Funct Integr Genomics (2000) 1:156–173
DOI 10.1007/s101420000019

ORIGINAL PAPER

Kui Zhang · Hongyu Zhao

Assessing reliability of gene clusters from gene expression data

Received: 4 May 2000 / Accepted: 5 July 2000 / Published online: 23 August 2000

Abstract The rapid development of microarray technologies has raised many challenging problems in experiment design and data analysis. Although many numerical algorithms have been successfully applied to analyze gene expression data, the effects of variations and uncertainties in measured gene expression levels across samples and experiments have been largely ignored in the literature. In this article, in the context of hierarchical clustering algorithms, we introduce a statistical resampling method to assess the reliability of gene clusters identified from any hierarchical clustering method. Using the clustering trees constructed from the resampled data, we can evaluate the confidence value for each node in the observed clustering tree. A majority-rule consensus tree can be obtained, showing clusters that only occur in a majority of the resampled trees. We illustrate our proposed methods with applications to two published data sets. Although the methods are discussed in the context of hierarchical clustering methods, they can be applied with other cluster-identification methods for gene expression data to assess the reliability of any gene cluster of interest.

Keywords Gene expression · Hierarchical clustering · Bootstrap · Consensus tree

Introduction

Recent development of microarray technologies has made it possible to simultaneously measure expression levels of tens of thousands of genes, and shifted our attention towards an integrated understanding of the genetic networks underlying complex biological phenotypes. Large-scale gene expression studies have been carried out to study cell cycle (Eisen et al. 1998), tumor tissues (DeRisi et al. 1996; Khan et al. 1999), drug targets (Debouck and Goodfellow 1999; Marton et al. 1998), and resequence and mutational analysis (Hacia 1999). Generally speaking, statistical methods can be developed to address three types of questions using microarray data, which are, in order of complexity:

Which genes are differently expressed among the samples studied? Which genes are expressed in a coordinated manner across a set of conditions? What are the global biological pathways?

Although the ultimate goal is to identify genetic network architectures (the third question), the amount of information required to achieve this goal may be, at this point, beyond experimental capacity for complex systems. As a first step towards this ultimate goal, many existing statistical procedures, most notably a variety of clustering algorithms, have been applied to analyze microarray data to identify genes expressed in a coordinated manner (the second question). These methods include hierarchical clustering algorithms (e.g., Eisen et al. 1998; Heyer et al. 1999), principal components analysis (e.g., Hilsenbeck et al. 1999; Raychaudhuri et al. 2000), multidimensional scaling methods (e.g., D'Haeseleer et al. 1998), self-organizing maps (Tamayo et al. 1999; Törönen et al. 1999), and graph-theoretic techniques (Ben-Dor and Yakhini 1999). They all fall into the unsupervised analysis category in contrast to the supervised learning algorithms, where there are some pre-defined classes, either for tissue samples (e.g., Golub et al. 1999) or for gene groups (e.g., Brown et al. 2000). These clustering methods have been found to work well in practice because genes with related functions were found to be enriched in particular clusters (Eisen et al. 1998), normal tissue samples and tumors can be classified with very high accuracy using gene expression data (Alon et al. 1999), and similarities and differences among tumors that cannot be recognized by traditional morphological examination can be identified through gene expression data (Anbazhagan et al. 1999; Golub et al. 1999).

Although the above approaches have proved valuable in gene expression pattern detection, most published

K. Zhang · H. Zhao (✉)
Department of Epidemiology and Public Health,
Yale University School of Medicine, New Haven, CT 06520, USA
e-mail: hongyu.zhao@yale.edu
Tel.: +1-203-7856271, Fax: +1-203-7856912

large-scale studies are quite elusive over the variations in measured gene expression levels among different samples and experiments. For glass slide arrays, up to two-fold differences among replicated experiments are commonly observed. Mir and Southern (1999) studied the effect of structure on nucleic acid heteroduplex formation by analyzing hybridization of tRNA to a complete set of complementary oligonucleotides ranging from single nucleotides to dodecanucleotides. They found that major determinants of hybridization lie in the structure of the RNA. Their finding is very relevant to gene expression studies using Affymetrix GeneChip microarrays, where 20 pairs of oligonucleotides corresponding to the same gene or EST are hybridized to the sample and a single expression level is derived from these 40 observations.

In essentially all published studies, the observed gene expression levels are treated as if they were an accurate measure of the true expression level, and the effects of measurement errors have seldom been addressed. However, it is not apparent how variations in the measurements might affect the conclusions drawn from these studies. For example, if we have identified a group of ten genes with similar expression profiles, we need to determine whether this cluster is a *real* cluster or a *superficial* one resulting from random variations in gene expression measurements. Therefore, it is both desirable and crucial to assess the reliability and statistical significance of an individual gene cluster of interest. In this article, as a first step towards an understanding of the effects of measurement errors on cluster identification, we propose a resampling method under the hierarchical clustering framework, the most commonly used approach in microarray analysis. Our procedure can be divided into three steps: we first generate a large number of resampled microarray data with information on the magnitude of measurement errors; we then use the majority-rule to construct a consensus tree; and lastly, we estimate the confidence value for each branch in the clustering tree from the original data set. In this article, we describe our methods and apply them to two published data sets. Although most of the discussion is within the context of hierarchical clustering algorithms, our methods can be applied to other unsupervised or supervised algorithms to assess the effects of measurement errors on cluster identification.

Materials and methods

Hierarchical clustering algorithms

There is a large volume of literature on cluster analysis in statistics (e.g., Hartigan 1975), and many methods are available in general statistics packages (e.g., S-Plus, SAS) and specialized programs (e.g., PHYLIP). The hierarchical clustering methods are commonly used because of their simplicity and fast running time. The first step of a hierarchical clustering algorithm is to select an appropriate mathematical description of similarity. There are many possible similarity measures that can be used, including euclidean distance, Pearson correlation coefficient, and rank correlation. The actual choice should reflect the nature of the biological question and the technology that was used to obtain the data. After calculating similarities among all the genes, the second step in a hierarchical clustering algorithm is to join the two most similar objects into a single cluster and recompute the similarity matrix. Three common options for this step are single linkage, average linkage, and complete linkage. These options differ in how the similarity matrix is recomputed among clusters. The process ends when all the objects agglomerate to a single cluster. We can use a binary tree to represent the result of a hierarchical clustering algorithm. It is not the purpose of this paper to survey all hierarchical clustering methods available, but rather to illustrate how to use the proposed resampling method to study the influence of random errors on cluster identification and make a statistical inference from the resulting clustering tree.

Resampling methods

Statistical resampling methods have been used extensively in genetic research. In the context of phylogenetic analysis, Felsenstein (1985) proposed using the bootstrap method to estimate the confidence value for each clade in a phylogenetic tree. Although there has been some criticism of this method (e.g., Hillis and Bull 1993), Efron et al. (1996) showed that Felsenstein's method is not biased, and it can be corrected to better agree with standard ideas of confidence levels and hypothesis testing at the expense of considerably more computation.

Because bootstrap has been well studied in statistics (e.g., Efron 1979; Efron and Tibshirani 1993), we describe this method only briefly here. Suppose we have an original data set with n observations and we want to make a statistical inference of a population parameter from this data set. We first compute a sample statistic to estimate the population parameter. In the typical bootstrap method (also called nonparametric bootstrap), we sample n observations at random with replacement from the observed sample. Using the bootstrap method, some data points will not be included in the bootstrap data set, some data points will be included only once, and still others will be included twice or more. We then compute the sample statistic based on this bootstrap data set. If we repeat the bootstrap procedure a large number of times, the amount of variation in the sample statistics calculated from the bootstrap data sets can be used to assess the uncertainty in our estimate of the population parameter.

Because of the nature of gene expression data, we adopt here the *parametric* bootstrap method. Suppose we know the uncertainty (standard error) in our measure of gene expression data. The standard errors can be estimated from replicated experiments in the study (e.g., Wen et al. 1998) or from previous experiments with similar conditions. Let x_{ij} denote the gene expression for gene i at condition j and let s_{ij} denote our estimated variation for this measure. In our resampling approach, we "bootstrap" the gene expression level for each gene under each experimental condition from the normal distribution $N(x_{ij}, s_{ij})$. Simulated data can also be generated assuming other distributions for the gene expression measurements. For example, we simulate data from log-normal distributions in the Discussion section. We then construct a binary tree (which will be rigorously defined in the next section) by using a hierarchical clustering algorithm. For a large number of resampled data sets, we can generate a set of binary trees. These trees can then be used to assess the reliability of the clusters observed in the original data set as discussed in the following. Note that if we assume the gene expression measurement follows a non-normal distribution, we will sample the "new" measurement from this distribution.

Consensus tree

From the binary trees constructed using the resampled data, we can define a consensus tree made up of all those nodes that appear in a majority of these resampled trees (Margush and McMorris 1981; Felsenstein 1985). The definition of the majority-consensus tree is given below following Margush and McMorris (1981).

First, we must give a more rigorous representation of a tree and a binary tree. If $S=\{1,2,...,n\}$ is a set of n objects (e.g., genes or tissue samples), then an n-tree is a type of hierarchical classification of S with the following definition:

Definition (Margush and McMorris 1981)

Let $\mathbf{P}(S)$ denote the set of all subsets of S. An n-tree is a subset T of $\mathbf{P}(S)$ satisfying the following three conditions:

1. $S\in T$, $\Phi\notin T$.
2. $\{i\}\in T$ for all $i\in S$.
3. If $A, B\in T$ with $A\cap B\neq\Phi$, then $A\subseteq B$ or $B\subseteq A$.

Definition

An n-tree T is a binary tree if for any subset $A\in B$ and $|A|\neq 1$, then

$\exists A_1\in T$ and $A_2\in T$ satisfy $A=A_1\cup A_2$.

Trees generated from hierarchical clustering algorithm are n-trees, especially they are binary trees. For example, the tree in Fig. 1 is an n-tree with $S=\{1,2,3,4,5\}$ and $T=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{2,3\},\{2,3,5\},\{1,4\},S\}$. The subsets $A\in T$ are called nodes of T. In this example, there are nine nodes: {1}, {2}, {3}, {4}, {5}, {2,3}, {2,3,5}, {1,4}, and S. A node is *nontrivial* if and only if it does not contain exactly one object in S nor all the objects in S, i.e. S itself. In this example, there are three nontrivial nodes. In general, there are $2|S|-1$ nodes and $|S|-2$ nontrivial nodes for a binary tree, where $|S|$ is the number of objects in the set S. The tree T given in Fig. 2 is an n-tree with $S=\{1,2,3,4,5,6\}$. Because there is a node $A=\{2,3,5,6\}$ in T with four objects that cannot be partitioned into two mutually exclusive sets, it is not a binary tree.

