BIOINFORMATICS

Vol. 18 no. 11 2002
Pages 1462–1469

Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data

Lisa M. McShane[1,*], Michael D. Radmacher[1], Boris Freidlin[1], Ren Yu[2,†], Ming-Chung Li[2] and Richard Simon[1]

[1]National Cancer Institute, Biometric Research Branch, DCTD, NIH, Bethesda, MD 20892-7434 and [2]The Emmes Corporation, Rockville, MD 20850, USA

Received on June 29, 2001; revised on February 6, 2002; accepted on April 17, 2002

ABSTRACT

Motivation: Recent technological advances such as cDNA microarray technology have made it possible to simultaneously interrogate thousands of genes in a biological specimen. A cDNA microarray experiment produces a gene expression 'profile'. Often interest lies in discovering novel subgroupings, or 'clusters', of specimens based on their profiles, for example identification of new tumor taxonomies. Cluster analysis techniques such as hierarchical clustering and self-organizing maps have frequently been used for investigating structure in microarray data. However, clustering algorithms always detect clusters, even on random data, and it is easy to misinterpret the results without some objective measure of the reproducibility of the clusters.

Results: We present statistical methods for testing for overall clustering of gene expression profiles, and we define easily interpretable measures of cluster-specific reproducibility that facilitate understanding of the clustering structure. We apply these methods to elucidate structure in cDNA microarray gene expression profiles obtained on melanoma tumors and on prostate specimens.

Availability: Software to implement these methods is contained in *BRB ArrayTools* microarray analysis package available from http://linus.nci.nih.gov./BRB-ArrayTools.html.

Contact: lm5h@nih.gov

INTRODUCTION

The cDNA microarray technology allows one to measure, for thousands of genes, the relative abundance of each gene's mRNA in a test sample compared to its abundance in a reference sample using a two-color fluorescent probe hybridization system (Schena *et al.*, 1995). The gene expression 'profiles' generated by this technique can be analyzed by clustering methods to try to identify novel subgroupings, or 'clusters', of specimens. Microarray analyses have already been useful in identifying tumor taxonomies (Khan *et al.*, 1998; Alizadeh *et al.*, 2000; Bittner *et al.*, 2000). Frequently used cluster analysis techniques include hierarchical clustering (Eisen *et al.*, 1998) and self-organizing maps (Tamayo *et al.*, 1999). However, clustering algorithms always detect clusters, even on random data, and it is imperative to conduct some statistical assessment of the strength of evidence for *any* clustering and to examine the reproducibility of individual clusters. Here we present statistical methods for testing for the existence of meaningful clustering, and we describe some easily interpretable cluster-specific reproducibility measures that we have developed and found useful for elucidating and clustering structure.

We demonstrate the methods by applying them to two different gene expression profile data sets. The first data set consists of gene expression profiles for 31 melanoma tumors (Bittner *et al.*, 2000), and the second data set consists of profiles obtained from 25 prostate specimens (Luo *et al.*, 2001). The gene expression profile obtained for a specimen consists of log transformed normalized expression ratios measured on the full set of genes represented on the microarray. For a given spot (e.g. gene) on an array, the expression ratio is formed by dividing the fluorescent signal measured for the test sample at that spot by the fluorescent signal measured from the reference sample. The test samples are fluorescently tagged cDNA samples derived from mRNA isolated from the tumors or other specimens of interest. In the examples we will consider, the reference sample is derived from a common pool of mRNA and tagged with a fluorescent dye, different from the dye used for the test samples. The reference sample used in the melanoma example was a pool of RNA from a non-tumorigenic revertant of a tumorigenic melanoma (Trent *et al.*, 1990). The reference sample used in the prostate example was composed of a pool of RNA

*To whom correspondence should be addressed.
† Present address: Human Genome Sciences Inc., Rockville, MD 20850, USA

from two benign prostatic hyperplasia samples. In general, the reference pool may be mRNA derived from normal tissue or a mixture of mRNA derived from a collection of tumor cell lines.

