BIOINFORMATICS

Vol. 19 no. 18 2003, pages 2381–2389
DOI: 10.1093/bioinformatics/btg330

Scoring clustering solutions by their biological relevance

I. Gat-Viks[1,*,†], R. Sharan[2,†] and R. Shamir[1]

[1]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel and
[2]International Computer Science Institute, 1947 Center St., Suite 600, Berkeley, CA 94704, USA

Received on February 5, 2003; revised on May 20, 2003; accepted on June 3, 2003

ABSTRACT

Motivation: A central step in the analysis of gene expression data is the identification of groups of genes that exhibit similar expression patterns. Clustering gene expression data into homogeneous groups was shown to be instrumental in functional annotation, tissue classification, regulatory motif identification, and other applications. Although there is a rich literature on clustering algorithms for gene expression analysis, very few works addressed the systematic comparison and evaluation of clustering results. Typically, different clustering algorithms yield different clustering solutions on the same data, and there is no agreed upon guideline for choosing among them.

Results: We developed a novel statistically based method for assessing a clustering solution according to prior biological knowledge. Our method can be used to compare different clustering solutions or to optimize the parameters of a clustering algorithm. The method is based on projecting vectors of biological attributes of the clustered elements onto the real line, such that the ratio of between-groups and within-group variance estimators is maximized. The projected data are then scored using a non-parametric analysis of variance test, and the score's confidence is evaluated. We validate our approach using simulated data and show that our scoring method outperforms several extant methods, including the separation to homogeneity ratio and the silhouette measure. We apply our method to evaluate results of several clustering methods on yeast cell-cycle gene expression data.

Availability: The software is available from the authors upon request.

Contact: iritg@post.tau.ac.il; rshamir@post.tau.ac.il; roded@icsi.berkeley.edu

INTRODUCTION

DNA microarray technology enables the monitoring of expression levels of thousands of genes simultaneously. This allows a global view on the transcription levels of many genes under specific cellular conditions. The applications of such technology range from gene functional annotation and genetic network reconstruction to diagnosis of disease conditions and characterization of effects of medical treatments.

A central step in the analysis of gene expression data is the identification of groups of genes that exhibit similar expression patterns. Clustering methods transform a large matrix of expression levels into a more informative collection of gene sets (or condition sets) which are assumed to share biological properties. Clustering gene expression data into homogeneous groups was shown to be instrumental in functional annotation, tissue classification, motif identification, and other applications [for a review see Sharan *et al.* (2002)].

Although there has been extensive research on clustering algorithms for gene expression analysis (Eisen *et al.*, 1998; Tamayo *et al.*, 1999; Ben-Dor *et al.*, 1999; Sharan and Shamir, 2000; Sharan *et al.*, 2003), very few works have been published on the systematic comparison and evaluation of clustering results. Typically, different clustering algorithms yield different clustering solutions on the same data, and often the same algorithm yields different results for different parameter settings, and there is no consensus on choosing among them.

Different measures for the quality of a clustering solution are applicable in different situations, depending on the data and on the availability of the true solution. In case the true solution is known, and we wish to compare it to another solution, one can use, e.g. the Minkowski measure (Sokal, 1977) or the Jaccard coefficient [cf. Everitt (1993)]. When the true solution is not known, there is no agreed-upon approach for evaluating the quality of a suggested solution. Several approaches evaluate a clustering solutions based on its intra-cluster homogeneity or inter-cluster separation (Hansen and Jaumard, 1997; Sharan *et al.*, 2003; Yeung *et al.*, 2001). However, the homogeneity and separation criteria are inherently conflicting, as an improvement in one will usually correspond to worsening of the other. One way of getting around this problem is to fix the number of clusters and seek a solution with maximum homogeneity. This is done, for example, by the classical *K*-means algorithm (MacQueen, 1965; Ball and Hall,

*To whom correspondence should be addressed.

†These authors contributed equally to this work.

1967). For methods that evaluate the number of clusters see, e.g. Hartigan (1975); Tibshirani *et al.* (2000); Ben-Hur *et al.* (2002); Pollard and van der Laan (2002); Dudoit and Fridlyand (2002); McLachlan (1987). Another way to overcome the problem is by presenting a curve of homogeneity versus separation (Ben-Dor, private communication). Such a curve can show that one algorithm dominates another if it provides better homogeneity for all separation values, but typically different algorithms will dominate in different value range. An alternative method suggested by Kaufman and Rousseeuw (1990), evaluates a solution using a numerical measure called the average silhouette. This method performs well in general, but fails to detect fine cluster structures (Pollard and van der Laan, 2002).

Clustering quality can also be visually assessed by using discriminant analysis [e.g. Stephanopoulos *et al.* (2002); McLachlan (1992)] or principal component analysis [e.g. Mendez *et al.* (2002)], that reduce data dimensionality. Single clusters can be scored based on prior biological knowledge, e.g. by checking for functional enrichment of genes in a cluster or searching for common motifs in their promoter regions (Tavazoie *et al.*, 1999). Clustering solutions can in some cases be assessed by applying standard statistical techniques. For high-dimensional data, multivariate analysis of variance (MANOVA) and discriminant analysis (Huberty, 1994; Mendez *et al.*, 2002) are appropriate if the data are normally distributed. For the case of non-normal data, there are several extensions that require the data to be either low-dimensional (Bishop *et al.*, 1975) or continuous (Katz and McSweeney, 1980). If attributes are independent one can also test the significance of the grouping for each dimension separately, and combine the resulting scores (Pesarin, 2001). None of these methods apply when wishing to test the significance of a clustering solution based on high-dimensional vectors of dependent biological attributes that do not necessarily follow a normal distribution and may even be discrete.

In this paper we devise a statistically based method for comparing clustering solutions according to prior biological knowledge. In our method, solutions are ranked according to their correspondence to prior knowledge about the clustered elements. Given a vector of (continuous or discrete) attributes for each element, our method tests the dependency between the attributes and the grouping of the elements. The test is applied simultaneously to all the attributes. In our application, elements are genes, clustered according to their expression patterns, and the attributes of a gene are binary indicators of its membership in specific functional classes. In this case, the method computes a quality score for the functional enrichment of these classes among each solution's clusters. At the heart of our method is a projection of the high-dimensional data to one dimension, to avoid the problem of applying MANOVA to the data. Using the one-dimensional data, the solutions are compared based on their score in a non-parametric ANOVA test.

