Validating clustering for gene expression data

K. Y. Yeung¹,* , D. R. Haynor² and W. L. Ruzzo¹

¹Computer Science and Engineering, Box 352350, University of Washington, Seattle, WA 98195, USA and ²Radiology, Box 357115, University of Washington, Seattle, WA 98195, USA

Received on August 23, 2000; revised on November 23, 2000; accepted on December 1, 2000

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 six 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


INTRODUCTION
In an attempt to understand complicated biological systems, large amounts of gene expression data have been generated by researchers (for example, DeRisi et al., 1997; 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 a variant of the hierarchical average-link 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). Given the same data set, different clustering algorithms can potentially generate very different clusters. 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. 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; Eisen et al., 1998). Our paper provides a quantitative data-driven framework to compare different clustering algorithms.

As a specific example, consider the Barrett’s esophagus data set (Barrett et al., 2000), discussed in more detail in Sections Data sets and Results and discussion. This data set contains samples of four tissue types, one neoplastic and three normal. We applied three clustering algorithms to this data, producing the same number of clusters with each. One goal was to identify clusters of genes having tissue specific profiles. Biologists previously had identified about twenty such genes. Two of the clustering algorithms assigned these genes to appropriate clusters, while the third separated them into different clusters, seemingly arbitrarily. Therefore, the biologists had little confidence in the third clustering. Unfortunately, this sort of prior biological knowledge is not always available. Hence, there is a great need for a systematic data-driven approach to compare the performance of different clustering algorithms. This paper offers such an approach.

In the above example, our methodology favors the same clusterings as the biologists, a determination made based solely on the expression array data itself, without reliance on additional biological information.

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 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 root mean square 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 FOMs 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 Figure of merit.

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.

**A real data set**

Figure 2 shows the adjusted FOMs 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 1, there is none the less a significant decline, strongly suggesting that this data does contain intrinsic patterns. Second, as in Figure 1, the graph reveals distinct differences among the algorithms, although different ones than in Figure 1. 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. It turns out that single-link tends to perform poorly over many different gene expression data sets as shown in the Results and discussion. Our methodology shows that single-link has poor performance in real data, but not in our synthetic data.
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). Milligan and Cooper (1985) evaluates the performance of 30 procedures for estimating the number of clusters on several artificial data sets. The gap statistic of Tibshirani et al. (2000) is an appealing recent attempt to estimate the number of clusters by comparing within-cluster dispersion to that of a reference null distribution. However, on purely philosophical grounds, it seems impossible to determine the ‘right’ number of clusters, or even to define the concept, in the absence of a well-grounded statistical model (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 Figure of merit), 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 2 is fairly typical. The relatively steep decline in the adjusted FOMs 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 must 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 External validation of FOM, 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.

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 root mean square deviation in the left-out condition $e$ of the individual gene expression levels relative to their cluster means:

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

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. (2000b) 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 $i$th class containing $a_i n$ genes, where $0 < a_i < 1$ and $\sum_{i=1}^{c} a_i = 1$. 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 $a_i k$ 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 $\sqrt{\frac{a-k}{n}} \tilde{\sigma}$, where $\tilde{\sigma}$ is a weighted average of the $\sigma_{i,e}$, independent of $k$.

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

If the assumptions in Theorem 1 are satisfied and round-off errors ($a_i k$ 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 $\sqrt{\frac{a-k}{n}}$. Define the adjusted figure of merit of $k$ clusters to be $FOM(e, k)/\sqrt{\frac{a-k}{n}}$. The figures in the previous section were shown using the adjusted figure of merit. In our results on real data (for example, Figure 2), 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.

**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 BE 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 ten conditions. Fresh biopsies of each tissue from 2 to 3 patients were pooled prior to RNA extraction. The ten conditions consist of four pools of esophageal squamous biopsies, four pools of BE 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 nine 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, seven of which are derived from normal tissues, four from blood samples, and the remaining 13 from ovarian cancers in various stages of malignancy. The 233 clones were sequenced, and they correspond to four different genes.

CLUSTERING ALGORITHMS
We implemented two partitional clustering algorithms: the Cluster Affinity Search Technique (CAST) (Ben-Dor and Yakhini, 1999) 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). A dendrogram is built bottom-up in each of the hierarchical algorithms until k subtrees were obtained. k clusters were obtained from the dendrogram by assuming that each of the k subtrees corresponds to a cluster. The details of the implementation can be found in (Yeung et al., 2000a).

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.

RESULTS AND DISCUSSION
In this section, we compared the performance of various clustering algorithms on the data sets described in Section Data sets. The results using another data set (Tamayo et al., 1999) (not shown here) can be found in Yeung et al. (2000b) and Yeung et al. (2000a). We did not show the results from all the clustering algorithms we implemented for clarity of figures. For details, please refer to Yeung et al. (2000a). The correlation coefficient was used to compute pairwise similarities of genes. In our experiments, the random clustering algorithm was repeated 1000 times, and the k-means algorithm with random initialization was run 30 times to obtain reliable FOMs. 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 Barret’s esophagus data set
Figure 3 shows the adjusted FOMs on the BE data set. When the number of clusters is above 30 (data not shown), the decline in the adjusted FOMs is very gentle, so only 1–30 clusters are shown. The FOMs of CAST and k-means with average-link initialization are comparable and are the best when the number of clusters is at least eight. Figure 3 also shows a sharp decline in FOM up to ten clusters, thus suggesting the number of clusters of interest is around 10.