Our focus is on hierarchical agglomerative clustering methods, although the same general principles could be applied to any of the numerous clustering methods available (Gordon, 1999; Jain and Dubes, 1988). In brief, a distance metric is defined between the profiles of each pair of specimens to be clustered. The hierarchical agglomerative algorithm proceeds by merging the two closest (most similar) specimens first, and then successively merging specimens or groups of specimens in order of greatest similarity. Two distance metrics commonly used in clustering gene expression profiles are Euclidean distance and one minus the Pearson correlation coefficient. Euclidean distance measures in absolute terms the closeness of two profiles, whereas correlation measures the similarity of patterns in the sense of how closely the values in one profile can be approximated by a linear function (scalar multiple or shift) of the values in the other profile. For example, if the expression ratio measurements for all genes for one tumor were exactly 3 times their counterpart ratios for another tumor, those two tumors would be considered distant using a Euclidean distance metric but close using the distance metric of one minus the correlation. See Gordon (1999, Chapter 2) for discussion of additional distance metrics.

The end result of a hierarchical clustering is a tree structure depicted by a dendrogram. An example dendrogram is presented in Figure 1a. The dendrogram in Figure 1a resulted from hierarchical cluster analysis, using the distance metric of one minus the Pearson correlation coefficient, applied to log expression ratios obtained from microarray experiments performed on 31 melanoma tumors (Bittner *et al.*, 2000). There were 3799 genes with measurements meeting the quality criteria used in these analyses. At the bottom of the tree, each of the original specimens constitutes its own cluster and, at the top of the tree, all specimens have been merged into a single cluster. The tree is 'rooted' at the top. Mergers between two specimens, or between two clusters of specimens, are represented by horizontal lines connecting them in the dendrogram. The height of each horizontal line represents the distance between the two groups it merges. See Gordon (1999, pp. 69–72) for a discussion of alternative dendrogram formats.

It is not obvious by looking at the dendrogram in Figure 1a what are the most meaningful clusters. Stopping the agglomerative process too early will result in a large number of small clusters. Allowing agglomeration to continue too long will result in fewer, larger clusters, potentially obscuring important structure or subgroups. The decision about where to stop the process is equivalent to where

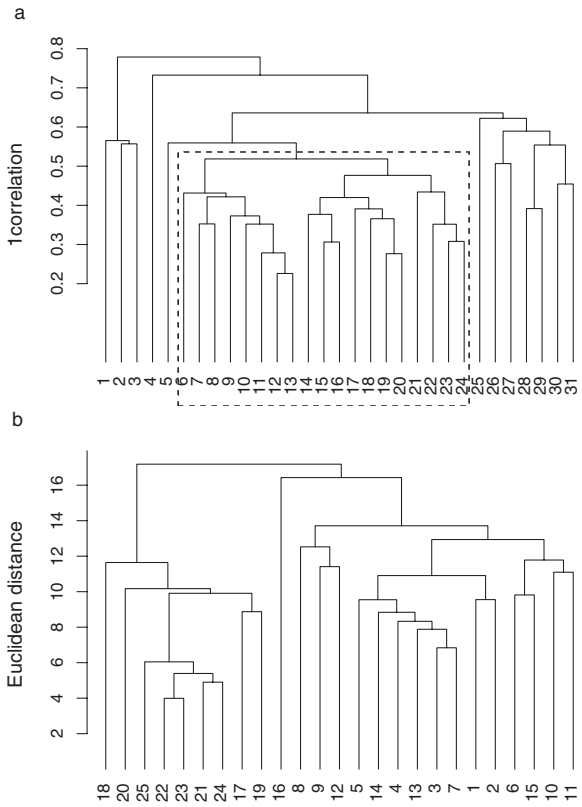

Fig. 1. (a) Dendrogram resulting from hierarchical agglomerative cluster analysis using average linkage and distance metric equal to one minus the Pearson correlation applied to melanoma data. Dashed box outlines the 19 tumor cluster of interest. (b) Dendrogram resulting from hierarchical agglomerative cluster analysis using complete linkage and Euclidean distance applied to prostate data.

to 'cut' the dendrogram. In viewing this dendrogram or the results from any other clustering technique applied to any other data, one must ask whether any of the observed clusters are believable, and if so, which ones.

