International Journal of Software Engineering
and Knowledge Engineering
Vol. 13, No. 6 (2003) 579–592

World Scientific
www.worldscientific.com

MODEL-BASED CLUSTERING IN GENE EXPRESSION MICROARRAYS: AN APPLICATION TO BREAST CANCER DATA

J. C. MAR and G. J. MCLACHLAN*

Department of Mathematics, University of Queensland, Australia
**gjm@maths.uq.edu.au*

In microarray studies, the application of clustering techniques is often used to derive meaningful insights into the data. In the past, hierarchical methods have been the primary clustering tool employed to perform this task. The hierarchical algorithms have been mainly applied heuristically to these cluster analysis problems. Further, a major limitation of these methods is their inability to determine the number of clusters. Thus there is a need for a model-based approach to these clustering problems. To this end, McLachlan *et al.* [7] developed a mixture model-based algorithm (EMMIX-GENE) for the clustering of tissue samples. To further investigate the EMMIX-GENE procedure as a model-based approach, we present a case study involving the application of EMMIX-GENE to the breast cancer data as studied recently in van 't Veer *et al.* [10]. Our analysis considers the problem of clustering the tissue samples on the basis of the genes which is a non-standard problem because the number of genes greatly exceed the number of tissue samples. We demonstrate how EMMIX-GENE can be useful in reducing the initial set of genes down to a more computationally manageable size. The results from this analysis also emphasise the difficulty associated with the task of separating two tissue groups on the basis of a particular subset of genes. These results also shed light on why supervised methods have such a high misallocation error rate for the breast cancer data.

Keywords: Microarray; mixture modelling; cluster analysis.

1. Introduction

The complexity and magnitude of DNA microarray data have inundated researchers with a flood of new bioinformatic challenges. The data generated by these experiments necessitate the use of specialised statistical tools in order to make reliable inferences about the data. In this paper we discuss the application of a model-based approach to cluster analysis for gene expression microarrays.

Cluster analyses have previously demonstrated their utility in the elucidation of unknown gene function, the validation of gene discoveries, and the interpretation of biological processes; see [1, 3, 5] for example. The aim of a typical cluster analysis is to organise genes or tissue samples (data produced by separate hybridisations) into clusters displaying similar patterns of gene expression. Initially, hierarchical (distance-based) methods were applied to these cluster analysis problems. These clustering algorithms are largely heuristically motivated and there exist a number of unresolved issues associated with their use, including how to determine the number of clusters. As commented by Yeung *et al.* [11], "in the absence of a well-grounded statistical model, it seems difficult to define what is meant by a 'good' clustering algorithm or the 'right' number of clusters."

To overcome these difficulties attention is now turning towards a model-based approach to the analysis of microarray data; see [4, 7, 8, 11]. For example, by adopting a finite mixture model for the distribution of each observation, Yeung *et al.* [11] were advocates of a model-based approach to the clustering of the genes on the basis of the tissue samples. The present paper considers the problem of clustering the tissues on the basis of the genes. This is a more challenging problem to consider in a mixture model framework, since the number of observations to be clustered (the tissue samples) is typically small relative to the number of genes in each tissue sample.

A recent application of microarray technology involves its use in the development of patient-tailored therapies to target complex, highly heterogeneous diseases. The work of van 't Veer *et al.* [10] used microarray experiments on three patient groups who had different classes of breast cancer tumours. The overall goal of the experiment was to identify a set of genes that could distinguish between the different tumour groups based on their gene expression information for a given tumour sample. We use the data set produced by van 't Veer *et al.* [10] as the basis for our analysis using the mixture model-based clustering algorithm, EMMIX-GENE [7].

The results of this analysis shed light on why it is such a difficult problem to distinguish between the two tissue groups (disease-free and metastases) and consequently why supervised methods have such a high error rate for this data set, as noted by Tibshirani and Efron [9].

2. EMMIX-GENE: A Mixture Model-based Clustering Algorithm

The EMMIX-GENE algorithm consists of three stages (see [7] for more specific details). The first is a filtering step designed to isolate the most informative genes to be considered in the cluster analysis. For all genes in the original data set mixtures of t distributions are fitted, and each gene is assigned a value $-2\log\lambda$ where λ is the likelihood ratio statistic that tests for the presence of a single component versus two components in the fitted mixture model. Clearly genes that display a strong differential expression across different tumour groups will have a significantly larger likelihood ratio statistic, whereas those genes bearing little change across tumour groups will receive a lower score. Thus values assigned to each gene for $-2\log\lambda$ form the basis of a filter wherein only genes with likelihood ratio scores above a user-specified threshold are retained for further analysis.

