Chapter 3

Clustering Microarray Data

The potential of clustering to reveal biologically meaningful patterns in microarray data was quickly realised and demonstrated in an early paper by Eisen et al. (1998), who used hierarchical clustering to identify functional groups of genes. As discussed in Chapter 2, hierarchical clustering is one of many conventional clustering methods that could be applied to microarray data and details of such methods may be found elsewhere (see Everitt et al., 2001, for a general introduction to cluster analysis). Whilst standard techniques can be effective in clustering microarray data, there are a number of common assumptions made by conventional methods that do not accommodate some of the features of gene expression.

Partitioning methods assume that each subject belongs to one group. This may not be appropriate for clustering genes, since genes can be involved in more than one active biological process or not be involved in any of the active processes. In the first case the underlying grouping structure may overlap and in the second case the grouping structure may not be exhaustive. Hierarchical clustering may be used to represent subgroups, but can not represent a partial overlap between groups as may be required. Redundant clusters may be avoided by filtering out genes with near-constant expression profiles, but such gene selection usually results in a large proportion of the data being thrown away on the basis of some ad hoc criteria. It is preferable to base an analysis on all of the data, using a clustering method that leaves genes with "uninteresting" patterns unclustered.

Conventional one-way clustering methods are based on similarity between subjects across all variables. However genes may be co-regulated under limited conditions and show little similarity outside these conditions. In this case, a group of

genes may be represented by a gene cluster and an associated subset of the samples which distinguishes the cluster. If the samples in the data set were taken over time, then gene clusters should be based on all the samples, but it may be more appropriate to use a clustering method that is designed for clustering time series, which clusters genes on the basis of key features of their expression profiles.

Assumptions of conventional clustering methods such as those described above may also be too restrictive for sample clustering. For example, if the samples represent patients diagnosed with a certain disease, it may not be appropriate to require every sample to belong to a cluster, as this does not allow for misdiagnosis.

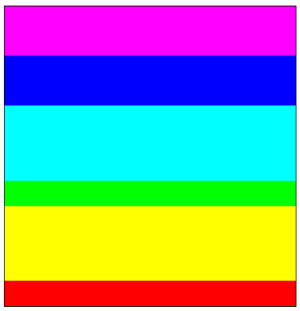
Such limitations of standard clustering techniques have motivated considerable research, leading to the development of several new methods for clustering microarray data. This chapter reviews some of these methods, paying particular attention to the extent to which they address the issues described above. The clustering methods will be compared according to the structure of the clusterings they produce; the nature of clusters they identify, and the search strategy they employ to find these clusters.

The review is divided into three sections: one-way clustering, two-way clustering and biclustering. One-way clustering methods may be used to find either gene clusters or sample clusters. Two-way clustering methods may be used to find both gene clusters and sample clusters in a combined approach. Biclustering methods may be used to find two-dimensional clusters, that is, gene clusters that are only defined over an associated sample cluster that is found simultaneously.

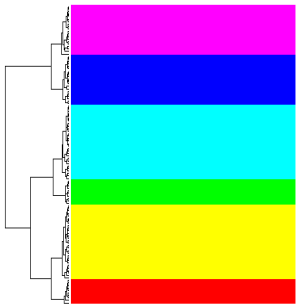
These categories are ordered by a general increase in flexibility of clustering structure. The classification also contrasts methods in the first two categories with methods in the third, from which the plaid model is selected for investigation in the remainder of this thesis.

3.1 One-way Clustering

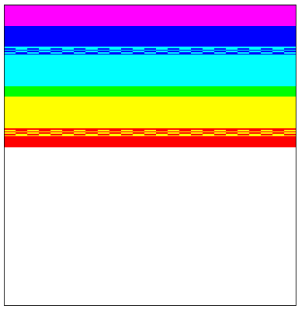
There are several one-way clustering methods that have been designed for the analysis of microarray data, usually motivated by the search for grouping structure in the genes. These methods will be described with reference to two keys: the first, in Figure 3.1, illustrates different clustering structures that one-way clustering methods may produce and Figure 3.2 illustrates different types of cluster.



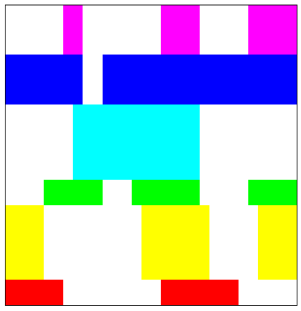
(a) Partition



(b) Partition obtained from hierarchy



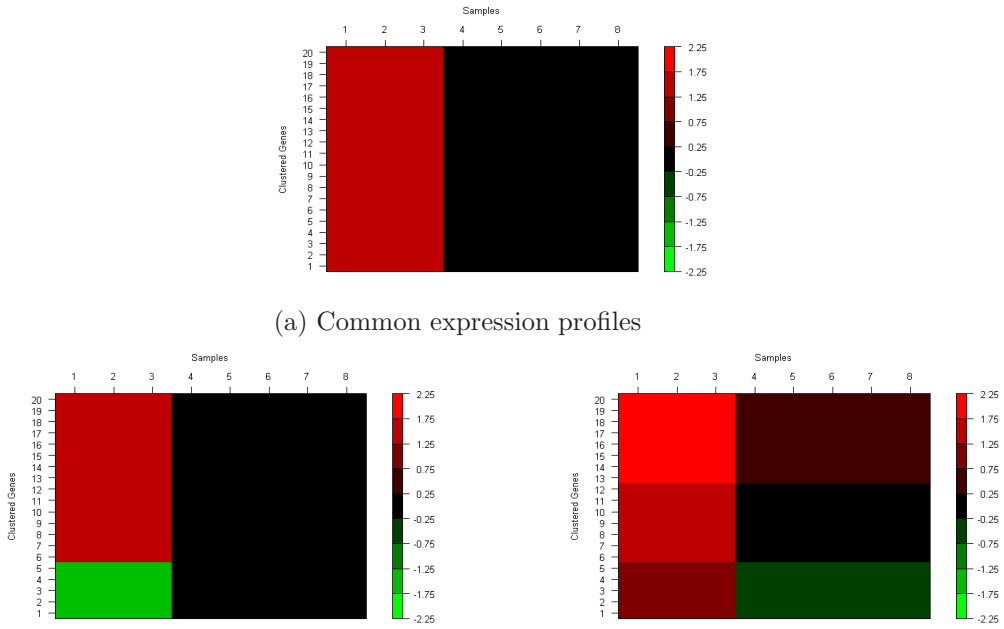
(c) Overlapping non-exhaustive clusters



(d) Context-specific clusters

Figure 3.1
One-way clustering structures. The images represent a gene by sample expression matrix in which the genes have been clustered into six clusters, identified by the different colours. Chequered blocks represent overlapping clusters.

Certain one-way clustering techniques allow for genes that are involved in multiple active processes, as well as genes that are not involved in any active process, by isolating possibly overlapping clusters from the data as in Figure 3.1(c). Percolation clustering, introduced by Šášik et al. (2001) is one such method, in which clusters are built up by connecting neighbouring points or clusters in a way that is similar to agglomerative hierarchical clustering, except that connections are made on a probabilistic basis and the clustering is repeated several times to produce an "average tree". This tree is used to filter out "uninteresting" clusters and the probability of membership to the remaining clusters is then calculated for each gene.



(a) Common expression profiles

(b) Mirrored expression profiles

(c) Common change in expression over samples

Figure 3.2
Types of one-way cluster. The images represent the (noiseless) expression levels of a cluster of twenty genes from a data set with eight samples.

Due to the probabilistic approach taken, the method is not greedy like ordinary hierarchical clustering, but more Monte-Carlo like. Since the clustering is based on a distance measure, genes in a cluster have a common profile as in Figure 3.2(a), but "uninteresting" clusters are filtered out on the basis of a minimum size.

