BMC Bioinformatics

BioMed Central

Methodology article

Statistical significance for hierarchical clustering in genetic association and microarray expression studies

Mark A Levenstien*, Yaning Yang and Jürg Ott

Address: Laboratory of Statistical Genetics Rockefeller University New York, NY 10021, United States

Email: Mark A Levenstien* - markl@linkage.rockefeller.edu; Yaning Yang - yyang@linkage.rockefeller.edu; Jürg Ott - ott@linkage.rockefeller.edu

* Corresponding author

Published: 11 December 2003

BMC Bioinformatics 2003, **4**:62

Received: 30 July 2003

Accepted: 11 December 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/62

Abstract

Background: With the increasing amount of data generated in molecular genetics laboratories, it is often difficult to make sense of results because of the vast number of different outcomes or variables studied. Examples include expression levels for large numbers of genes and haplotypes at large numbers of loci. It is then natural to group observations into smaller numbers of classes that allow for an easier overview and interpretation of the data. This grouping is often carried out in multiple steps with the aid of hierarchical cluster analysis, each step leading to a smaller number of classes by combining similar observations or classes. At each step, either implicitly or explicitly, researchers tend to interpret results and eventually focus on that set of classes providing the "best" (most significant) result. While this approach makes sense, the overall statistical significance of the experiment must include the clustering process, which modifies the grouping structure of the data and often removes variation.

Results: For hierarchically clustered data, we propose considering the strongest result or, equivalently, the smallest p-value as the experiment-wise statistic of interest and evaluating its significance level for a global assessment of statistical significance. We apply our approach to datasets from haplotype association and microarray expression studies where hierarchical clustering has been used.

Conclusion: In all of the cases we examine, we find that relying on one set of classes in the course of clustering leads to significance levels that are too small when compared with the significance level associated with an overall statistic that incorporates the process of clustering. In other words, relying on one step of clustering may furnish a formally significant result while the overall experiment is not significant.

Background

Hierarchical clustering is an information theoretical method that sequentially merges samples based on the pair-wise similarity of a given measurement to form common groups until all samples are contained in a single group [1]. The method has many applications and is widely used in the analysis of biological data. For example, researchers testing for association between haplotypes and disease have employed hierarchical clustering as a means to reduce a large number of haplotypes to a manageable number of haplotype classes with the aim to increase statistical power [2]. The alleles present at multiple genetic marker loci across a given chromosome form a haplotype [3]. With an increasing number of marker loci, the number of possible haplotypes grows exponentially so that many of these haplotypes tend to have low frequency. In comparisons of haplotype frequencies between case and control individuals, the corresponding contingency tables are often sparse and difficult to interpret. Hierarchical clustering then allows researchers to merge haplotypes into classes that are easier to handle. Variability within a haplotype class is generally considered unimportant (random noise) so that the researcher can focus on the "larger picture", that is, whether some of the haplotype classes exhibit a statistically significant difference in frequency between case and control individuals. Typically, the statistical significance (computed with exact tests [4]) for an initial, sparse contingency table is lower than for tables obtained by clustering the haplotype classes present in the initial table. Often the process of clustering incorrectly removes the variation within a class, and in these cases the increase in statistical significance is entirely due to the clustering process. Here we propose an analysis method that properly takes clustering into account. We achieve this by defining the strongest result or, equivalently, the smallest *p*-value, occurring in the course of clustering as the statistic of interest and computing its associated (experiment-wise) statistical significance.

Another example of hierarchical clustering is its application in microarray analyses [5-7]. Often clustering of arrays based on microarray expression data is utilized to distinguish tumor subclasses, which have clinical implications [8,9]. In many of these studies involving microarray expression data from tumor specimens, researchers are interested in examining survival information for the subjects who contributed the samples and comparing the survival curves between groups formed by the hierarchical clustering procedure [10-13]. The methods developed in this paper will be applied to three previously published datasets in which hierarchical clustering has been employed. One of these datasets involves a haplotype association analysis while the other two datasets refer to survival analyses of groups of individuals determined by microarray expression measurements.