The second stage involves grouping the retained set of genes into a user-specified number of clusters. The genes are clustered into groups, using Euclidean distance with a view to representing the genes within a group by their mean. If a clustering is sought on the basis of the totality of the genes, then it can be obtained by fitting a mixture model to these group means.

However, it may be that the number of group means N is too large to fit a normal mixture model with unrestricted component-covariance matrices. In this circumstance EMMIX-GENE has a third and optional step allowing for the fitting of mixtures of factor analyzers. The use of mixtures of factor analyzers reduces the number of parameters by imposing the assumption that the correlations between the genes can be expressed in a lower space by the dependence of the tissues on $q(q < N)$ unobservable factors.

In addition to clustering the tissues on the basis of all of the genes, there may be interest in seeing whether the different groups of genes lead to different clustering of the tissues when each is considered separately. For example, only a subset of the genes may be useful in identifying certain subtypes of the cancer being studied.

3. Description of the Experimental Data Set

In van 't Veer *et al.* [10], microarray experiments were performed on 98 primary breast cancers acquired from three groups of patients: 44 representing a good prognosis group, (i.e. those who remained metastasis free after a period of more than 5 years), 34 from a poor prognosis group (those who developed distant metastases within 5 years), and 20 representing a hereditary form of cancer, due to a BRCA1 (18 tumours) or BRCA2 (2 tumours) germline mutation.

Each microarray experiment involved an initial set of 24,881 genes. To reduce the number of genes to something more computationally manageable, van 't Veer *et al.* [10] applied a pre-processing filter in which only genes with both a $P-$value of less than 0.01 and at least a two fold difference in more than five out of the ninety-eight tissues for the gene were retained. This filter effectively reduced the initial set of genes to 4,869. The current paper makes use of the same pre-processing filter, working with the 4,869 retained genes.

The focus of the study by van 't Veer *et al.* [10] was to identify a subset of genes that would be useful in predicting the disease outcome of any given tissue sample. They anticipated that this gene signature could be applied as a diagnostic screen to select patients that would benefit from certain therapies over others.

4. Clustering Genes on the Basis of Tissue Samples Using EMMIX-GENE

As can be seen by the heat map displayed in Fig. 1, the task of discerning an underlying class structure in the data on the basis of the full set of 4,869 genes is extremely difficult.

For the present breast cancer data set, the heat maps of the genes in a group tend to mainly support the same breakup of the 98 tissues. To illustrate this, we list in Figs. 2 to 4 the heat maps for the top three groups G_1, G_2, and G_3, which contain 146, 93, and 61 genes, respectively. Important features to note from these heat maps are that they each indicate a change in gene expression is apparent between the

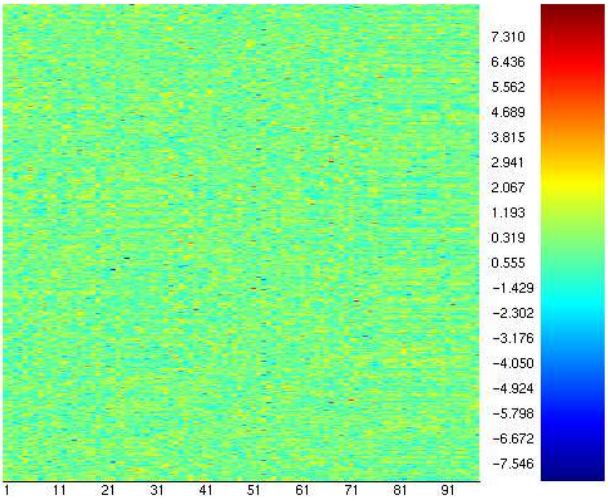

Fig. 1. Heat map displaying the initial set of 4,869 genes on the 98 breast cancer tumours. (Each row refers to a gene and each column to a tumour.)