Rather than simply looking for similar genes, the gene shaving approach of Hastie et al. (2000) searches for coherent clusters with high between-sample variance, ignoring genes involved in constantly activated processes as well as those involved in none of the active processes. The gene shaving algorithm finds a series of nested clusters on the basis of correlation with the leading principal component, such that each nested cluster has the maximum variance of the cluster mean, given the cluster size. The nested cluster with the largest difference between its R^2 value and its expected R^2 value is selected for the final result. Once a cluster has been selected, the data is orthogonalised with respect to the cluster centroid in order to search for a further cluster. Clusters are sought until a pre-specified number of clusters is

reached. The gene shaving approach is underpinned by the following model for the gene expression level, Y_{ij}, $i = 1, \ldots, n; j = 1, \ldots, p$ of the ith gene in the jth sample

$$Y_{ij} = \sum_{k=1}^{K} \rho_{ik} e_{jk} + \epsilon_{ij}$$

where $\rho_{ik} \in \{0, 1\}$ indicates whether the ith gene is in the kth cluster, such that $\sum_{k=1}^{K} \rho_{ik} \geqslant 0$; e_{jk} is the jth element of the kth eigenvector, and $\epsilon_{ij} \sim N(0, \sigma^2)$ is the error. Gene shaving clusters usually cover a small proportion of the data and as they are found independently, may overlap, giving the structure of Figure 3.1(c). In addition, correlation can be positive or negative, so clusters can include profiles of opposite sign as in Figure 3.2(b).

Other one-way clustering methods allow clusters to be based on a subset of the attributes. Friedman and Meulman (2004) use this approach in the context of clustering samples, allowing for the situation where only a small proportion of the genes are useful in distinguishing a particular cluster. This type of cluster may be difficult to uncover by giving equal weight to all the genes. Their procedure, Clustering Objects on Subsets of Attributes (COSA) computes distances between samples, giving the expression levels gene- and sample-specific weights. These distances are then passed to a distance-based clustering algorithm, such as hierarchical clustering, to cluster the samples. This will identify clusters characterised by a common profile as illustrated (for a gene cluster) in Figure 3.2(a) . The structure of the clustering will depend on the method used, for example, hierarchical clustering would give a structure as illustrated (for gene clustering) in Figure 3.1(b). Thus the form and structure of the clusters is conventional, but the importance of each gene in the discovery of a sample cluster can be quantified and relevant genes can be isolated.

Barash and Friedman (2002) propose a context-specific Bayesian clustering technique in which an explicit subset of relevant attributes is determined for each cluster, as illustrated in Figure 3.1(d). They focus on clustering genes, but also demonstrate their method on clustering samples. Expression data may be analysed in conjunction with other sources of data, so that attributes such as the occurrence of putative binding sites in the promoter region of the genes can also be included. The clustering is exhaustive so may not be suitable for clustering full sets of genes, but clusters may overlap and the number of clusters is determined automatically. The clusters

are estimated using a structural EM algorithm, based on the probabilistic model

$$P(Y_1, \ldots, Y_p|G) = \Big(\prod_{j \notin G} P(Y_j) \Big) \sum_{k}^{K} \Big(P(C = k) \prod_{j \in G} P(Y_j|C = k) \Big)$$

where the Y_j are attributes, G is the union of all relevant attributes, C is the cluster variable and

$$P(Y_j|C = k) = \begin{cases} P(Y_j|C = k_l) & \text{if } k = k_l \in L_j \\ P(Y_j) & \text{otherwise} \end{cases}$$

in which L_j is the group of clusters for which attribute Y_j is relevant. An attribute will be relevant to a cluster if the clustered objects have unusual values for that attribute. The (conditional) probability distributions of continuous attributes, such as the expression level on a particular array, are modelled as Gaussian. For discrete attributes, multinomial distributions are used. Therefore clustered genes will have a common expression profile, as shown for a gene cluster in Figure 3.2(a).

Several methods have been proposed for clustering gene expression time series. They differ from other methods by using the actual time values to model the gene expression profiles. In this way dependencies between time points are accounted for, which is particularly important when the sampling rate is not uniform. Ramoni et al. (2002) use a clustering model in which the temporal profiles are represented by autoregressive (AR) equations. In this model, Y_{ij} the expression level of gene i in sample j, is given by

$$Y_{ij} = \sum_{k=1}^{K} \rho_{ik} (\beta_{ik0} + Y_{i(j-1)} \beta_{ik1} + \ldots + Y_{i(j-p)} \beta_{ikp}) + \epsilon_{ijk}$$

where $\rho_{ik} \in \{0, 1\}$ indicates whether gene i is in cluster k, such that $\sum_{k=1}^{K} \rho_{ik} \geqslant 0$; the β_{bik} are regression coefficients; p is the order of the model, and the $\epsilon_{ijk} \sim N(0, \sigma_k^2)$ model the error. This model is estimated using a Bayesian agglomerative hierarchical method, which seeks the model with maximum posterior probability at each stage and stops when it has found a set of clusters that cannot be merged without reducing the marginal likelihood. This gives a partition of the data with associated hierarchical relationships within each cluster as illustrated in Figure 3.1(b). The autoregressive model suggests that expression levels of profiles in the same cluster

change over time in a similar way, but the actual expression levels need not be the same. This type of cluster is illustrated by Figure 3.2(c). The disadvantages of the approach of Ramoni et al. (2002) are that the final clusters can be greatly affected by small errors made early on and that AR models have several limitations, for example in the interpretation of the coefficients when time points are not evenly spaced (Luan and Li, 2003).

More flexibility is given by using spline models of the form

$$Y_{ij} = \sum_{k=1}^{K} \rho_{ik} \Big(\sum_{l=1}^{p} \beta_{kl} S_l(t_{ij}) + \sum_{l=1}^{q} \gamma_{il} \bar{S}_l(t_{ij}) \Big) + \epsilon_{ijk}$$

where $\rho_{ik} \in \{0, 1\}$ indicates whether the ith gene is in the kth cluster, , such that $\sum_{k=1}^{K} \rho_{ik} \geqslant 0$; the β_{kl} are coefficients of the spline basis $\bar{S}$ for the kth cluster; the γ_{il} are normal random coefficients of the spline basis S with mean zero and covariance matrix $Cov(\gamma_i) = \Gamma_k$, and $\epsilon_{ijk} \sim N(0, \sigma^2$ is residual error. This model represents a standard partition of the data, as in Figure 3.1(a). The inclusion of random gene effect curves implies that clustered profiles have the same pattern of expression, as in Figure 3.2(c), not necessarily similar expression levels, though the degree of similarity will depend on the covariance matrix Γ_k.

The clustering method of Bar-Joseph et al. (2002) is based on this spline model, using the same cubic spline basis vector to model the fixed cluster effect curve and the random gene effect curve. Their clustering algorithm finds a pre-specified number of clusters and the probabilities for each gene of belonging to each cluster. They use an EM algorithm, which alternates between updating the membership probabilities and updating the parameters of the expression profile models. Luan and Li (2003) propose a similar approach, using different B-spline bases for the fixed and random effect curves. To avoid too many parameters in the model they assume a common covariance matrix for the γ_i across clusters. They also consider determining the number of clusters and allowing for genes that don't belong to any cluster.

Heard et al. (to appear) use a spline model without a random gene effect curve, so clustered profiles have a common expression profile as in Figure 3.2(a). They choose to use a truncated power spline basis to model the cluster profiles. By adopting conjugate priors on the coefficients of the regression splines they obtain an analytical expression for the marginal likelihood, which they then seek to optimise through

agglomerative hierarchical clustering. The number of clusters that maximises the posterior distribution is selected for the final result, giving a partition with an associated hierarchy as in Figure 3.1(b). Thus the method of optimisation and the structure of the clustering is similar to the method of Ramoni et al. (2002), but the underlying model is more flexible, in particular it allows for non-stationary time series.

Finally Wakefield et al. (2003) consider the generic time series model

$$Y_{ij} = \sum_{k=1}^{K} \rho_{ik} f(\theta_k, t_{ij}) + \epsilon_{ijk}$$

in which $\rho_{ik} \in \{0, 1\}$ indicates whether gene i is in cluster k, such that $\sum_{k=1}^{K} \rho_{ik} \geqslant 0$, and f is a function of time with parameters θ_k that depends on the experimental context. For example, for periodic data, they propose the random effects model

$$Y_{ij} = \sum_{k=1}^{K} \rho_{ik} (A_{ik} \sin(\omega t_{ij}) + B_{ik} \cos(\omega t_{ij})) + \epsilon_{ijk}$$

where the A_{ik} and B_{ik} are assumed to be bivariate normal and ω is fixed on the basis of prior knowledge of the period. Priors for the distribution of A_{ik}, B_{ik} are also determined from the data and then the clustering model is estimated using MCMC. The result is a standard partition of the data as in Figure 3.1(a) in which clustered profiles have a common expression profile as illustrated in Figure 3.2(a).