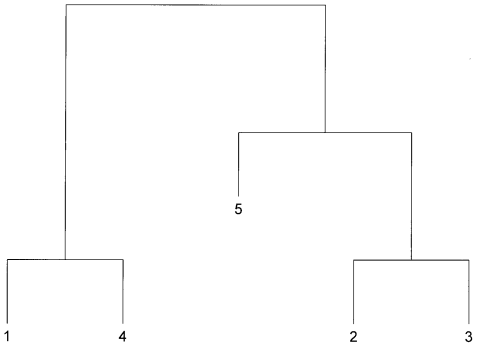


Fig. 1 A binary tree with five objects. Using the notation introduced in the text, this tree has $S=\{1,2,3,4,5\}$ and $T=\{\{1\},\{2\},\{3\},\{4\},\{5\},\{2,3\},\{2,3,5\},\{1,4\},S\}$

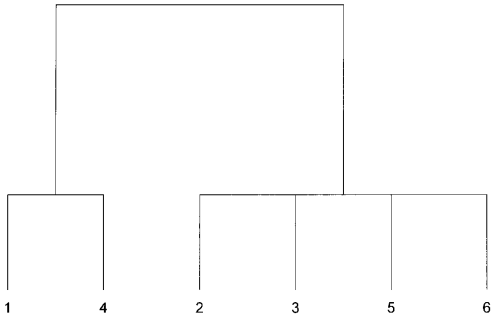


Fig. 2 An n-tree with six objects. This tree is *not* a binary tree

After resampling the original data set m times, we get a set of m trees. Let $\mathbf{T}=\{T_1, T_2,...,T_m\}$ be a collection of these trees. We use the majority rule to extract the common features among these trees.

Definition (Margush and McMorris 1981)

The majority rule of $\mathbf{T}$, denoted by $M(\mathbf{T})$, is the subset of $\mathbf{P}(S)$ where $A\in M(\mathbf{T})$ if and only if $A\in T_i$ for more than half of the T_i.

Now we can define the *consensus* tree as $M(\mathbf{T})$ as follows:

Theorem (Margush and McMorris 1981)

If $\mathbf{T}=\{T_1,T_2,...,T_m\}$ is a collection of n-trees, then $M(\mathbf{T})$ is an n-tree, we call it the consensus tree of $\mathbf{T}$. If we let $d(T_1,T_2)$ be the number elements in the symmetric difference of T_1 and T_2 [that is, $d(T_1,T_2)$ counts the number of elements T_1 and T_2 disagree on], then the set of all n-trees together with d form a metric space, and a nice feature of the consensus tree is that $M(\mathbf{T})$ satisfies the following condition:

$$\sum_{T\in\mathbf{T}} d(M(\mathbf{T}),T)=\min_{M:\mathbf{Missan-tree}}\sum_{T\in\mathbf{T}} d(M,T). \quad (1)$$

Furthermore, if $|\mathbf{T}|$ is odd, then $M(\mathbf{T})$ is the unique n-tree which satisfies the above condition.

This theorem indicates that the consensus tree $M(\mathbf{T})$ is also a hierarchical classification of S. Furthermore, under the metric d, $M(\mathbf{T})$ itself can be considered an optimal consensus classification of a collection of classifications which are estimated from the resampled data sets.

Confidence values in the clustering tree

The consensus tree has clusters that only show up in the majority of n-trees obtained through resampling. In addition to this consensus tree, we can define confidence values in the original clustering tree (Felsenstein 1985) through the resampled trees.

Definition

Let $\mathbf{T}=\{T_1,T_2,...,T_m\}$ be a collection of n-trees. The confidence value for a node A in the original tree T_0 is the percentage of times that A is also a node in T_i, for $i=1,2,...,m$, the set of n-trees obtained through resampling.

Efron et al. (1996) showed that Felsenstein's method provides a reasonable first approximation to the actual levels of the observed clades. They also discussed possible corrections that can be made to better agree with standard ideas of confidence levels and hypothesis testing in statistics.

Results

In this section, we apply the resampling method to two published data sets. The goal is to examine the effects of possible measurement errors on the clusters identified in the two articles in which these two data sets were originally analyzed. Before performing the bootstrap procedure, we must know the variability of the measured gene expression levels. In the first example, this variability can be estimated from the replicated experiments. However, if there is a lack of knowledge on the degree of variability, we may vary the magnitude of variability and examine the effects on the consensus tree constructed as well as the confidence level for each node in the original

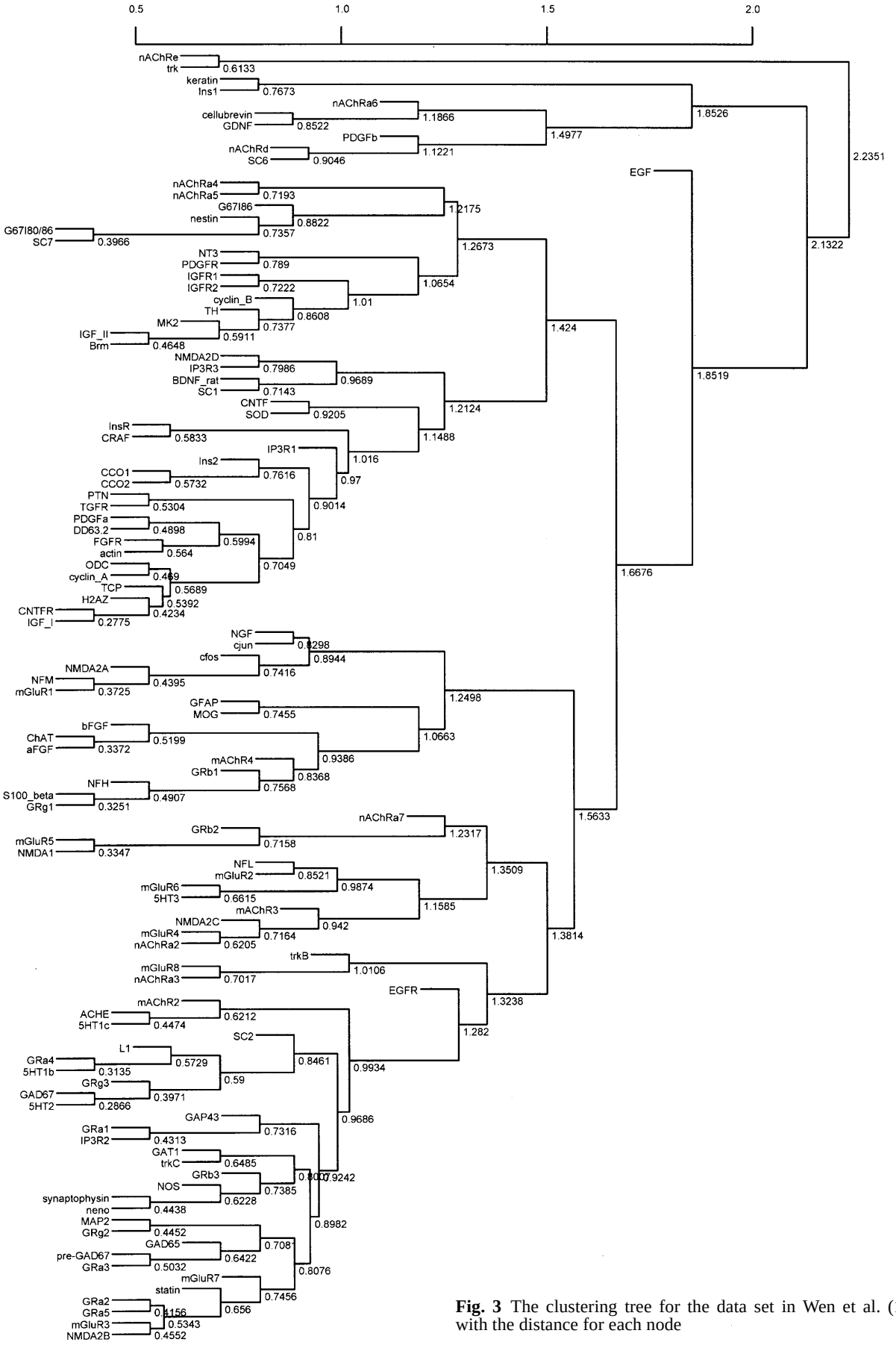


Fig. 3 The clustering tree for the data set in Wen et al. (1998), with the distance for each node

tree. This approach will be utilized in our analysis of the second data set.

Temporal gene expression mapping of central nervous system development

In their study of rat central nervous system development, Wen et al. (1998) used reverse transcription-coupled PCR (RT-PCR) to produce a temporal map of fluctuations in mRNA expression of 112 genes. Using distance matrices for the pairwise comparison of these genes, they distinguished six gene clusters. They noted that genes belonging to distinct functional classes and gene families map to particular expression files. Their data set is available at http://rsb.info.nih.gov/molphysiol/PNAS/GEMtable.html. In the table provided by the authors, there are raw ratio-metric RT-PCR data from triplicate experiments and the standard error for each gene. The gene expression levels used in our analysis were the average from the three replicated experiments and the standard errors were also estimated from the same three experimental measurements. This data set suits our illustration purpose well because of the availability of the estimate of variation in the observed gene expression levels.