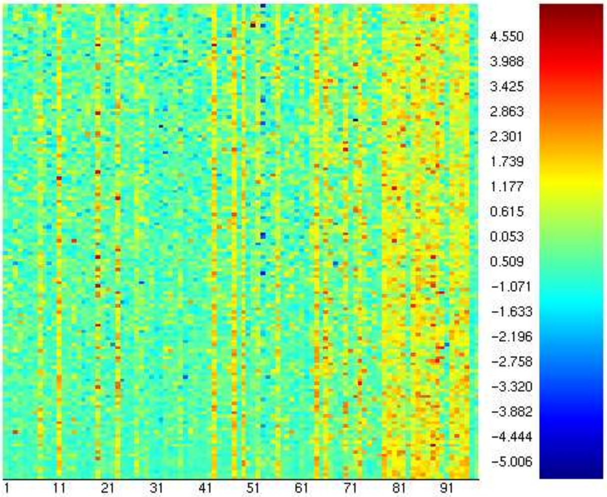

Fig. 2. Heat map of genes in group G_1.

sporadic (first 78 tissue samples) and hereditary (last 20 tissue samples) tumours. For instance, in Fig. 2, the genes in this cluster are generally down-regulated for the former group of tumours, and up-regulated in the latter. Genes in G_2 were largely constant in expression across the sporadic tumours but notably down-regulated for the hereditary tumours.

Additionally, the final two tissue samples, which represent the two BRCA2 tumours show consistent patterns of expression in each of the clusters that are different

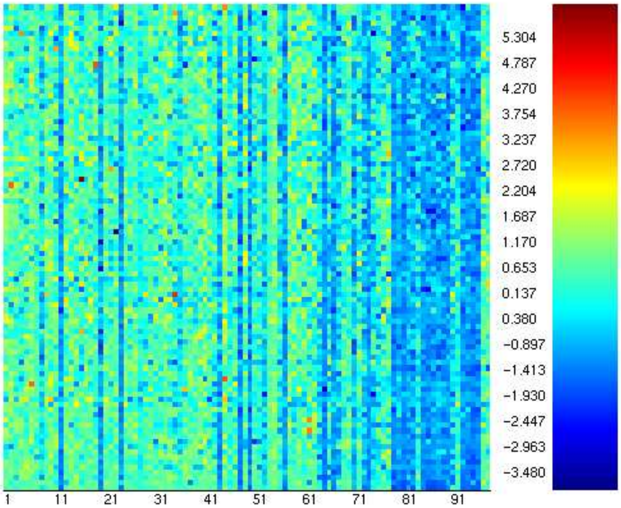

Fig. 3. Heat map of genes in group G_2.

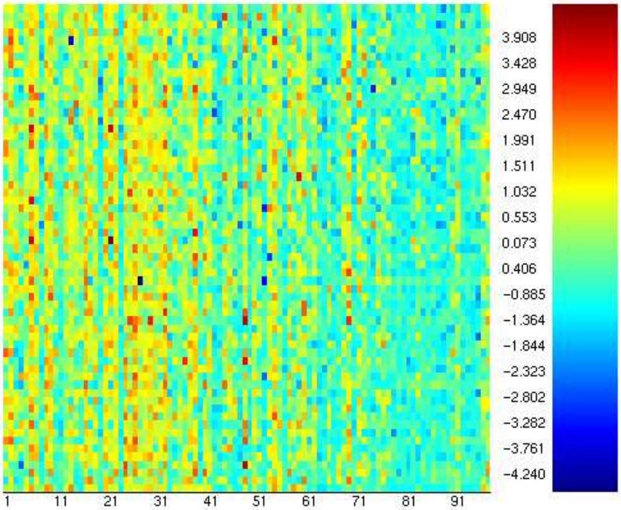

Fig. 4. Heat map of genes in group G_3.

from those exhibited by the set of BRCA1 tumours.

It can be seen from these groups that the problem of trying to distinguish between the two classes, patients who were disease-free after 5 years Π_1 and those with metastases within 5 years Π_2, is not straightforward on the basis of the gene expressions.

5. An Unsupervised Classification Analysis Using EMMIX-GENE

The first step of the EMMIX-GENE algorithm was used to select the most relevant genes from this filtered set of 4,869 genes, further reducing the number to 1,867. The 1,867 retained genes were clustered into forty groups using the second step of the EMMIX-GENE algorithm, and the majority of gene groups produced were reasonably cohesive and distinct. Based upon these forty group means, the tissue samples were clustered into two and three components using a mixture of factor analyzers model with $q = 4$ factors.

6. Investigating the Usefulness of the Selection of Relevant Genes

In clustering the genes, van 't Veer *et al.* [10] relied upon an agglomerative hierarchical algorithm to organise the genes into dominant genes groups. Two of these clusters were highlighted in the paper and the genes contained in these two groups correspond to biologically significant features. We denote Cluster A as the group of genes van 't Veer *et al.* [10] have identified as containing genes co-regulated with the ER-α gene (ESR1) and Cluster B as the group containing "co-regulated genes that are the molecular reflection of extensive lymphocytic infiltrate, and comprise a set of genes expressed in T and B cells". Both of these clusters contain 40 genes.