The problem of testing group differences sequentially is in the framework of multiple testing. Historically, both genetic association studies and microarray studies have been plagued with multiple testing problems. In the case of association studies, multiple testing occurs because researchers perform tests of association for large numbers of haplotypes, alleles, or genotypes across entire chromosomes or genomes [14]. In the case of microarray data analysis, researchers sequentially test thousands of genes for differential expression. Testing at each different clustering step within a hierarchical structure also represents a form of multiple comparisons; therefore, the experiment-wise type I error is inflated. Various correction methods such as Bonferroni, step-up, and step-down have been employed to adjust for the multiplicity of testing [15]. These procedures appear to work well only when the tests in the sequence are independent or weakly correlated. Since the tests within the hierarchy possess a nested structure, these procedures are inappropriate for our situation. As mentioned above, here we propose an alternative solution by defining a single test statistic, for which we evaluate the experiment-wise statistical significance.

Methods

Local p*-values*

Consider multiple steps in hierarchical clustering. For each of n steps of the hierarchy, we calculate our statistic of interest depending on the application. In the case of haplotype association tests, we compute the Pearson χ^2 [4] for a $2 \times s$ contingency table (case/control individuals versus s haplotypes or haplotype classes) while in the case of survival analyses, we compute the log-rank statistic [16]. We represent these statistics as a vector, $\vec{X} = (X_1, X_2, ..., X_n)$, where X_i represents the statistic obtained at the i^{th} step in the clustering process. To make statistics from different steps comparable, we compute the significance level, p_i, associated with X_i and call this a local *p*-value. We approximate these local empirical significance levels via permutation analysis. These permutation methods involve randomly permuting labels for each individual as follows. For haplotype association tests, we permute the case/control labels [17,18] while for survival analyses, we permute failure times and censorship statuses jointly. For each permutation of the dataset, we cluster the permuted samples as illustrated by the dendrogram and calculate a null statistic based on the permuted samples at each step in order to generate the null distribution for the statistic. We can represent the collection of null statistics calculated from each of m permutations of the data at each of n steps within the hierarchy as the matrix,

$$\mathbf{X_{null}} = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix},$$

where the entry appearing in the i^{th} row and the j^{th} column, X_{ij}, is the statistic of interest computed from the i^{th} permutation of the data at the j^{th} step in the hierarchy. At each step of the hierarchy, by comparing the statistic we computed from the data with the null statistics we computed from the m permutations, we calculate a local *p*-value, p_j, as the proportion of permutation samples with a null statistic at least as large as the observed statistic. We represent the local *p*-values as the vector, $\vec{p} = (p_1, p_2, ..., p_n)$.

Permutation (randomization) samples allow one to conveniently approximate the sampling distribution of test statistics under the null hypothesis (the "null distribution"). Ideally, permutation tests are based on the total of all permutations but in practice we usually can only collect a random sample from these permutations. The number m of permutation samples should be large enough to adequately represent the sample space of permutations. For the haplotype data (example 1), at each step we compared approximated p-values obtained with different values of m with exact p-values calculated with the aid of the statistical software package *StatXact 5*. For the first few steps in the hierarchy, values of m on the order of 10,000 were sufficient to provide p-values very close to the correct ones. However, at later steps, agreement was only obtained with m = 100,000, presumably because at early steps the total number of permutations is much smaller than at later steps. The calculations for the two survival analyses (examples 2 and 3) were also performed with m = 100,000.

Global p*-value*

In order to gain an empirical significance assessment for the entire experiment, we define a single statistic, that is, the smallest of the local p-values, $min_i(p_i)$ [19]. To assess the empirical significance level (global p-value), p_{min}, associated with this statistic, we generate the null distribution of $min_i(p_i)$from the matrix of null statistics, $\mathbf{X_{null}}$. In this matrix, we consider each row (replicate) in turn as observed data and evaluate these data based on the remaining m - 1 null data as described above for m null data. That is, for each of these "null observed" permutation samples a minimum p-value is obtained at whatever step it occurs. This leads to a set of m null values for $min_i(p_i)$. The proportion of these values at least as small as the observed $min_i(p_i)$represents the global significance level, p_{min}, associated with our single experiment-wise statistic. Since this approach requires that the p-values be ordered, starting with the most significant, it could be considered a step-down p-value adjustment procedure similar to the procedure developed by Westfall and Young [20]. If $p_{min} \leq 0.05$ then we say that the experiment (at least one of the steps in the clustering process) is significant at the 5% level.