As a first step in the investigation of the clustering structure, we recommend that a global test of clustering be performed to determine the strength of evidence for any clustering. It is expected *a priori* that expression profiles for genes would cluster because, for example, there are classes of genes known to be co-regulated. In contrast, clustering of an arbitrarily selected set of specimen profiles, is not necessarily expected. The finding of clusters of expression profiles for specimens of morphologically and histologically similar tumors is a potentially important observation, but one which requires statistical verification. If the significance tests provide substantial evidence for clustering, we employ two cluster reproducibility measures that can aid in assessing the meaningfulness of individual

clusters of interest. The first measure we refer to as the robustness (R) index, and the second measure we call the discrepancy (D) index. The fundamental idea behind both of these methods is to identify clusters likely to be preserved if new data were collected on the same specimens. In effect, we want to assess the stability of the observed clusters in the background of experimental noise. Details of the methods are described in the next section.

METHODS

Global statistical test of clustering

We test whether the gene expression profiles are consistent with having arisen from a single multivariate Gaussian distribution, i.e. that there is no meaningful clustering. The test we propose is based on examination of the Euclidean distances between specimens in principal components space. If the distance metric of interest is one minus the Pearson correlation, the test should be applied to the standardized expression profiles. Converting the problem to one involving Euclidean distances allows us to use methods based on inter-event distances that have been derived and studied in the context of Euclidean distance (Diggle, 1983, Section 2.2). This conversion is justified because the one minus Pearson correlation distance metric is proportional to the square of the Euclidean distance metric computed on the standardized expression profiles. The standardization is applied by mean centering and standard deviation scaling the expression levels of each specimen prior to applying the principal components transformation.

The test can be described as follows. The log ratio or standardized log ratio profile data are first transformed to the principal components space (Johnson and Wichern, 1988) to simplify subsequent calculations by adjusting for the effects of correlations among genes. The maximum number of principal components that can be calculated is the smaller of the number of genes and one less than the number of specimens. Typically in microarray studies the number of genes assessed far outnumbers the number of specimens analyzed, so this maximum would be one less than the number of specimens. We base our test on only the first three principal components. This is to avoid data sparseness in the high dimensional space that would lead to instability in the properties of the test. We found through a variety of simulation studies that using three principal components led to tests with good properties. We expect that many important clustering patterns could be detected with examination of only a few principal components, and this tends to makes the results of the test consistent with the three-dimensional principal components visualization commonly used for microarray data. In a related context, Silverman (1986, pp. 93–94) has recommended that one could reasonably perform non-parametric density estimation with a few dozen to several dozen observations of two- or three-dimensional data. However, a few hundred to several hundred observations would be required for adequate density estimation for four- or five-dimensional data. Therefore, detecting statistically significant clustering in high dimensional data. Therefore, detecting statistically significant clustering in high dimensions would likely require profiling greater numbers (many hundreds or even thousands) of specimens than is common.

If the profile data have an approximate Gaussian distribution, then the principal components will also have approximate Gaussian distributions. The global test we consider compares the distribution of nearest neighbor distances for the observed data in the space formed by the first three principal components to the distribution of nearest neighbor distances simulated under a Gaussian distribution (corresponding to the null hypothesis) in that space. We compute the mean and standard deviation for each of the coordinates in the three-dimensional principal components space. We generate each coordinate of the simulated data as normally distributed with mean and standard deviation as estimated from the observed principal components. The ability to generate the points in the principal components space one coordinate at a time is a by-product of the orthogonality of the principal components. A collection of many Gaussian data sets is generated in this way.