3.2 Two-way Clustering

A simple way to cluster both genes and samples is to apply a one-way clustering method to each dimension and then to relate the two analyses with the aid of an ordered plot of expression values (see e.g. Alon et al., 1999; Alizadeh et al., 2000). However if there are relationships between gene and sample clusters, it may be more appropriate to use a method in which the clustering of one dimension is dependent on the clustering of the other. This form of clustering is what is meant by two-way clustering here.

Generic two-way clustering techniques obtain two-way clusters by applying a one-way clustering method in a sequential manner. Such methods will be considered

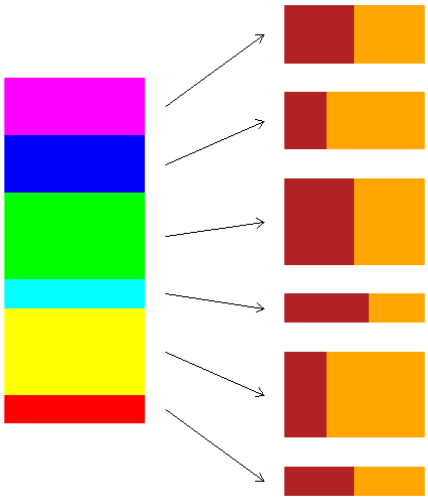
first, before reviewing simultaneous two-way clustering methods in Section 3.2.1. A diagrammatic guide to generic two-way clustering techniques techniques is given in Figure 3.3. For these methods the type of clusters obtained will not usually be described since it will depend on the one-way clustering method used. However where a particular method is described reference will be made to Figure 3.4 which gives a key to types of two-way cluster.

Two-way clustering may be used to identify a subset in one dimension that is useful for clustering the other dimension. This is the idea behind the interrelated two-way clustering method of Tang et al. (2001). In this method, the genes are clustered and each gene cluster is used to cluster the samples, as illustrated in Figure 3.3(a). The results of the different sample clusterings are used to filter out "irrelevant" genes. The reduced gene set is then used to initiate another iteration of gene and sample clustering and the process is repeated until the sample clusterings reach a certain level of similarity or the number of genes reaches a pre-specified threshold. At this point the remaining genes are used to find a final set of sample clusters. The process of reducing the gene set assumes that each iteration partitions the genes into a fixed number of clusters, say k, and each sample clustering is a partition into a fixed number of clusters, say l. Therefore the method used to cluster the genes and the samples can be any one-way partitioning method for which the number of clusters can be specified, e.g. k-means or self-organising maps (SOM).

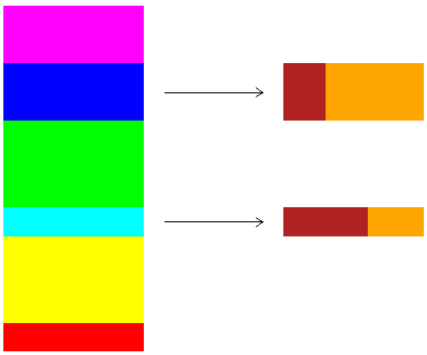
The interrelated two-way clustering method assumes that there is only one meaningful or interesting way to cluster the samples and all gene clusters should be related to this sample grouping. However it may be that different gene clusters reveal different ways of grouping the samples. McLachlan et al. (2002) suggest first clustering the genes; ranking the gene clusters by their potential for clustering the samples and then using selected gene clusters to cluster the samples, as illustrated in Figure 3.3(b). To cluster the genes, McLachlan et al. (2002) consider the p-vector of expression levels for gene to be a realisation of the random vector Y with mixture probability density function

$$f(y; \pi_k, \mu_k, \Sigma_k) = \sum_{k=1}^{K} \pi_k \phi(y; \mu_k, \Sigma_k)$$

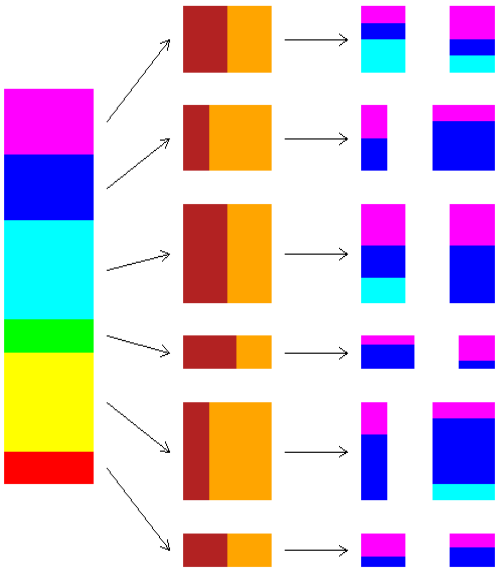
where $\phi(y; \mu_k, \Sigma_k)$ is the p-variate normal density probability function with mean



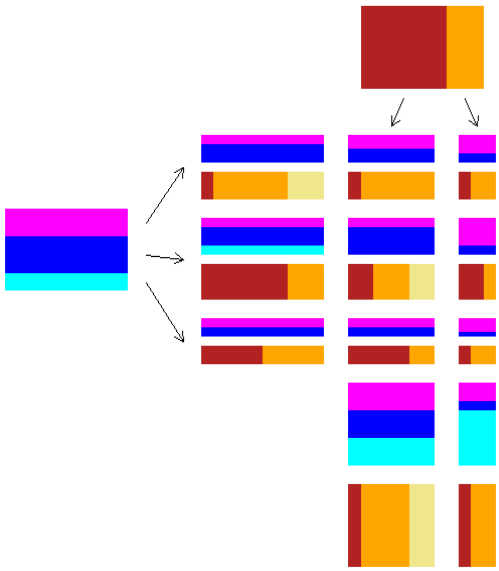
(a) Gene clustering → sample clustering within gene clusters



(b) Gene clustering → sample clustering within selected gene clusters



(c) Gene clustering → sample clustering within gene clusters → gene clustering within sample clusters



(d) Gene and sample clustering → gene and sample clustering within submatrices defined by all pairs of gene and sample clusters from previous iteration

Figure 3.3
Diagrammatic representation of generic two-way clustering techniques, assuming a one-way partitioning method is used to cluster the genes and samples as described.

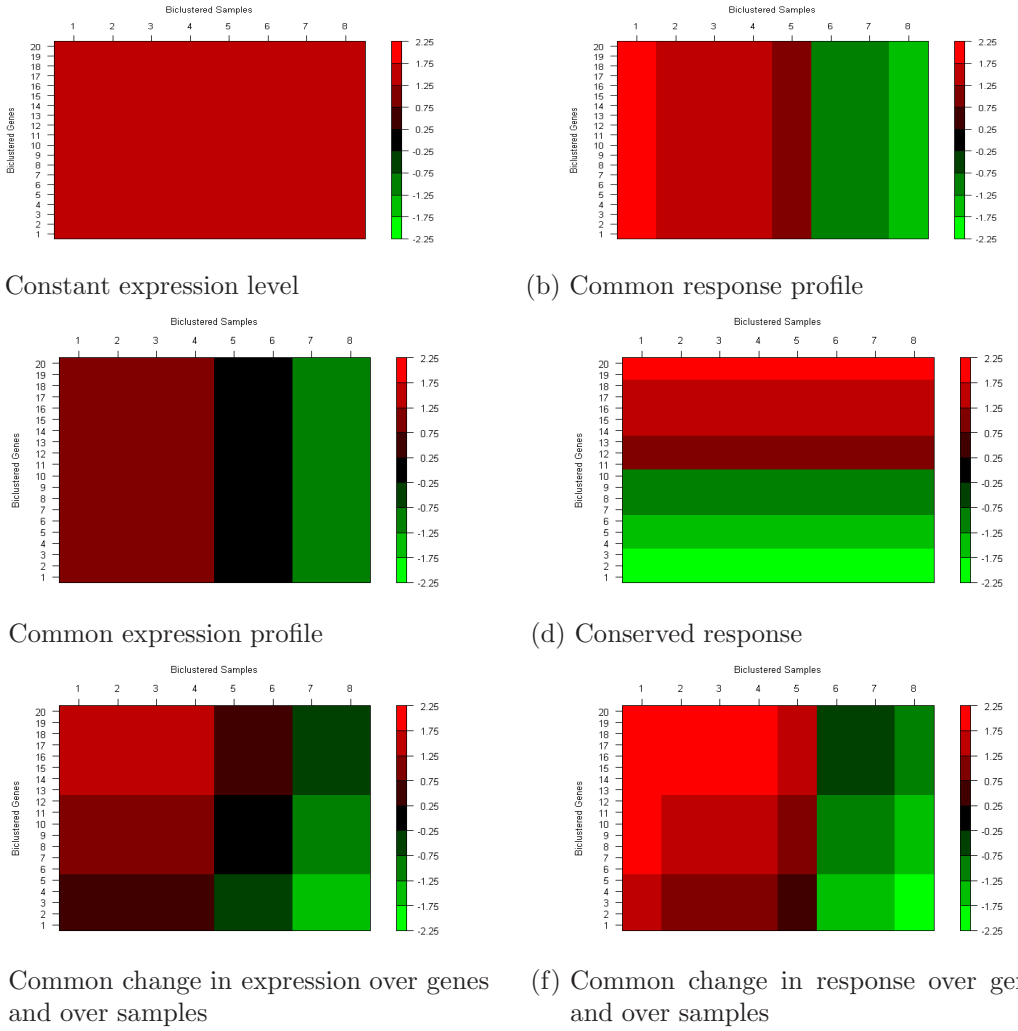
μ_k and covariance matrix Σ_k, and the π_k are the mixing proportions. On the basis of this model the genes are partitioned into K clusters, giving the type of structure shown in Figure 3.1(a). To cluster the samples, McLachlan et al. (2002) propose a mixture of factor analyzers, since the number of genes is typically much greater than the number of samples. This model is similar to the normal mixture model, except that the covariance matrix is given by