In the rest of the paper, after providing some background, we describe our method, and give results on its performance on simulated and real data.

PRELIMINARIES

The input to a clustering problem consists of a set of elements and a characteristic vector for each element. A measure of (dis)similarity is defined between pairs of such vectors. (In gene expression, elements are usually genes, and the vector of each gene contains its expression levels under each of the monitored conditions. Dissimilarity between vectors can be measured, e.g. by their Euclidean distance.) The goal is to partition the elements into subsets, which are called *clusters*, so that two criteria are satisfied: homogeneity – elements in the same cluster are similar to each other; and separation – elements from different clusters are dissimilar.

Let N be a set of n elements and let $\mathcal{C} = \{C_1, \ldots, C_l\}$ be a partition of these elements into l clusters. We call two elements from the same cluster *mates* (with respect to $\mathcal{C}$). A common procedure for evaluating a clustering solution given the true solution, is to compute its *Jaccard coefficient* [see, e.g. Everitt (1993)], which is the proportion of correctly identified mates out of the sum of the correctly identified mates plus the total number of disagreements (pairs of elements that are mates in exactly one of the two solutions). Hence, a perfect solution has score 1, and the higher the score – the better the solution. When the true solution is not known, a solution can be evaluated by its homogeneity and separation. The *homogeneity* of $\mathcal{C}$ is the average distance between mates, and the *separation* of $\mathcal{C}$ is the average distance between non-mates (Hansen and Jaumard, 1997; Sharan *et al.*, 2003). Another popular measure is the average silhouette (Kaufman and Rousseeuw, 1990), which is computed as follows: define the *silhouette* of element j as $(b_j - a_j)/\max(a_j, b_j)$, where a_j is the average distance of element j from other elements of its cluster, b_{jk} is the average distance of element j from the members of cluster C_k, and $b_j = \min_{\{k: j \notin C_k\}} b_{jk}$. The *average silhouette* is the mean of this ratio over all elements.

Our main focus is the evaluation of clustering solutions using external information. The setup for the problem is as follows: we are given an $n \times p$ *attribute matrix* A. The rows of A correspond to elements, and the ith row vector is called the *attribute vector* of element i. We are also given a clustering $\mathcal{C} = \{C_1, \ldots, C_l\}$ of the elements, where $s_i = |C_i|$. For convenience, we shall also index the attribute vectors by the clustering, i.e. use $a_{ij} = (a_{ij}^1, \ldots, a_{ij}^p)$ as the vector of element j in cluster i. Typically $\mathcal{C}$ is obtained without using the information in A. Our goal is to evaluate $\mathcal{C}$ with respect to A.

When $p = 1$, there are established statistical tests for the problem. Such tests will serve as building blocks in our method. In the case that the attribute is normally distributed, and under the assumption that the variances of the l population distributions are identical, we can use standard

analysis of variance (ANOVA) methods to test the significance of the grouping [see, e.g. Sokal and Rohlf (1995)]: suppose that the attribute of element j in cluster i has value a_{ij}. Let $\bar{a}_i$ denote the mean of the elements in cluster i, and let $\bar{a}$ denote the total mean of all n elements. When ANOVA is carried out, the null hypothesis is that the groups do not differ in location, i.e. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_l$, where μ_i is the expectation of group i. The test statistic typically used is the ratio of variance estimator, i.e. the ratio of the hypothesis (or between-groups) mean square (MSH) to the error mean square (MSE):

$$F_H = \frac{\text{MSH}}{\text{MSE}} = \frac{\text{SSH}/(l-1)}{\text{SSE}/(n-l)} \tag{1}$$

where the hypothesis sum of squares is $\text{SSH} = \sum_{i=1}^{l} s_i(\bar{a}_i - \bar{a})^2$ and the error sum of squares is $\text{SSE} = \sum_{i=1}^{l}\sum_{j=1}^{s_i}(a_{ij} - \bar{a}_i)^2$. Under certain data conditions the F_{H} statistic has a (central) F distribution with $l-1$ and $n-l$ degrees of freedom.

In case the attribute (or some transformation of it) does not follow a normal distribution, one can use the Kruskal–Wallis (KW) test [cf. Sokal and Rohlf (1995)] as a non-parametric ANOVA test. The test assumes that the clusters are independent and have similar shape. We shall denote by $P^{\text{KW}}(\mathcal{C}, A)$ the p-value obtained by the KW test for a clustering $\mathcal{C}$ using the attribute $A : N \to R$. For the multidimensional case ($p > 1$), the MANOVA test [cf. Sokal and Rohlf (1995)] applies the same objective function F_{H}, but it applies only if the attribute matrix is multinormally distributed.

METHOD

Our goal is to evaluate a clustering solution given an attribute vector for each element, which represents the prior biological knowledge about the element. To this end, the MANOVA test is particularly appealing, as the numerator in Equation (1) (MSH) measures the separation (normalized by the number of clusters) and the denominator (MSE) measures the (normalized) homogeneity. However, the distribution of attribute vectors does not necessarily meet the requirements of MANOVA test. Such is the case, in particular, when attributes are binary. Thus, we propose to project the high-dimensional attribute vectors onto the real line using a linear combination of the attributes. Then, the solution $\mathcal{C}$ is scored by a non-parametric one-way ANOVA test on the one-dimensional data. We refer to the result as the *CQS* (Clustering Quality Score) of the clustering. CQS is computed as follows:

