Biostatistics (2008), **9**, 3, *pp.* 467–483
doi:10.1093/biostatistics/kxm046
Advance Access publication on December 18, 2007

Complementary hierarchical clustering

GEN NOWAK*

Department of Statistics, Stanford University, Stanford, CA 94305, USA
gnowak@stanford.edu

ROBERT TIBSHIRANI

Department of Health Research and Policy and Department of Statistics,
Stanford University, Stanford, CA 94305, USA

SUMMARY

When applying hierarchical clustering algorithms to cluster patient samples from microarray data, the clustering patterns generated by most algorithms tend to be dominated by groups of highly differentially expressed genes that have closely related expression patterns. Sometimes, these genes may not be relevant to the biological process under study or their functions may already be known. The problem is that these genes can potentially drown out the effects of other genes that are relevant or have novel functions. We propose a procedure called complementary hierarchical clustering that is designed to uncover the structures arising from these novel genes that are not as highly expressed. Simulation studies show that the procedure is effective when applied to a variety of examples. We also define a concept called relative gene importance that can be used to identify the influential genes in a given clustering. Finally, we analyze a microarray data set from 295 breast cancer patients, using clustering with the correlation-based distance measure. The complementary clustering reveals a grouping of the patients which is uncorrelated with a number of known prognostic signatures and significantly differing distant metastasis-free probabilities.

Keywords: Hierarchical clustering; Microarray; Principal components; Relative gene importance.

1. INTRODUCTION

Hierarchical clustering algorithms are very useful tools for analyzing microarray data. They provide a very simple and appealing way of displaying the organizational structure of the data using a tree diagram called a dendrogram. An example of the application of hierarchical clustering algorithms to microarray data is given in Eisen *and others* (1998). To motivate the procedure described in this paper, let us suppose we are clustering RNA samples based on their gene expression profiles. What is often observed is that the clustering pattern is dominated by a small group of highly differentially expressed genes with very related expression patterns. These “strong” genes can potentially drown out “weaker” genes that are not as highly expressed. A problem arises when these weaker genes are responsible for structures among the data that have important biological relevance as this information cannot be discerned from the clustering pattern. The goal of the complementary hierarchical clustering procedure is to uncover the structure arising from these weaker genes.

*To whom correspondence should be addressed.

Before delving into the details of this procedure, we will give a very brief review of clustering methods. Further details on cluster analysis and various clustering methods can be found in Hastie *and others* (2001) and Gordon (1999). Cluster analysis is an unsupervised learning procedure with the goal of grouping data into clusters, with members within a cluster being closer to each other than to members outside that cluster. In order to quantify how close one data point is to another, a distance measure is required. A typical distance measure used with microarray data is one minus the correlation between the gene expression profiles of RNA samples.

Most clustering methods fall into 2 categories: partitioning methods and hierarchical methods. Partitioning methods try to find the most optimal grouping of the data into a predetermined number of clusters. A well-known example is the K-means algorithm. Hierarchical methods will produce clusters of a hierarchical nature. The lowest level of the hierarchy consists of each individual data point, and at each level, the clusters are obtained by merging clusters from the previous lower level. As mentioned above, the hierarchical nature enables the clustering pattern to be displayed as a dendrogram. Hierarchical methods also require the definition of an intercluster distance measure. One example of such a measure is the maximum pairwise distance between an element from one cluster and an element from the other cluster. There exist many hierarchical clustering algorithms, and they can differ in aspects such as the intercluster distance measure or whether the hierarchy is constructed in a top–down or a bottom–up manner.

Some more recent work on the applications of clustering to microarray data include model-based clustering, spectral clustering, and biclustering. Model-based clustering methods are based on the assumption that the microarray data are generated from some underlying probabilistic model. A common example is to assume that the gene expression profiles of the RNA samples are generated from a mixture of normal distributions, with each component of the mixture corresponding to a cluster. Other examples of model-based clustering can be found in Yeung *and others* (2001), Pan *and others* (2002), and Pan (2006). Spectral clustering is a technique where the microarray data are treated as a graph with a set of vertices and edges (with corresponding weights) and attempts to find an optimal partition of the vertices. The problem is solved via an eigenvector algorithm involving the matrix of weights. Applications to microarray data are given in Higham *and others* (2007), Kluger *and others* (2003), and Xing and Karp (2001). Biclustering methods attempt to simultaneously cluster both the samples and the genes with the goal of finding "biclusters," subsets of genes that seem to be closely related for a given subset of samples. For more details on biclustering, including both model-based and spectral approaches, see Cheng and Church (2000), Madeira and Oliveira (2004), Turner *and others* (2005), and Sheng *and others* (2003).

Complementary hierarchical clustering is a procedure that can be applied using any hierarchical clustering algorithm, as the only requirement is the ability of the clustering pattern to be represented as a dendrogram. The main idea behind this procedure is to use the information contained in the dendrogram to remove the main structural features from the data and subsequently uncover the structure arising from the weaker genes. The procedure can be broken down into 3 steps. First, we perform an "initial" clustering on the original data. Second, the original data are modified, and third, we perform a "complementary" clustering on the modified data. The key to uncovering the structure lies in the modification of the original data in the second step. In this paper, we describe in detail the complementary hierarchical clustering procedure. The procedure is motivated and outlined in Section 2, with computational details included in Section 3. Some simulation studies and an extension to the procedure are given in Sections 4 and 5, respectively. An analysis of a breast cancer data set is presented in Section 6, and a discussion is given in Section 7.

2. Complementary hierarchical clustering

2.1 *Motivation*

Consider a simple example of microarray data, shown in panel (a) of Figure 1, consisting of 4 RNA samples and 2 genes. Applying a hierarchical clustering algorithm to the data, we would obtain the

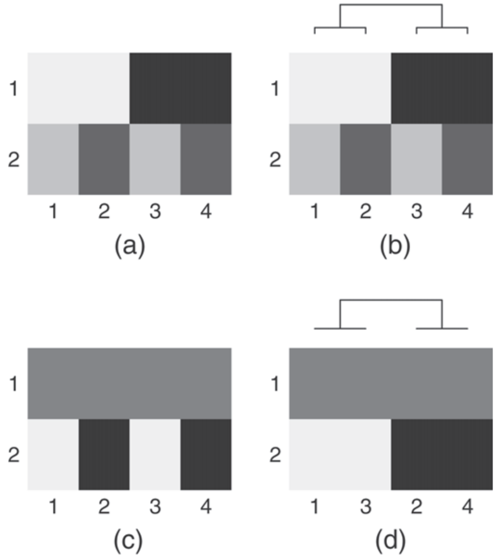


Fig. 1. Motivating example of microarray data with 4 RNA samples and 2 genes.

dendrogram displayed in panel (b). It's clear that the clustering pattern is dominated by Gene 1. However, although we know that samples 1 and 3 are related based on their expression levels of Gene 2, it is impossible to discern this relationship from the dendrogram. Ideally, we would like to modify the data by removing the effect of Gene 1. An example of such modified data is shown in panel (c). Thus, applying a hierarchical clustering algorithm to these modified data would uncover the structure arising from Gene 2, generating the dendrogram in panel (d).