With our FOM analysis in mind, we produced clustering results on the full data set (all ten conditions) with ten clusters using CAST, average-link and k-means with average-link initialization. Then, we compared the clusters in light of prior biological knowledge about the data. 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 ten 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, manual inspection and 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

Figures 4 and 5 show the adjusted FOMs for 1–15 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 5, hierarchical complete-link and average-link achieve lower adjusted FOMs 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 FOMs obtained by complete-link on this data.

Figure 4 shows that both initialization methods (random and average-link) of k-means achieve comparable FOMs to CAST. Figure 5 shows that k-means initialized with average-link results achieve lower FOMs 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.

The yeast cell cycle data set

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

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

The ovary data set

Figure 6 shows the adjusted FOMs 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 FOMs on this data. Both algorithms show a steep decline of FOMs up to around 4–6 clusters. The 233 clones in this data set correspond to four 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 FOMs produce higher quality clusters.

EXTERNAL VALIDATION OF FOM

In addition to the comments from the previous section, 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 Figure 1, 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 comparing clusters across different hierarchy levels (i.e. different numbers of clusters). 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 present results using the adjusted Rand index. Similar
Comparing clustering algorithms on gene expression data

Fig. 3. Adjusted FOMs of clustering algorithms on the BE data set (795 genes).

Fig. 4. Adjusted FOMs of partitional clustering algorithms on the rat CNS data (112 genes).

Fig. 5. Adjusted FOMs of hierarchical clustering algorithms on the rat CNS data set (112 genes).

Fig. 6. Adjusted FOMs of clustering algorithms on the ovary data set (233 clones).

Fig. 7. The adjusted Rand index (based on Wen et al.’s four categories) against $FOM(0, 4)$ on the rat CNS data set.

Fig. 8. The adjusted Rand index (based on five stages of cell cycle) against $FOM(0, 5)$ on the yeast cell cycle data.
results were also obtained for other external indices; see (Yeung et al., 2000b) 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 \( \text{FOM}(e, k) \) 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 \( \text{FOM}(e, k) \) 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 \( \text{FOM} \) for comparing clustering algorithms—rising \( \text{FOM} \) 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 7 shows the result of plotting the adjusted Rand index against \( \text{FOM}(0, 4) \). Three trials of \( k \)-means with random initializations are shown in Figure 7. 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 FOMs and higher adjusted Rand indices.

As shown in Figure 7, the adjusted Rand index for random partitions reveals the expected lack of correlation to Wen et al.’s (1998) categorization, whereas 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.’s (1998) 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; 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 FOMs 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 8.

The general picture is the same as seen in the rat CNS data. There is a downward trend in Figure 8. The \( k \)-means (with random and average-link initializations) and CAST algorithms show significantly lower FOMs and higher adjusted Rand indices than random. Again, successive iterations of \( k \)-means and CAST tend to show lower FOMs and higher adjusted Rand indices, and all are significantly better than the random controls.

**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.

**CONCLUSIONS AND FUTURE WORK**

**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 BE and ovary data sets shows that clustering algorithms recommended by our FOM methodology actually produce relatively high quality clusters.

**Summary of comparisons**

Although comparison between specific clustering algorithms is not our primary focus, our results lead to some
conclusions on the relative performance of clustering algorithms used in this study. Our results 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 FOMs 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.

Limitations
Our methodology takes a predictive approach, i.e. our model assumes that the left-out experimental condition contains information from the experiments that are used to produce clusters. In other words, our approach compares the relative strength in predictive power of clustering algorithms given the related information in the conditions used to produce clusters. Our approach is not applicable to all situations: if all the experimental conditions contain independent information, no predictive approach is possible. Despite the limitations, we believe that our method is applicable to many gene expression data sets. We successfully applied our method to data sets with varying degree of dependence including time series data (the yeast cell cycle data; (Cho et al., 1998) and the rat data (Wen et al., 1998)) and data sets with different types of tissue samples (the BE data, Barrett et al., 2000; and the ovary data, Schummer, 2000).

Another limitation of our approach is that with our definition of FOM, it is not safe to compare clustering results with different numbers of clusters or different similarity metrics.

Future work
An interesting direction of further research would be definitions of figures of merit 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 FOMs that depend on the normalization methods of the data.

The nature of our methodology to leave out each experiment in turn and repeat clustering makes it computationally intensive for large data sets with lots of experiments. A direction of future work is to leave out groups of experiments at a time for large data sets.

Another direction of future work is to compare our predictive approach to other measures of cluster validation (for example, the within-cluster variance of all the experiments).

We compared the performance of three hierarchical and two partitional clustering algorithms. We would be interested to compare the performance of other clustering algorithms, for example, the Self-Organizing Map (SOM) algorithm (Tamayo et al., 1999). However, SOM has many tunable parameters in addition to the number of clusters. We would like to investigate the stability and performance of SOM with respect to different parameters in the future.

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.

ACKNOWLEDGEMENTS
We would like to thank Mike Barrett at the Fred Hutchinson Cancer Research Center for the BE data set, and Michel Schummer from the Department of Molecular Biotechnology at University of Washington for the ovary data set. We would also like to thank Richard M. Karp at University of California, Berkeley for suggesting the iterative algorithm, and 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. Finally, we would like to thank the Whitehead Institute for use of their CLUSTER and TREE VIEW softwares. This work is partially supported by NSF grant DBI-9974498.

REFERENCES
APPENDIX

Figure 9 illustrates a fictitious data set in two-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.

Fig. 9. Example illustrating merits of our FOM approach.