1. *Computing a linear combination of the attributes*. Each element is assigned a real value, which is a weighted sum of its attributes. An attribute's *weight* is its coefficient in the linear combination. Intuitively, we would like to weight the attributes such that they will contribute to the solution score according to their 'importance'. Usually, we do not know in advance the desired weighting of the attributes. In such cases, we propose to use weights that maximize the ability to discriminate between the clusters using the one-dimensional data. Finding the weights will be done in the same manner as in Linear Discriminant Analysis (LDA) (Huberty, 1994). The procedure for weight finding does not require any assumptions on the distribution of A. LDA creates such a linear combination by maximizing the ratio of between-groups-variance to within-groups-variance, as follows: let w be some p-dimensional vector of weights. The statistic being maximized is the ratio of MSH to MSE:

$$F(w) = \frac{\sum_{i=1}^{l} s_i(w \cdot \bar{a}_i - a \cdot \bar{a})^2/(l-1)}{\sum_{i=1}^{l}\sum_{j=1}^{s_i}(w \cdot a_{ij} - w \cdot \bar{a}_i)^2/(n-l)} \tag{2}$$

where $\bar{a}_i$ is the mean vector of cluster i, and $\bar{a}$ is the total mean vector. When introducing an additional constraint of a unit denominator, the maximum value of $F(w)$ is proportional to the greatest root of the equation $|H - \lambda E| = 0$. Here, H is a $p \times p$ matrix containing the between-groups sum of square $H_{rs} = \sum_{i=1}^{l} s_i(\bar{a}_i^r - \bar{a}^r)(\bar{a}_i^s - \bar{a}^s)$, and E is a $p \times p$ matrix of the sum of squared errors $E_{rs} = \sum_{i=1}^{l}\sum_{j=1}^{s_i}(a_{ij}^r - \bar{a}_i^r)(a_{ij}^s - \bar{a}_i^s)$, where $\bar{a}_i^r$ is the mean of attribute r in cluster i and $\bar{a}^r$ is the total mean of attribute r. Thus, the desired combination w is the eigenvector corresponding to the greatest root. This result holds without assuming any prior distribution on the attributes.

2. *Projection*. Apply the linear combination w to the attribute vectors, thereby projecting these vectors onto the real line. That is, $z_{ij} = \sum_t a_{ij}^t w^t$.

3. *Computing CQS using the projected values*. We now evaluate the clustering vis-á-vis the projected attributes using the KW test. We define CQS as $-\log p$, where $p = P^{\text{KW}}(\mathcal{C}, Z)$, i.e. the p-value assigned to the clustering by the KW test. Note that p is not the probability of observing the original attributes allocation randomly, since the vector data was first projected to maximize the variance ratio. Rather, the p-value is the probability that all values in this particular projection have been taken from the same population. Hence, CQS favors clustering solutions whose best discriminating weights enable significant grouping.

4. *Estimating confidence*. In order to estimate the accuracy of the scores and the significance of differences between the scores of distinct solutions, we evaluate the sensitivity of CQS to small modifications of the clustering solution. Intuitively, the larger the influence of small perturbations in the clustering on the CQS value, the smaller the confidence we have in the CQS. Specifically, for a given original solution we generate a group of alternative clustering solutions. Each alternative solution is obtained by introducing k exchanges of random pairs of elements from different clusters of the original solution (k is typically small, such as 2% of the elements). The *CQS confidence* is the standard deviation of CQS for the group of alternative clustering solutions.

The overall procedure is as follows:

1. Find the eigenvector w corresponding to the greatest root of the system of equations $|H - \lambda E| = 0$.
2. For each attribute vector a_{ij} set $z_{ij} = \sum_t a_{ij}^t w^t$.
3. Compute $p = P^{\mathrm{KW}}(\mathcal{C}, Z)$; let $\mathrm{CQS}(\mathcal{C}, A) = -\log p$.
4. Estimate the statistical confidence of the result by perturbations on $\mathcal{C}$.

Our scoring scheme can be applied in several ways and for several purposes. Our focus in this study is the evaluation of clustering solutions given external biological attributes, that were not used in the clustering process. Another application of our score is internal validation of solutions based on the same attributes that were used in generating the clustering. This can help in choosing among different clustering algorithms, as well as in optimizing the parameters of a specific algorithm (for example, choosing the number of clusters for K-means).

RESULTS

The score calculation was implemented in Perl under linux, using MATLAB. Running time for a data set of 750 clustered elements and 80 attributes is about a minute, on a standard 800 MHz PC. Below we report on the performance of our method on simulated and real data.

Simulations

We validate our method by conducting a series of tests on simulated data. We tested the effect of the one-dimensional projection of the attribute vector, the sensitivity of CQS to the solution accuracy, and the ability of CQS to pinpoint the right number of clusters and to detect fine clustering structures.

The data were generated as follows: profiles of 80 binary attributes were generated for five groups of $n = 50$ genes each. (We use the term 'genes' for uniformity. The simulations test the score irrespective of the nature of the clustered elements.) For each attribute we randomly selected one group in which its frequency will be r, and in the other four groups its frequency was set to r_0. The set of r (r_0) genes with that attribute was randomly selected from the relevant groups. $r_0 = 5$ was used throughout. Since we randomly select for each attribute the single group with frequency r, the overall density of the attribute vectors should be about the same for all elements, and the distinction must be based on individual attributes. Clearly, the larger the difference between r and r_0, the easier the distinction between the groups.

A. *The effect of one-dimensional projection*. First, we wished to examine the effect of reducing the attribute dimension to 1. We simulated data sets with $r = 6, 10, 15, 20$ and 25. For each data set we computed the ratio of separation to homogeneity of the true clustering on the original data (S/H) and on the projected data (S^*/H^*). This procedure was repeated 10 times. The results are shown in Figure 1B. Clearly, the

