Journal of Biomedicine and Biotechnology • 2005:2 (2005) 215–225 • DOI: 10.1155/JBB.2005.215

REVIEW ARTICLE

Finding Groups in Gene Expression Data

David J. Hand and Nicholas A. Heard

Department of Mathematics, Faculty of Physical Sciences, Imperial College, London SW7 2AZ, UK

Received 11 June 2004; revised 24 August 2004; accepted 24 August 2004

The vast potential of the genomic insight offered by microarray technologies has led to their widespread use since they were introduced a decade ago. Application areas include gene function discovery, disease diagnosis, and inferring regulatory networks. Microarray experiments enable large-scale, high-throughput investigations of gene activity and have thus provided the data analyst with a distinctive, high-dimensional field of study. Many questions in this field relate to finding subgroups of data profiles which are very similar. A popular type of exploratory tool for finding subgroups is cluster analysis, and many different flavors of algorithms have been used and indeed tailored for microarray data. Cluster analysis, however, implies a partitioning of the entire data set, and this does not always match the objective. Sometimes pattern discovery or bump hunting tools are more appropriate. This paper reviews these various tools for finding interesting subgroups.

INTRODUCTION

Microarray gene expression studies are now routinely used to measure the transcription levels of an organism's genes at a particular instant of time. These mRNA levels serve as a proxy for either the level of synthesis of proteins encoded by a gene or perhaps its involvement in a metabolic pathway. Differential expression between a control organism and an experimental or diseased organism can thus highlight genes whose function is related to the experimental challenge.

An often cited example is the classification of cancer types (Golub et al [1], Alizadeh et al [2], Bittner et al [3], Nielsen et al [4], Tibshirani et al [5], and Parmigiani et al [6]). Here, conventional diagnostic procedures involve morphological, clinical, and molecular studies of the tissue, which both are highly subjective in their analysis and cause inconvenience and discomfort to the patient. Microarray experiments offer an alternative (or additional), objective means of cell classification through some predetermined functionals of the gene expression levels for a new tissue sample of an unknown type. Whilst potentially very powerful, the statistical robustness of these methods is still hampered by the "large p, small n" problem; a microarray slide can typically hold tens of thousands of gene fragments whose responses here act as the predictor variables (p), whilst the number of patient tissue samples (n) available in such studies is much less (for the above examples, 38 in Golub et al, 96 in Alizadeh et al, 38 in Bittner et al, 41 in Nielsen et al, 63 in Tibshirani et al, and 80 in Parmigiani et al).

More generally, beyond such "supervised" classification problems, there is interest in identifying groups of genes with related expression level patterns over time or across repeated samples, say, even within the same classification label type. Typically one will be looking for coregulated genes showing similar expression levels across the samples, but equally we may be interested in anticorrelated genes showing diametric patterns of regulation (see, eg, Dhillon et al [7]) or even genes related through a path of genes with similar expression (Zhou et al [8]). In the case of classification of cancer, these "unsupervised" studies can give rise to the discovery of new classifications which may be morphologically indistinguishable but pathogenetically quite distinct. In general, they may shed light on unknown gene functions and metabolic pathways.

A common aim, then, is to use the gene expression profiles to identify groups of genes or samples in which the members behave in similar ways. In fact, that task description encompasses several distinct types of objectives.

Firstly, one might want to partition the data set to find naturally occurring groups of genes with similar expression patterns. Implicit in this is the assumption that there do exist groups such that members of a given group have similar patterns which are rather different from the patterns exhibited by members of the other groups. The aim,

Correspondence and reprint requests to David J. Hand, Department of Mathematics, Faculty of Physical Sciences, Imperial College, London SW7 2AZ, UK, E-mail: d.j.hand@imperial.ac.uk

then, is to "carve nature at the joints," to identify these groups. Statistical tools for locating such groups go under the generic name of cluster analysis, and there are many such tools.

Secondly, one might simply want to partition the data set to assign genes to groups such that each group contains genes with similar expression profiles, with no notion that the groups are "naturally occurring" or that there exist "joints" at which to carve the data set. This exercise is termed dissection analysis (Kendall [9]). The fact that the same tools are often used for cluster analysis and dissection analysis has sometimes led to confusion.

Thirdly, one might simply want to find local groups of genes which exhibit similar expression profiles, without any aim of partitioning the entire data set. Thus there will be some such local groupings, but many, perhaps most, of the genes will not lie in any of these groups. This sort of exercise has been termed *pattern discovery* (Hand et al [10]).