2.2 *Complementary hierarchical clustering procedure*

The complementary hierarchical clustering procedure essentially implements panels (c) and (d) of Figure 1: modifying the data and uncovering that structure arising from the weaker genes. To establish some notation, we will denote the expression data by a $p \times n$ matrix $\boldsymbol{X}$ corresponding to p genes and n samples.

1. Perform the initial hierarchical clustering on the data $\boldsymbol{X}$.
2. Given the dendrogram of the initial hierarchical clustering, define a random variable H that is uniformly distributed on $(0, h)$, where h is the total height of the dendrogram.
   Suppose we cut the dendrogram at height H, resulting in a natural grouping of the samples.
   Define $z^{(H)}$ to be the corresponding vector of group labels arising from the cut at H.
   Define $\boldsymbol{R}^{(H)}$ to be the matrix of residuals corresponding to the linear regression of each row of $\boldsymbol{X}$ onto $z^{(H)}$.
   Define $\boldsymbol{X}' = E(\boldsymbol{R}^{(H)})$, where the expectation is taken over the random variable H.
3. Perform the complementary hierarchical clustering on the modified data $\boldsymbol{X}'$.

A key observation is that the residuals $\boldsymbol{R}^{(H)}$ remain constant for all H falling between 2 consecutive merges in the dendrogram. Therefore, an equivalent representation of $\boldsymbol{X}'$ is

$$\boldsymbol{X}' = \sum_{m=1}^{n-1} \boldsymbol{R}^{(h_m)} \frac{h_m - h_{m-1}}{h}, \tag{2.1}$$

where h_m is the height of the mth merge (with $h_0 = 0$). For notational purposes, we will assume that a cut at h_m is equivalent to a cut at any point between h_m and h_{m-1}. We see from this representation that more weight is given to residuals when there is a large jump between consecutive merges. This seems intuitive as the distance between consecutive merges should somewhat reflect the strength of a particular grouping in the initial clustering. Thus, the complementary hierarchical clustering procedure will consider many groupings present in the initial clustering, while at the same time focusing more on removing the structures arising from the strong genes.

2.3 *Theoretical example*

The following concrete example demonstrates in detail the steps of the procedure. Taking the example shown in Figure 1, suppose our microarray data are given by the following matrix:

$$\boldsymbol{X} = \begin{bmatrix} c & c & -c & -c \\ d & -d & d & -d \end{bmatrix},$$

where $c \gg d$. We will apply complete linkage agglomerative hierarchical clustering to these data, with squared Euclidean distance as the distance measure. Complete linkage indicates that the maximum pairwise distance is used as the intercluster distance. Letting $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_4$ denote the 4 columns of $\boldsymbol{X}$, we have the following distance relationships:

$$\begin{aligned} d(\boldsymbol{x}_1, \boldsymbol{x}_2) &= d(\boldsymbol{x}_3, \boldsymbol{x}_4) = 4d^2, \\ d(\boldsymbol{x}_1, \boldsymbol{x}_3) &= d(\boldsymbol{x}_2, \boldsymbol{x}_4) = 4c^2, \\ d(\boldsymbol{x}_1, \boldsymbol{x}_4) &= d(\boldsymbol{x}_2, \boldsymbol{x}_3) = 4(c^2 + d^2). \end{aligned}$$

Thus, the first 2 merges occur at height $h_1 = h_2 = 4d^2$, where we simultaneously merge $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ into one cluster and $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$ into another. The only choice for the third merge is to merge these 2 clusters, at height $h_3 = 4(c^2 + d^2)$. The dendrogram for the initial clustering is displayed in Figure 2.

Applying (2.1) to calculate $\boldsymbol{X}'$ and noting that $h_1 = h_2$, we have

$$\begin{aligned} \boldsymbol{X}' &= \boldsymbol{R}^{(h_1)} \frac{h_1 - h_0}{h} + \boldsymbol{R}^{(h_2)} \frac{h_2 - h_1}{h} + \boldsymbol{R}^{(h_3)} \frac{h_3 - h_2}{h} \\ &= \boldsymbol{R}^{(h_1)} \frac{d^2}{c^2 + d^2} + \boldsymbol{R}^{(h_3)} \frac{c^2}{c^2 + d^2}. \end{aligned}$$

For any cut between $h_0 = 0$ and $h_1 = 4d^2$, we generate the vector of labels $\boldsymbol{z}^{(h_1)} = (1, 2, 3, 4)$. It follows immediately that $\boldsymbol{R}^{(h_1)}$, the residual matrix corresponding to the linear regression of each row of $\boldsymbol{X}$ onto $\boldsymbol{z}^{(h_1)}$, is exactly zero. For any cut between $h_2 = 4d^2$ and $h_3 = 4(c^2 + d^2)$, we generate the vector of labels $\boldsymbol{z}^{(h_3)} = (1, 1, 2, 2)$. Letting $\hat{\boldsymbol{X}}^{(h_3)}$ denote the fitted values of the linear regression of each row of $\boldsymbol{X}$ onto $\boldsymbol{z}^{(h_3)}$,

$$\begin{aligned} \boldsymbol{R}^{(h_3)} &= \boldsymbol{X} - \hat{\boldsymbol{X}}^{(h_3)} \\ &= \begin{bmatrix} c & c & -c & -c \\ d & -d & d & -d \end{bmatrix} - \begin{bmatrix} c & c & -c & -c \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ d & -d & d & -d \end{bmatrix}. \end{aligned}$$

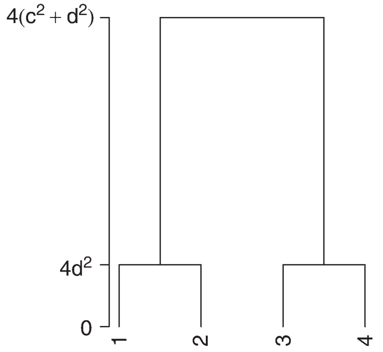


Fig. 2. Initial clustering dendrogram for theoretical example.

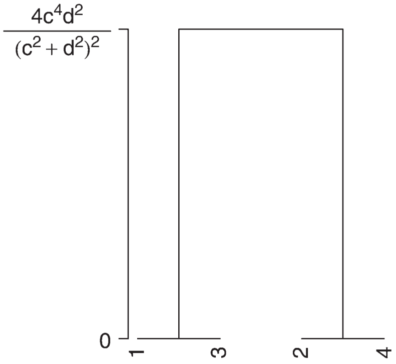


Fig. 3. Complementary clustering dendrogram for theoretical example.