Wen et al. (1998) normalized the expression level of each gene at each condition according to the highest expression level for this gene under all nine conditions studied. For each gene, they further derived eight slopes (the difference between gene expression levels) between each pair of consecutive time points. This results in 17 observations (between –1 and 1) for each gene, including the nine observations and eight slopes. Wen et al. (1998) then calculated the Euclidean distance based on these 17 points and used the FITCH program in PHYLIP (Felsenstein 1993) to identify clusters among these 112 genes. In our analysis, we created the same distance matrix as that used by Wen et al. (1998). We then used the hierarchical clustering method with the average-linkage option in tree construction. The resulting clusters are similar to the clusters generated by Wen et al. (1998). This original tree with distances is plotted in Fig. 3. The coefficient of variation (standard error/mean) varies among the 112 genes and nine time points. When the coefficients of variation of all nine time points are averaged for each gene, *nAChRa3* has the largest average value (0.40). The average over all 112 genes is 0.15. Using the standard errors from the data set, we generated 1,000 resampled data sets and the corresponding 1,000 binary trees. The consensus tree is plotted in Fig. 4. We can see that after considering random variations in gene expression levels, only a few nodes in the original clustering tree show up in the majority of the resampled trees. The confidence value for each node in the original tree is shown in Fig. 5. Many nodes have very low confidence values, and this suggests that the clusters identified are not very reliable and extra caution is needed to interpret the biological meaning of the identified clusters. We have also

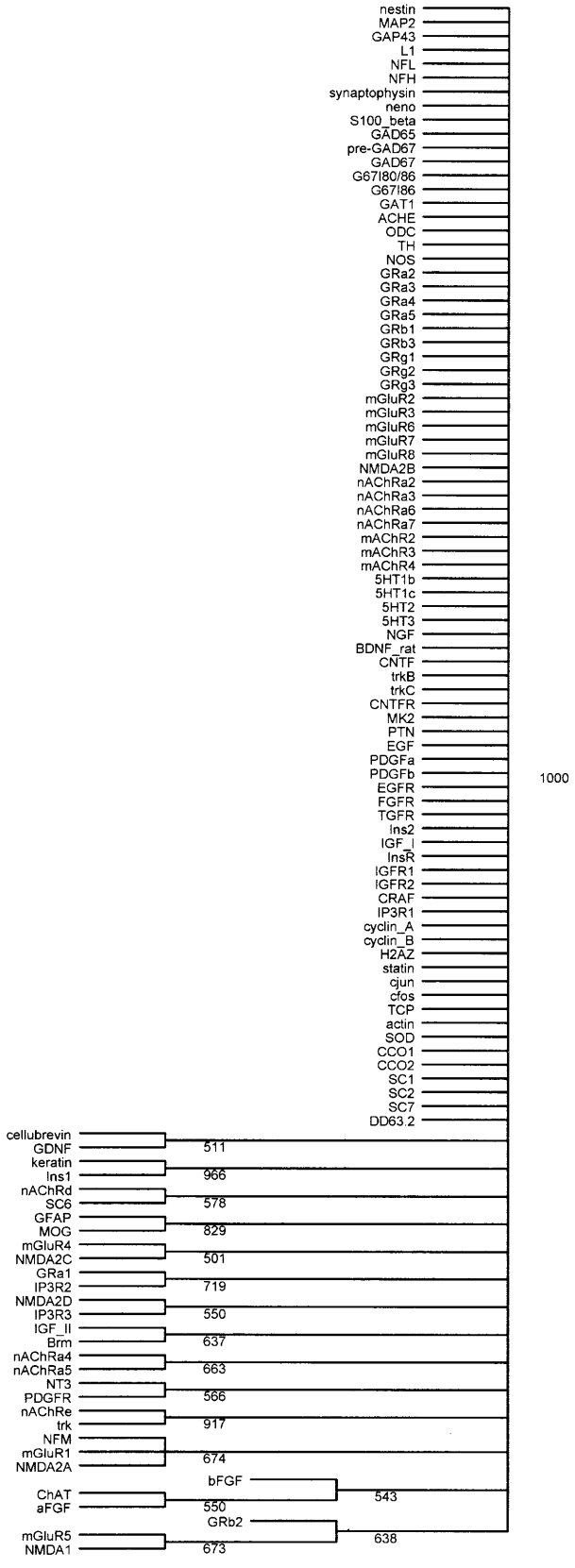


Fig. 4 The majority-rule consensus tree derived from 1,000 resampled trees for the data set in Wen et al. (1998). The number on each node in the tree represents the number of times (out of a total of 1,000) that node was present in the resampled trees

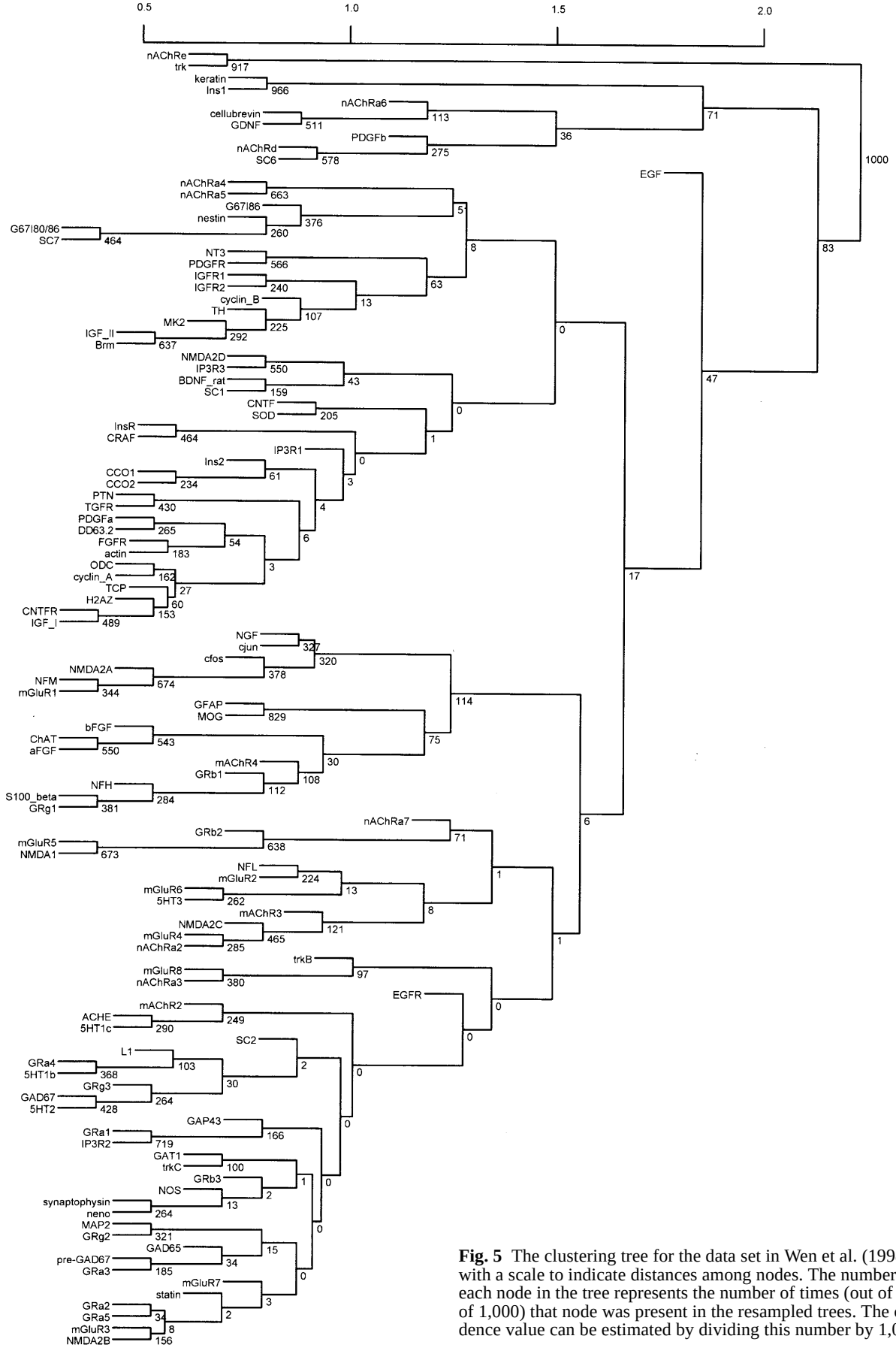


Fig. 5 The clustering tree for the data set in Wen et al. (1998) with a scale to indicate distances among nodes. The number on each node in the tree represents the number of times (out of a total of 1,000) that node was present in the resampled trees. The confidence value can be estimated by dividing this number by 1,000

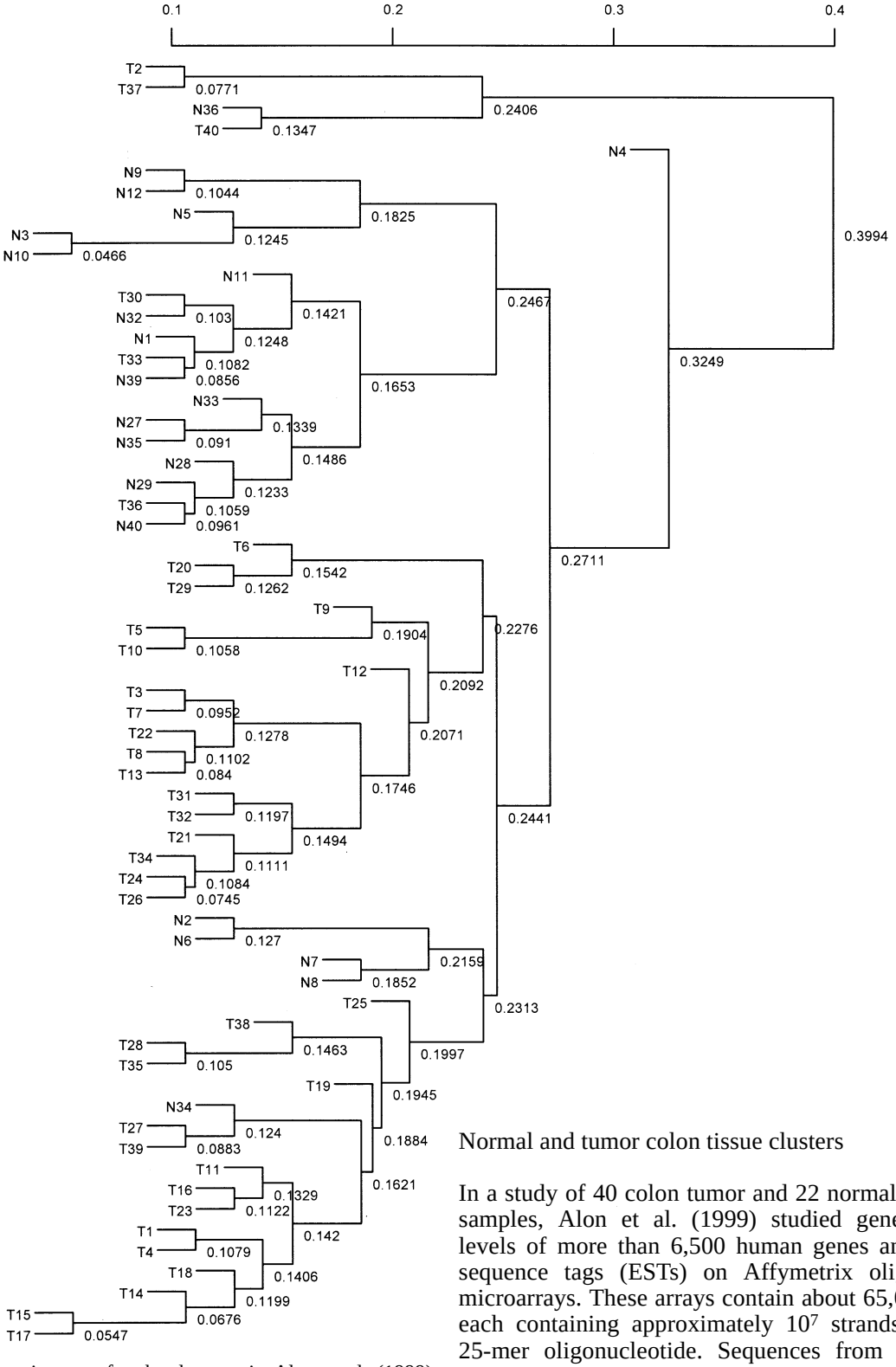


Fig. 6 The clustering tree for the data set in Alon et al. (1999) with the distance for each node

applied the resampling method using the FITCH program in PHYLIP. The results were essentially the same as those obtained using the hierarchical clustering algorithm.