$$\Sigma_k = B_k B_k^T + D_k$$

where B_k is a matrix of factor loadings and D_k is a diagonal matrix. McLachlan et al. (2002) fit the normal mixture model and the factor analyzers mixture model by maximum likelihood estimation using EM algorithms; the factor analyzers mixture model requiring a variant of the EM algorithm called Alternating Expectation-Conditional Maximisation (AECM). Either mixture model will produce clusters characterised by a common expression profile, as illustrated in Figure 3.4(c).

The concept of clustering samples within gene clusters is included in the general framework proposed by Pollard and van der Laan (2002), in which a "simultaneous" clustering function is defined as a composition of gene and sample clustering functions. This produces a hierarchical clustering method in which one-way clustering is performed within (possibly two-way) clusters from the previous stage. Examples of such simultaneous clustering methods are illustrated by Figure 3.3(a) (gene clustering followed by sample clustering) and Figure 3.3(c) (gene clustering followed by sample clustering followed by gene clustering).

Getz et al. (2000) propose a related procedure, Coupled Two-Way Clustering (CTWC), in which a one-way clustering method is applied to both genes and samples within submatrices defined by gene and sample clusters from previous iterations. In this case however, the clustering is not strictly hierarchical. The submatrices considered are all possible pairs of gene and sample clusters, including both the full set of genes and the full set of samples, as illustrated (for two iterations) in Figure 3.3(d). The process is terminated when all of the clusters found in an iteration fail to meet pre-specified criteria (which may be with respect to size or stability for example). Clusters obtained at any stage of the process that pass a stability criterion are selected for the final set of clusters. CTWC can be used with any one-way clustering method, the choice of which will determine the nature of the clusters



(a) Constant expression level

(b) Common response profile

(c) Common expression profile

(d) Conserved response

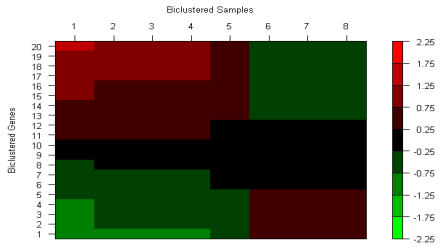
(e) Common change in expression over genes and over samples

(f) Common change in response over genes and over samples

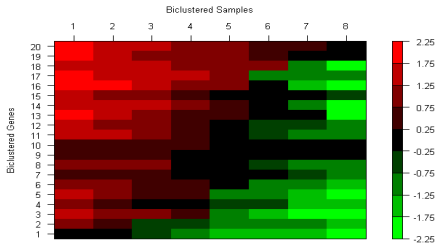
Figure 3.4
Types of two-way cluster. The images represent the (noiseless) expression levels of a two-way cluster of twenty genes and eight samples. Unless otherwise specified, the patterns are described in terms of the gene expression profiles. A "response" is defined an up-regulation or down-regulation, i.e. a non-zero expression level.

discovered.

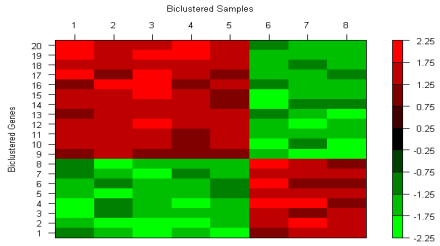
Methods that iteratively cluster genes and samples have the potential to reveal local correlations as shown in the use of CTWC by Getz et al. (2003), but with many iterations there is a danger that the search will become too aggressive as the analysis is based on smaller and smaller data sets (Pollard and van der Laan, 2002).



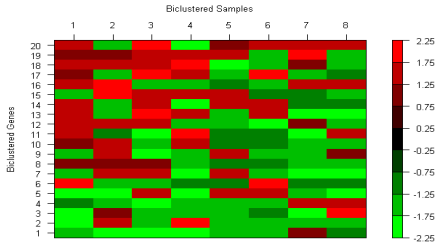
(g) Mirrored expression profiles with different scales



(h) Common rank expression profile



(i) Mirrored response profiles



(j) Arbitrary response

Figure 3.4
Types of two-way cluster (contd.). The images represent the (noiseless) expression levels of a two-way cluster of twenty genes and eight samples. The patterns are described in terms of the gene expression profiles. A "response" is defined an up-regulation or down-regulation, i.e. a non-zero expression level.

3.2.1 Co-clustering

Co-clustering is a form of two-way clustering in which both dimensions are clustered simultaneously. The concept was introduced by Dhillon (2001) in the context of document-keyword analysis and Kluger et al. (2003) proposed a similar method for co-clustering gene expression data. Although Kluger et al. (2003) refer to their method as spectral "biclustering", it is differentiated from biclustering methods here, as the clusters are dependent on the full expression profile of genes or samples, and the clustering results in exhaustive, non-overlapping clusters.

In the method of Kluger et al. (2003), clustering is based on the singular vectors of the row- and column-centred data. Pairs of left and right singular vectors that are approximately piecewise constant indicate a block structure in the expression levels of the original matrix, so the singular vectors are examined to find the best-partitioning pair in this sense. These singular vectors indicate a partitioning of both the genes and the samples, as well as an ordering of the genes and samples within

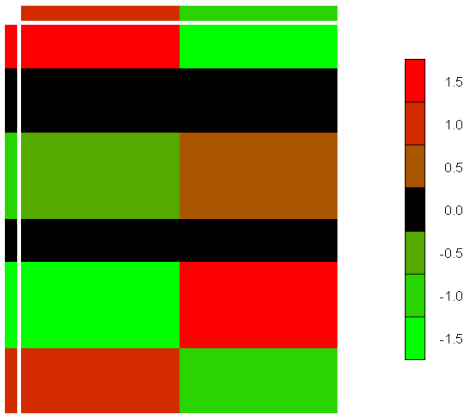


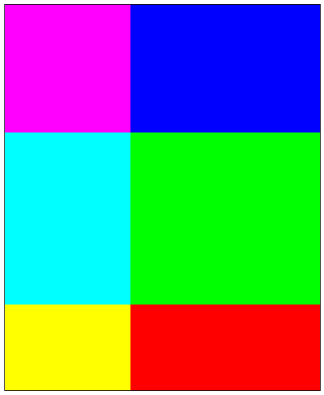
Figure 3.5
Spectral biclustering structure and form. The coloured bars on the left side and the top side of the plot represent the gene clusters and the sample clusters derived from the best-partitioning left and right singular vectors respectively. The image plot represents the values given by the inner product of these singular vectors, with corresponding block structure.

these clusters according to the actual values of the eigenvectors. The inner product of the best-partitioning singular vectors may be used to represent the block structure in the original matrix, as illustrated in Figure 3.5. Further pairs of singular vectors may be used to enhance the partitioning of either genes or samples, for example by projecting the data into the space of the first two best-partitioning "eigengenes" and clustering the data in that space, using a conventional clustering algorithm such as k-means.