Of these 80 genes, the first step of the EMMIX-GENE algorithm select-genes retains only 47 genes (24 from Cluster A, 23 from Cluster B). When compared to the 40 groups that the cluster-genes step of the EMMIX-GENE algorithm produces, subsets of these 47 genes appeared inside several of these 40 groups (see Table 1 below).

Table 1. Comparing clusters constructed by a hierarchical algorithm with those produced by the EMMIX-GENE algorithm.

	Cluster index (EMMIX-GENE)	Number of genes matched	Percentage matched (%)
Cluster A	2	21	87.5
	3	2	8.33
	14	1	4.17
Cluster B	17	18	78.3
	19	1	4.35
	21	4	17.4

The motivation behind select-genes is to isolate the most informative genes to be used for the cluster analysis. For any clustering algorithm, genes that lack distinctive expression pattern changes across different tumour groups only serve to confuse the clustering algorithm and increase the number of misallocation errors made.

The 21 genes that appear in Cluster A have been grouped in the second cluster constructed by EMMIX-GENE. In Fig. 5 (below), these genes demonstrate clear expression changes for the three groups of tumours (indicated by the vertical blue lines).

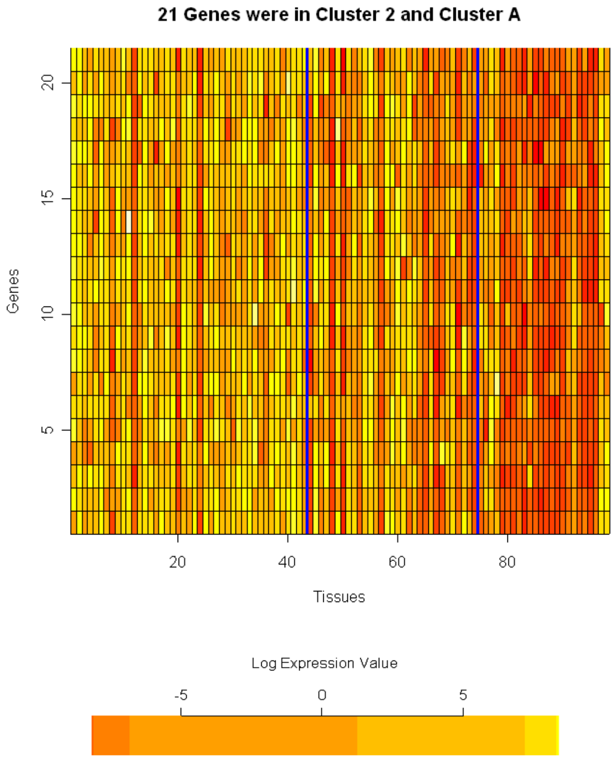

Fig. 5. Genes retained by EMMIX-GENE appearing in Cluster A.

For the remaining sixteen genes that were rejected by select-genes but belong to Cluster A, it is evident from Fig. 6 that these genes bear very little information in distinguishing between the tumour groups.

The heat maps displayed in Figs. 7 and 8 display the corresponding information for the genes in Cluster B. The genes in Fig. 7 (those retained by EMMIX-GENE) show much variation across the tumour groups. In contrast, the genes in Cluster B (those rejected by EMMIX-GENE) show little variation between the tumour groups.

The expression profile of the gene that received the highest $-2\log\lambda$ value is shown in Fig. 9. This gene is notably up-regulated for the disease-free tumour group and the metastases tumour group, and down-regulated in the hereditary tumour group.

An expression profile is shown in Fig. 10 for a gene which appeared in Cluster A, but whose value of $-2\log\lambda$ was not high enough for it to be retained by the select-genes step. The overall expression of the gene is essentially unchanging, however excessively large values for the seventeenth disease-free patient in the first tumour group and the sixth BRCA patient in the third tumour group appear to dominate the expression profile. These outliers seem to account for this gene's inclusion in Cluster A.