Normal and tumor colon tissue clusters

In a study of 40 colon tumor and 22 normal colon tissue samples, Alon et al. (1999) studied gene expression levels of more than 6,500 human genes and expressed sequence tags (ESTs) on Affymetrix oligonucleotide microarrays. These arrays contain about 65,000 features, each containing approximately 10^7 strands of a DNA 25-mer oligonucleotide. Sequences from about 3,200 full-length human cDNAs and 3,400 ESTs that have some similarity to other eukaryotic genes are represented on a set of chips (Alon et al. 1999). They did two-way clustering in their analysis: classified genes into functional groups and classified tissues based on their expression similarity. Alon et al. (1999) found coherent patterns of genes whose expression is correlated, and ar-

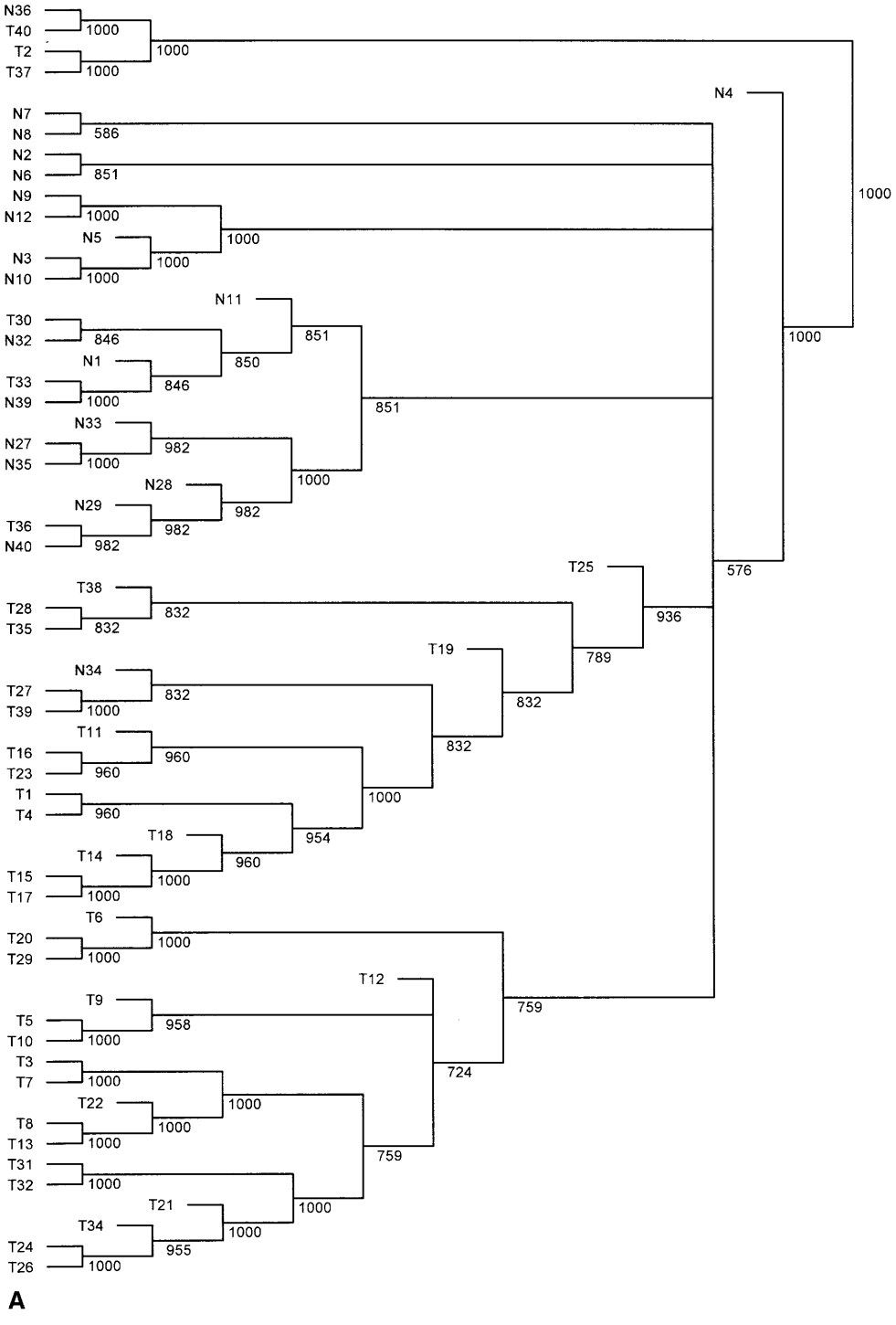
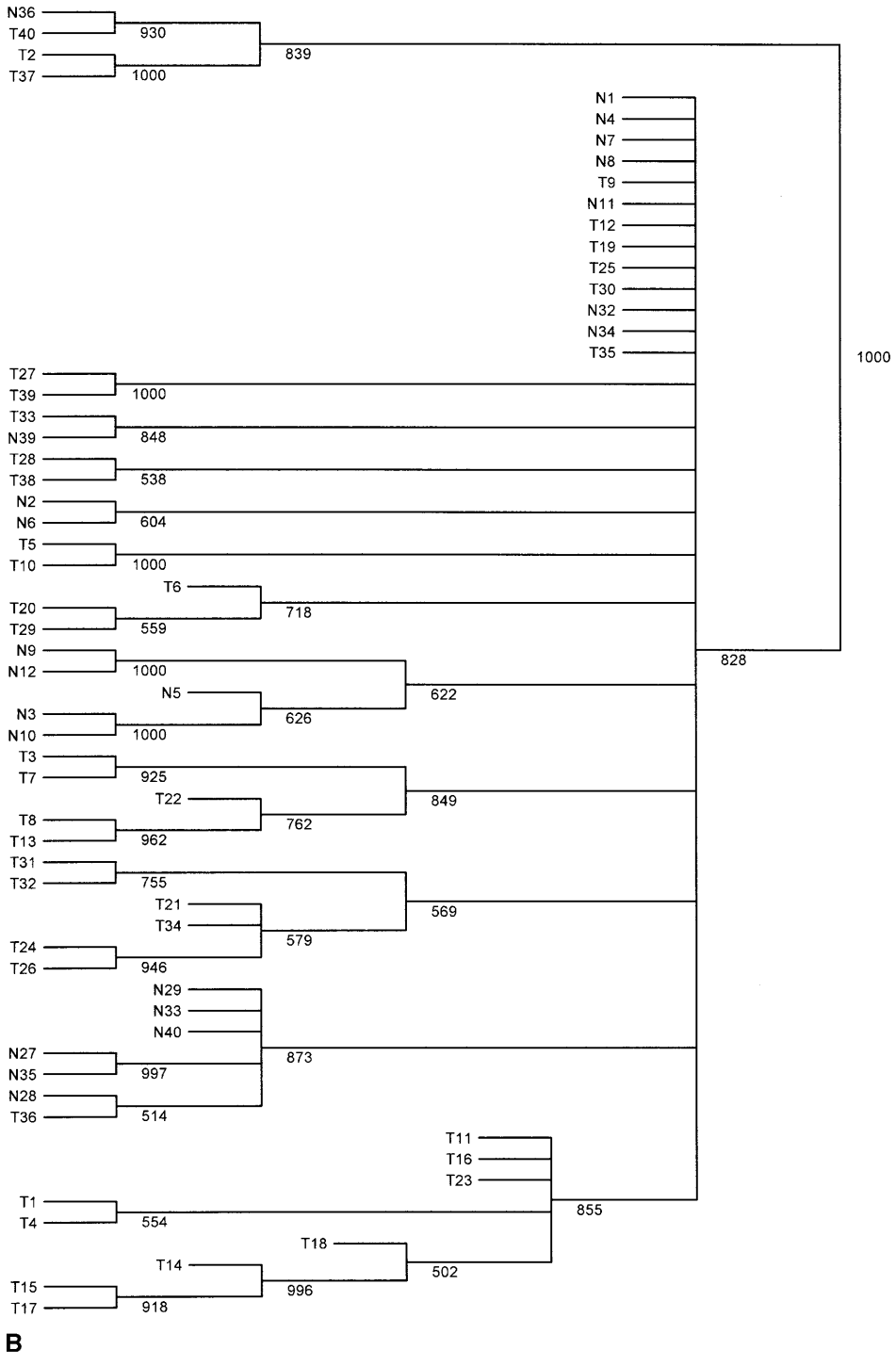


Fig. 7A–D The majority-rule consensus tree derived from 1,000 resampled trees for the data set in Alon et al. (1999) when the value of f is set at different levels. The number on each node in the tree represents the number of times (out of a total of 1,000) that node was present in the resampled trees. **A** $f=0.01$; **B** $f=0.10$; **C** $f=0.20$; **D** $f=0.40$

gued that this suggests a high degree of organization underlying gene expression in these tissues. Two thousand genes with the highest minimal intensity across the samples are available from the web (http://www.molbio.princeton.edu/colondata). Among the 2,000 genes, three ESTs (*HSAC07*, *UMGAP*, and *i*) were listed four times with the same name and identical expression levels. After removing nine duplicates for these three ESTs, we got a data matrix with 62 columns and 1,991 rows. Every column in this matrix represents a tumor or normal tissue sample, and every row corresponds to a gene or an EST. Alon et al. (1999) used the deterministic-annealing algorithm in their analysis. When we used the correlation coefficient as the similarity measure and the average-linkage hierarchical clustering method to classify the normal and tumor colon tissues, we obtained essentially the same tree structure. This original tree with distances is shown in Fig. 6.