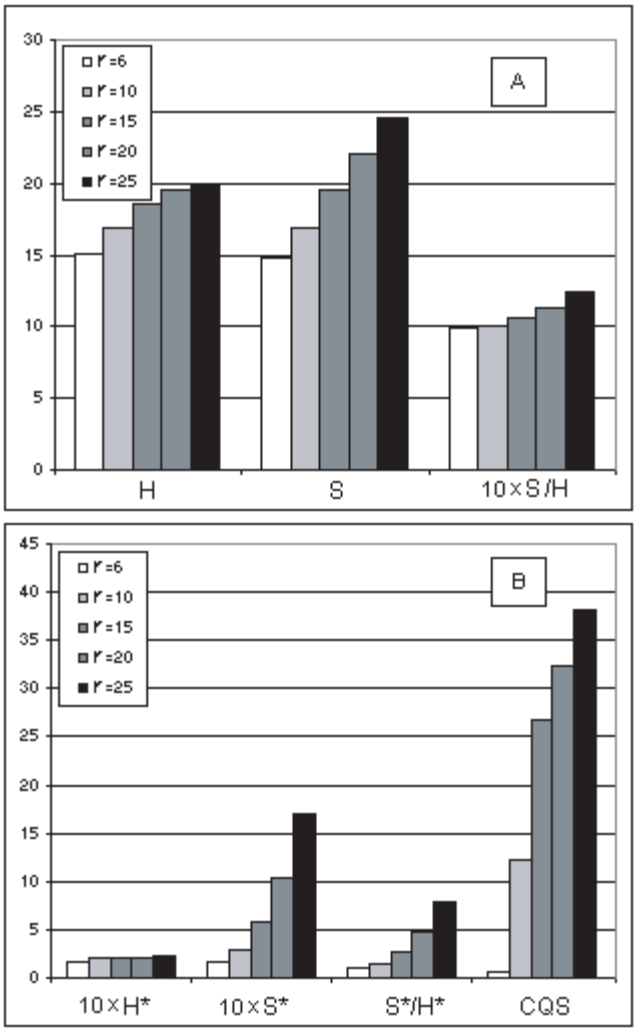


Fig. 1. Clustering parameters on simulated data. Y-axis: scores of five simulation setups $r = 6, 10, 15, 20, 25$ in different gray scale colors. (**A**) Scores are Homogeneity (H), separation (S) and their ratio on the original data. (**B**) Scores are Homogeneity (H^*), separation (S^*), their ratio and CQS on the projected (reduced) data. Numbers are average of 10 runs.

monotonicity of the homogeneity, separation and their ratio as a function of r, which is manifested on the original data, is preserved on the reduced data. The same monotonicity was observed in each of the 10 repetitions. Also, as expected, CQS improves monotonically with r.

The projected data for two simulations with $r = 6$ and $r = 25$ are visualized in Figure 2. For $r = 6$, the clusters look very similar, even though the data were reduced using the best separating linear combination. On the other hand, for $r = 25$, inter-cluster separation of most clusters is clearly visible.

B. *The effect of solution accuracy on CQS*. To test the sensitivity of CQS to the clustering solution, we simulated data with $r = 25$, and compared CQS of the true partition with that of

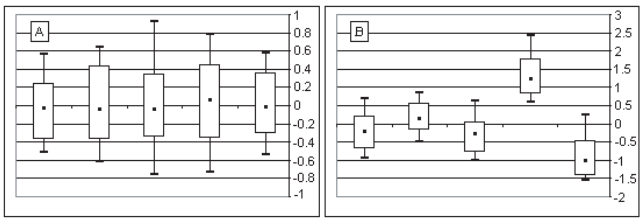


Fig. 2. Box plots for the projection of five simulated clusters with $r = 6$ (**A**) and $r = 25$ (**B**) after dimensionality reduction. The y-axis is the real-valued projection of the elements. Each box-plot depicts the median of the distribution (dot), 0.1 and 0.9 distribution quantiles (white box), and the maximum and minimum values.

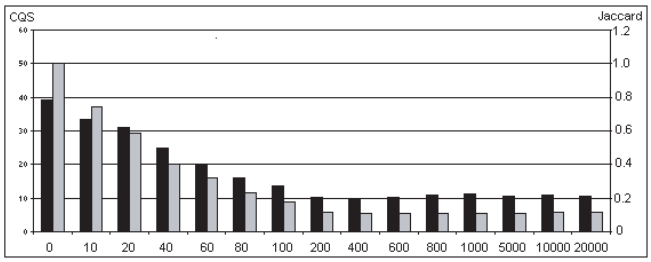


Fig. 3. Effect of the solution accuracy on CQS. The accuracy of different clustering solutions is measured by the number of inter-cluster exchanges introduced in the original solution. X-axis: number of exchanges. Y-axis: CQS (black bars, left scale) and Jaccard coefficient (gray bars, right scale).

other, similar and remote partitions. Those were produced by starting with the true solution and repeatedly exchanging a randomly chosen pair of elements from different clusters. As evident from the results in Figure 3, CQS is highest for the true partition and decreases with the number of exchanges applied (200 exchanges generate an essentially random partition, so further exchanges have no effect). We also computed for each intermediate solution its Jaccard score. As expected, the Jaccard coefficients of these solutions decrease with the number of exchanges.

C. *Sensitivity of CQS to the number of clusters*. Our next goal was to test the sensitivity of CQS with respect to the number of clusters. A robust score is essential for comparing solutions with different number of clusters. To this end we tested how CQS changes when splitting or merging clusters. For the splitting test we simulated data with $r = 25$. We compared the true 5-cluster solution with a 25-cluster solution obtained by randomly splitting each of the 5 clusters into 5 equal-size sub-clusters. This test was repeated 10 times. The parameters of the solutions before and after the splitting, averaged over 10 runs, are shown in Figure 4A. In all runs, as well as on the average, we observe a decrease of the clustering quality measures. The decrease of S/H is maintained (and even made more pronounced) in CQS and on the reduced data (S^*/H^*).

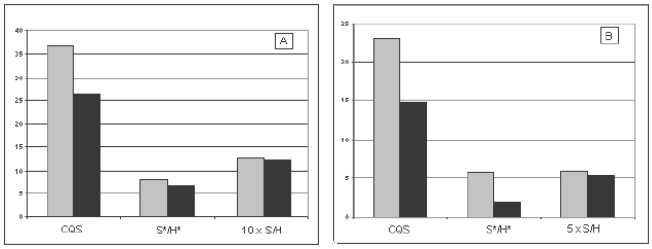


Fig. 4. Comparison between clustering solutions on simulated data after splitting clusters (**A**) or merging clusters (**B**). Each diagram shows CQS, S^*/H^* and S/H (y-axis) of the true solution (gray) and the modified solution (black). Numbers are average of 10 runs.