To quantify the clustering pattern in the real and simulated data sets, we examine the distribution of 'nearest neighbor' (NN) distances. For each specimen represented as a point in the three-dimensional principal components space, we compute the Euclidean distance from that specimen to the nearest other specimen. We then compute the empirical distribution function (EDF) of these distances. For any distance d, the EDF is the proportion of NN distances that are less than or equal to d. We compare the nearest neighbour empirical distribution function (NN EDF) for the observed data to that expected under the null hypothesis, and we quantify the difference by a squared difference discrepancy measure (Diggle, 1983, Equation 2.3.2). Specifically, let $\hat{G}_1(y)$ be the NN EDF computed from the observed data, and let $\hat{G}_i(y)$: $i = 2, \ldots, s$ be the NN EDF's computed from the s-1 data sets simulated under the null distribution (distribution corresponding to the null hypothesis). Calculate $\overline{G}_i(y) = (s-1)^{-1} \sum_{j \neq i} \hat{G}_j(y)$ to serve as an estimate of the expected NN EDF under the null distribution that is independent of the ith simulated EDF. The squared difference discrepancy measure $u_i = \int \{\hat{G}_i(y) - \overline{G}_i(y)\}^2 \, \mathrm{d}y$ is a measure of how different the ith NN EDF is from that expected under the null distribution. We compute this integral numerically by evaluating the integrand at 30 equally spaced points along the range of the nearest-

neighbor distances and computing the Riemann sum. If u_1 is unusually large compared to the distribution of the $s - 1$ u_i values that were generated under the null distribution, then there is evidence that the observed data were generated under a distribution different than that null distribution. The Monte Carlo p-value for a test of clustering (relative to the expected NN distance pattern under the null distribution) is obtained as the proportion of the s u_i values that are at least as big as that computed from the observed data. Typically, we take $s = 10\,000$ in order to obtain a high degree of accuracy on the Monte Carlo p-value. For example, if 3% of the s u_i values are at least as large as u_1 (which is calculated from the observed data), the calculated Monte Carlo p-value would be 0.03. The Monte Carlo p-value is interpreted as any other p-value, so small values such as values less that 0.05 are commonly called significant.

Calculation of the *R*-index and *D*-index

For assessing cluster-specific reproducibility, we utilize the general approach of data perturbation to assess clustering stability, a technique that has been used by others in different settings (Rand, 1971; Gnanadesikan *et al.*, 1977; Fowlkes and Mallows, 1983). We simulate 'new data' by adding artificial experimental error in the form of Gaussian white noise to the existing log ratio measurements. Error distributions other than Gaussian could be used, but the Gaussian error assumption is a useful approximation in many settings. Wolfinger *et al.* (2001) have reported that they have found Gaussian assumptions to be reasonable for several data sets they examined. In practice, one can check assumptions of Gaussian error on replicate data sets using statistical tests for normality, assessing skewness and kurtosis, and examining graphical displays such as normal quantile–quantile plots. An appropriate variance to use in generating this Gaussian experimental error can be estimated from the data. Our estimate is based on an assumption that a majority of the genes are not truly differentially expressed across tumors. Any differences observed in non-differentially expressed genes would be due to experimental noise. We calculate the variance of the log ratio across experiments for each gene in the data set and use the median (50th percentile) of the observed distribution of variances as the experimental variance estimate. The median should be robust to contamination by modest numbers of large standard deviation estimates that reflect true tumor-to-tumor differences rather than experimental noise. A lower percentile such as 10th or 25th may be a good choice if larger numbers of differentially expressed genes are expected.

The result is a new set of 'perturbed' data. We then re-cluster the perturbed data and compute our indices to measure how much the clustering has changed. We repeat the perturbation–clustering cycle numerous times and estimate the stability of the original clustering to data perturbations. Our R-index measures the proportion of pairs of specimens within a cluster for which the members of the pair remain together in the re-clustered perturbed data. Our D-index measures the number of discrepancies (additions or omissions) comparing an original cluster to a best-matching cluster in the re-clustered perturbed data.