In addition, we are interested in assessing whether clustering has provided a benefit for the data analysis. In order to achieve this, we compare the statistical significance for the entire experiment involving tests at each step created by clustering with the statistical significance for an experiment where no clustering has been applied; that is, we compare the global p-value, p_{min}, with the significance level, p_0, of the statistic prior to clustering. (We are able to make this exact comparison for the haplotype association application; however, for the microarray expression studies, we must compare the global significance with the statistical significance at step 2 in the hierarchy since permuting the data at steps 0 and 1 creates a collection of null statistics possessing a very small variance.) Clustering is only beneficial when $p_{min} < p_0$ (or p_2 for the microarray datasets) since the results are more significant when clustering is applied. It may well happen that the smallest p-value, $min_i(p_i)$, at one of the steps in the course of clustering is smaller than p_0 (or p_2 for the microarray datasets), but the clustering process is such that this smallest p-value has a high probability of occurring by chance. In that case, one will find that $p_{min} > p_0$ (or p_2 for the microarray datasets).

Statistics of interest

As mentioned above, in the case of association studies between haplotypes and disease we employ the Pearson χ^2 to test each step of the hierarchy for association [4]. However, in the case of survival analyses, our statistic of interest is the log-rank statistic [16]. It provides an overall comparison of the Kaplan-Meier survival curves for two or more groups of subjects. For r groups, the log-rank statistic asymptotically follows a χ^2 distribution with r - 1 degrees of freedom under the null hypothesis of equality of survival curves.

Results

To demonstrate our approach on real data, we re-analyze the following three previously published datasets.

Example 1 (haplotype data)

The first dataset consists of 52 statistically predicted haplotypes in 172 African-American study participants (137 case and 35 control individuals) [2]. The aim of that case-control study was to test for association between haplotypes at 25 single nucleotide polymorphism (SNP) loci in the human μ opioid receptor gene (OPRM1) and substance dependence. The large number of haplotypes was difficult to interpret and appeared to create a situation with insufficient power to detect association. Thus, hierarchical clustering was applied to the 52 haplotypes. These were sequentially grouped according to the procedure CLUSTER (method = BAVERAGE, measure = SEUCLID) from the SPSS software package for Windows [2]. For each step of the resulting dendrogram, the hierarchical clustering procedure designates which haplotypes are clustered to form haplotype classes. At each step of the hierarchy an association test was performed between haplotype classes and disease status. As the clustering progressed, the number of classes became smaller and smaller.

Using the same clustering methods and resulting hierarchical structure, we apply our algorithm for assessing local and global p-values in this dataset. Our p-values differ somewhat from the ones previously published [2] but the