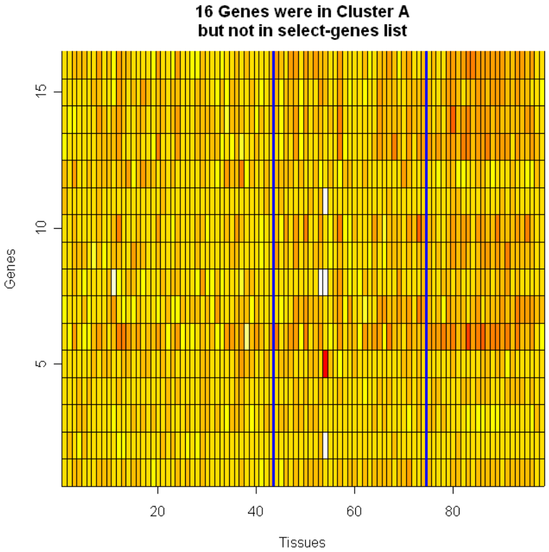

Fig. 6. Genes rejected by EMMIX-GENE appearing in Cluster A.

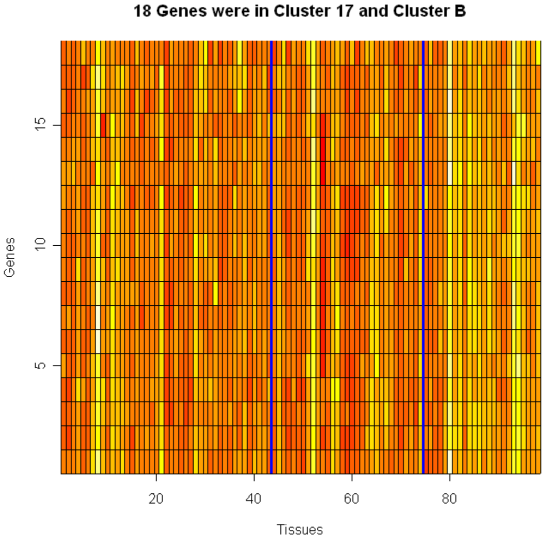

Fig. 7. Genes retained by EMMIX-GENE appearing in Cluster B.

7. Clustering Tissue Samples on the Basis of Gene Groups Using EMMIX-GENE

Turning now to the problem of clustering tissues on the basis of gene expression, we investigate the clusters constructed by the EMMIX-GENE algorithm in light of the genuine tissue grouping.

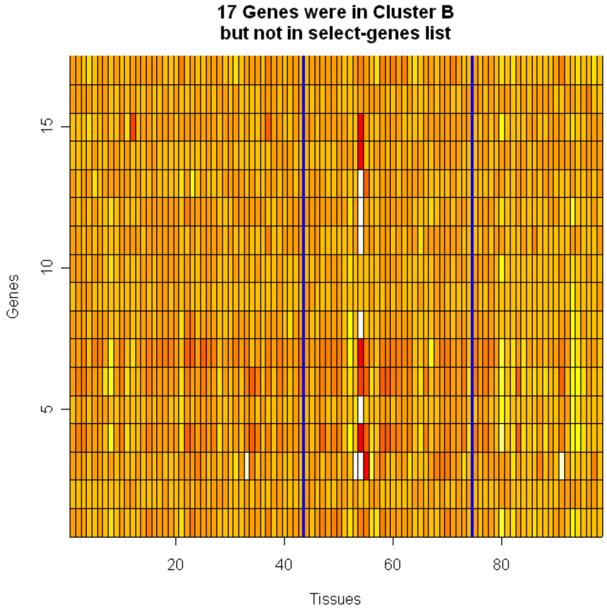

Fig. 8. Genes rejected by EMMIX-GENE appearing in Cluster B.

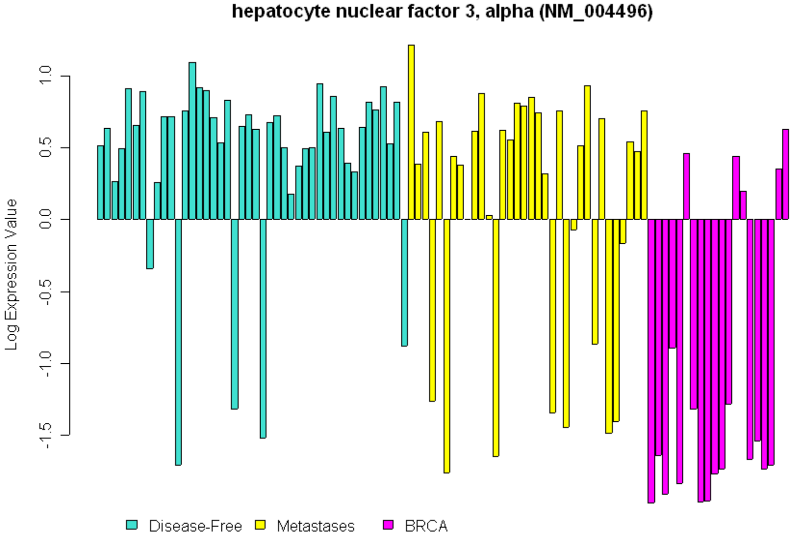

Fig. 9. Expression profile for the gene with the highest $-2\log\lambda$ value.