Thus, $\boldsymbol{X}'$ reduces to

$$\boldsymbol{X}' = \frac{c^2}{c^2+d^2}\begin{bmatrix} 0 & 0 & 0 & 0 \\ d & -d & d & -d \end{bmatrix},$$

with the following distance relationships between its columns:

$$d(\boldsymbol{x}'_1, \boldsymbol{x}'_3) = d(\boldsymbol{x}'_2, \boldsymbol{x}'_4) = 0,$$

$$d(\boldsymbol{x}'_1, \boldsymbol{x}'_2) = d(\boldsymbol{x}'_3, \boldsymbol{x}'_4) = d(\boldsymbol{x}'_1, \boldsymbol{x}'_4) = d(\boldsymbol{x}'_2, \boldsymbol{x}'_3) = \frac{4c^4d^2}{(c^2+d^2)^2}.$$

Therefore, when applying complete linkage agglomerative hierarchical clustering to $\boldsymbol{X}'$, the first 2 merges occur at height 0, where we simultaneously merge $\boldsymbol{x}'_1$ and $\boldsymbol{x}'_3$ into one cluster and $\boldsymbol{x}'_2$ and $\boldsymbol{x}'_4$ into another. The third merge then merges these 2 clusters, at height $\frac{4c^4d^2}{(c^2+d^2)^2}$. The dendrogram for the complementary clustering is displayed in Figure 3.

This simple theoretical example, although completely deterministic, effectively illustrates how the complementary hierarchical clustering procedure can successfully uncover structure that would otherwise be hidden in an initial cluster analysis.

2.4 *Relative gene importance*

Typically, when looking at the initial and complementary clusterings produced by this procedure, we would like to have an idea of which genes were important in influencing the patterns present in these clusterings. To quantify this concept more clearly, we define a measure called the relative gene importance, analogous to R^2, the coefficient of determination in linear regression.

Letting $\hat{\boldsymbol{X}}^{(h_m)}$ denote the fitted values from the regression of each row of $\boldsymbol{X}$ onto $\boldsymbol{z}^{(h_m)}$, $\Delta h_m \equiv h_m - h_{m-1}$, and using the fact that $\boldsymbol{R}^{(h_m)} = \boldsymbol{X} - \hat{\boldsymbol{X}}^{(h_m)}$, we can manipulate (2.1) to obtain

$$\boldsymbol{X} = \hat{\boldsymbol{X}} + \boldsymbol{X}', \tag{2.2}$$

where

$$\hat{\boldsymbol{X}} = \sum_{m=1}^{n-1} \hat{\boldsymbol{X}}^{(h_m)} \frac{\Delta h_m}{h}. \tag{2.3}$$

Recall that

$$R^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}}.$$

Thus for the initial clustering, we define the relative gene importance of the ith gene to be

$$I_i = \frac{\sum_j (\hat{x}_{i,j} - \bar{x}_i)^2}{\sum_j (x_{i,j} - \bar{x}_i)^2},$$

where $\hat{x}_{i,j}$ is the (i, j)th element of $\hat{\boldsymbol{X}}$ and $\bar{x}_i$ is the average of the elements in the ith row of $\boldsymbol{X}$.

If we apply the complementary hierarchical clustering procedure to $\boldsymbol{X}'$, we can write

$$\boldsymbol{X}' = \hat{\boldsymbol{X}}' + \boldsymbol{X}'' \tag{2.4}$$

with $\hat{\boldsymbol{X}}'$ and $\boldsymbol{X}''$ defined appropriately. Thus for the complementary clustering, the relative gene importance is defined similarly as

$$I_i' = \frac{\sum_j (\hat{x}_{i,j}' - \bar{x}_i')^2}{\sum_j (x_{i,j}' - \bar{x}_i')^2}.$$

Both I_i and I_i' are between 0 and 1. Intuitively, a gene which strongly influences the clustering pattern will be highly differentially expressed. Since the complementary hierarchical clustering procedure attempts to remove the effect of this gene, a larger proportion of its variation across the samples will tend to be explained by the fitted values described in (2.2) and (2.4). Thus, genes which strongly influence the clustering pattern will have a relative gene importance closer to 1 compared to genes that do not have much influence.

3. COMPUTATION

Efficient algorithms for performing hierarchical clustering are readily available, so we will focus on finding an efficient method for calculating $\boldsymbol{X}'$, the expected residual matrix. From (2.2) and (2.3), recall that

$$\boldsymbol{X}' = \boldsymbol{X} - \sum_{m=1}^{n-1} \hat{\boldsymbol{X}}^{(h_m)} \frac{\Delta h_m}{h},$$

where $\hat{\boldsymbol{X}}^{(h_m)}$ were the fitted values from the regression of each row of $\boldsymbol{X}$ onto $z^{(h_m)}$. Since $z^{(h_m)}$ is a factor variable, for a given row of $\hat{\boldsymbol{X}}^{(h_m)}$, the fitted values are simply averages of certain elements from the corresponding row of $\boldsymbol{X}$. The main advantage of this is that $\hat{\boldsymbol{X}} = \sum_{m=1}^{n-1} \hat{\boldsymbol{X}}^{(h_m)} \frac{\Delta h_m}{h}$ can be obtained by applying a linear transformation to $\boldsymbol{X}$,

$$\hat{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{A}, \tag{3.1}$$

where $\boldsymbol{A}$ is the appropriate transformation matrix.

To calculate $\boldsymbol{A}$, we will first define some new notation. For a given merge m, let

$$c_m(i, j) = \begin{cases} 1, & \text{if samples } i \text{ and } j \text{ are in same group, as defined by } z^{(h_m)}, \text{ or if } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$C_m(j) = \sum_{i=1}^{n} c_m(i, j) = \text{ size of group, as defined by } z^{(h_m)}, \text{ to which } j \text{ belongs.}$$