Fourthly, one might wish to identify groups of genes with high variations over the different samples or perhaps dominated by one label type in a supervised classification setting. Methods for identifying such groups which start with a set of genes and sequentially remove blocks of the genes until some criterion is optimised have been termed by Hastie et al [11] as "gene shaving."

Fifthly, in *pattern matching* one is given a gene a priori, with the aim being to find other genes which have similar expression profiles. Technically, solutions to such problems are similar to those arising in nucleotide sequencing, with more emphasis on imprecise matches.

Sixthly, in supervised classification, there is the case described above where samples of genes are provided which belong to each of several prespecified classes, and the aim is to construct a rule which will allow one to assign new genes to one of these classes purely on the basis of its expression profile (see Golub et al [1]).

Of these objectives, cluster analysis and pattern discovery both seek to say something about the intrinsic structure of the data (in contrast to, eg, dissection and pattern matching) and both are exploratory rather than necessarily being predictive (in contrast to, eg, supervised classification). This means that these problems are fundamentally open ended: it is difficult to say that a tool will never be useful under any circumstances. Perhaps partly because of this, a large number of methods have been developed. In the body of this paper we describe tools which have been developed for cluster analysis and pattern discovery, since these are intrinsically concerned with finding natural groups in the data, and we summarise their properties. We hope that this will be useful for researchers in this area, since two things are apparent: (i) that the use of such methods in this area is growing at a dramatic rate and (ii) that often little thought is given about the appropriateness of the choice of methods. An illustration of the last point is given by the fact that different cluster analysis algorithms are appropriate for detecting different kinds of cluster structure, and yet it is clear that often the choice of methods has been a haphazard one, perhaps based on software availability or programming ease, rather than an informed one.

In the "microarray experiments" section we give an introduction to microarray technology and discuss some of the issues that arise in its analysis. The "cluster analysis" and "pattern discovery" sections detail clustering and pattern discovery methods, respectively; in the case of clustering, examples are given of situations where these techniques have been applied to microarray data, and for pattern discovery we suggest how these methods could carry across to this area. Finally some conclusions are given.

MICROARRAY EXPERIMENTS

There are two main microarray technologies, complementary DNA (cDNA) and oligonucleotide, though both work on the same principle of attaching sequences of DNA to a glass or nylon slide and then hybridising these attached sequences with the corresponding DNA or (more commonly) RNA in a sample of tissue through "complementary binding." The two technologies differ according to the type of "probe" molecules used to represent genes on the array. With cDNA microarrays genes are represented by PCR-amplified (polymerase chain reaction) DNA sequences spotted onto a glass slide; with oligonucleotide arrays, between 16 and 20 complementary subsequences of 25 base pairs from each gene are attached to the chip by photolithography, together known as the perfect match (PM), along with the same sequences altered slightly at the middle bases, known as the mismatch (MM), for factoring out nonspecific binding. As it is difficult to measure the amount of PCR product in the former case, it follows that for cDNA microarrays we can only achieve relative expression levels of two or more samples against one another, whereas for oligonucleotide arrays absolute measurements are taken, such as mean or median of PM-MM or log(PM/MM).

After hybridisation a fluorescence image of the microarray is produced using a scanner, which is usually a laser confocal microscope using excitation light wavelengths to generate emission light from appropriate fluors. This image is pixelated and image analysis techniques are used to measure transcript abundance for each probe and hence give an overall expression score. Finally a normalisation procedure (Dudoit et al [12], Yang et al [13], Irizarry et al [14], and Bolstad et al [15]) is used to remove any systematic variations in the expression scores such as background correction and allow for the effect of location of the probe on the slide (smudges). Besides the difficulties of these procedures, there are many other sources of error such as noncomplementary binding, instability from small expression levels of a gene in both samples, and missing values (Troyanskaya et al [16]).

The resulting low signal-to-noise ratio of microarray experiments means most interest is focused on multiple slide experiments, where each hybridisation process

is performed with tissue samples possibly from the same (replicate data) or different experimental conditions, allowing us to "borrow strength." For life-cycle processes time-course experiments are also popular, where expression levels of an experimental subject are measured at a sequence of time points to build up a temporal profile of gene regulation.

A microarray experiment can measure the expression levels of tens of thousands of genes simultaneously. However, they can be very expensive. Therefore when it comes to data analysis, there is a recurring problem of high dimension in the number of genes and only a small number of cases. This is a characteristic shared by spectroscopic data, which additionally have high correlations between neighbouring frequencies; analogously for microarray data, there is evidence of correlation of expression of genes residing closely to one another on the chromosome (Turkheimer et al [17]). Thus when we come to look at cluster analysis for microarray data, we will see a large emphasis on methods which are computationally suited to cope with the high-dimensional data.