For the merging test, we simulated two 5-cluster data sets with $r = 25$ and $r = 6$ as above, using $n = 25$. We then combined these data sets into a single data set whose true solution consists of 10 equal size clusters with 25 genes each. We next merged pairs of clusters, one from each original data set, to form in total five clusters with 50 genes each. These five clusters comprised the alternative (merged) solution. Figure 4B shows the parameters of the resulting partition before and after the merging, averaged over 10 runs. As in the splitting test, all the measures decrease due to the merging, and this is observed in all runs, as well as on the average. The decrease of S/H is maintained and enlarged in S^*/H^* and CQS.

Next, we tested the agreement of CQS with Jaccard coefficient: we simulated 5-cluster data with $r = 25$ and applied K-means (MacQueen, 1965; Ball and Hall, 1967) to the data, with $K = 2, \ldots, 15$. Since K-means seeks a clustering solution with K clusters, we expect the solution's quality to decline as the difference $|K-5|$ increases. A good score should manifest such trend. We computed CQS and Jaccard coefficient for each clustering solution, as well as S/H. The results are shown in Figure 5. CQS behaves as the Jaccard coefficient and S/H, with a maximum at $K = 5$, the true number of clusters. Moreover, the ranking of all 14 solutions according to the Jaccard score (which is based on the true solution) and according to CQS (which is based on the attributes only) are virtually identical. The ratio score also does quite well, with a maximum at $K = 5$. However, the ranking of solutions by this score does not agree with the Jaccard score.

D. *CQS ability to detect fine clustering structures*. Our next goal was to test the ability of CQS to identify fine structures in the data. Profiles of 30 binary attributes were generated for four clusters of $n = 50$ genes each. For each attribute, its frequencies in clusters 1, 2, 3 and 4 were set to 2, b, $50-b$ and 48, respectively. We simulated data sets with $b = 3, 5, 10, 15, 20$. For each data set, we scored two clustering solutions: the original 4-cluster solution, and a 2-cluster solution obtained by merging cluster 1 with 2 and merging cluster 3 with 4. Thus, for large values of b we expect the 4-cluster solution to

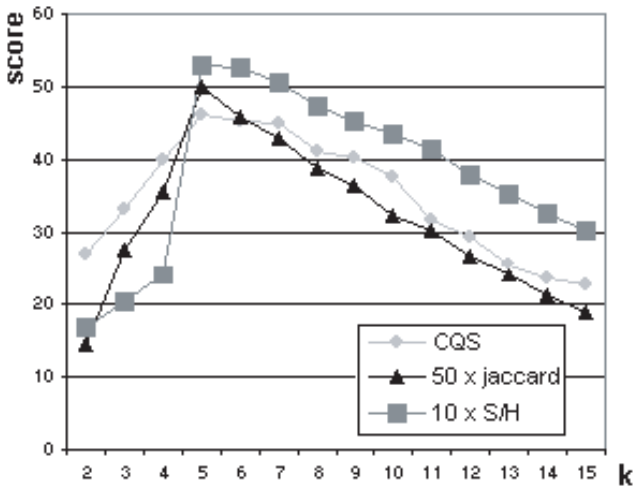


Fig. 5. Comparison of quality measures on solutions of various accuracies. Scores are plotted for different K-means' solutions. X-axis: K-means' solutions with $K = 2, \ldots, 15$. Y-axis: CQS (light gray), Jaccard (black) and S/H (gray) scores. The true number of clusters is 5.

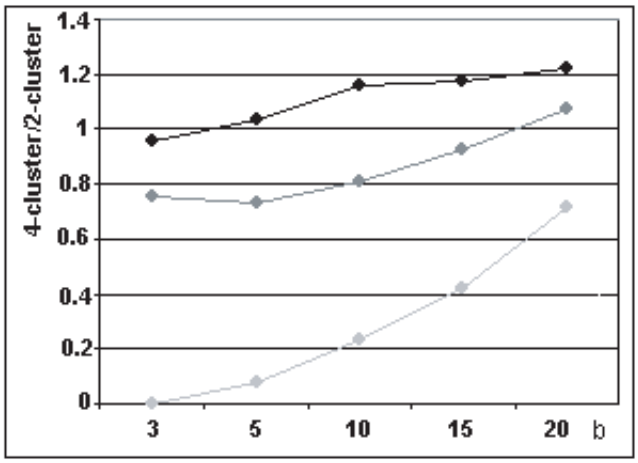


Fig. 6. Ability of the different scores to distinguish similar clusters. We simulated 4-cluster data with attribute frequencies $2, b, 50-b, 48$, and used different values for b. We obtained a 2-cluster solution by merging cluster 1 with 2 and 3 with 4. X axis: value of b in the simulation. Y-axis: the ratio of the scores for the 4-cluster and 2-cluster solutions. The scores are silhouette (gray), S/H (dark gray) and CQS (black).

score higher than the 2-cluster solution. Note that unlike the previous simulations, where distributions of individual attributes were designed to differ between clusters, here it is only the overall attribute density which is directly controlled. This design is the binary equivalent to the Gaussian clusters with different means that appears, e.g. in Pollard and van der Laan (2002). For each data set and each of the two solutions, we computed S/H, CQS and the average silhouette score.

The ratios of the 4-cluster to 2-cluster scores, averaged over 10 runs, are presented in Figure 6. As expected, the ratios are increasing with b in all scores. The silhouette for the 2-cluster solution is always greater than for the corresponding 4-cluster solution. Similarly, for $b = 3, 5, 10, 15$, S/H is greater for the 2-cluster solution. In all those cases, the scores would prefer the incorrect, 2-cluster solution. In contrast, CQS is able to identify the fine structure in the data: for all b values except $b = 3$, CQS rates the 4-cluster solution above the 2-cluster solution, as desired. For $b = 3$, the 2-cluster CQS is higher than the 4-cluster CQS, since there is almost no difference between the clusters with 2 or 3 occurrences of attributes, and between the clusters with 47 or 48 occurrences.