The tissue samples can be subdivided into two groups corresponding to the 78 sporadic tumours and 20 hereditary tumours. Figure 11 shows the two-cluster assignment produced by EMMIX-GENE with respect to this genuine grouping (pink vertical lines denote the three tumour groups; black denotes the hereditary tumour

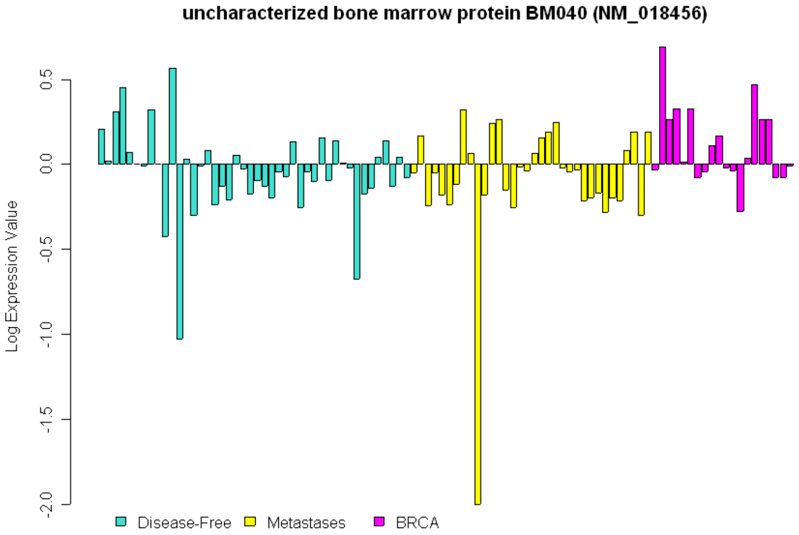

Fig. 10. Example of a gene rejected by select-genes but retained by Cluster A.

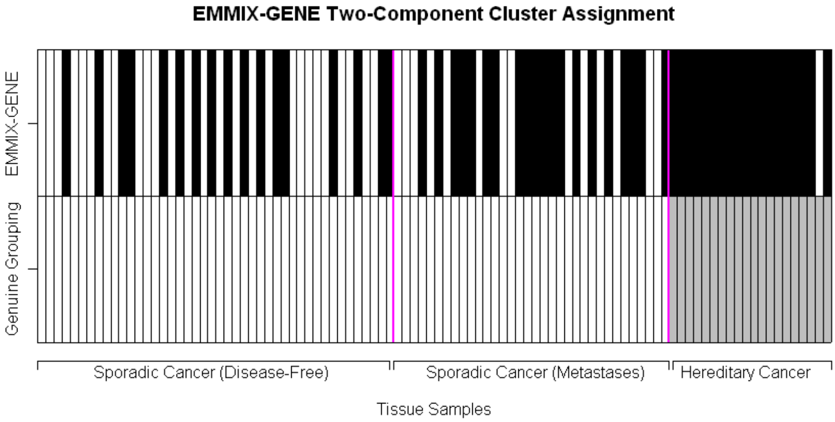

Fig. 11. Comparing EMMIX-GENE cluster assignments with the genuine two-group structure.

cluster, white denotes the sporadic tumour cluster; grey distinguishes the genuine grouping).

Clearly EMMIX-GENE has correctly clustered the majority of the hereditary tumours (misallocation error of 1/20), although 37 of the sporadic tumours were incorrectly assigned to the cluster of hereditary tumours.