Consider calculation of the R-index for the set of k clusters resulting from a cut of a dendrogram. We perturb the data by adding to the log ratio measurements independent, normally distributed random numbers with mean zero and variance equal to the estimated experimental noise variance. After perturbing, the data is re-clustered to obtain k clusters. If a cluster i of the original data contains n_i specimens, it can be viewed as containing $m_i = n_i(n_i - 1)/2$ pairs of specimens. If the clusters are robust, then members of a pair should fall in the same cluster in the re-clustered data. Let c_i denote the number of these m_i pairs with members falling in the same cluster in the re-clustered perturbed data. Then $r_i = c_i/m_i$ is a measure for the robustness of the ith cluster in the original data set. An overall measure of the set of k clusters is $R = (c_i + c_2 + \cdots + c_k)/(m_1 + m_2 + \cdots + m_k)$. Note that this overall measure is a weighted average of the cluster-specific measures, weighted by cluster size. In computing the overall measure, we exclude singleton clusters in the original data. The robustness indices can be averaged over a large number of cycles of perturbations and re-clusterings. For a singleton cluster in the original data, it can be informative to record the proportion of times it remains a singleton as opposed to being merged into another cluster in the perturbed data.

The D-index is computed somewhat differently. For each cluster of the original data, determine the cluster of the perturbed data that is the 'best match', defined as the one having the greatest number of elements in common with the original cluster. (Ties are broken by choosing the match with the least number of added elements.) The discrepancy can be subdivided into one of two types—either specimens in the original cluster that are not in the best match perturbed cluster (omissions), or elements in the best match cluster that were not in the original cluster (additions). It can be helpful to keep track of these two types separately, and this is one potential advantage of the D-index compared to the R-index. An overall measure of discrepancy is the summation of cluster-specific discrepancy indices. These indices can also be averaged over a large number of cycles of perturbations and re-clusterings. In computing the discrepancy index, we have found it useful to consider cuts of the perturbed data tree with similar, in addition to identical, numbers of clusters as in the original data, and to report the D-index as the minimum over the several cuts considered.

RESULTS

We first applied these cluster assessment methods to the melanoma data of Bittner *et al.* (2000) described previously. For the global test of clustering, we obtained a p-value of 0.003. Figure 2 (ftp://linus.nci.nih.gov/pub/techreport/TechReport2_Fig2.pdf) shows the observed NN EDF (nearest neighbor empirical distribution function) plotted versus that expected under the Gaussian null distribution. Our observed data exhibit a significant excess of small NN distances compared to a single multivariate Gaussian distribution. We interpret this as evidence of clustering pattern, and we proceed with examination of cluster-specific reproducibilities.

For examination of individual clusters, we computed the robustness and discrepancy indices. Of particular interest to Bittner *et al.* (2000) was the 19-member cluster containing tumors 6–24 as shown in the dashed box in Figure 1a. That group of 19 tumors occurs as a stand-alone cluster at cuts of 8, 9 and 10 clusters. The estimated experimental noise standard deviation (square root of median estimated variance) estimate was 0.16 (log base 10 scale) for this data. Table 1 presents cluster-specific reproducibility measures for clusters formed at cuts of 7 and 8 clusters. Cutting the tree at 7, we see that the cluster with elements 5–24 (composed of Bittner *et al.*'s major cluster and one additional tumor) is highly reproducible. With the exception of the cluster containing tumors 2 and 3, all of the other individual clusters formed at this cut of the tree are reproducible (robustness > 0.90). Cutting the tree at 8 clusters, the cluster containing tumors 2 and 3 continues to have very poor reproducibility (robustness $= 0.001$). Furthermore, it is evident that, on average, there is an addition of one member to the collection 6–24. The observation of a large average number of additions (17) to the singleton cluster containing tumor 5, along with inspection of several of the perturbed data trees (not shown), reveals that tumor 5 is frequently merged with tumors 6–24 when cutting to obtain 8 clusters. Thus, there appears to be strong evidence for reproducibility of a large cluster containing tumors 5–24. These results support Bittner *et al.*'s identification of subsets of melanoma within the data set, though they suggest a minor refinement with tumor 5 being included in the major cluster of non-invasive melanomas (Bittner *et al.*, 2000).