Fig. 7B



Unlike the data by Wen et al. (1998), where the standard error for the average gene expression was estimated from triplicates, there was no indication of the level of variation in the observed expression level for each gene in each tissue sample in the data set. We assessed the effects of measurement errors on the resulting tree structure by varying the standard errors to generate the resampled data sets. Our assumption was that the standard error s_{ij} for the *i*th EST in the *j*th condition is proportional to the observed signal x_{ij}, i.e. $s_{ij}=f\times x_{ij}$, where f can be varied at many levels. To be less arbitrary, we began with small f values in the resampling and increased it in relatively small increments. For each fixed f value, we resampled the data set and constructed a binary tree 1,000 times. The consensus trees produced with different f are plotted in Fig. 7A–D for f values ranging from 0.01, 0.1, 0.2, to 0.4. The corresponding confidence values for the nodes in the original tree were estimated and shown in Fig. 8A–D.

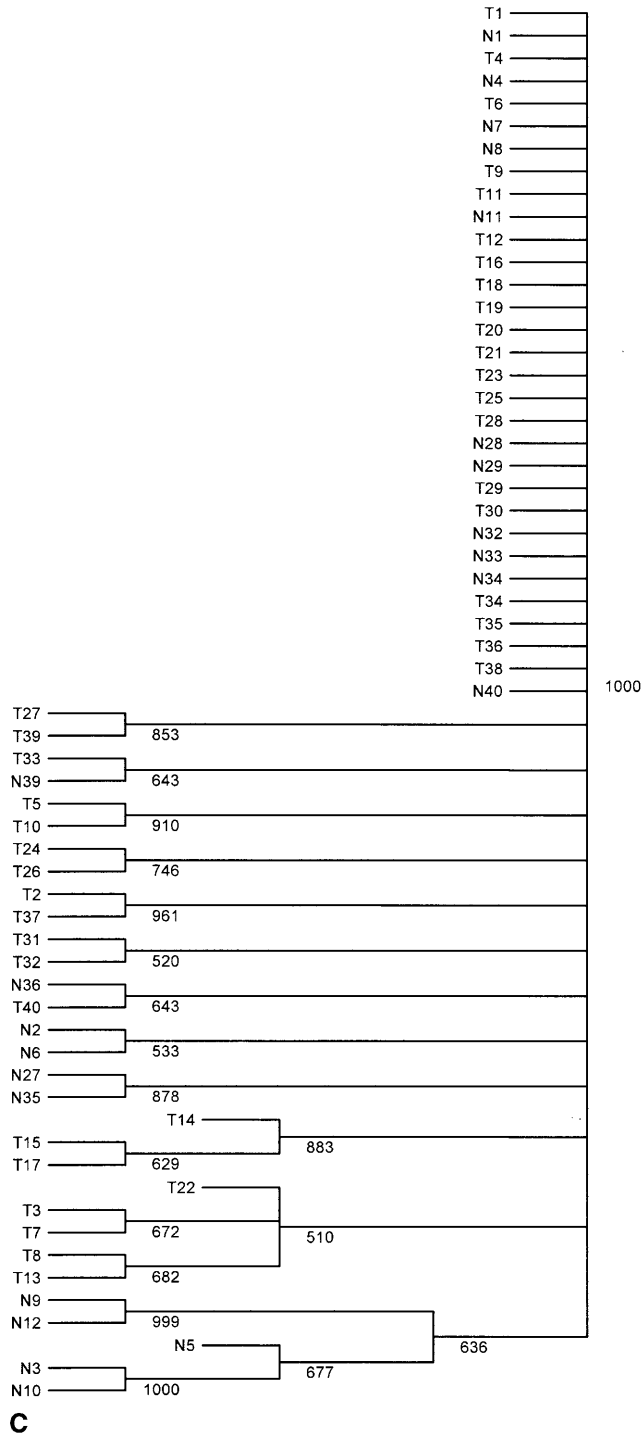


Fig. 7C

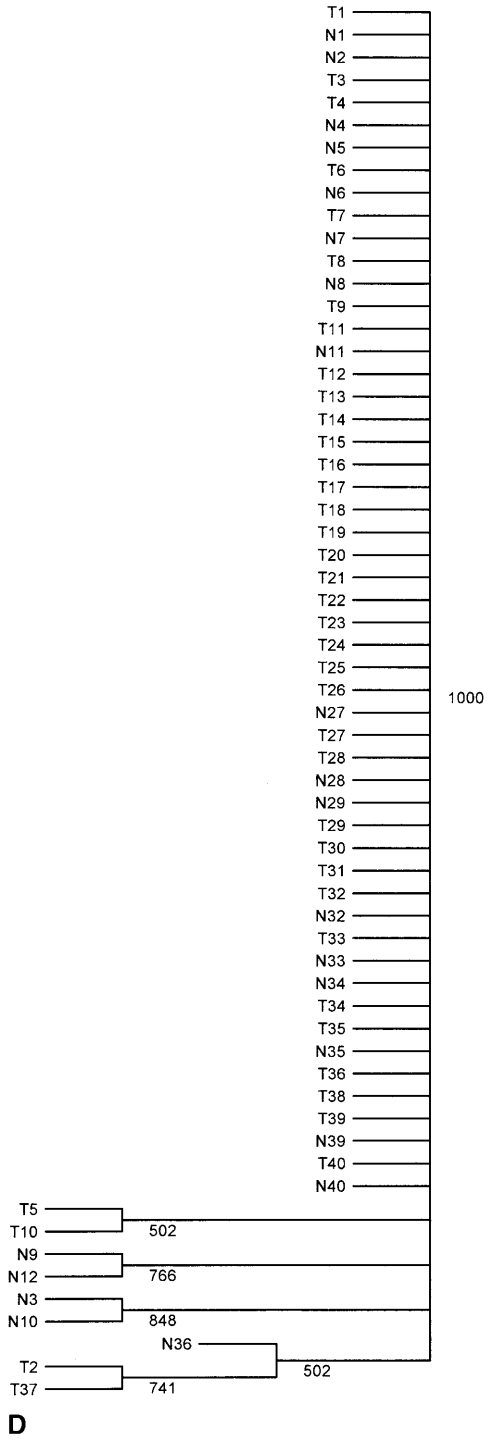


Fig. 7D

When f is relatively small (f<0.10), it can be seen that most clusters have only tumor tissues or normal tissues, with the exception of tissue samples labeled *n36*, *T33*, and *T36*. These samples have abnormal muscle indexes described by Alon et al. (1999). As we increased the magnitude of errors in the observed gene expression levels, this tree became less reliable. When f>0.70, the consensus tree had only the trivial nodes. Therefore, it is useful to get the whole spectrum of outcomes for possible variation in the data by varying the magnitude of f values. As discussed above, we can measure the difference between the original tree and the consensus tree by $d(T_0, M)$ defined as the number of elements the original tree T_0 and the consensus tree M disagree on. For example, through 1,000 resamplings, $d(T_0, M)$ is 33 when f=0.1 and is 42 when f=0.2.