Yeast cell-cycle data

We also tested our approach on clustering solutions computed on the yeast cell-cycle data set of Spellman *et al.* (1998). The data set contains 72 expression profiles from yeast cultures synchronized by four independent methods: α factor arrest, arrest of a cdc15 temperature sensitive mutant, arrest of a cdc28 temperature sensitive mutant and elutriation. [As in Tamayo *et al.* (1999), an additional 90 min data point in the cdc15 experiment was not used.] Spellman *et al.* (1998) identified in these data 800 genes that are cell-cycle regulated. We used the expression levels of 698 out of those 800 genes, which have up to three missing entries, over the 72 conditions. The missing entries in each gene were completed with the average of its present entries. Each row of the 698×72 matrix was normalized to have mean 0 and variance 1.

Based on the analysis conducted by Spellman *et al.* (1998), we expect to find in the data five main clusters, each one corresponding to genes peaking in one of the cell cycle phases (G1, S, G2, M and M/G1). The 698×72 data set was clustered using four clustering methods: K-means (MacQueen, 1965; Ball and Hall, 1967), SOM (Kohonen, 1997; Tamayo *et al.*, 1999), CAST (Ben-Dor *et al.*, 1999) and CLICK (Sharan and Shamir, 2000; Sharan *et al.*, 2003). The solutions of K-means, SOM and CLICK were obtained using the EXPANDER software (Sharan *et al.*, 2003). CAST's solution was produced by the authors of the software and is the same as reported in (Shamir and Sharan, 2002). The K-means algorithm was executed with $K = 5$. The SOM algorithm was executed on a 2×3 grid and produced six clusters. The CAST solution has five clusters. CLICK was executed with default parameters and generated a solution with six clusters and 23 singletons. Each singleton was subsequently assigned to its closest cluster in order to produce a solution with no singletons. The similarity measure used in all cases was Pearson correlation coefficient. Another solution that we included in the analysis is the one reported in Spellman *et al.* (1998), which was generated by manually dividing the genes into five groups using their peak of expression, in order

to approximate the five cell-cycle phases. We shall refer to it as the 'true' solution.

To evaluate the five solutions, we used as gene attributes the GO classes (The Gene Ontology Consortium, 2000) at level 5 of the ontology, including process, function and component attributes. In addition, we used the MIPS annotation (Mewes *et al.*, 2002) at level 4. We removed attributes indicating that the functional class of the gene is still unknown and used only attributes that occur in at least four of the genes. Overall we used 51 GO process attributes, 37 GO function attributes, 27 GO component attributes and 59 MIPS attributes. We applied the analysis to 370 genes that had at least one attribute. CQS was computed three times, using the GO process attributes only, all GO attributes, and the MIPS attributes only. The results are depicted in Figure 7. For comparison purpose, we also scored a random clustering of the data into five equal-size clusters.

The random solution consistently obtained the lowest scores in all annotation categories. Using the process GO annotation (Fig. 7A), the CLICK, CAST and SOM solutions achieved the highest scores. Notably, they are scored higher than the 'true', K-mean and random solutions. When using all GO annotations (Fig. 7B), a similar pattern of scores is observed. Qualitatively, we got the same results when using GO annotations at level 4 of the hierarchy (data not shown). When evaluating all solutions based on MIPS level 4 annotations (Fig. 7C), CAST achieved the highest score. This exemplifies the fact that different biological attributes lead to different evaluations of clustering solutions.

In a different test, we ran SOM with $2, 3, \ldots, 8$ clusters on the same data set and calculated CQS of each solution. Clear best results were obtained for 5 and 6 clusters, as expected, (28 ± 1, 29 ± 1 respectively, with all other cluster solutions scored below 23).

Next, we present an analysis of CQS for the CLICK solution using all 115 GO attributes. Figure 8A is a scatter plot of weight versus enrichment for each attribute, using this solution. The *enrichment* of a k-cluster solution for a given attribute, is defined by $-\log p$ where p is the p-value of the G-test of independence (Sokal and Rohlf, 1995) with a $2 \times k$ table. It tests independence between element attributes and the partition into clusters. Note that the G-test enables us to evaluate functional enrichment for more than a single cluster. The frequently used hyper-geometric Fisher exact test of independence (Sokal and Rohlf, 1995) tests functional enrichment of a single cluster only.

As expected, the highest ranking attributes (both using the weights and the enrichment) are related to cell cycle. Notably, there is a correlation between the enrichment of an attribute and the absolute value of its assigned weight (Fig. 8B). This correlation is expected, since more enriched attributes can contribute more to our ability to discriminate between the clusters and, thus, they are expected to have higher weights. However, we do not expect a perfect correlation between the

