuitively, combining circles 1 and 3 is appealing because they overlap. For the clustering result with circles 1 and 3 combined, the FOM is lower but the average Euclidean distance is higher than the other clustering result. Even though this example is highly contrived and does not reflect the complexities of gene expression data, it illustrates a situation in which one might want to consider predictive power when assessing clustering results in addition to or instead of using measures of compactness.

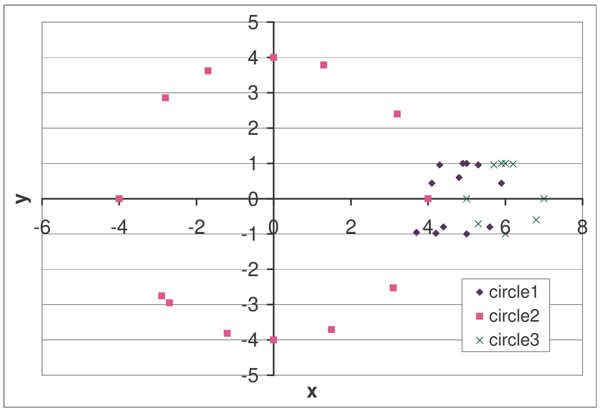

Figure 11: Example illustrating merits of our FOM approach.

References

[Anderberg, 1973] Anderberg, M. R. (1973) *Cluster analysis for applications*. Academic Press.

[Banfield and Raftery, 1993] Banfield, J. D. and Raftery, A. E. (1993) Model-based gaussian and non-gaussian clustering. *Biometrics*, **49**, 803–821.

[Barrett *et al.*, 2000] Barrett, M. T., Delrow, J., Yeung, K. Y., Blount, P. L., Reid, B. J. and Rabinovitch, P. S. (2000) Will be submitted.

[Ben-Dor and Yakhini, 1999] Ben-Dor, A. and Yakhini, Z. (1999) Clustering gene expression patterns. In *RECOMB99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, Lyon, France.

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

[DeRisi *et al.*, 1997] DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

[Efron, 1982] Efron, B. (1982) *The jackknife, the bootstrap, and other resampling plans*. Society for Industrial and Applied Mathematics.

[Eisen *et al.*, 1998] Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science USA*, **95**, 14863–14868.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of Classification*, 193–218.

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

[Jain and Moreau, 1987] Jain, A. K. and Moreau, J. V. (1987) Bootstrap technique in cluster analysis. *Pattern Recognition*, **20**, 547–568.

[Kohavi, 1995] Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*.

[Michaels *et al.*, 1998] Michaels, G. S., Carr, D. B., Askenazi, M., Fuhrman, S., Wen, X. and Somogyi, R. (1998) Cluster analysis and data visualization of large-scale gene expression data. In *Pacific Symposium on Biocomputing 3*.

[Milligan and Cooper, 1986] Milligan, G. W. and Cooper, M. C. (1986) A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, **21**, 441–458.

[Milligan *et al.*, 1983] Milligan, G. W., Soon, S. C. and Sokol, L. M. (1983) The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **5**, 40–47.

[Schummer, 2000] Schummer, M. (2000) Manuscript in preparation.

[Schummer *et al.*, 1999] Schummer, M., Ng, W. V., Bumgarner, R. E., Nelson, P. S., Schummer, B., Bednarski, D. W., Hassell, L., Baldwin, R. L., Karlan, B. Y. and Hood, L. (1999) Compartive hybridization of an array of 21500 ovarian cdnas for the discovery of genes overexpressed in ovarian carcinomas. *An International Journal on Genes and Genomes*, **238**, 375–385.

[Tamayo *et al.*, 1999] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Science USA*, **96**, 2907–2912.

[Wen *et al.*, 1998] Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. and Somogyi, R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Science USA*, **95**, 334–339.

[Yeung *et al.*, 2000] Yeung, K. Y., Haynor, D. R. and Ruzzo, W. L. (2000) Validating clustering for gene expression data. Tech. Rep. UW-CSE-00-01-01, Dept. of Computer Science and Engineering, University of Washington.