Then, we can express any given row of $\hat{\boldsymbol{X}}$ as

$$\begin{aligned}
\sum_{m=1}^{n-1} (\hat{x}_1^{(h_m)}, \ldots, \hat{x}_n^{(h_m)}) \frac{\Delta h_m}{h} &= \left(\sum_{m=1}^{n-1} \hat{x}_1^{(h_m)} \frac{\Delta h_m}{h}, \ldots, \sum_{m=1}^{n-1} \hat{x}_n^{(h_m)} \frac{\Delta h_m}{h} \right) \\
&= \left(\sum_{m=1}^{n-1} \frac{\sum_{i=1}^{n} x_i c_m(i, 1)}{C_m(1)} \cdot \frac{\Delta h_m}{h}, \ldots, \sum_{m=1}^{n-1} \frac{\sum_{i=1}^{n} x_i c_m(i, n)}{C_m(n)} \cdot \frac{\Delta h_m}{h} \right) \\
&= \left(\sum_{i=1}^{n} x_i \sum_{m=1}^{n-1} \frac{c_m(i, 1)}{C_m(1)} \cdot \frac{\Delta h_m}{h}, \ldots, \sum_{i=1}^{n} x_i \sum_{m=1}^{n-1} \frac{c_m(i, n)}{C_m(n)} \cdot \frac{\Delta h_m}{h} \right) \\
&= (x_1, \ldots, x_n)\boldsymbol{A},
\end{aligned}$$

where

$$\boldsymbol{A} = \begin{bmatrix} \sum_{m=1}^{n-1} \frac{c_m(1,1)}{C_m(1)} \cdot \frac{\Delta h_m}{h} & \cdots & \sum_{m=1}^{n-1} \frac{c_m(1,n)}{C_m(n)} \cdot \frac{\Delta h_m}{h} \\ \vdots & & \vdots \\ \sum_{m=1}^{n-1} \frac{c_m(n,1)}{C_m(1)} \cdot \frac{\Delta h_m}{h} & \cdots & \sum_{m=1}^{n-1} \frac{c_m(n,n)}{C_m(n)} \cdot \frac{\Delta h_m}{h} \end{bmatrix} = \sum_{m=1}^{n-1} \frac{\Delta h_m}{h} \begin{bmatrix} \frac{c_m(1,1)}{C_m(1)} & \cdots & \frac{c_m(1,n)}{C_m(n)} \\ \vdots & & \vdots \\ \frac{c_m(n,1)}{C_m(1)} & \cdots & \frac{c_m(n,n)}{C_m(n)} \end{bmatrix}.$$

Algorithm 1 describes how to calculate $\boldsymbol{A}$. As we cycle through m, not all positions of $\boldsymbol{A}$ are incremented. Hence at each step, we only need to consider the positions for which their corresponding group size changes, further reducing the number of necessary calculations for determining $\boldsymbol{A}$.

ALGORITHM 1 (Calculating $\boldsymbol{A}$)

```
1. Initialize $\boldsymbol{A}$ to be a $n \times n$ matrix of zeros.
2. for $m = 1$ to $n - 1$ do
3.     For each pair of samples $(x_i, x_j)$, determine the size of the group (as defined by $z^{(h_m)}$) to which
       they both belong.
4.     If the size is nonzero, increment position $(i, j)$ of $\boldsymbol{A}$ by $\frac{\Delta h_m}{h \times \text{size of group}}$.
5. end for
```

4. Simulation study

We tested the complementary hierarchical clustering procedure on a variety of simulated microarray data sets. The data sets were generated according to the matrix displayed in Figure 4. Each data set has 2 effects, corresponding to 2 sets of significant genes. The first effect is represented by the first p_{e} rows, and the second effect is represented by the last p_{e} rows. An example of the initial and complementary clusterings produced by applying the procedure to one of the simulated data sets is displayed in Figure 5. To investigate the performance of the procedure under different conditions, we focused on 4 particular scenarios. In the first scenario, we varied the ratio of the strengths of the 2 effects. The second scenario involved varying the level of background noise. In the third and fourth scenarios, we looked at how the balance structure of the data and the dimension of the data, respectively, affected the procedure. In each scenario, we generated 1000 data sets for each particular combination of parameters and looked at the effects identified by the initial and complementary clusterings. The first bifurcation of the dendrogram was used to determine whether a clustering had identified a particular effect. The results are tabulated in Tables 1–4.

We see from Table 1 that provided the first effect was stronger than the second effect, the initial clustering identified the first effect and the complementary clustering identified the second effect. When the 2 effects were of equal strength, approximately half the time the initial clustering identified the second effect. The counts in Table 2 indicate that the procedure is relatively robust to the level of background noise. Only when the signal-to-noise ratio (with respect to the second effect) was almost 1, did the complementary clustering begin to fail in identifying the second effect. Table 3 indicates that the procedure is independent of whether the first effect is balanced. Finally, Table 4 shows that the proportion of significant genes corresponding to the second effect needs to approach 1% to be consistently identified by the complementary clustering.

5. Backfitting

As first alluded to in Section 2.4, a natural extension to the complementary hierarchical clustering procedure is to repeatedly apply the procedure, obtaining a sequence of hierarchical clusterings. The hope would be that each successive clustering would be uncovering different uncorrelated structures in the data. Suppose, for example, we repeat the procedure 3 times. Utilizing (2.2) and (2.4), we can express the 3 procedures by the following 3 equations:

$$\boldsymbol{X} = \hat{\boldsymbol{X}} + \boldsymbol{X}',$$
$$\boldsymbol{X}' = \hat{\boldsymbol{X}}' + \boldsymbol{X}'',$$
$$\boldsymbol{X}'' = \hat{\boldsymbol{X}}'' + \boldsymbol{X}''',$$

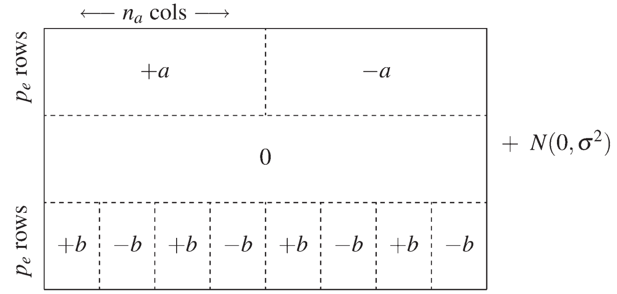


Fig. 4. The simulated data sets were generated by adding $N(0,\sigma^2)$ Gaussian noise to the $p \times n$ matrix described on the left. The top p_{e} rows correspond to the first effect (the first n_a columns were assigned $+a$), and the bottom p_{e} rows correspond to the second effect (odd-numbered columns were assigned $+b$).