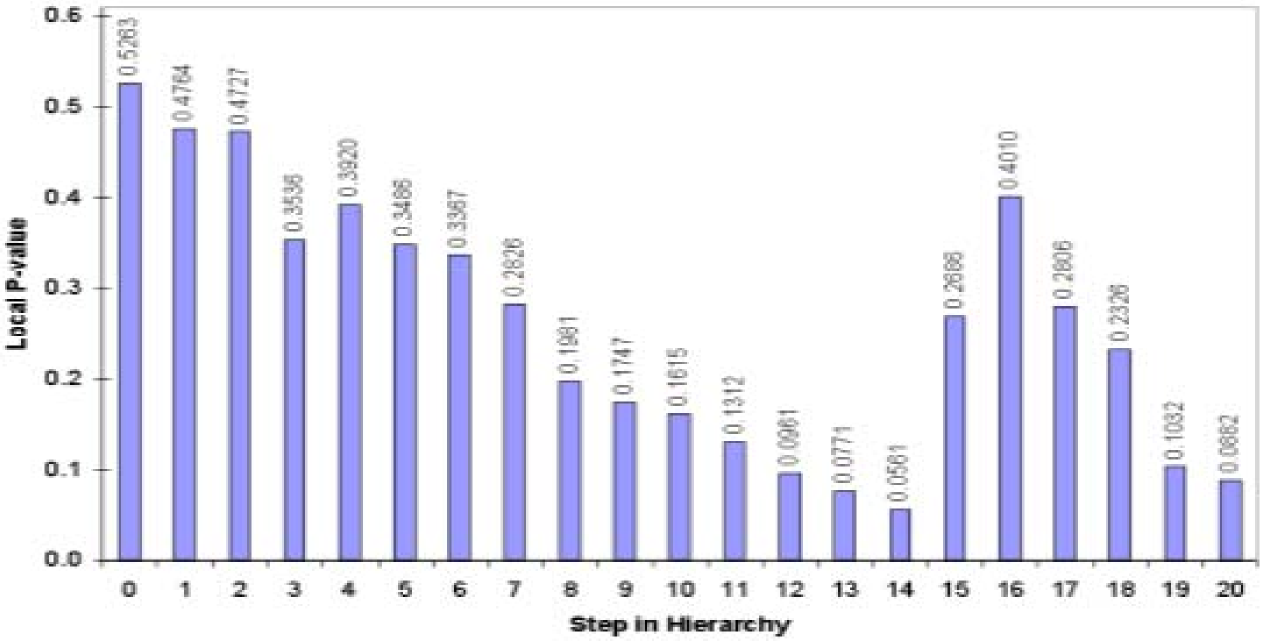

Figure 1
Results from haplotype association tests applied to all steps of the hierarchical structure formed by clustering data from Hoehe et al. [2]. This bar graph presents the local *p*-values computed by our group at all steps within the hierarchical structure.

patterns of the local *p*-values across the clustering steps shown in Figure 1 and in [2], respectively, are highly comparable. Based on m = 100,000 permutation samples (see Methods), we calculate local *p*-values for hierarchical clustering steps zero through 20, where zero represents the step with un-clustered haplotypes and 20 represents the step where only two haplotype groups remain. We find the smallest *p*-value, $min_i(p_i)$= 0.0561, at step 14 (Figure 1). Thus, one is tempted to declare this result borderline significant at the 5% level. However, the (global) significance level associated with this smallest *p*-value turns out to be p_{min} = 0.6918; that is, there is almost a 70% random chance (unrelated to association between haplotypes and disease) to find at any step in the hierarchy a minimum *p*-value at least as small as the value of 0.0561 found for the observed data. This sobering result leaves the experiment statistically non-significant. Since clustering failed to produce an experiment-wise significance level of p_{min} less than the initial pre-clustering significance level of p_0 = 0.5263, the clustering process did not provide any benefit for this dataset.

Example 2 (lung cancer data)

This dataset contains expression levels for 835 unique genes represented by 918 cDNA clones in tissues harvested from lung cancer patients and normal individuals [11]. Specifically, expression levels are measured in 41 adenocarcinomas (ACs), 16 squamous cell carcinomas (SCCs), five large cell lung cancers (LCLCs), five small cell lung cancers (SCLCs), five normal lung samples, and one normal fetal lung sample. Based on the Complete Linkage method and Pearson's correlation coefficient as a measure of similarity in the CLUSTER software, hierarchical cluster analysis was performed to group the samples according to the degree of similarity present in the gene expression data. In the resulting dendrogram, the AC samples appeared in three distinct clusters. The aim of the study was to examine whether the groups of AC samples created by the hierarchical clustering procedure correlated with clinical outcomes of the AC patients, that is, whether the Kaplan-Meier survival curves differed for these groups [11].

Again, using the same clustering methodology as in the publication [11], we apply this technique to their AC data [21] and work with the resulting hierarchical structure for assessing the local and global *p*-values. The dendrogram in Figure 2A details the hierarchical clustering of the data (for the 24 AC samples from patients with reported survival information) for steps zero through 22. For each step in the hierarchy we calculate a log-rank statistic and the corresponding local *p*-value (m = 100,000 permutation