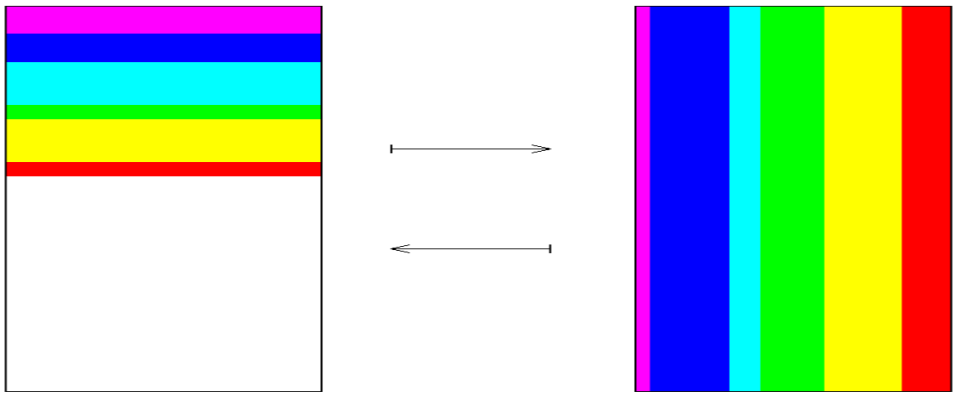
Other co-clustering methods have been proposed for the analysis of microarray data, in which the type and structure of the co-clusters are not so closely interrelated. These methods will be described with reference to the key to co-clustering structures given by Figure 3.6 and the keys to cluster types given by Figures 3.2 and 3.4.

Cho et al. (2004) propose a co-clustering method which simultaneously partitions the genes and the samples, so that the data is partitioned into a grid of co-clusters as illustrated in Figure 3.6(a). They propose two alternative co-clustering models, in which co-clusters are either modelled by their mean or a two-way additive model. In the second case, the expression level Y_{ij} for the ith gene and the jth sample is given by

$$Y_{ij} = \sum_{k=1}^{K} \rho_{ik}\kappa_{jk}(\mu_k + \alpha_{ik} + \beta_{jk}) + \epsilon_{ij}$$



(a) Two-way partition



(b) Conjugate gene and sample clusters

Figure 3.6
Co-clustering structures. The images represent a gene by sample expression matrix in which six co-clusters are identified by six different colours.

where $\rho_{ik} \in \{0, 1\}$ indicates whether gene i is in cluster k, $\kappa_{jk} \in \{0, 1\}$ indicates whether sample j is in cluster k, and μ_k, α_{ik} and β_{jk} are the mean, gene and sample effects of cluster k respectively. The two-way partition is given by the constraints $\sum_k \kappa_{jk} = l$ and $\sum_k \rho ik = m$, where l and m are the number of gene and sample partitions respectively. The mean-only model will identify blocks of similar expression level as in Figure 3.4(a), whereas the two-way model will identify blocks of similar expression pattern, as illustrated in Figure 3.4(e). The full co-clustering model is estimated using an alternating least squares procedure. A disadvantage of the approach of Cho et al. (2004) is that the number of gene clusters and the number of sample clusters must both be specified. However the modelling framework offers

potential for extensions such as the inclusion of anti-correlated genes in clusters.

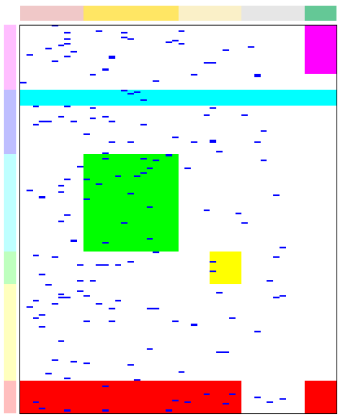
The co-clustering methods described so far partition the genes and samples. As discussed earlier, it may be that some genes and samples are not particularly informative and may be better left unclustered. The Double Conjugated Clustering (DCC) method of Busygin et al. (2002), allows for this possibility. DCC alternates between clustering genes and samples using a node-driven clustering method such as SOM. After each iteration the nodes of the current clustering space are mapped to conjugate nodes of the other clustering space. The end result is a set of gene clusters and a conjugate set of sample clusters as illustrated in Figure 3.6(b). A gene cluster is interpreted as those genes which can be used to distinguish the conjugate sample cluster from the other samples. "Uninteresting" genes that do not discriminate enough between the samples are clustered by nodes having no samples in the sample clustering space and are therefore ignored (treated as unclustered). Sample clusters may be interpreted in a similar way with respect to the genes. Although each cluster has a conjugate in the opposite clustering space, the conjugate is not taken as a subset of attributes on which the clustering is based, rather the full data set is used in each clustering cycle. Thus the resultant clusters are one-way and are characterised by a common expression profile as illustrated for a gene cluster in Figure 3.2(a).

3.3 Biclustering

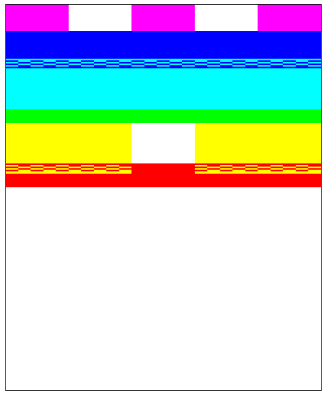
A bicluster is a cluster of the genes and an associated cluster of the samples over which the genes are co-regulated. Biclustering does not seek to cluster all the genes or all the samples; the aim is to identify possibly overlapping submatrices of the data that exhibit interesting patterns - leaving the remaining data unclustered. Thus biclustering may be viewed as an extension of context-specific one-way clustering. It combines the features of iterative two-way clustering and co-clustering, in that local dependencies can be discovered, but the analysis is based on the full expression matrix and the genes and samples are clustered simultaneously.

Biclusters are of course a type of two-way cluster, so Figure 3.5 continues to provide a reference for cluster type. A key to biclustering structures is given by Figure 3.7.

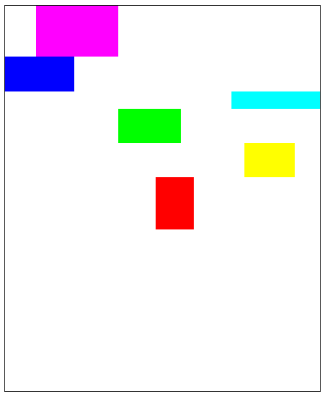
Segal et al. (2001) use probabilistic relational models (PRMs) to model the de-



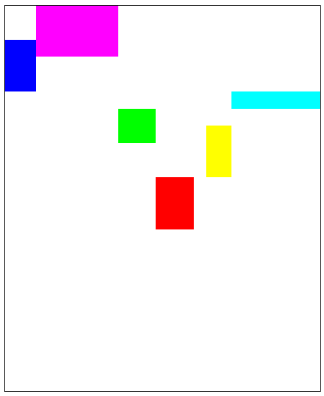
(a) Biclusters dependent on latent gene and array clusters as well as other attributes



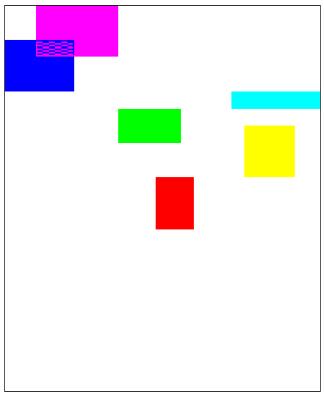
(b) Clusters dependent on latent processes, some of which are not active in all arrays



(c) Biclusters with no overlap in gene membership



(d) Biclusters with no overlap in sample membership



(e) Biclusters with unrestricted membership

Figure 3.7
Biclustering structures. The images represent a gene by sample expression matrix in which six clusters are identified by six different colours. Chequered blocks represent overlapping clusters. In plot (a), latent clusters are represented in the sidebars by blocks of muted colour.

pendency of expression levels on gene and array attributes, which may include gene and array clusters. The dependencies are represented by a binary tree with conditions on the attributes at the nodes and expression levels at the leaves. Thus a bicluster may be represented by a leaf with the parents "Gene cluster = 3" and "Sample cluster = 4", for example. However the PRM may also identify one-way clusters, that depend only on a gene cluster, say, or subsets of the data that are not

aligned with the (latent) gene and array clusters since the expression levels depend on the values of other variables, such as the functional annotation of the genes or the presence of binding sites for certain transcription factors. An illustration of the type of clustering that may be defined by such a PRM is shown in Figure 3.7(a). Since each cluster is represented by a single expression level, clustered expression levels will be similar in value, as illustrated for a bicluster in Figure 3.4(a). The structure of the PRM and parameters of the associated conditional probability distributions are estimated using a variant of the structural EM algorithm. The method of Segal et al. (2001) is very flexible in the type of grouping that can be identified, but the assumption of a latent clustering of the genes and arrays restricts the gene and array clusters that can be used in the model.