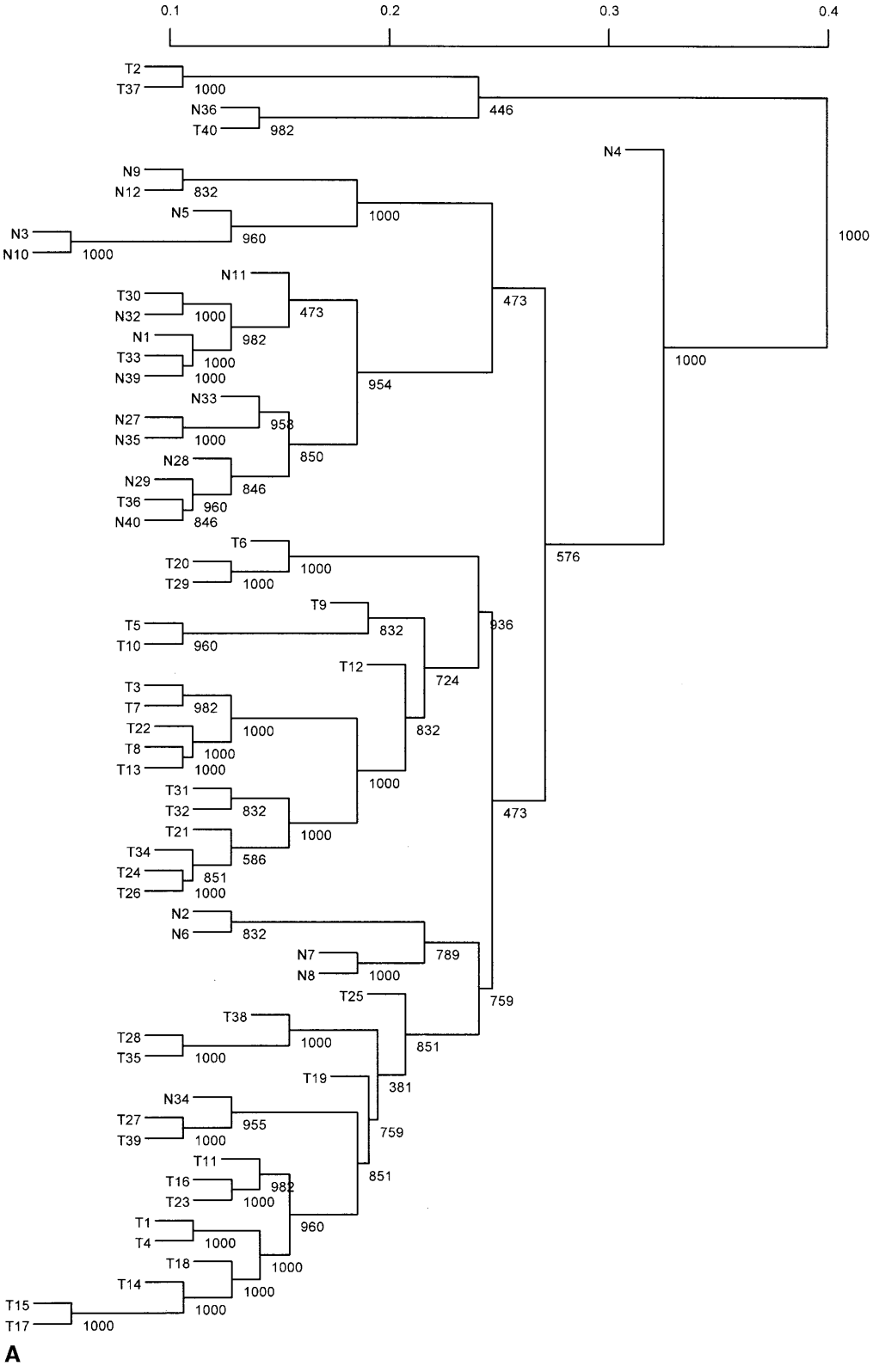
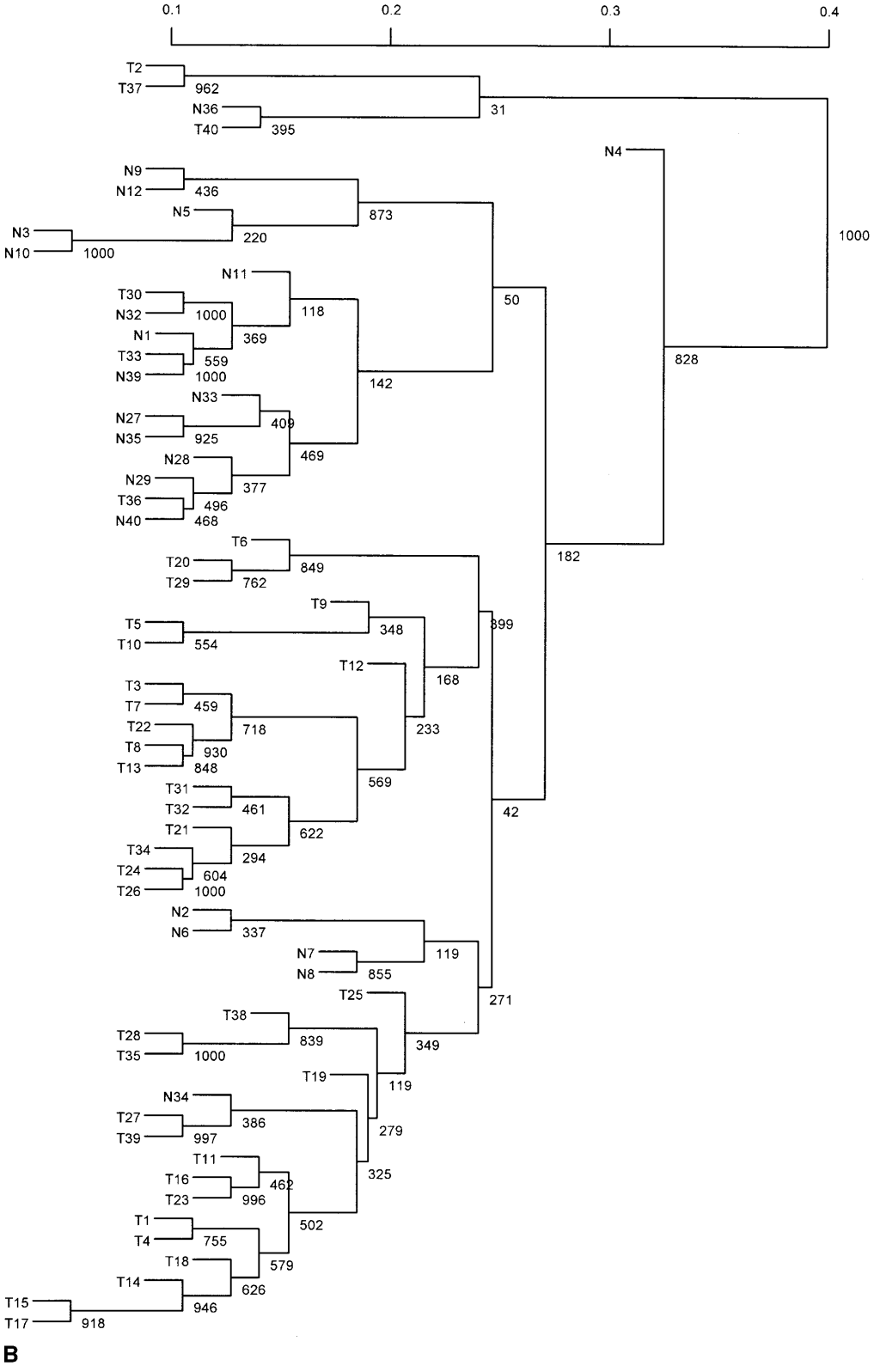


Fig. 8A–D The clustering tree for the data set in Alon et al. (1999) when the value of f is set at different levels. The scale indicates distances among nodes. The number on each node in the tree represents the number of times (out of a total of 1,000) that node was present in the resampled trees. The confidence value can be estimated by dividing this number by 1,000. **A** *f=0.01*; **B** *f=0.10*; **C** *f=0.20*; **D** *f=0.40*

Discussion

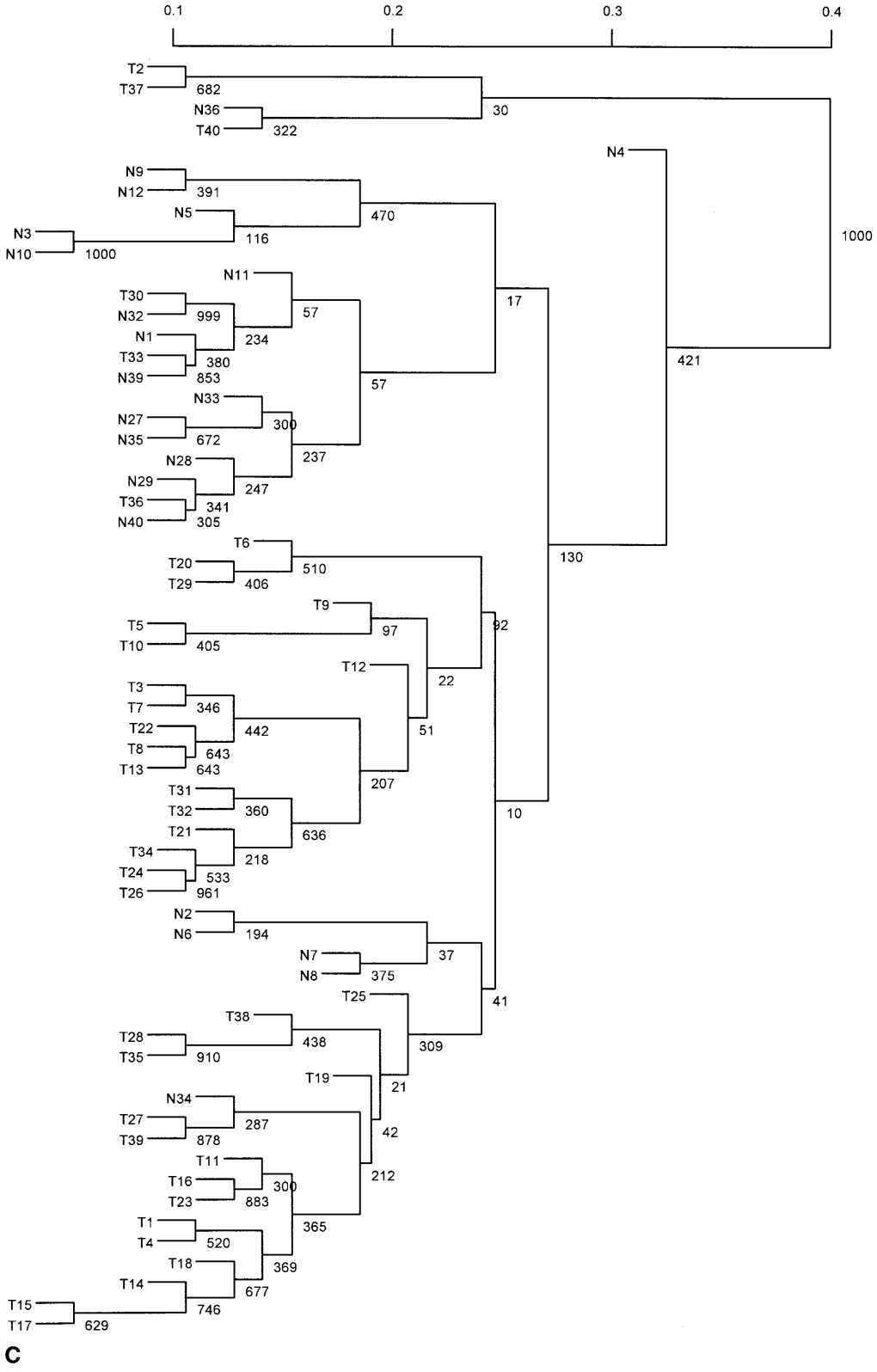
With microarray-based genomic surveys becoming more feasible, numerous methods have been developed to mine the potential information in these massive data sets. One area that has often been ignored in microarray data analysis is the variability in gene expression levels. After many gene clusters are identified from large-scale gene expression data, we must be able to investigate the reliability of the observed clusters. In this article, we have