The second data set to which we applied our cluster assessment methods is that of Luo *et al.* (2001) which consists of gene expression profiles obtained from 25 prostate specimens, 16 of which were prostate cancer and 9 of which were Benign Prostatic Hyperplasia (BPH). These prostate expression profiles were obtained using cDNA microarrays consisting of 6500 human genes. Quality scores were provided for each log ratio measurement, with a score of zero indicating that the log ratio was deemed unreliable and should not be used. We did not include in the analysis any genes having quality scores of zero in more that 7 of the 25 specimens. This left 2817 genes for analysis. Then, we imputed any remaining missing values using a k-nearest neighbors algorithm (KNNimpute, Troyanskaya *et al.*, 2001): a missing log-ratio for gene j in specimen i was imputed with a weighted average of log ratios from 10 other genes in specimen i, where the 10 genes used were those whose expression profiles across specimens were closest (in Euclidean distance) to the profile of gene j, and the inverse Euclidean distance was used as the weight in averaging.

Figure 1b presents the dendrogram for the prostate data. Applying the global test, we find evidence that these expression profiles do not arise from a single Gaussian distribution ($p = 0.037$). Figure 3 (ftp://linus.nci.nih.gov/pub/techreport/TechReport2_Fig3.pdf) shows the observed NN EDF plotted versus that expected under the Gaussian null distribution. Our observed data exhibits an excess of small NN distances compared to a single multivariate Gaussian distribution. The above results were obtained using log ratios which had been median centered within each array. Without the median centering, the calculated p-value for the global test was $p = 0.0057$.

Table 2 presents cluster-specific reproducibility measures for clusters formed at cuts of 2, 3 and 4 clusters (reproducibility continues to deteriorate for larger number of clusters). The estimated experimental noise standard deviation (log 10 scale) for this data was 0.13. Cutting the tree at 3, we see that all clusters are highly reproducible, including the singleton cluster containing specimen #16. Specimens 1–16 were the 16 prostate cancer specimens 17–25 were all BPH specimens. Cutting the tree at four clusters, the discrepancies begin to increase, suggesting that any claims based on this data that there are two subtypes of prostate cancer would not be strongly supported by this data set. Our reproducibility assessment supports the conclusions of Luo *et al.* (2001) that the prostate cancers appear biologically distinct from the BPH specimens, but it also directs attention to the possibility that there is something unique about cancer specimen #16. We were not able to identify any obvious data quality problems with this array (such as an unusually large number of bad spots) that might explain this finding. Further investigation would be needed to determine if this specimen has any particular biological significance.

DISCUSSION

We have suggested a two-step approach to the evaluation of sample clustering. First we perform a global test for clustering, and if pattern is suggested, we follow with examination of cluster-specific reproducibilities. We have assessed the global test by simulation under a

Table 1. Cluster-specific reproducibility measures for the melanoma data[a]

Cluster	Tumor members	Robustness[c]	Omissions	Additions
Cut at 7 clusters	R-index = 0.993, D-index = 1.421[b]			
1	1	0.927	0.000	0.000
2	2–3	0.121	0.863	0.000
3	4	1.00	0.000	0.000
4	5–24	1.00	0.000	0.013
5	25	0.997	0.000	0.000
6	26–27	0.984	0.016	0.180
7	28–31	0.911	0.248	0.101
Cut at 8 clusters	R-index = 0.991, D-index = 19.590[b]			
1	1	0.999	0.000	0.001
2	2–3	0.001	0.894	0.001
3	4	0.968	0.000	0.000
4	5	0.000	0.000	17.29
5	6–24	1.00	0.001	0.910
6	25	1.00	0.000	0.000
7	26-27	0.966	0.064	0.011
8	28–31	0.902	0.411	0.003

[a] A hierarchical agglomerative clustering algorithm using average linkage and distance metric equal to one minus the Pearson correlation was applied. One thousand simulated perturbed data sets were generated using a noise SD = 0.16. [b] The D-index, omissions, and additions computed here allow searching over numbers of clusters in the perturbed data ranging from two less to two more than the number of clusters considered in the original data. [c] The reported robustness measure for a singleton cluster is the proportion of perturbed data clusterings for which it remained a singleton in the perturbed data clustering.