Also using PRMs, Segal et al. (2003) propose an approach for identifying cellular processes and the biclusters in which they are active. The approach is based on the following probability model for the expression level Y_{ij} of the ith gene in the jth sample

$$P(Y_{ij}) = exp\left(\frac{Y_{ij} - \sum_{k=1}^{K} \rho_{ik}\beta_{jk}}{2\sigma_j^2}\right)$$

in which $\rho_{ik} \in \{0, 1\}$ indicates whether gene i participates in process k, such that $\sum_{k=1}^{K} \rho_{ik} \geqslant 0$; μ_{jk} is the activity level of process k in sample j, and σ_j is the standard deviation of all expression measurements in sample j. The set of genes participating in a given process may be viewed as a cluster. Since the activity of a process is modelled separately for each sample, clusters are characterised by a common expression profile, as illustrated in Figure 3.4(b). A bicluster may be represented by a set of genes belonging to a process with zero activity in some samples. Thus the PRM may represent a number of overlapping one-way clusters and biclusters, as illustrated in Figure 3.7(b). Segal et al. (2003) use an EM algorithm to determine which genes belong to each process and the activity level of each process in each sample, for a given number of processes.

MacKay and Miskin (2001) use a similar clustering technique in which expression levels are modelled by latent variables, which have different strengths for each gene and sample. More specifically, the expression level Y_{ij} of the ith gene in the jth sample is modelled as

$$Y_{ij} = \sum_{k=1}^{K} \alpha_{ik}\beta_{jk} + \epsilon_{ijk}.$$

where K is the number of latent variables; α_{ik} and β_{jk} are the effects of latent variable k on the expression levels of gene i and sample j respectively, and ϵ_{ijk} is the error. Thus the membership indicators ρ_{ik} in the model of Segal et al. (2003) are replaced by multiplicative gene effects, α_{ik}. Clusters are represented by processes for which some of the α_{ik} and/or some of the β_{jk} are (approximately) zero. The clustering structure is therefore the same as that induced by the model of Segal et al. (2003) (illustrated in Figure 3.7(b)), but the clustered expression profiles need only be the same up to a multiplicative factor, as illustrated in Figure 3.4(g). MacKay and Miskin (2001) use a variational approach to estimate this clustering model. A drawback of this model is that it is designed for modelling untransformed expression levels, rather than the commonly used log-transformed expression levels. For log-transformed data, multiplicative gene effects may not be appropriate. A further disadvantage of this method, and the related method of Segal et al. (2003), is that they may have a tendency to produce "fuzzy biclusters", which include genes and samples that appear to have a weak involvement in a process.

Sheng et al. (2003) model the expression levels within a bicluster by a set of multinomial distributions, one for each sample in the bicluster. Thus the biclustered genes are assumed to share the same state of expression within each of the samples, but this state of expression may differ across the samples, giving the type of structure illustrated in Figure 3.4(c). In order to distinguish unusual distributions of expression levels, the background expression levels are modelled by a single multinomial distribution. Biclusters are found one at a time using a Gibbs sampling approach. Once genes are assigned to a bicluster they are masked from further analysis, so genes can only belong to one bicluster. This leads to the type of the biclustering structure illustrated in Figure 3.7(c). Since the expression levels are modelled by multinomial distributions, the data must first be discretised to use this method. Sheng et al. (2003) suggest dividing the data into three bins, which may over-simplify the expression pattern and magnify the effect of noise on some expression levels.

Tanay et al. (2002) introduced the Statistical Algorithmic Method for Bicluster Analysis (SAMBA), in which the data are modelled as a bipartite graph whose two parts correspond to the genes and the samples, with edges representing significant changes in expression. Edges and non-edges are weighted by likelihood scores derived from a probabilistic model for the bipartite graph. A bicluster is defined as a

heavy subgraph, where the weight of a subgraph is the sum of the weights of the corresponding edges and non-edges. SAMBA identifies the most significant biclusters under simplifying conditions, then searches for local improvements in a heuristic manner. Working with an unsigned graph, the algorithm will produce biclusters of genes that jointly respond to the conditions represented by the biclustered samples. The response may be up-regulation or down-regulation, and need not be the same across the genes or across the samples, as illustrated in Figure 3.4(j). Such biclusters may present difficulties for interpretation. Working with a signed graph, the algorithm will produce biclusters in which every two conditions have either a similar effect or the opposite effect on the biclustered genes, as illustrated in Figure 3.4(i). In either case, biclusters may overlap and need not cover the data matrix, as illustrated in Figure 3.7(e). To use SAMBA, the expression levels must first be converted to up-regulated (1), down-regulated (-1) or unchanged (0) levels, so this method carries the disadvantages of discretisation described above. However, the SAMBA algorithm is being developed to allow greater sensitivity using multiple response levels (Tanay et al., 2002).

Order preserving submatrix clustering (Ben-Dor et al., 2002) defines a bicluster as a cluster of genes with the same rank profile across the biclustered samples. In some sense, this definition is quite broad, since the expression levels of the biclustered profiles may be quite different and even the pattern of the biclustered profiles may vary quite considerably, as illustrated in Figure 3.4(h). On the other hand, the definition may be viewed as over-prescriptive, since it is possible for a group of genes to have a common pattern of expression without having exactly the same rank profile. Ben-Dor et al. (2002) do suggest a variant of their approach in which each value in the submatrix corresponding to the bicluster is exempt from the ordering condition with some probability π. Whilst this allows for slight variations in rank profile, it will of course broaden the definition further, perhaps producing biclusters that are too heterogeneous. Ben-Dor et al. (2002) propose a heuristic algorithm for discovering one order preserving submatrix at a time, which finds a solution for all possible values for s, the number of samples in the bicluster, and selects the most significant of these solutions. For a given value of s, the algorithm iteratively builds sample permutations, choosing the best l permutations in each iteration, until the permutations are of length s, when the best is selected and the probabilities of each gene belonging to the bicluster are calculated. The biclusters retrieved by

the algorithm may overlap, producing the clustering structure illustrated in Figure 3.7(e).

The biclustering method of Cheng and Church (2000) identifies biclusters of genes that have similar or opposite patterns of expression. The biclustered profiles may include near-zero expression levels as shown in Figure 3.4(e) for a bicluster of genes with similar expression pattern only. Cheng and Church (2000) propose a node-deletion algorithm in which a set number of biclusters are found one at a time; after each bicluster is identified, the data corresponding to the bicluster are replaced by random numbers generated uniformly over the range of the full data set. This allows biclusters to overlap, giving the structure illustrated in Figure 3.7(e). The algorithm searches for biclusters of the form

$$Z_{ij} = \delta_{ik}(\mu_k + \alpha_{ik} + \beta_{jk})$$

where Z_{ij} is the expression level or random number corresponding to gene i and sample j in bicluster k; μ_k, α_{ik}, and β_{jk} are mean, gene and sample effects respectively, and $\delta_{ik} \in \{-1, 1\}$ is the sign of the ith profile. Since this model will fit biclusters of genes with near constant expression profiles, the method identifies several trivial biclusters before discovering biclusters of unusual expression patterns.

This issue is addressed in the FLOC algorithm proposed by Yang et al. (2003), which is based on the work of Cheng and Church (2000). In the FLOC algorithm, a lower bound may be set for the variance of biclustered gene expression profiles, to reject trivial biclusters. Moreover, FLOC estimates a set number of biclusters simultaneously, avoiding the problem of random interference caused by masking discovered biclusters in the algorithm of Cheng and Church (2000). FLOC is a probabilistic move-based algorithm, which performs one move per gene and sample in each iteration. The moves in or out of biclusters are made sequentially in a random order, with probabilities determined by the "gain" of each move. Gain is measured in terms of the relative reduction in the mean of the sum of squared residuals over the bicluster and the relative increase in the size of the bicluster. Unlike the algorithm of Cheng and Church (2000), there is no step in the algorithm to include anti-regulated genes, so the δ_{ik} are all 1, giving biclusters of the form illustrated in Figure 3.4(e).