Fig. 8B



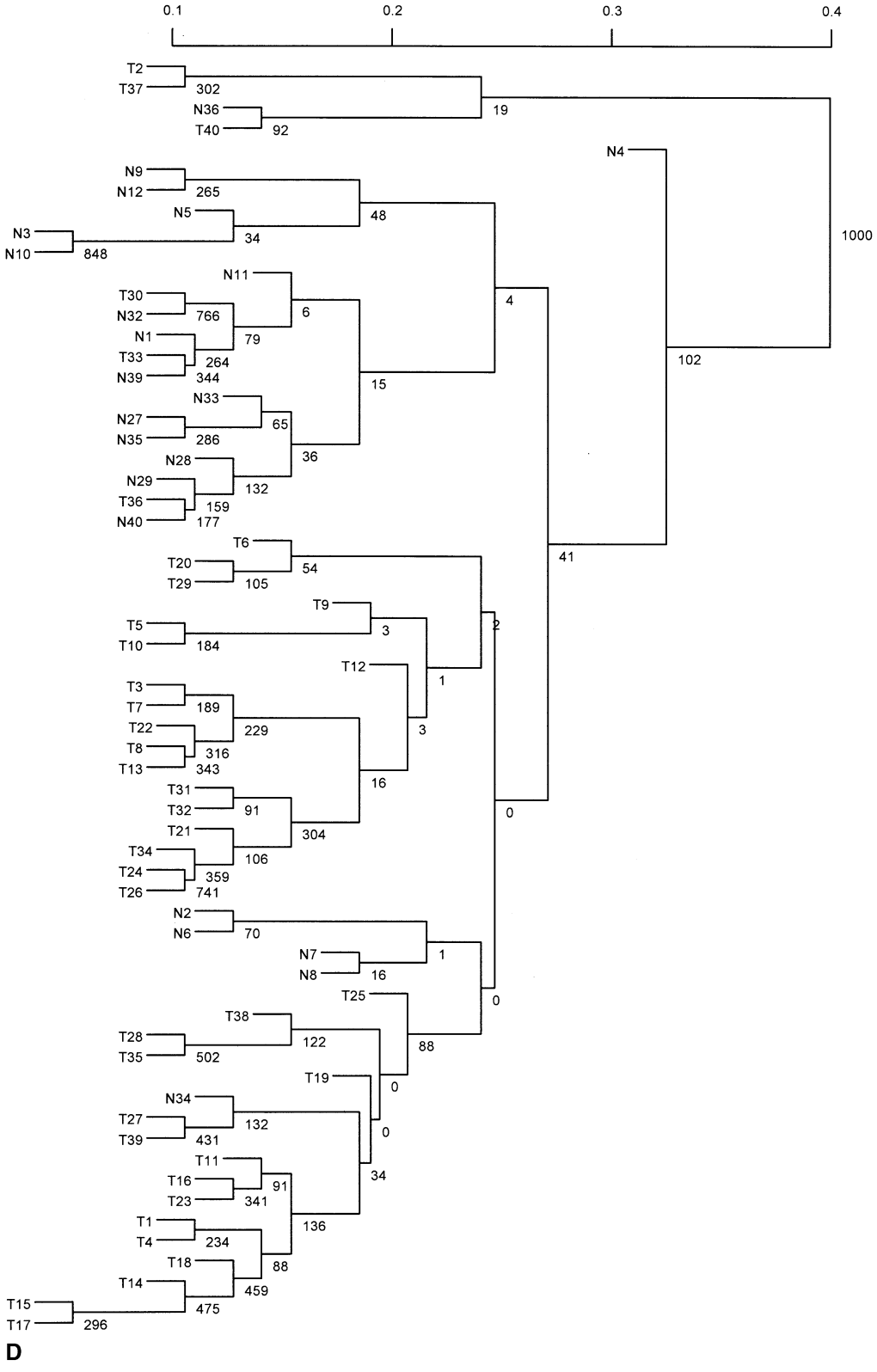
proposed a resampling method that fits this need. To apply this method, some estimate of the uncertainty in gene expression measurement is needed. Such estimates can either be obtained from repeated experiments or from pilot studies on the variation of the measurements. For each resampling, a set of "new" observations are generated by replacing the true observation for each gene under each condition with a random variable sampled using the observed expression level and the estimated uncertainty in gene expression measurement. Each "new" data

Fig. 8C



set leads to a "new" tree describing the relationships among the genes. A consensus tree can be constructed from these resampled trees using the majority rule. If there is a need to study a particular gene cluster, the significance of this cluster can be estimated using the percentile method and a confidence value can be assigned to each node in the original clustering tree. Our proposal is in spirit similar to the bootstrap method proposed by Felsenstein (1985) for phylogenetic analysis. We have developed a computer program to cluster genes, resam-

Fig. 8D



ple data, construct consensus tree, and assess the confidence value of each node in the tree. For the analysis of Alon's data set with 62 conditions and 1,991 genes, it took less than 10 min for 1,000 resamplings on a personal computer with Windows NT system having one Pentium III 500 MHz processor and 256 megabytes memory.

If a more accurate estimation of a statistic of interest, such as the confidence value, is desired, we must increase the number of resamplings. However, from our

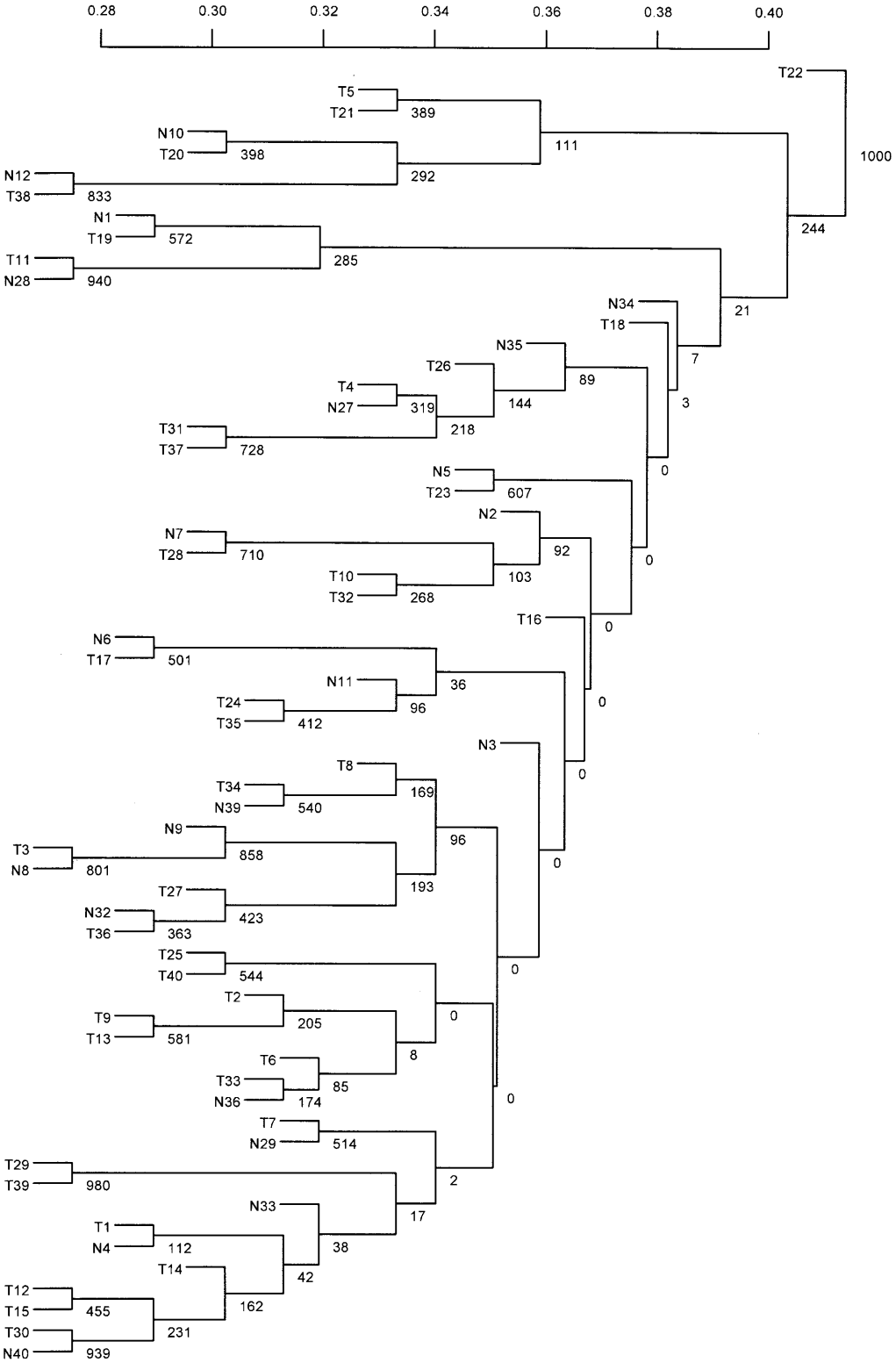


Fig. 9 Clustering tree for the randomly permuted data set in Alon et al. (1999) when the value of f is set at 0.10. The scale indicates distances among nodes. The number on each node in the tree represents the number of times (out of a total of 1,000) that node was present in the resampled trees. The confidence value can be estimated by dividing this number by 1,000

experience, resampling more than 1,000 times is usually not necessary for most practical purposes. This is consistent with the observations by Efron et al. (1996). Although this resampling method is not biased (Efron et al. 1996), the estimate of confidence values can be corrected to better correspond to the standard ideas of confidence levels and hypothesis testing. Detailed discussion on this method was described in Efron et al. (1996). The basic idea is to bootstrap on the bootstrap samples. Although this bias-correction method is very