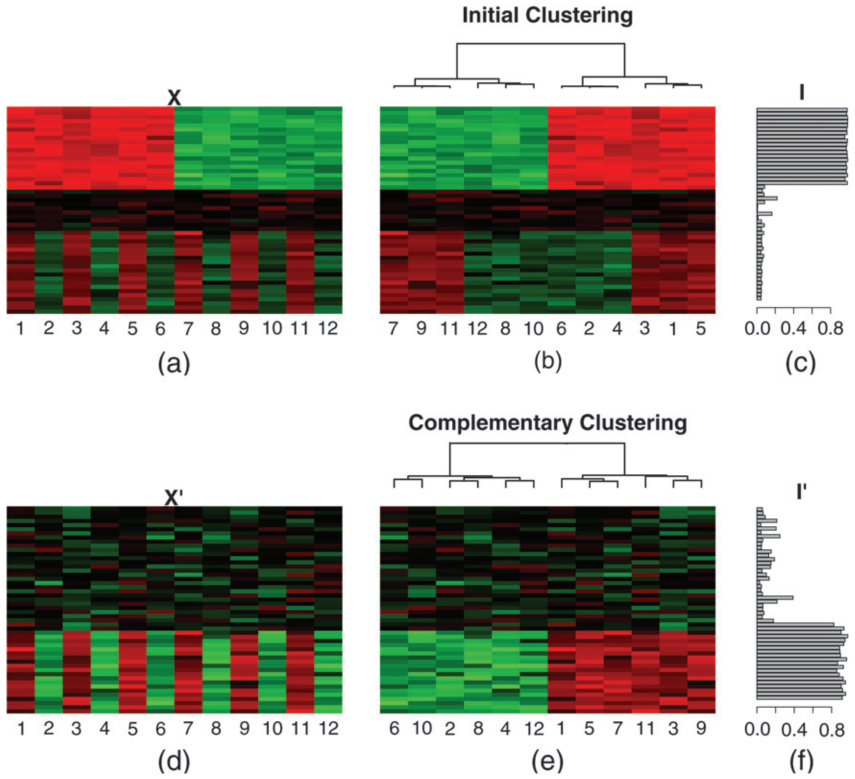


Fig. 5. The initial and complementary clusterings for a 50 × 12 simulated data set with $p_{\mathrm{e}} = 20$, $n_a = 6$, $a = 6$, $b = 3$, and $\sigma = 1$. Also displayed are the heat maps of the clustered data and the relative gene importance plots for each clustering.

Table 1. *The effects identified by the initial and complementary clusterings for varying values of b. For example, the first row indicates out of* 1000 *realizations, the initial and complementary clustering identified the first and second effects, respectively,* 936 *times, and the first effect and neither effect, respectively,* 64 *times*

Parameters							Initial/complementary		
p	n	p_{e}	n_a	a	b	σ	Effect 1/effect 2	Effect 2/effect 1	Effect 1/neither
50	12	20	6	6	1	1	936	0	64
					2		1000	0	0
					3		1000	0	0
					4		1000	0	0
					5		1000	0	0
					6		485	515	0

or, equivalently, by 1 equation:

$$X = \hat{X} + \hat{X}' + \hat{X}'' + X'''. \tag{5.1}$$

To increase the likelihood that each successive clustering is uncovering new uncorrelated structure, an alternative way to fit the model described in (5.1) would be to use a backfitting algorithm, where we

Table 2. *The effects identified by the initial and complementary clusterings for varying values of σ*

Parameters							Initial/complementary		
p	n	p_e	n_a	a	b	σ	Effect 1/effect 2	Effect 1/neither	Neither/neither
50	12	20	6	6	3	1	1000	0	0
						2	991	9	0
						2.5	862	138	0
						3	563	436	1
						4	62	921	17
						5	4	879	117
						6	1	684	315

Table 3. *The effects identified by the initial and complementary clusterings for varying values of n_a*

Parameters							Initial/complementary
p	n	p_e	n_a	a	b	σ	Effect 1/effect 2
50	12	20	6	6	3	1	1000
			8				1000
			10				1000

iteratively readjust for the fitted matrices. The backfitting algorithm and its application to fitting additive models are described in Hastie and Tibshirani (1990). Before describing how we apply the backfitting algorithm, we will define some notation. Recall from (3.1) that

$$\hat{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{A}_{\boldsymbol{X}},$$

where the derivation of $\boldsymbol{A}_{\boldsymbol{X}}$ is described in Section 3. Note that the matrix $\boldsymbol{A}_{\boldsymbol{X}}$ directly depends on the hierarchical clustering of $\boldsymbol{X}$, and this dependence is indicated by the subscript.

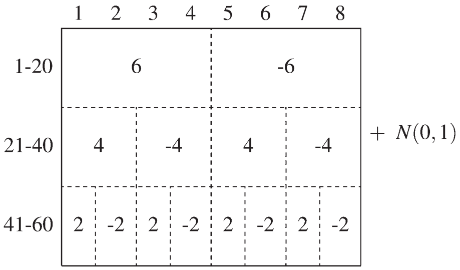
ALGORITHM 2 (Backfitting algorithm for complementary hierarchical clustering)

```
Initialize: $\hat{X} = XA_X$; $\hat{X}' = (X - \hat{X})A_{X-\hat{X}}$; $\hat{X}'' = (X - \hat{X} - \hat{X}')A_{X-\hat{X}-\hat{X}'}$
repeat
    $\hat{X} = (X - \hat{X}' - \hat{X}'')A_{X-\hat{X}'-\hat{X}''}$
    $\hat{X}' = (X - \hat{X} - \hat{X}'')A_{X-\hat{X}-\hat{X}''}$
    $\hat{X}'' = (X - \hat{X} - \hat{X}')A_{X-\hat{X}-\hat{X}'}$
until $\hat{X}$, $\hat{X}'$, and $\hat{X}''$ don't change.
```

Algorithm 2 describes how to apply the backfitting algorithm. Letting $\hat{\boldsymbol{X}}_{\text{old}}$ and $\hat{\boldsymbol{X}}_{\text{new}}$ refer to successive updates of $\hat{\boldsymbol{X}}$, we typically iterate the "repeat" steps until $\|\hat{\boldsymbol{X}}_{\text{old}} - \hat{\boldsymbol{X}}_{\text{new}}\|/\|\hat{\boldsymbol{X}}_{\text{old}}\|$ is below some threshold, say 10^{-5}, and similarly for $\hat{\boldsymbol{X}}'$ and $\hat{\boldsymbol{X}}''$. The operator for obtaining each fit is very nonlinear, as it depends directly on a hierarchical clustering of a matrix, and as such we do not have any convergence proofs. In certain situations, the algorithm may potentially oscillate between the fits, and thus convergence is not guaranteed. However, in our experience, provided the number of fitted matrices is set to be less than or equal to the true number of uncorrelated structures present in the data, convergence occurred very quickly (10–20 iterations). Upon convergence, we use the notation $\hat{\boldsymbol{X}}_{bf}$, $\hat{\boldsymbol{X}}'_{bf}$, and $\hat{\boldsymbol{X}}''_{bf}$ for the fitted matrices in

Table 4. *The effects identified by the initial and complementary clusterings for* 10 000 × 50 *data for varying values of* p_{e}