Table 2. Cluster-specific reproducibility measures for the prostate data[a]

Cluster	Tumor members	Robustness[c]	Omissions	Additions
Cut at 2 clusters	R-index = 0.946, D-index = 2.621[b]			
1	1–16	0.938	0.528	0.574
2	17–25	0.973	0.198	1.329
Cut at 3 clusters	R-index = 0.984, D-index = 2.589[b]			
1	1–15	1.00	0.098	0.719
2	16	1.00	0.000	0.000
3	17–25	0.938	0.236	1.536
Cut at 4 clusters	R-index = 0.923, D-index = 2.939[b]			
1	1–7, 10, 11, 13–15	0.905	0.698	0.461
2	8, 9, 12	0.899	0.182	0.635
3	16	1.00	0.000	0.000
4	17–25	0.958	0.286	0.677

[a] A hierarchical agglomerative clustering algorithm using complete linkage with Euclidean distance was applied. One thousand simulated perturbed data sets were generated using a noise SD = 0.13. [b] The D-index, omissions, and additions computed here allow searching over numbers of clusters in the perturbed data ranging from two less to two more than the number of clusters considered in the original data. [c] The reported robustness measure for a singleton cluster is the proportion of perturbed data clusterings for which it remained a singleton in the perturbed data clustering.

variety of situations. We generated data from multivariate Gaussian distributions in the original high-dimensional gene space using a variety of covariance matrices and means, and we found that the percent of test rejections was consistent with the nominal 0.05 level or less. Also, the test had good power in the several multiple cluster situations we examined unless the clusters were very close to one another, were extremely elongated, or contained very few members. Even under a few non-Gaussian multiple cluster situations we examined, the test maintained good properties. We speculate that this robustness may be due to the fact that even somewhat non-Gaussian data may appear approximately Gaussian in the principal components space because the principal components are formed by taking linear combinations over a very large number of genes. The global test can

be sensitive to outliers, but one should be able to identify these situations by following up with the cluster specific reproducibility assessment, and such outliers may be of interest in their own right. We emphasize that the cluster-specific reproducibility assessment is an important step in the interpretation process.

A number of methods have been proposed for detecting the 'optimal' number of clusters. Milligan and Cooper (1985) provide an extensive review and comparison of methods. Tibshirani *et al.* (2001) has proposed the 'gap' statistic for estimating an optimal number of clusters. The optimal number of clusters is chosen where the gap function shows a drop of larger than one standard deviation, where the standard deviation is determined by Monte Carlo simulation under a uniform null distribution. It is not designed to control the probability of falsely declaring the presence of multiple clusters. In contrast, we take a hypothesis testing approach based on the difference between the observed nearest neighbor distribution and that expected under two different null distributions. We generate the null distribution of the test statistic and do not rely on use of a standard deviation. Our tests are not directed at estimating the number of clusters, and so the methods are complementary. Yeung *et al.* (2001) propose a jackknife-type approach in which they successively leave out experimental conditions (arrays) to compute a figure-of-merit function that they plot to estimate an optimal number of clusters. Their particular interest was in clustering genes and they relied on the independence of the experimental conditions for justification of their jackknife approach. It is not clear if their method can be applied to clustering specimens, as it is not reasonable to assume that genes are independent. Golub *et al.* (1999) suggest a cross-validation method for assessing clustering results that involves building a predictor for the observed clusters and qualitatively assessing whether the predictor provides a high probability of an array being a member of one cluster or another. This appears to provide useful information when used for classifying a new set of expression profiles. Without an independent data set, however, it may be problematic. With thousands of candidate predictors, it may be possible to develop a predictor that appears to clearly distinguish among even random clusters. Ben-Dor *et al.* (2001) have developed a computationally intensive method for finding clusters and have proposed measures for assessing the strength of the results obtained from their clustering procedure. The utility of their measures in the context of other clustering algorithms is not yet established, however. Kerr and Churchill (2001) recently proposed the use of an index equivalent to our overall R-index, but they generated experimental error perturbations by bootstraping residuals obtained from an ANOVA model. Their methods require replicate profile measurements from at least some specimens. Replicates were not available in the data sets we considered. Also, the type of replicates available must be consistent with the totality of sources of experimental error that one wishes to account for in the reproducibility assessment. In our experience with microarray data, when replicates are available they often incorporate only some sources of the total experimental variation, for example hybridization of a single sample to the multiple arrays, but not replication at the level of re-sampling a tumor or re-isolating mRNA. In settings where appropriate replicates are available, bootstrap re-sampling could be readily incorporated into all of our cluster reproducibility assessement methods.