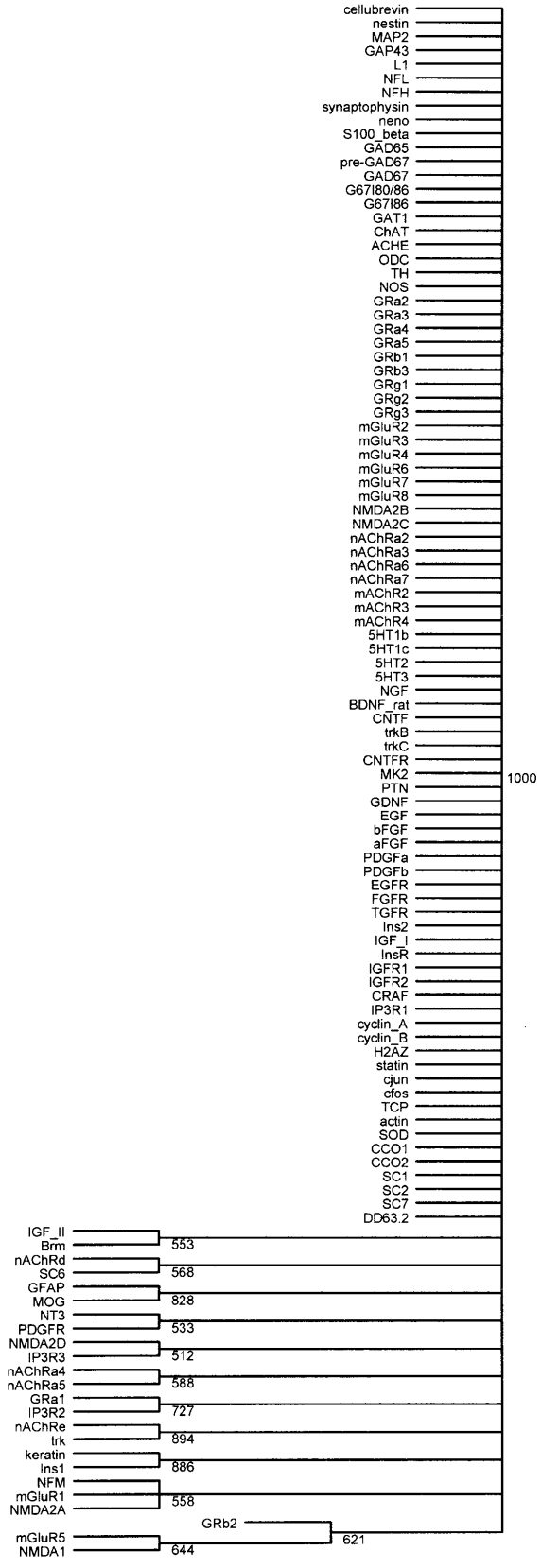


Fig. 10 The majority-rule consensus tree derived from 1,000 resampled trees for the data set in Wen et al. (1998) when the measurements were assumed to follow log-normal distributions. The number on each node in the tree represents the number of times (out of a total of 1,000) that node was present in the resampled trees

appealing, a large amount of extra computation is needed.

To understand the structure of the consensus tree when there are no inherent clusters in the data, the initial gene expression measurements can be randomized across all conditions (samples) or across all genes and then be analyzed using the same procedure. When we apply this method to the normal and tumor colon tissue data (Alon et al. 1999), for each gene, we randomly permuted the gene expression levels among the samples independently from other genes. This procedure will break any clustering in the data. One such example is shown in Fig. 9, where the original tree constructed from the permuted data is shown with confidence values estimated through 1,000 resampings when the error rate was set at f=0.1. It is apparent that the confidence values are much lower than those in Fig. 7B. In addition, the normal and tumor colon tissue samples are no longer clustered. The structure from this permuted tree further suggests that the clusters observed in the original tree represent true clusters.

To investigate how sensitive the analysis is to the normality assumption, we performed simulations assuming gene expression measurements follow log-normal distributions. That is, the log transformed gene expression levels follow normal distributions. For the data set in Wen et al. (1998), we select the parameters in the log-normal distribution so that the mean and standard deviation of the log-normal distribution are equal to the observed mean x_{ij} and the observed s_{ij}, respectively. The consensus tree and the confidence values are estimated by using the same procedure as described in the previous section. When we "bootstrap" from the log-normal distribution for each gene under each condition, the obtained consensus tree is shown in Fig. 10, and the original tree with the estimated confidence values are shown in Fig. 11. They are all essentially the same as those using the normality assumption for this particular data set.

In summary, we have proposed a parametric bootstrap resampling method to incorporate information on variations in gene expression levels to assess the reliability of gene clusters identified from large-scale gene expression data. Our approach can distinguish gene clusters with high confidence values from those with low confidence values. Although our discussion has focused on the hierarchical clustering methods, this resampling method can also be combined with other methods (e.g., *k*-means and self-organizing-maps) to prioritize gene clusters according to strength of evidence in the data for further biological studies of the functions of these genes. The generalizations of our resampling methods to other clustering algorithms will be reported in a future study.

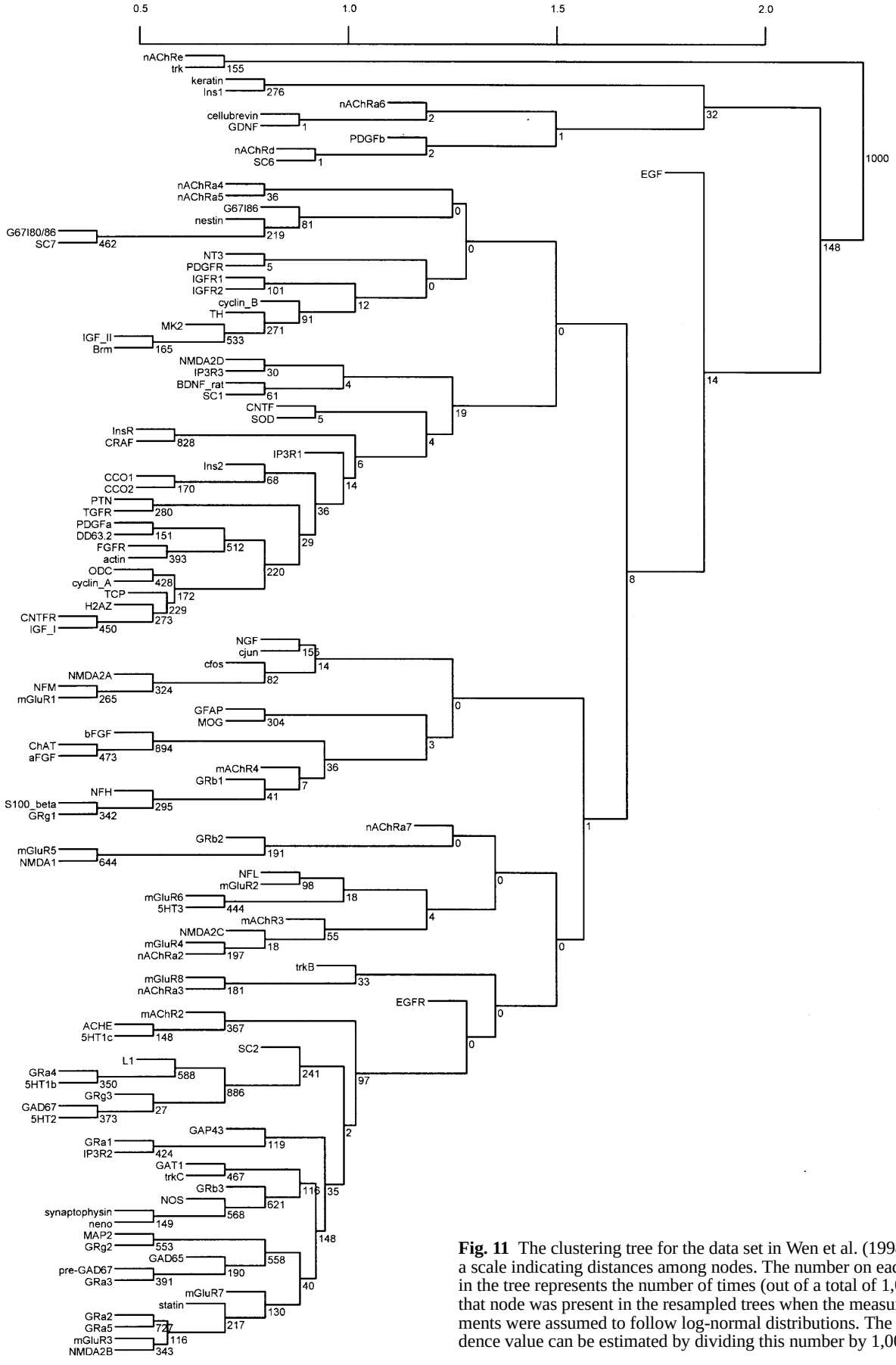


Fig. 11 The clustering tree for the data set in Wen et al. (1998) with a scale indicating distances among nodes. The number on each node in the tree represents the number of times (out of a total of 1,000) that node was present in the resampled trees when the measurements were assumed to follow log-normal distributions. The confidence value can be estimated by dividing this number by 1,000

Acknowledgements We would like to thank Shuanglin Zhang, Jinming Li, and two anonymous referees for their helpful comments. This work was supported in part by grants MG59507 and HD36834 from the National Institutes of Health and Research Grant FY98–0752 from the March of Dimes Birth Defects Foundation.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745–6750

Anbazhagan R, Tihan T, Borman DM, Johnston JC, Saltz JH, Wegering A, Piantadosi S, Gabrielson E (1999) Classification of small cell lung cancer and pulmonary cacinnoid by gene expression profiles. Cancer Res 59:5119–5122

Ben-Dor A, Yakhini Z (1999) Clustering gene expression patterns. J Comput Biol 6:281–287

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler D (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 97:262–267

Debouck C, Goodfellow PN (1999) DNA microarrays in drug discovery and development. Nature 21:48–55

DeRisi J, Oenaland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen YD, Su YA, Trent JM (1996) Use of a cDNA microarray to analyze gene expression patterns in human cancer. Nat Genet 14:457–460

D'Haeseleer P, Wen X, Fuhrman S, Somogyi R (1998) Mining the gene expression matrix: inferring gene relationships from large scale gene expression data. In: Paton RC, Holcombe M (eds) Information processing in cells and tissues. Plenum, New York

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, New York

Efron B, Halloran E, Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci USA 93:13429–13434

Eisen MB, Spellman PT, Brown PO, Bostein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package), version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligluri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537

Hacia JG (1999) Resequence and mutational analysis using oligonucleotide microarrays. Nat Genet 21:42–47

Hartigan JA (1975) Clustering algorithms. Wiley, New York

Heyer LJ, Kruglyak S, Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. Genome Res 9:1106–1115

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42:182–192

Hilsenbeck SG, Friedrichs WE, Schiff R, O'Connell P, Hansen RK, Obserne CK, Fuqua SAW (1999) Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. J Natl Cancer Inst 91:453–459

Khan J, Saal LH, Bittner ML, Chen YD, Trent JM, Meltzer PS (1999) Expression profiling in cancer using cDNA microarrays. Electrophoresis 20:223–229

Margush T, McMorris FR (1981) Consensus *n*-trees. Bull Math Biol 43:239–244

Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai HY, Bassett DE, Hartwell LH, Brown PO, Friend SH (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4:1293–1301

Mir KU, Southern EM (1999) Determining the influence of structure on hybridization using oligonucleotide arrays. Nat Biotechnol 17:788–792

Raychaudhuri S, Stuart JM, Altman RB (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 5:452–463

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

Törönen P, Kolehmainen M, Wong G, Castrén E (1999) Analysis of gene data using self-organizing maps. FEBS Lett 451:142–146

Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R (1998) Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci USA 95:334–339