Parameters							Initial/complementary	
p	n	p_{e}	n_a	a	b	σ	Effect 1/effect 2	Effect 1/neither
10 000	50	20	25	6	3	1	0	1000
		40					0	1000
		60					578	422
		80					999	1



	1	2	3	4	5	6	7	8
1-20	6				-6			
21-40	4		-4		4		-4	
41-60	2	-2	2	-2	2	-2	2	-2

$+\ N(0,1)$

Fig. 6. Matrix used to generate the simulated example of microarray data with 3 main effects.

order to differentiate them from the fits obtained from single applications of the complementary hierarchical clustering procedure. We then present the hierarchical clusterings of $\boldsymbol{X} - \hat{\boldsymbol{X}}'_{bf} - \hat{\boldsymbol{X}}''_{bf}$, $\boldsymbol{X} - \hat{\boldsymbol{X}}_{bf} - \hat{\boldsymbol{X}}''_{bf}$, and $\boldsymbol{X} - \hat{\boldsymbol{X}}_{bf} - \hat{\boldsymbol{X}}'_{bf}$ for the initial, complementary, and tertiary clusterings, respectively.

To illustrate the backfitting algorithm, we applied it to a simulated example with 3 main effects corresponding to 3 sets of significant genes. The matrix from which the example was generated is shown in Figure 6. Displayed in Figure 7 are the initial, complementary, and tertiary clusterings obtained when using the backfitting algorithm. For comparison, the corresponding clusterings produced by single applications of the complementary hierarchical clustering procedure are shown in Figure 8. The backfitting procedure performs very well and tends to produce a much cleaner, sharper set of clusterings. However, the single applications of the complementary hierarchical clustering procedure still effectively uncover the correct structure at each clustering, and we would tend to favor this latter approach due to its straightforward nature and interpretability.

6. Breast cancer data

The data set analyzed consists of the gene expression profiles across approximately 25 000 human genes for tumor samples taken from 295 consecutive patients with breast cancer. The tumor samples were selected from the fresh-frozen tissue bank of the Netherlands Cancer Institute. Detailed information regarding the patients and the microarray data are described in van de Vijver *and others* (2002) and van't Veer *and others* (2002). The complementary hierarchical clustering procedure was applied to the data using complete linkage agglomerative hierarchical clustering with the correlation-based distance measure. The initial and complementary clusterings are displayed in Figure 9. In each clustering, the patients were divided into 2 groups as determined by the first bifurcation of the dendrogram. The leaves of each dendrogram have been colored to indicate the 2 groups. Also displayed in Figure 9 is the complementary clustering dendrogram with the leaves colored according to the initial clustering grouping. Listed

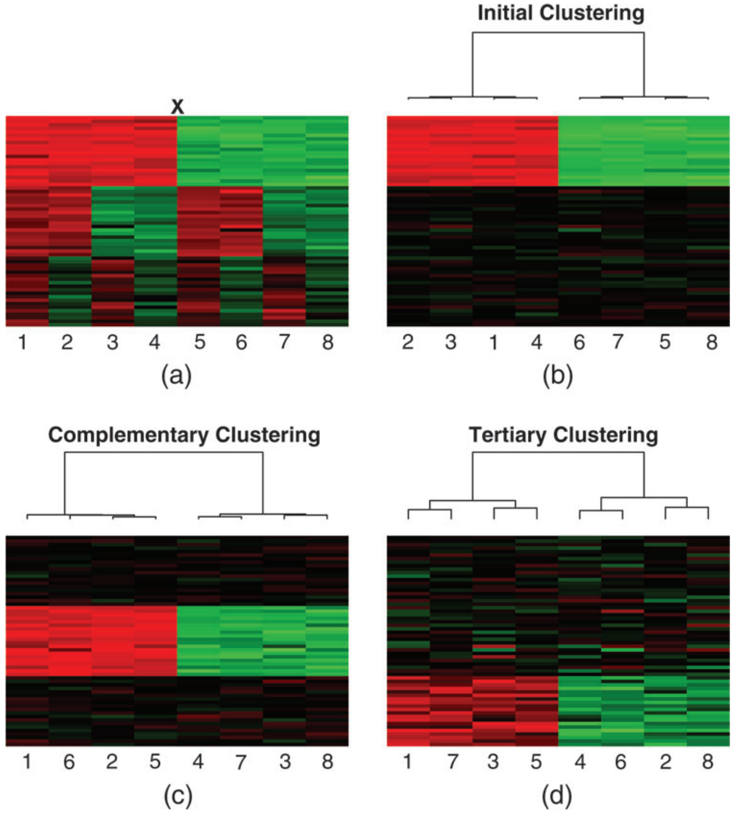


Fig. 7. The clusterings produced by applying the backfitting algorithm to the simulated example with 3 main effects. The heat maps in panels (b), (c), and (d) correspond to $\boldsymbol{X} - \hat{\boldsymbol{X}}'_{bf} - \hat{\boldsymbol{X}}''_{bf}$, $\boldsymbol{X} - \hat{\boldsymbol{X}}_{bf} - \hat{\boldsymbol{X}}''_{bf}$, and $\boldsymbol{X} - \hat{\boldsymbol{X}}_{bf} - \hat{\boldsymbol{X}}'_{bf}$, respectively.

in Table 5 are the cross-tabulated counts of the number of patients falling into each group of the initial and complementary clusterings. Table 5, along with the far-right dendrogram in Figure 9, gives a good indication that the complementary clustering grouping differs from the initial clustering grouping.

For further analysis, we compared the relationship between the complementary clustering grouping and the 3 established prognostic signatures, the “wound-response”, the “intrinsic gene”, and the “70-gene prognosis” signatures. We found the complementary clustering grouping to be uncorrelated with all 3 signatures, indicating that we may have found a novel grouping of the patients. The results are summarized in Table 6. The wound-response signature is a quantitative score which has been shown to be correlated with overall survival and distant metastasis-free survival (see Chang *and others*, 2005, for details). We performed a 2-sample t-test to compare the mean wound-response signature in each group and concluded that the means do not differ ($p = 0.329$). The intrinsic gene signature, described in Sørlie *and others* (2003), classifies malignant breast tumors into 5 subtypes: basal, ERBB2, luminal A, luminal B, and normal. Details on how the intrinsic gene signature was assigned to the 295 tumor samples are described in Chang *and others* (2005). A chi-square test for independence indicated that the complementary clustering grouping was uncorrelated with the intrinsic gene signature ($p = 0.241$). The 70-gene prognosis signature, described in van de Vijver *and others* (2002), classifies patients with breast cancer into 2 categories: “good” prognosis and “poor” prognosis. Again, from the chi-square test for independence,