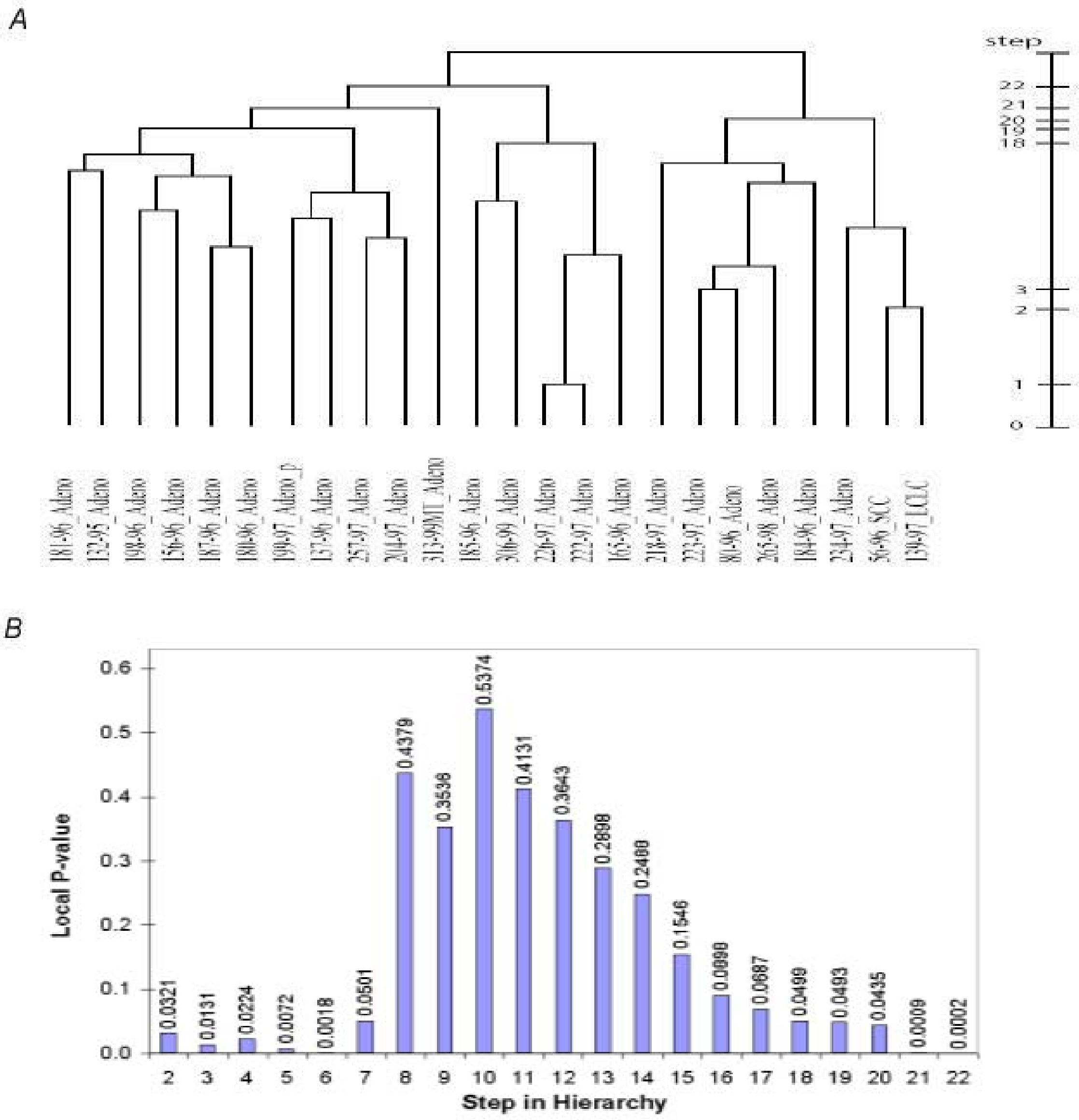

Figure 2

Results from log-rank tests applied to steps of the hierarchical structure formed by clustering data from Garber et al. [11]. **A**, This schematized dendrogram reflects the process of clustering microarray samples according to the similarity of their gene expression profiles as measured by the Pearson correlation coefficient. Distances between array sample clusters are approximated (not to scale) by the vertical axis. Along the bottom of the dendrogram are the microarray tissue samples from individuals for which survival data was available [11]. **B**, This bar graph displays the local *p*-values we compute at each step within the structure created by hierarchical clustering.

samples). Figure 2B graphically presents these local *p*-values. We exclude the first two clustering steps (0 and 1) from the figure and further assessments because insufficient variability in the log-rank statistic at these steps does not permit meaningful calculation of local *p*-values. At step 22, we observe the minimum local *p*-value of 0.0002, and we calculate the global *p*-value for this dataset to be 0.0040. Thus, the experiment shows a statistically significant result, and clustering was effective. It reduced the initial *p*-value of 0.0321 at step 2 to the global significance level of $p_{min} = 0.0040$.

Example 3 (lymphoma data)

The third dataset contains expression levels of cDNA clones from genes expressed in germinal center B-cells for 47 samples of diffuse large B-cell lymphoma (DLBCL) [12]. Hierarchical clustering was performed with the CLUSTER program and the Pearson correlation coefficient as its similarity measure to group the samples by similarity of gene expression levels for all genes expressed in germinal center B-cells. The resulting dendrogram shows two main branches, one containing samples with expression patterns similar to those of germinal center B-cells and one containing samples with expression patterns similar to those of activated B-cells. To examine the clinical relevance of this subdivision of DLBCL, a Kaplan-Meier survival analysis for the two groups of patients was performed based on the dendrogram's penultimate clustering step [12].