The set of sporadic tumours have been divided into good and poor prognosis groups (i.e. 44 patients who continued to be disease-free after 5 years, and 34 patients who developed metastases within 5 years, respectively). Hence we also considered the partitioning of the tissues into three clusters, corresponding to the

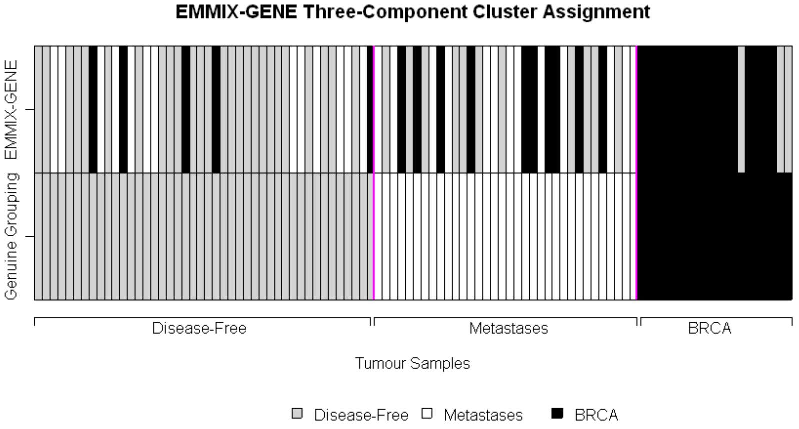

Fig. 12. Comparing EMMIX-GENE cluster assignments with a genuine three-group structure.

disease-free, metastases, and hereditary groups. Figure 12 shows the tissue samples rearranged according to the three cluster assignments allocated by EMMIX-GENE when a mixture of factor analyzers model with $q = 4$ factors.

Using a mixture of factor analyzers model with $q=$ 8 factors, we would misallocate 7 out of the 44 members of Π_1 and 24 out of the 34 members of Π_2; one member of the 18 BRCA1 samples would be misallocated.

The misallocation rate of 24/34 for the second class Π_2 is not surprising given the gene expressions as summarized in the groups of genes (see Figs. 2 to 4). Also, one has to bear in mind that we are classifying the tissues in an unsupervised manner without using the knowledge of their true classification. But even when such knowledge was used (supervised classification) in van 't Veer *et al.* [10], the reported error rate was approximately 50% for members of Π_2 when allowance was made for the selection bias in forming a classifier on the basis of an optimal subset of the genes [2]. Further analysis of this data set in a supervised context by Tibshirani and Efron [9] confirms the difficulty in trying to discriminate between the disease-free class Π_1 and the metastases class Π_2.

8. Assessing the Number of Tissue Groups

We also considered the choice of the number of components g to be used in our normal mixture. The likelihood ratio statistic λ was adopted for this purpose, and we used the resampling approach of McLachlan [6] to assess the $P-$value. This is because the usual chi-squared approximation to the null distribution of $-2\log\lambda$ is not valid for this problem, due to the breakdown in regularity conditions. We proceeded sequentially, testing the null hypothesis H_0: $g = g_0$ versus the alternative hypothesis H_1: $g = g_0 + 1$, starting with $g_0 = 1$ and continuing until a non-significant result was obtained. We concluded from these tests that $g = 3$ components were adequate for this data set.

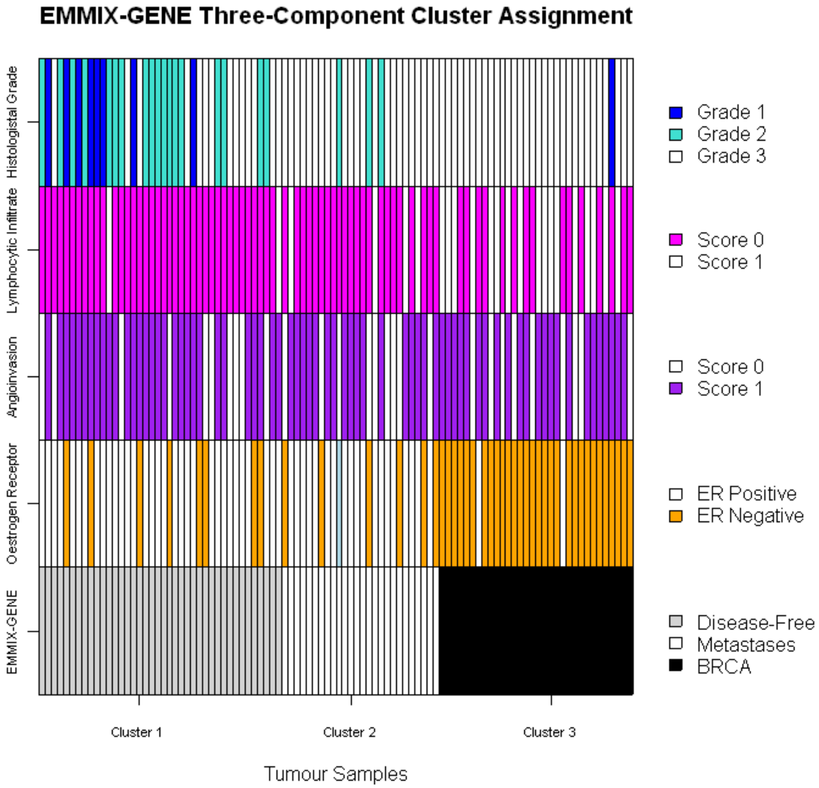

Fig. 13. Comparing EMMIX-GENE cluster assignments with other clinical indicators.