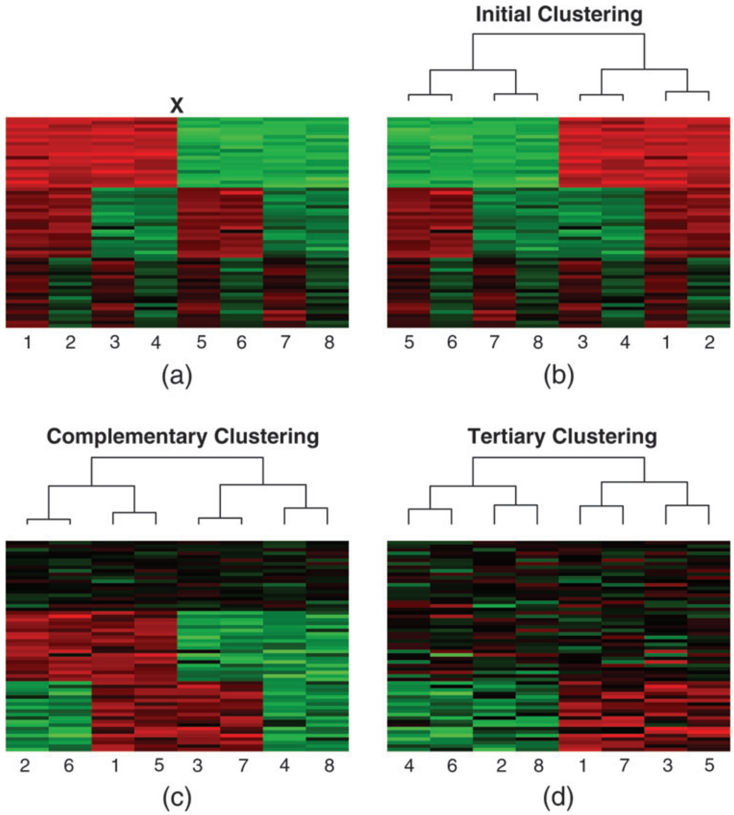


Fig. 8. The clusterings produced by single applications of the complementary hierarchical clustering procedure to the simulated example with 3 main effects. The heat maps in panels (b), (c), and (d) correspond to $\boldsymbol{X}$, $\boldsymbol{X}'$, and $\boldsymbol{X}''$, respectively.

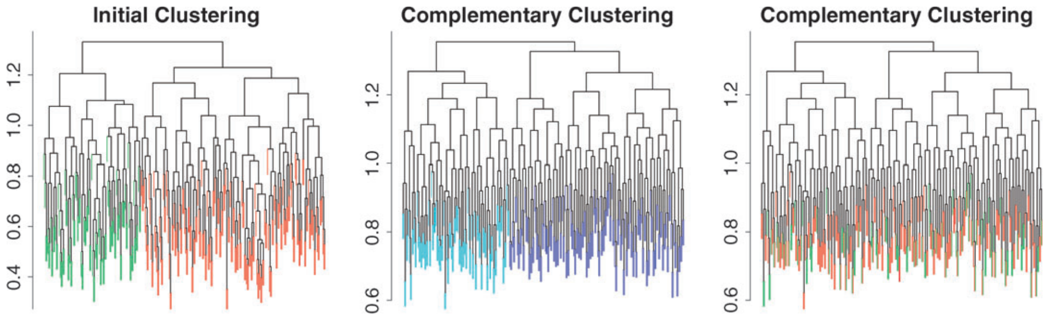


Fig. 9. The initial and complementary clusterings of the breast cancer data. The dendrogram on the far-right is the complementary clustering with the leaves colored according to the initial clustering grouping.

we concluded that this prognostic signature was also uncorrelated with the complementary clustering grouping ($p = 0.234$).

We also conducted some survival-type analyses. Displayed in Figure 10 are the Kaplan–Meier survival curves for overall survival and distant metastasis-free probability (DMFP) for the 2 groups of

Table 5. *The number of patients falling into each of the* 2 *groups identified by the initial and complementary clusterings*

	Complementary clustering	
	Group 1 (navy)	Group 2 (cyan)
Initial clustering		
Group 1 (red)	118	75
Group 2 (green)	65	37

Table 6. *The relationships between the complementary clustering grouping and the* 3 *prognostic signatures, wound response, intrinsic gene, and* 70*-gene prognosis. The wound-response signature is a continuous variable, thus listed are the averages (standard deviations) of the variable in each group. The p-value corresponds to a* 2*-sample t-test. The intrinsic and* 70*-gene prognosis signatures are classification variables with* 5 *and* 2 *classes, respectively. Listed are the number of patients falling into each class and each group of the complementary clustering, with the p-values corresponding to a chi-square test for independence*

	Group 1 (navy)	Group 2 (cyan)	p-value
Wound-response signature	0.0107 (0.18)	−0.0105 (0.18)	0.329
Intrinsic gene signature			
Basal	34	12	
ERBB2	27	22	
Luminal A	53	35	0.241
Luminal B	47	34	
Normal	22	9	
70-gene prognosis signature			
Good	66	49	0.234
Poor	117	63	

patients identified by the initial and complementary clusterings. As described in Chang *and others* (2005), overall survival was defined by death from any cause and DMFP was defined by distant metastasis as a first recurrence event. Each set of curves have been colored to correspond with the cluster dendrograms in Figure 9. The "log-rank test" was used to test for differences between the survival curves, with the p-values indicated on each figure. When looking at overall survival, there was a significant difference between the 2 groups in the initial clustering but not in the complementary clustering ($p = 3.93 \times 10^{-4}$ and 0.176, respectively). However, for DMFP, there was a significant difference between the 2 groups in both the initial and the complementary clusterings ($p = 2.8 \times 10^{-5}$ and 0.0414, respectively). As the complementary clustering grouping was shown to be generally uncorrelated with the prognostic signatures wound response, intrinsic gene, and 70-gene prognosis, the complementary clustering may be uncovering some novel structure in the data which is responsible for a difference in DMFP among the patients. All computations were performed using the software package R 2.5.1.