As with the other datasets, we cluster the data [22] with the same method as published [12] and use the resulting hierarchical structure for calculations of log-rank statistics and associated local *p*-values (m = 100,000 permutation samples) at different steps in the hierarchy. The dendrogram in Figure 3A provides the order of clustering (for the 40 DLBCL samples from patients with reported survival information) for steps zero through 38 while Figure 3B graphically presents local *p*-values at the different clustering steps. As in example 2, we observe a very small variance in the log-rank statistic at the first two clustering steps and, therefore, exclude these steps from further analysis. At step 6, we observe the minimum local *p*-value of 0.0010, with an associated global *p*-value of $p_{min} = 0.0364$. This result is statistically significant at the 5% level, but clustering has not contributed to an increase in significance because un-clustered or only minimally clustered data show higher significance (lower *p*-value). These are not mutually exclusive ideas. The implication is that if clustering is applied, the result of testing over the entire hierarchy is significant; however, the results would be more significant had the experimenter never clustered the data in the first place.

Discussion

In hierarchical clustering, evaluating the minimum local *p*-value in isolation, outside of the context of the larger hierarchical structure used to organize the data, can drastically affect the interpretation of test results. For example, even though the haplotype data show an apparently significant result with a minimum *p*-value of 0.0561, our analysis demonstrates that clustering the same data but without association between haplotypes and disease has a high chance of obtaining such a "significant" result. In fact, that chance is $p_{min} = 0.6918$, which represents the actual significance level of the experiment. On the other hand, as example 2 shows, clustering can improve the significance of a result. The global significance level, p_{min}, can be viewed as a version of $min_i(p_i)$ corrected for multiple testing within the hierarchy. In all three examples presented above, applying the Bonferroni correction to $min_i(p_i)$ provides an upper bound for p_{min} in agreement with theory. In fact, it is interesting that for examples 2 and 3 our procedure which accounts for the correlated nature of the tests yields a p_{min} value only slightly smaller than the Bonferroni correction applied to $min_i(p_i)$.

How can we explain that in some cases clustering is beneficial while in other cases it is not? Presumably, some datasets possess an underlying heterogeneity; that is, such datasets are composed of samples from multiple distinct populations. If the information used for clustering (haplotypes for example 1 and gene expression patterns for examples 2 and 3) is related to the information used to perform the statistical test (in our examples, proportions of cases to controls and survival times), hierarchical clustering will detect the heterogeneity. Otherwise, the clustering process is random, and any heterogeneity detected is artificial. Our approach allows one to distinguish between these two situations. If the clustering process is random because the information used for clustering and calculating the test statistic are unrelated (or because the dataset is homogeneous), a large p_{min} will result indicating that any small local *p*-values probably occurred only by chance. Whereas if the clustering process is directed by a measurement strongly related to the test statistic, a small p_{min} will result indicating that any heterogeneity found within the hierarchy is most likely real.

Often when hierarchical clustering is applied to a dataset, it is of interest to determine the true number of classes present. This situation commonly arises in the analysis of microarray data. For instance, as in examples 2 and 3, in the study of human cancers, researchers often utilize microarray expression data to cluster samples. From the hierarchical structure created by clustering, it may be of interest to distinguish the optimum number of tumor subclasses that are most clinically relevant. Several statistics-based methods have been utilized to estimate the true number of groups from such microarray expression datasets [23,24]. However, such methods rely solely on the expression data itself. Alternatively, it may prove practical in such microarray expression studies to consider additional information available, such as survival data, on each sample for distinguishing clinically relevant subclasses. Employing our procedure of calculating the local *p*-values for a test statistic at multiple steps within the hierarchy and then selecting the step where the minimum of