We feel that the more detailed reproducibility assessment methods we present here have several distinct advantages for interpreting the results of clustering biological specimens (e.g. tumors) on the basis of microarray data. First, the measures here have natural interpretations: robustness measured by proportion of preserved pairings, or discrepancy measured by numbers of additions or omissions. Second, the ability to examine cluster-specific reproducibility greatly enchanges understanding of the structure of the data. This was clearly seen in the melanoma data example. Had only measures for determining 'optimal' numbers of clusters been applied, they would likely have lacked sensitivity to shifting around of a few tumors, for example splitting of tumors 2 and 3 and cleaving of tumor 5 from the 5–24 cluster. Our cluster specific reproducibility measures very clearly indicated what was going on.

There are several other examples of situations in which ability to examine cluster-specific reproducibility will be important. Suppose a set of profiles fell into three distinct clusters, each of which could be further separated into two subclusters. The mindset of an 'optimal' number of clusters encourages one to choose three or six clusters when in fact, the appropriate conclusion would be that there is multi-level structure in the data. This multi-level structure would be elucidated by examination of cluster-specific reproducibilities at multiple cuts of the tree. Another interesting example is one in which there is a single very tight cluster that is surrounded by, but separated from, many 'noise' elements. Any method searching for an optimal number of clusters would have difficulty because clustering of the noise points would be essentially random and lacking in reproducibility, hence obscuring the fact that there was a tight, reproducible cluster in the middle. A method searching for an optimal number of clusters would likely conclude that there is one cluster, but this could not be distinguished from the situation of 'no clusters'. Applying the cluster specific reproducibility measures to this situation, one would find the tight cluster emerging as cuts of the tree corresponding to higher numbers of clusters are considered.

In summary, we feel it is important that objective

measures be used to interpret patterns of clustering and assess the reproducibility. Although relating observed clusters to known biology is one way to 'validate' observed clusters, a great hope in conducting microarray studies is that new biological features will be uncovered. Application of objective measures such as the ones we have described here should help to distinguish novel and potentially important biological findings from spurious findings.

REFERENCES

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Ben-Dor,A., Friedman,N. and Yakhini,Z. (2001) Class discovery in gene expression data. (http://citeseer.nj.nec.com/387748.html).

Bittner,M. *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Diggle,P.J. (1983) *Statistical Analysis of Spatial Point Patterns*. Academic Press, Orlando, FL, pp. 16–18.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fowlkes,E.B. and Mallows,C.L. (1983) A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, **78**, 553–569.

Gnanadesikan,R., Kettenring,J.R. and Landwehr,J.M. (1977) Interpreting and assessing the results of cluster analysis. *Bulletin of the International Statistical Institute*, **47**, 451–463.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gordon,A.D. (1999) *Classification*, 2nd edn, Chapman and Hall/CRC Press, London.

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

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

Johnson,R.A. and Wichern,D.W. (1988) *Applied Multivariate Statistical Analysis*, 2nd edn, Prentice-Hall, Englewood Cliffs, NJ, pp. 340–345.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Khan,J. *et al.* (1998) Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.*, **58**, 5009–5013.

Luo,J. *et al.* (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.*, **61**, 4683–4688.

Milligan,G.W. and Cooper,M.C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.

Rand,W.M. (1971) Objective criteria for evaluating clustering methods. *J. Amer. Stat. Assoc.*, **66**, 846–850.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Silverman,B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Tamayo,P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tibshirani,R., Walther,G. and Hastie,T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. Royal Statist. Soc.* B, (to appear).

Trent,J.M. *et al.* (1990) Tumorigenicity in human melanoma cell lines controlled by introduction of human chromosome 6. *Science*, **247**, 568–571.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Wolfinger,R.D. *et al.* (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**, 625–637.

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