7. Discussion

Although far from exhaustive, the simulation analysis covered a range of interesting examples of microarray data and showed that the complementary hierarchical clustering procedure successfully uncovers the

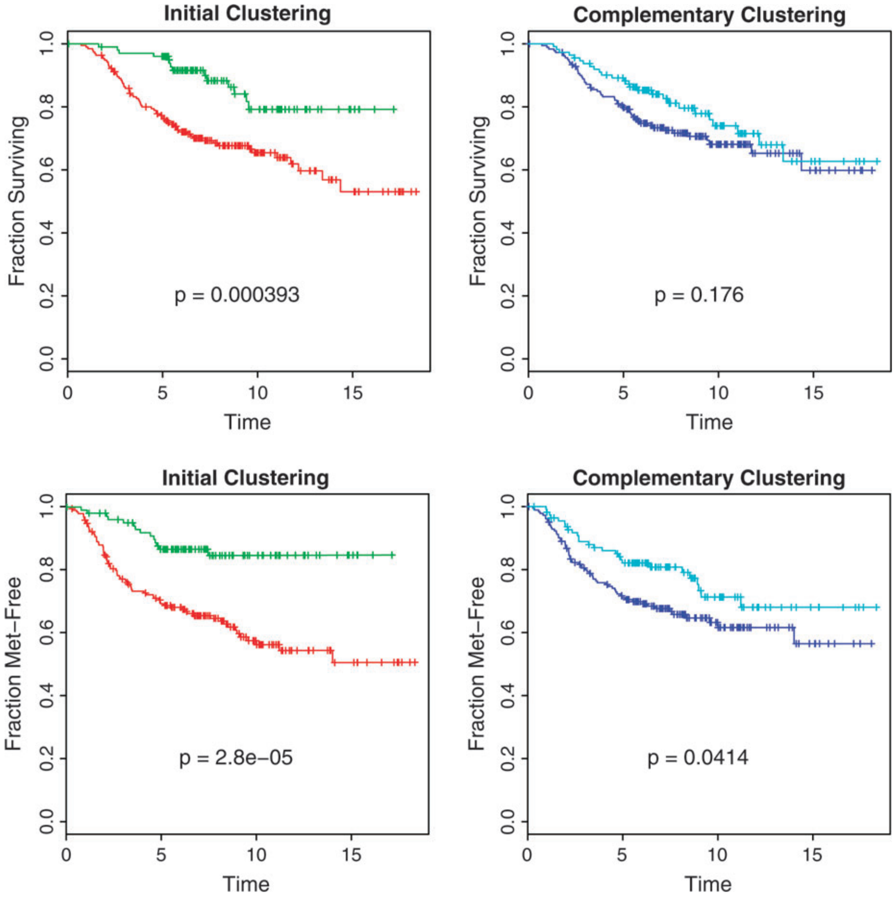


Fig. 10. The Kaplan–Meier survival curves for overall survival (top 2 panels) and DMFP (bottom 2 panels) for the 2 groups of patients identified by the initial and complementary clusterings of the breast cancer data.

structures arising from the weaker genes. The procedure seemed fairly robust to both the effect strength of the weaker gene set and the level of background noise. Using the backfitting algorithm to estimate the fitted matrices was an effective extension to the procedure and produced favorable results. However, the backfitting algorithm somewhat complicates the interpretation of the initial, complementary, and further clusterings. The analysis of the breast cancer data demonstrated that the complementary clustering was identifying potentially biologically important structure in the data. When the patients were grouped according to the first bifurcation of the complementary clustering dendrogram, there was a significant difference in DMFP between the 2 groups. Further, this grouping appeared to be uncorrelated with 3 previously known prognostic signatures. For this particular data set, it may be of interest to examine the genes which are important in the complementary clustering, as identified by the relative gene importance. General further work regarding the complementary hierarchical clustering procedure could involve looking deeper into the application of the backfitting algorithm.

An advantage of the complementary hierarchical clustering procedure is that it is an automatic procedure. After performing the initial clustering, there is no need to decide how many groups should be considered or where to cut the dendrogram. Also, the procedure is very versatile. We can use any clustering procedure that produces a nested sequence of clusters that can be represented as a dendrogram, for example, tree-structured vector quantization (Gersho and Gray, 1992) or hybrid hierarchical clustering (Chipman and Tibshirani, 2006). Eventually, we hope that this procedure can be used routinely to quickly and easily uncover interesting patterns in data that would otherwise not be revealed with a typical cluster analysis. An R package, compHclust, will soon be available on both authors' website.

Acknowledgments

The authors thank Trevor Hastie for helpful comments. The authors would also like to thank the editors and reviewers for helpful comments that led to improvements in the paper. *Conflict of Interest:* None declared.

Funding

National Science Foundation (DMS-9971405 to R.T.); National Institutes of Health (N01-HV-28183 to R.T.).

References

Chang, H. Y., Nuyten, D. S. A., Sneddon, J. B., Hastie, T., Tibshirani, R., Sorlie, T., Dai, H., He, Y. D., van't Veer, L. J., Bartelink, H. *and others* (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3738–3743.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)* **8**, 93–103.

Chipman, H. and Tibshirani, R. (2006). Hybrid hierarchical clustering with applications to microarray data. *Biostatistics* **7**, 286–301.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

Gersho, A. and Gray, R. M. (1992). *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Publishers.

Gordon, A. D. (1999). *Classification*. London: Chapman and Hall.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. New York: Springer.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Higham, D. J., Kalna, G. and Kibble, M. (2007). Spectral clustering and its use in bioinformatics. *Journal of Computational and Applied Mathematics* **204**, 25–37.

Kluger, Y., Basri, R., Chang, J. T. and Gerstein, M. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research* **13**, 703–716.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE Transactions on Computational Biology and Bioinformatics* **1**, 24–45.

Pan, W. (2006). Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics* **22**, 795–801.

Pan, W., Lin, J. and Le, C. T. (2002). Model-based cluster analysis of microarray gene-expression data. *Genome Biology* **3**, research0009.1–0009.8.

SHENG, Q., MOREAU, Y. AND DE MOOR, B. (2003). Biclustering microarray data by Gibbs sampling. *Bioinformatics* **19**, ii196–ii205.

SØRLIE, T., TIBSHIRANI, R., PARKER, J., HASTIE, T., MARRON, J. S., NOBEL, A., DENG, S., JOHNSEN, H., PESICH, R., GEISLER, S. *and others* (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 8418–8423.

TURNER, H. L., BAILEY, T. C., KRZANOWSKI, W. J. AND HEMINGWAY, C. A. (2005). Biclustering models for structured microarray data. *IEEE Transactions on Computational Biology and Bioinformatics* **2**, 316–329.

VAN DE VIJVER, M. J., HE, Y. D., VAN'T VEER, L. J., DAI, H., HART, A. A., VOSKUIL, D. W., SCHREIBER, G. J., PETERSE, J. L., ROBERTS, C., MARTON, M. J. *and others* (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine* **347**, 1999–2009.

VAN'T VEER, L. J., DAI, H., VAN DE VIJVER, M. J., HE, Y. D., HART, A. A., MAO, M., PETERSE, H. L., VAN DER KOOY, K., MARTON, M. J., WITTEVEEN, A. T. *and others* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

XING, E. P. AND KARP, R. P. (2001). CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics* **17**, S306–S315.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E. AND RUZZO, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987.

[*Received July 16, 2007; revised November 12, 2007; accepted for publication November 12, 2007*]