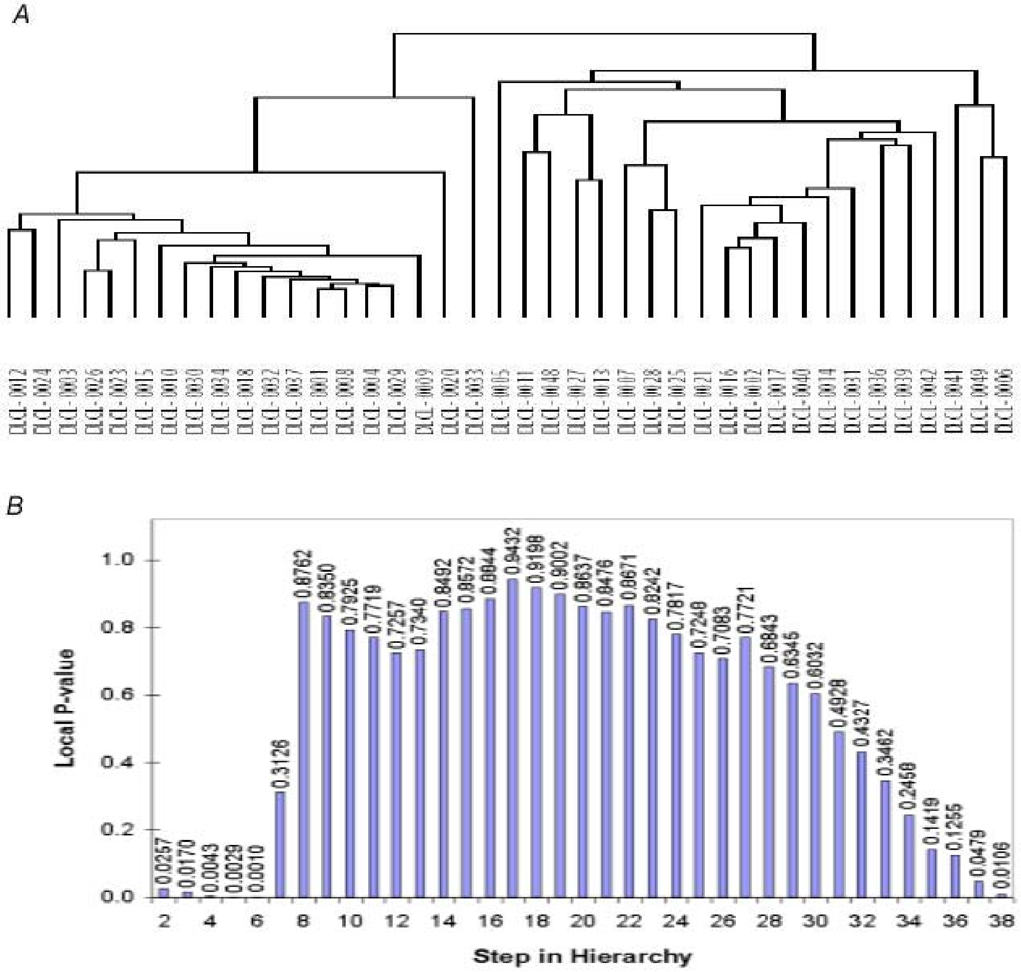

Figure 3

Results from log-rank tests applied to steps of the hierarchical structure formed by clustering data from Alizadeh et al. [12]. **A**, This schematized dendrogram reflects the process of clustering microarray samples according to the similarity of their gene expression profiles as measured by the Pearson correlation coefficient. Distances between array sample clusters are approximated (not to scale) by the vertical axis. Along the bottom of the dendrogram are the microarray tissue samples from individuals for which survival data was available [12]. **B**, This bar graph displays the local *p*-values we compute at each step within the structure created by hierarchical clustering.

these *p*-values occurs as the basis for determining the true number of classes which exist for a given dataset may provide an advantage over existing methods. Of course, if such a method for determining the true number of classes is applied, the global *p*-value will provide an assessment of its significance. However, applying our procedure to some datasets, such as the data in example 3, results in determining a large number of true classes. In fact, the number of classes determined may be so large that the use of these expression-based tumor subclasses in clinical diagnosis may not provide a benefit. Therefore, in order to increase the practicality of our method, it may prove necessary to eliminate some of the lower steps in the hierarchy from eligibility for selecting the minimum local *p*-value and the calculation of its significance.

Besides determining subclasses for biological samples, hierarchical clustering is often employed in the context of microarray expression studies in order to identify groups of genes that are regulated in a similar manner. In these cases, the clustering is performed on the genes rather than on the samples. Our method relies on two sets of data – one for clustering and a second for the statistical test. Since the samples possess both expression data across genes and survival data, our method is applicable to hierarchies created by clustering on samples. However, genes only possess expression data across samples, and, consequently, our method is inappropriate for analyzing the significance of hierarchies created by clustering on genes.