Ambler (2003) propose the following clustering model for the gene expression

levels Y_{ij}

$$Y_{ij} = \sum_{k=1}^{K} \rho_{ik}\kappa_{jk}\mu_k + \epsilon_{ij}$$

in which $\rho_{ik} \in \{0, 1\}$ indicates whether gene i is in bicluster k, such that $\sum_i \rho_{ik} \geqslant 0$; $\kappa_j \in \{0, 1\}$ indicates whether sample j is in bicluster k such that $\sum_j \kappa_{jk} \geqslant 0$, and μ_k is the effect of bicluster k. In this model each bicluster represents a cluster of genes that are expressed at a common level over the biclustered samples, as illustrated in Figure 3.4(a). Genes and samples may belong to more than one bicluster or not belong to any bicluster, giving the structure illustrated in Figure 3.7(e). Ambler (2003) estimate this model using a Bayesian MCMC approach. A drawback of this method is that it can be hard to summarise the posterior distributions of the parameters in a meaningful way. Whilst the number of biclusters and the associated bicluster means can usually be identified, it is difficult to associate genes and samples with these biclusters.

Murali and Kasif (2003) define the"interesting" states of expression for each gene as the subintervals of its range of expression levels, within the experiment, that contain more observations than would be expected by chance (assuming a uniform distribution over the range). They then identify biclusters in which the genes are conserved in one of their interesting states over the samples. Thus, the biclustered genes are not necessarily expressed at a similar level, but jointly respond as illustrated in Figure 3.4(d). All possible subintervals are considered as potential states, so this method is not as restrictive as methods that discretise the data. However it is possible for a state to be statistically significant without being biologically significant, so prior gene selection may be required to remove uninformative genes. The biclusters are discovered using a simple algorithm, that looks for conserved genes over seed samples and randomly selected "discriminating subsets" of the samples, to form the basis of candidate biclusters. The algorithm returns the best bicluster out of these candidates, then the samples from this bicluster are removed from the data set and the process is repeated. Therefore genes may belong to more than one bicluster, but samples may not, as illustrated in Figure 3.7(d).

Plaid model clustering (Lazzeroni and Owen, 2002) models the usual expression level for each gene and sample, then models the additional effects of biclusters which show an unusual pattern of expression. For the expression level Y_{ij} of the ith gene

and the jth sample, the plaid model is given by

$$Y_{ij} = \mu_0 + \alpha_{i0} + \beta_{j0} + \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk}) \rho_{ik} \kappa_{jk} + \epsilon_{ij}$$

where μ_0, α_{i0} and β_{j0} are the "usual" or background mean, gene and sample effects; the μ_k, α_{ik} and β_{jk} are the additional mean, gene and sample effects for bicluster k, and the $\rho_{ik} \in \{0, 1\}$ and $\kappa_{jk} \in \{0, 1\}$ indicate respectively whether the ith gene or jth sample belongs to bicluster k, such $\sum_i \rho_{ik} \geqslant 0$ and $\sum_j \kappa_{jk} \geqslant 0$. Thus states of expression are identified as being unusual or interesting in the context of the full data set, rather than within each gene's expression profile, as in the method of Murali and Kasif (2003). The biclustered genes have similar patterns of expression, as illustrated in Figure 3.4(f). Lazzeroni and Owen (2002) use an alternating least squares algorithm to estimate one bicluster at a time. The actual expression levels are used, so information is not lost through pre-processing. The number of biclusters is determined automatically and biclusters may overlap, producing the type of structure illustrated in Figure 3.7(e).

3.4 Selection of the Plaid Model for Further Study

Despite the number of clustering methods developed for microarray analysis, few are able to address all of the issues discussed at the beginning of this chapter.

The one-way clustering methods address a range of issues collectively, but individually they tend to address single issues. Percolation clustering (Šášik et al., 2001) introduces flexible cluster membership, allowing membership of more than one cluster or none at all, but clustering is based on the full expression profile. COSA (Friedman and Meulman, 2004) and context-specific Bayesian clustering (Barash and Friedman, 2002) allow clusters to be based on subsets of the attributes, but the clustering is exhaustive. Methods developed for time series data are obviously focused on this particular issue, but the method of Luan and Li (2003) can also accommodate genes that don't belong to any cluster.

In general, the two-way clustering methods address the issue of a limited number of attributes being relevant for clustering. The methods of Tang et al. (2001) and McLachlan et al. (2002) use gene clusters as the attributes for clustering the samples,

but their methods assume that the selected genes are useful for clustering all of the samples. Local dependencies can be identified through iterative two-way clustering, as in the methods of Getz et al. (2000) and Pollard and van der Laan (2002), but there is a danger that such methods can be too aggressive. In co-clustering methods, there is an overall dependence between the clustering of one dimension and the other. A cluster of genes can then be interpreted as a group of genes with similar expression levels in each of the sample clusters. The Double Conjugated Clustering method of Busygin et al. (2002), identifies conjugate pairs of gene and sample clusters, in which the expression levels are most homogeneous and also allows for genes or samples that don't belong to any cluster. However neither of the co-clustering methods reviewed here allow clusters to overlap.

The concept of a bicluster offers great potential for accommodating the features of microarray data, since biclusters can represent local dependencies between genes and samples. However not all biclustering methods realise the full potential of this concept. Some methods place restrictions on the bicluster membership, for example, Segal et al. (2001) assume a latent clustering of the genes and samples on which the biclusters must be based and the method of Sheng et al. (2003) does not allow biclusters to overlap. The latent models of Segal et al. (2003) and MacKay and Miskin (2001) only include biclusters as a special case, usually producing gene clusters based on all the samples. Other methods have a tendency to include non-informative genes in the biclusters, such as the method of Cheng and Church (2000) and possibly the method of Murali and Kasif (2003). In some cases, there are drawbacks in the way a bicluster is defined, for example Ben-Dor et al. (2002) require biclusters of genes to have the same (or approximately the same) rank profile across the biclustered samples, which is not a necessary or sufficient basis for identifying homogeneous profiles, whereas the use of SAMBA with an unsigned graph (Tanay et al., 2002) may allow too much flexibility to identify meaningful groups of genes.

The methods of Yang et al. (2003), Ambler (2003) and Lazzeroni and Owen (2002) all allow biclusters to overlap. The practical utility of allowing biclusters to overlap was demonstrated by Tanay et al. (2002), who presented an example in which tissue samples in a certain class were uniquely characterised by the overlap between two biclusters. Therefore it is in an important feature to allow both genes and samples to belong to more than one bicluster. The methods of Yang et al. (2003), Ambler (2003) and Lazzeroni and Owen (2002) also have the advantage of

using the actual gene expression levels for analysis, rather than requiring the data to be discretised as in Tanay et al. (2002).

The FLOC algorithm (Yang et al., 2003) has the disadvantage that it requires the number of biclusters to be specified. In addition, the biclusters found using the FLOC algorithm include any samples in which the biclustered genes are co-regulated, even if the genes are not expressed at an unusual level. The plaid model, on the other hand, models the background expression level of genes and samples and seeks biclusters that represent a substantial departure from this model. The background expression levels are modelled as part of the clustering process, which is preferable to removing global effects prior to clustering, since this can add noise to the data.

The plaid model may also be preferred over the Bayesian clustering model proposed by Ambler (2003), as it seeks biclusters of genes with a similar expression pattern over the samples (Figure 3.7(c)), rather than a common expression level (Figure 3.7(a)). Lazzeroni and Owen (2002) illustrate the practical utility of allowing gene and sample effects in their plaid model analysis of yeast gene expression data. The first bicluster covers samples taken from two yeast strains during sporulation and includes many genes involved in the cell cycle. The sample effects range from 0.72 to 1.41 and the top 12 gene effects range from 1.51 to 3.34. This shows that genes with a common function may jointly respond under certain conditions, but the gene expression levels are not necessarily the same for all the genes across all the samples. Allowing gene and sample effects enables such biclusters to be discovered.

The features of the plaid model make it a particularly attractive method for clustering microarray data. This method addresses all of the issues discussed at the beginning of this chapter. As a biclustering method, each gene cluster is associated with a sample cluster over which the genes are co-regulated, allowing for limited co-regulation. The biclusters represent unusual patterns of expression, so that uninteresting expression profiles are not clustered. Genes involved in more than one active biological process can be accommodated through overlapping biclusters.