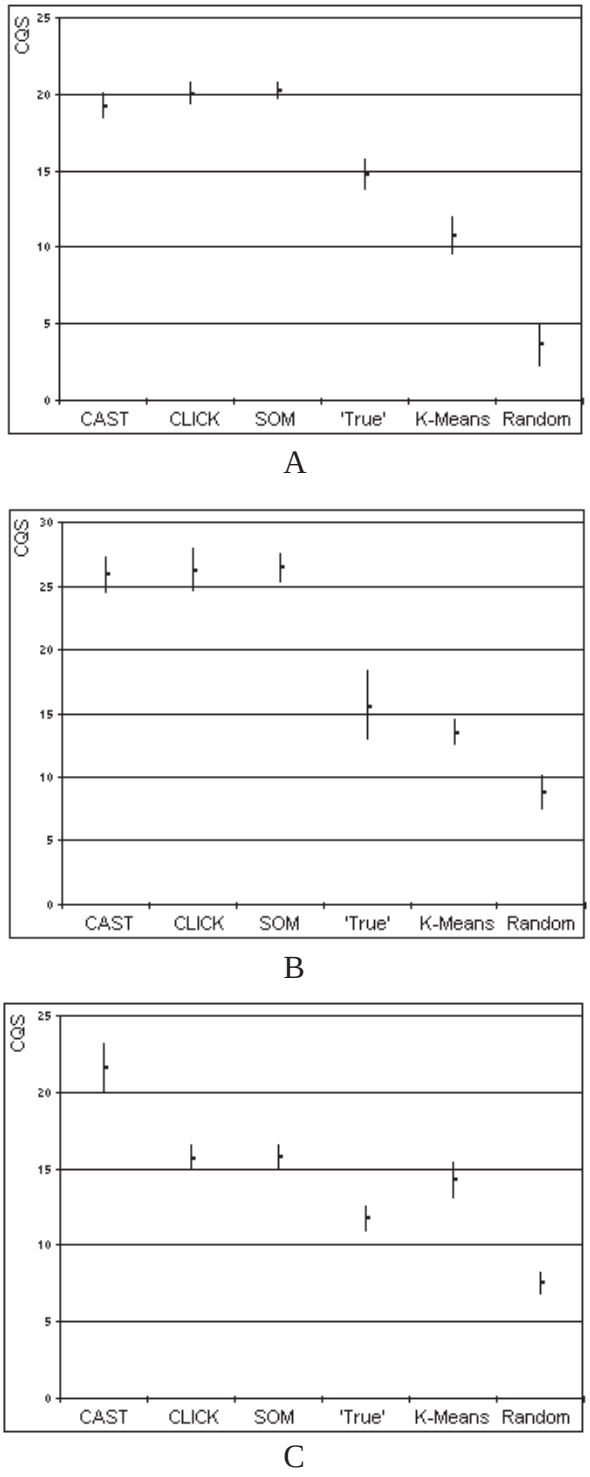
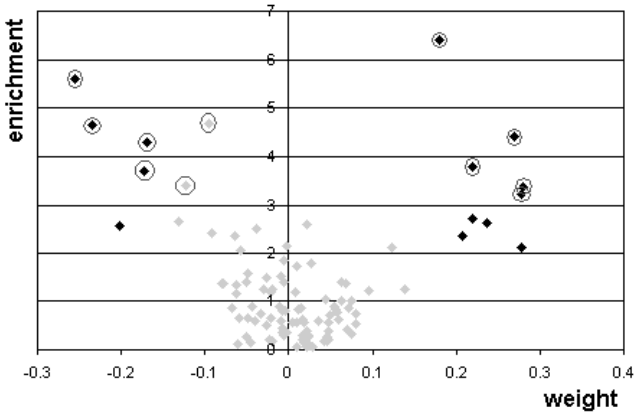


Fig. 7. CQS of six clustering solutions for the yeast cell cycle data of Spellman *et al.* (1998). CQS is computed using GO level 5 process attributes only (**A**), all GO level 5 attributes (**B**) and MIPS level 4 attributes (**C**). Y-axis: CQS. X-axis: clustering solutions. CQS for each clustering solution is presented along with its confidence, by computing the standard deviation of 10 other solutions achieved by seven random pair exchanges in the original solution.



A

weight	enrichment	Go Name	GO number
0.18	6.40	DNA metabolism	GO:0006259
-0.25	5.61	DNA replication	GO:0006260
-0.10	4.70	sensory perception	GO:0007600
-0.24	4.67	nucleoplasm	GO:0005654
0.27	4.41	amino acid metabolism	GO:0006520
-0.17	4.31	ATP dependent DNA helicase	GO:0004003
0.22	3.79	chromatin	GO:0005717
-0.17	3.70	chromatin binding	GO:0003682
-0.12	3.41	hexose transporter	GO:0015149
0.28	3.38	microtubule organizing center	GO:0005815
0.28	3.24	spindle	GO:0005819
0.22	2.69	DNA binding	GO:0003677
-0.13	2.63	cell wall	GO:0005618
0.24	2.62	chromosome organization	GO:0007001
0.02	2.58	nucleotidyltransferase	GO:0016779
-0.20	2.55	cytoplasm	GO:0005737
-0.04	2.51	transport	GO:0006810
-0.09	2.41	monosaccharide transport	GO:0015749
0.21	2.35	structural constituent of cytoskeleton	GO:0005200
-0.06	2.33	zygote formation (sensu Fungi)	GO:0030462
0.00	2.12	endoplasmic reticulum	GO:0005783
0.28	2.11	organelle organization and biogenesis	GO:0006996

B

Category/Cluster	1	2	3	4	5	6
DNA Metabolism (i)	23	11	2	1	1	0
DNA Replication (ii)	17	1	3	2	0	6
Chromosome organization (iii)	4	10	0	1	1	0
(i)+(ii)	10	0	1	1	0	0
(i)+(iii)	4	9	0	0	0	0
(ii)+(iii)	0	0	0	0	0	0
Total genes in clusters	101	76	45	94	32	22

C

Fig. 8. Attribute weights and enrichment values in the CLICK solution to the cell cycle data of Spellman *et al.* (1998), using all GO attributes. (**A**) A scatter plot of enrichment (y-axis) versus weight (x-axis), for each GO attribute. Attributes with high absolute weights (>0.15) are marked in black. Attributes with high enrichment (>3) are circled. (**B**) The 22 most enriched attributes. High attribute values in enrichment (>3) or weight (>0.15) are highlighted. Note that the 14 top weighted attributes are contained in the 22 most enriched attributes. (**C**) The distribution and co-occurrence of the attributes 'DNA metabolism', 'DNA replication' and 'Chromosome organization and biogenesis' in the six clusters of the CLICK solution.

two measures, since the goals of the attributes weighting and the enrichment measure are different and, more importantly, because the G-test takes into consideration each attribute separately, while the weights are computed by considering all attributes together and, thus, they reflect relations between attributes. For example, consider the 'DNA metabolism' attribute, which deviates significantly from the correlation (Fig. 8C). The enrichment of 'DNA metabolism' in clusters 1 and 2 overlaps to a large extent with that of 'DNA replication' and 'Chromosome organization and biogenesis', and this is partially reflected in their weights. Therefore, the weight of 'DNA metabolism' is lower than expected.