9. Investigating Underlying Signatures with Other Clinical Indicators

For each of the tumour samples in this data set, additional clinical predictors containing information about histological grade, angioinvasion and lymphocytic infiltrate was included. We investigated whether the three clusters constructed by EMMIX-GENE followed patterns according to these biological indicators. The tumour samples have been ordered in Fig. 13 according to the three clustered groups.

Tumours assigned to Cluster 3 appear to match tumours labelled ER negative, while the majority of tumours in Clusters 1 and 2 were ER positive. A close association was also noted between tumours assigned to Cluster 1 and a histological grade of 1, and, to a lesser degree, a grade of 2, while the tumours in Clusters 2 and 3 were more likely to have a histological grade of 3. Some association was visible between Clusters 1 and 2 and the lymphocytic infiltrate score, where the majority of tumours in these clusters had scores of 0, while tumours in Cluster 3 had scores of 1. Indicators related to angioinvasion did not bear a strong association with the EMMIX-GENE clusters. These observations were consistent with those reported by van 't Veer *et al.* [10].

10. Discussion

In this study, we have demonstrated how the model-based algorithm EMMIX-GENE can be applied to cluster a limited number of tissue samples (98 breast cancer tumours) on the basis of subsets of genes selected from an initial set of 24, 881 genes. The filtered set of 4,869 genes was further reduced by using the selection option of EMMIX-GENE to eliminate those genes that showed little variation across the 98 breast cancer tumours. The 1,867 genes so retained were then clustered into forty groups. Based on the means of these forty groups, the tissue samples were clustered into two and three clusters using a mixture of factor analyzers model with $q = 4$ factors.

Identification of the clusters produced by EMMIX-GENE with the externally existing classes Π_1(disease-free group), Π_2(metastases group), and Π_3(BRCA), gives an error rate that is not small. However, this clustering is consistent with the gene expressions as displayed in the heat maps for the 40 groups of similar genes. For example, in the first three groups given in Figs. 2 to 4, it can be seen that those tissues of class $\Pi_2$that have been misallocated to $\Pi_1(\Pi_3)$ have similar gene expression patterns to those of the majority of the tissues in class $\Pi_1(\Pi_3)$. Likewise, the tissues of class $\Pi_1$that have been misallocated to $\Pi_2$have similar gene expression patterns to those of the majority of the tissues in class Π_2. This comparison provides some insight into why even in a supervised context there is difficulty in trying to discriminate between the disease-free class Π_1 and the metastases class $\Pi_{2.}$

References

1. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", *Nature* **403** (2000) 503–511.
2. C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on basis of microarray gene expression data", *Proc. National Academy of Sciences USA* **99** (2002) 6562–6566.
3. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proc. National Academy of Sciences USA* **95** (1998) 14863–14868.
4. D. Ghosh and A. M. Chinnaiyan, "Mixture modelling of gene expression data from microarray experiments", *Bioinformatics* **18** (2002) 275–286.
5. V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. Hudson, M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown, "The transcriptional program in the response of human fibroblasts to serum", *Science* **283** (1999) 83–87.
6. G. J. McLachlan, "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture", *Applied Statistics* **36** (1987) 318–324.
7. G. J. McLachlan, R. W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data", *Bioinformatics* **18** (2002) 413–422.
8. W. Pan, J. Lin, and C. T. Le, "Model-based cluster analysis of microarray gene expression data", *Genome Biology* **3** (2002) Research 0009.1-0009.8.
9. R. J. Tibshirani and B. Efron, "Pre-validation and inference in microarrays", *Statistical Applications in Genetics And Molecular Biology* **1** (2002) 1.

10. L. J. van 't Veer, H. Dai, M. van de Vijver, Y. D. He, A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer", *Nature* **415** (2002) 530–536.
11. K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data", *Bioinformatics* **17** (2001) 977–987.