The plaid model also has the potential to be adapted to take into account further structural information, by extending or modifying the underlying model. Examples of such additional structure are given by the data sets introduced in Chapter 1: the grouping structure of the Infectious Disease data set and the three-way structure of

the TB Susceptibility data set.

The problem of estimating a set of biclusters is acknowledged to be NP-hard (Cheng and Church, 2000; Yang et al., 2003; Lazzeroni and Owen, 2002), so it is unlikely that an algorithm will find the globally optimal solution, but should return a "good" local optimum. The algorithm proposed by Lazzeroni and Owen (2002) uses alternating least squares which is a standard approach to model estimation with a clearly defined optimisation criterion. As such, their approach seems less ad hoc than some of the other algorithms discussed in this chapter, particularly with regard to two-way clustering and biclustering methods.

Since the plaid model addresses all of the issues discussed at the beginning of this chapter and compares favourably with other methods for clustering gene expression data, the remainder of this thesis will focus on this technique.

3.4.1 Summary

Out of the clustering methods reviewed in this chapter, the plaid model has been selected for further study. Plaid model clustering does not suffer from the drawbacks of conventional clustering techniques that were discussed at the beginning of this chapter. In addition the plaid model has potential for further extension, which may be particularly useful for three-way gene expression data sets such as the TB Susceptibility data set introduced in Chapter 1.

The alternating least squares algorithm proposed by Lazzeroni and Owen (2002) for fitting the plaid model seems a reasonable approach. However, this algorithm is considered in more detail in the next chapter and some drawbacks of the method are identified.

The remaining chapters of this thesis are organised as follows. The next chapter considers the problem of estimating the plaid model, first examining the original algorithm as discussed above, then reviewing algorithms of related methods, leading to the proposal of an alternative approach. Chapter 5 compares the performance of the alternative algorithm to the original algorithm and the proposed algorithm is adopted as a result. Some refinements of the algorithm are introduced in Chapter 6, then some extensions of plaid model clustering are proposed and investigated in Chapters 7 and 8.

Bibliography

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, Jr., J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., Staudt, L. M., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403 (6769), 503 – 511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J., June 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96 (12), 6745–6750.

Ambler, G., 2003. Bayesian two-way clustering for gene expression data, `http://www.bgx.org.uk/presentations.html`.

Bar-Joseph, Z., Gerber, G., Gifford, D. K., Jaakkola, T. S., Simon, I., 2002. A new approach to analyzing gene expression time series data. In: Myers, G., Hannenhalli, S., Sankoff, D., Istrail, S., Pevzner, P., Waterman, M. (Eds.), Proceedings of the Sixth Annual International Conference on Computational Biology (RECOMB-02). ACM Press, New York, NY, pp. 39 – 48.

Barash, Y., Friedman, N., 2002. Context-specific bayesian clustering for gene expression data. J. Comput. Biol. 9 (2), 169–191.

Ben-Dor, A., Chor, B., Karp, R., Yakhini, Z., 2002. Discovering local structure in gene expression data: the order-preserving submatrix problem. In: Myers, G., Hannenhalli, S., Sankoff, D., Istrail, S., Pevzner, P., Waterman, M. (Eds.), Pro-

ceedings of the Sixth Annual International Conference on Computational Biology (RECOMB-02). ACM Press, New York, NY, pp. 49–57.

Busygin, S., Jacobsen, G., Krämer, E., 2002. Double conjugated clustering applied to leukemia microarray data, 2nd SIAM ICDM, Workshop on clustering high dimensional data and its applications, Arlington, VA, `http://www.busygin.dp.ua`.

Cheng, Y., Church, G. M., 2000. Biclustering of expression data. In: Bourne, P., Gribskov, M., Altman, R., Jensen, N., Hope, D., Lengauer, T., Mitchell, J., Scheeff, E., Smith, C., Strande, S., Weissig, H. (Eds.), Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB-2000). Vol. 8. AAAI Press, pp. 93–103.

Cho, H., Dhillon, I. S., Guan, Y., Sra, S., 2004. Minimum sum-squared residue co-clustering of gene expression data. In: Proc. SIAM Int. Conf. Data Mining. `http://www.siam.org/meetings/sdm04/proceedings/`.

Dhillon, I. S., 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In: Provost, F., Srikant, R. (Eds.), Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining. ACM Press, New York, NY, pp. 269–274.

Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95 (25), 14863–14868.

Everitt, B. S., Landau, S., Leese, M., 2001. Cluster analysis, 4th Edition. Hodder Arnold, London, UK.

Friedman, J. H., Meulman, J. J., 2004. Clustering objects on subsets of attributes. J. R. Statist. Soc. B 66 (4), 815–849.

Getz, G., Gal, H., Kela, I., Notterman, D. A., Domany, E., 2003. Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. Bioinformatics 19 (9), 1079–1089.

Getz, G., Levine, E., Domany, E., 2000. Coupled two-way clustering analysis of gene microarray data. Proc. Natl. Acad. Sci. USA 97 (22), 12079–12084.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., Brown, P., 2000. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. 1(2), research0003.1–0003.21.

Heard, N. A., Holmes, C. C., Stephens, D. A., to appear. A quantitative study of gene regulation involved in the immune response of anopheline mosquitos: An application of bayesian hierarchical clustering of curves. J. Am. Statist. Assoc.

Kluger, Y., Basri, R., Chang, J. T., Gerstein, M., 2003. Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res. 13 (4), 703–716.

Lazzeroni, L., Owen, A., 2002. Plaid models for gene expression data. Statist. Sinica 12 (1), 61–86.

Luan, Y., Li, H., 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics 19 (4), 474–482.

MacKay, D. J. C., Miskin, J., 2001. Latent variable models for gene expression data. Tech. rep., Cavendish Lab., Cambridge Univ., UK, `http://www.inference.phy.cam.ac.uk/mackay/abstracts/icagenes.html`.

McLachlan, G. J., Bean, R. W., Peel, D., 2002. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18 (3), 413–422.

Murali, T. M., Kasif, S., 2003. Extracting conserved gene expression motifs from gene expression data. In: Pacific Symposium on Biocomputing. Vol. 8. `http://helix-web.stanford.edu/psb03/`, pp. 77–88.

Pollard, K. S., van der Laan, M. J., 2002. Statistical inference for simultaneous clustering of gene expression data. Math. Biosci. 176 (1), 99–121.

Ramoni, M. F., Sebastiani, P., Kohane, I. S., 2002. Cluster analysis of gene expression dynamics. Proc. Natl. Acad. Sci. USA 99 (14), 9121–9126.

Šášik, R., Hwa, T., Iranafar, N., Loomis, W., 2001. Percolation clustering: a novel approach to the clustering of gene expression patterns in *Dictyostelium* devel-

opment. In: Pacific Symposium on Biocomputing. Vol. 6. http://helix-web.stanford.edu/psb01/, pp. 335–347.

Segal, E., Battle, A., Koller, D., 2003. Decomposing gene expression into cellular processes. In: Pacific Symposium on Biocomputing. Vol. 8. http://helix-web.stanford.edu/psb03/, pp. 89–100.

Segal, E., Taskar, B., Gasch, A., Friedman, N., Koller, D., 2001. Rich probabilistic models for gene expression. Bioinformatics 17 (Suppl. 1), S243–S252.

Sheng, Q., Moreau, Y., De Moor, B., 2003. Biclustering microarray data by Gibbs sampling. Bioinformatics 19 (Suppl. 2), ii196 – ii205.

Tanay, A., Sharan, R., Shamir, R., 2002. Discovering statistically significant biclusters in gene expression data. Bioinformatics 18 (Suppl. 1), S136–S144.

Tang, C., Zhang, L., Zhang, A., Ramanathan, M., 2001. Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. In: 2nd IEEE International Symposium on Bioinformatics and BioEngineering (BIBE 2001). IEEE Computer Society, Los Alamitos, CA, pp. 41–48.

Wakefield, J. C., Zhou, C., Self, S. G., 2003. Modelling gene expression data over time: Curve clustering with informative prior distributions. In: Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., West, M. (Eds.), Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting. Oxford University Press, Oxford, UK, pp. 721–732.

Yang, J., Wang, H., Wang, W., Yu, P., 2003. Enhanced biclustering on expression data. In: 3rd IEEE International Symposium on BioInformatics and BioEngineering (BIBE 2003). IEEE Computer Society, Los Alamitos, CA, pp. 321–327.