Our approach may be viewed as a contribution to the problem of multiple testing. We address this problem by defining a single experiment-wise statistic whose associated empirical significance level represents the overall significance of the experiment. For the cases we have examined, the experiment refers to performing a test at each step in a hierarchy created by clustering. However, the meaning of experiment can be expanded to reflect other practices adopted by researchers. For example, researchers may apply several clustering algorithms involving various combinations of clustering methods and distance measures before finalizing their choice of clustering algorithm. Since this practice introduces an additional test at each step within each of the trial hierarchies, it compounds the effect of multiple testing. Additionally, in some situations researchers may be interested in testing for heterogeneity among groups with multiple measurements. For instance, when searching for clinically relevant subclasses of cancer, researchers may examine groups for differences in survival times, as well as, differences in physical characteristics of the tumor cells. Both sets of information may be clinically relevant; however, to correct for the additional testing, the meaning of experiment in calculating p_{min} must be expanded to reflect the entire process employed by the researcher. Of course, it is possible that the process of hierarchical clustering forms medically relevant groups that do not display heterogeneity for any of the measurements collected. In this case, our strategy will not find these groups as the true grouping structure for the samples.

Several other methods addressing multiple comparison problems have been proposed and are in current use. In particular, as an alternative to the classical significance level, p, the False Discovery Rate (FDR) has become rather popular [15]. However, it is important to keep in mind that p and FDR are not really comparable – p is the conditional probability of a significant test result given the null hypothesis is true (the expected proportion of false positive results among all "false" results, i.e., results obtained under the null hypothesis) while FDR is the conditional probability of the null hypothesis being true given a significant test result (the expected proportion of false positive results among all "positive" results, i.e., significant test results). Future research will have to determine which of these various approaches to eliminate the effects of multiple testing is most effective.

Authors' Contributions

ML developed and implemented the algorithm, carried out analyses, and drafted the manuscript. JO suggested the problem and drafted the manuscript. YY provided support for statistical issues and critically read the manuscript.

Acknowledgements

We are grateful to Dr. Margret Hoehe for providing the dataset used as example 1 and to Chad Haynes for programming support. This work was supported by NIH grants MH59492 and HG00008.

References

1. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction New York, Springer*; 2001.
2. Hoehe MR, Kopke K, Wendel B, Rohde K, Flachmeier C, Kidd KK, Berrettini WH, Church GM: **Sequence variability and candidate gene analysis in complex disease: association of mu opioid receptor gene variation with substance dependence.** *Hum Mol Genet* 2000, **9:**2895-2908.
3. Ott J: *Analysis of Human Genetic Linkage* 3rd editionth edition. *Baltimore, The Johns Hopkins University Press*; 1999.
4. Agresti A: *An Introduction to Categorical Data Analysis. Wiley Series in Probability and Statistics NewYork, John Wiley and Sons*; 1996.
5. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
6. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Natl Acad Sci USA* 1999, **96:**6745-6750.
7. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11:**4241-4257.
8. Chung CH, Bernard PS, Perou CM: **Molecular portraits and the family tree of cancer.** *Nat Genet* 2002, **Suppl 32:**533-540.
9. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
10. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98:**10869-10874.
11. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I: **Diversity of gene expression in adenocarcinoma of the lung.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98:**13784-13789.
12. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J., Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511.
13. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci USA* 2001, **98:**13790-13795.
14. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273:**1516-1517.

15. Reiner A, Yekutieli D, Benjamini Y: **Identifying differentially expressed genes using false discovery rate controlling procedures.** *Bioinformatics* 2003, **19:**368-375.
16. Kalbfleisch JD, Prentice RL: *The Statistical Analysis of Failure Time Data New York, John Wiley and Sons*; 1980.
17. Zhao JH, Curtis D, Sham PC: **Model-free analysis and permutation tests for allelic associations.** *Hum Hered* 2000, **50:**133-139.
18. Zhao JH, Sham P: **Faster haplotype frequency estimation using unrelated subjects.** *Hum Hered* 2002, **53:**36-41.
19. Hoh J, Wille A, Ott J: **Trimming, weighting, and grouping SNPs in human case-control association studies.** *Genome Res* 2001, **11:**2115-2119.
20. Westfall PH, Young SS: *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment New York, John Wiley & Sons*; 1993.
21. **Lung Adenocarcinomas > Home** [http://genome-www.stanford.edu/lung_cancer/adeno/index.shtml]
22. **Lymphoma/Leukemia Molecular Profiling Project (LLMPP)** [http://llmpp.nih.gov/lymphoma/]
23. Horimoto K, Toh H: **Statistical estimation of cluster boundaries in gene expression profile data.** *Bioinformatics* 2001, **17:**1143-1151.
24. Dudoit S, Fridlyand J: **A prediction-based resampling method for estimating the number of clusters in a dataset.** *Genome Biol* 2002, **3:**RESEARCH0036.