DISCUSSION

Clustering is a central tool in gene expression analysis. Different clustering methods usually produce different solutions, of which one has to pick one or few preferred solutions. We propose here a method called CQS for evaluating a clustering solution based on its biological relevance. Our method can be applied to compare the functional enrichment of many biological attributes simultaneously in different clustering solutions. In addition, it may be applied to optimize the parameters of a clustering algorithm (e.g. to determine the number of clusters). The method is based on using attributes of the clustered elements, which are available independently from the data used to generate the clusters.

We empirically validated CQS using a variety of simulations. Our scoring method was shown to outperform previous numeric methods for clustering evaluation, including the separation to homogeneity ratio and the average silhouette measure. We also applied CQS to compare between different clustering solutions of the cell cycle data set of Spellman *et al.* (1998) using binary attributes from the GO and MIPS annotation databases.

According to our results, CQS is sensitive to small modification of the clustering solution and to changes in the simulation setting. In order to evaluate the significance of the difference in CQS between clustering solutions, we use a CQS confidence measure. For example, the CAST, CLICK and SOM solutions in Figure 7A and B, cannot be meaningfully ranked by their scores. We may only conclude that CAST, CLICK and SOM have higher scores than the 'True', Random and K-means solutions. According to the results, although the 'True' solution was hand crafted in order to approximate the cell cycle phases, the solutions produced by CAST, CLICK and SOM are more aligned with the biological attributes. We note that these results should be treated with caution since the database annotations are incomplete and may be biased.

The attribute weights were computed using information about all the attributes together, without assuming that the attributes are independent. Frequently, the functional enrichment of each attribute in each cluster, is computed separately

[e.g. Tavazoie *et al.* (1999)]. In such cases, since the attributes might be dependent (as we exemplify in Fig. 8B), the real fraction of functionally enriched attributes might be over estimated.

CQS can be applied to a wide range of other attribute types. For example, one can use continuous attributes corresponding to sequence motifs, that represent the likelihood of having that motif. CQS has the advantage that it can use such continuous data without any assumption on the data distribution.

ACKNOWLEDGEMENTS

This study was supported in part by a research grant from the Ministry of Science and Technology, Israel. I.G.-V. was supported by the Colton Foundation. R. Sharan was supported by a Fulbright grant.

REFERENCES

Ball,G. and Hall,D. (1967) A clustering technique for summarizing multivariate data. *Behav. Sci.*, **12**, 153–155.

Ben-Dor,A., Shamir,R. and Yakhini,Z. (1999) Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–297.

Ben-Hur,A., Elisseeff,A. and Guyon,I. (2002) A stability based method for discovering structure in clustered data. *Pac. Symp. Biocomput.*, 6–17.

Bishop,Y., Fienberg,S. and Holland,P. (1975) *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

Dudoit,S. and Fridlyand,J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.*, **3**, 0036.1–0036.21.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *PNAS*, **95**, 14863–14868.

Everitt,B. (1993) *Cluster Analysis*. 3rd edn. Edward Arnold, London.

Hansen,P. and Jaumard,B. (1997) Cluster analysis and mathematical programming. *Math. Program.*, **79**, 191–215.

Hartigan,J. (1975) *Clustering Algorithms*. Wiley, New York.

Huberty,C. (1994) *Applied Discriminant Analysis*. Wiley, New York.

Katz,B. and McSweeney,M. (1980) A multivariate Kruskal–Wallis test with post hoc procedures. *Multivariate Behavioral Res.*, **15**, 281–297.

Kaufman,L. and Rousseeuw,P. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley New York.

Kohonen,T. (1997) *Self-Organizing Maps*. Springer, Berlin.

MacQueen,J. (1965) Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. pp. 281–297.

McLachlan,G.J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Stat.*, **36**, 318–324.

McLachlan,G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.

Mendez,M., Hoedar,C., Vulpe,C., Gonzales,M. and Cambiazo,V. (2002) Discriminant analysis to evaluate clustering of gene expression data. *FEBS Lett.*, **522**, 24–28.

Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acid Res.*, **30**, 31–4.

Pesarin,F. (2001) *Multivariate Permutation Tests*. Wiley, New York.

Pollard,K. and van der Laan,M. (2002) A method to identify significant clusters in gene expression data. In *Sixth World Multiconference on Systemics, Cybernetics, and Informatics*, to appear.

Shamir,R. and Sharan,R. (2002) Algorithmic approaches to clustering gene expression data. In Jiang,T., Smith,T., Xu,Y. and Zhang,M. (eds.), *Current Topics in Computational Biology*. MIT Press, Cambridge, MA, pp. 269–299.

Sharan,R., Elkon,R. and Shamir,R. (2002) Cluster analysis and its applications to gene expression data. In Mewes,H.-W., Seidel,H. and Weiss,B. (eds.), *Bioinformatics and Genome Analysis*. Springer, Berlin, pp. 83–108.

Sharan,R., Maron-Katz,A. and Shamir,R. (2003) Click and expander: a system for clustering and visualizing gene expression data. *Bioinformatics*, in press.

Sharan,R. and Shamir,R. (2000) CLICK: a clustering algorithm with applications to gene expression analysis. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*. pp. 307–316.

Sokal,R. and Rohlf,F. (1995) *Biometry*. Freeman, San Fransisco.

Sokal,R.R. (1977) Clustering and classification: background and current directions. In Van Ryzin,J. (ed.), *Classification and Clustering*. Academic Press, London, pp. 1–15.

Spellman,P.T., Sherlock,G., Zhang,H.Q., Iyer,V.R., Andres,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stephanopoulos,G., Hwang,D., Schmitt,W., Misra,J. and Stephanopoulos,G. (2002) Mapping physiological states from microarray expression measurments. *Bioinformatics*, **18**, 1054–1063.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Gene.*, **22**, 281–285.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Gene.*, **25**, 25–29.

Tibshirani,R., Walther,G. and Hastie,T. (2000) Estimating the number of clusters in a dataset via the gap statistics. Technical report, Stanford University, Stanford.

Yeung,K., Haynor,D. and Ruzzo,W. (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.