CLUSTER ANALYSIS

The need to group or partition objects seems fundamental to human understanding: once one can identify a class of objects, one can discuss the properties of the class members as a whole, without having to worry about individual differences. As a consequence, there is a vast literature on cluster analysis methods, going back at least as far as the earliest computers. In fact, at one point in the early 1980s new ad hoc clustering algorithms were being developed so rapidly that it was suggested there should be a moratorium on the development of new algorithms while some understanding of the properties of the existing ones was sought. In fact, without this hiatus in development occurring, a general framework for such algorithms was gradually developed.

Another characteristic of the cluster analysis literature, apart from its size, is its diversity. Early work appeared in the statistics literature (where it caused something of a controversy because of its purely descriptive and noninferential nature—was it really statistics?), the early computational literature, and, of course, the biological and medical literature. Biology and medicine are fundamentally concerned with taxonomies and diagnostic and prognostic groupings.

Later, nonheuristic, inferential approaches to clustering founded on probability models would appear. These models fit a mixture probability model to the data to obtain a "classification likelihood," with the similarity of two clusters determined by the change in this likelihood that would be caused by their merger. In fact, most of the heuristic methods can be shown to be equivalent to special cases of these "model-based" approaches. Reviews of model-based clustering procedures can be found in Bock [18], Bensmail et al [19], and Fraley and Raftery [20]. Model-based clustering approaches allow the choice of clustering method and number of clusters to be recast as a statistical model choice problem, and, for example, significance tests can be carried out.

More recently, research in machine learning and data mining has produced new classes of clustering algorithms. These various areas are characterised by their own emphases—they are not merely reinventing the clustering wheel (although, it has to be said, considerable intellectual effort would be saved by more cross-disciplinary reading of the literature). For example, data mining is especially concerned with very large data sets, so that the earlier algorithms could often not be applied, despite advances in computer storage capabilities and processing speed. A fundamental problem is that cluster analysis is based on pairwise similarities between objects and the number of such distances increases as the square of the number of objects in the data set does.

Cluster analysis is basically a data exploration tool, based solely on the underlying notion that the data consist of relatively homogeneous but quite distinct classes. Since cluster analysis is concerned with partitioning the data set, usually each object is assigned to just one cluster. Extensions of the ideas have been made in "soft" clustering, whereby each object may partly belong to more than one cluster. Mixture decomposition, of course, leads naturally to such a situation, and an early description of these ideas (in fact, one of the earliest developments of a special case of the expectation-maximisation (EM) algorithm) was given by Wolfe [21].

Since the aim of cluster analysis is to identify objects which are similar, all such methods depend critically on how "similarity" is defined (note that for model-based clustering this follows automatically from the probability model). In some applications the raw data directly comprise a dissimilarity matrix (eg, in direct subjective preference ratings), but gene expression data come in the form of a gene $\times$ variable data matrix, from which the dissimilarities can be computed. In many applications of cluster analysis, the different variables are not commensurate (eg, income, age, and height when trying to cluster people) so that decisions have to be made about the relative weight to give to the different components. In gene expression data, however, each variable is measured on the same scale. One may, nonetheless, scale the variables (eg, by the standard deviation or some robust alternative) to avoid variables with a greater dispersion playing a dominant role in the distance measure. Note that in the case of model-based clustering methods, however, the reverse is true; different levels of variability for each variable are easy to incorporate into the models whereas the likelihood will be much harder to write down for data rescaled in this way. Likewise, it is worthwhile considering transforming the variables to remove skewness, though, in the case of gene expression data based on the log of a ratio, this may not be necessary or appropriate. Reviews of distance measures are given in Gower [22] and Gordon [23].

Briefly, distance metrics used to define cluster dissimilarity are usually either geometric or correlation based. Variations of the former theme include Euclidean, Manhattan, and Chebychev distances, and of the latter Pearson and Spearman correlations, for example. In the field of gene expression clustering, Eisen et al [24] used an uncentred correlation-based distance measure which takes into account both the "shape" of the gene expression profile and each gene's overall level of expression, and this measure has now been widely adopted.

As noted above, certain types of gene expression clustering problems, such as clustering tissue samples on the basis of gene expression, involve relatively few data points and very large numbers of variables. In such problems especially, though also more generally, one needs to ask whether all the variables contribute to the group structure—and, if not, whether those that do not contribute serve to introduce random variation such that the group structure is concealed (see, eg, Milligan [25], DeSarbo and Mahajan [26], and Fowlkes et al [27]). One can view the problem as that of choosing the distance measure so that the irrelevant variables contribute nothing to the distance, that is, such variables are given a weight of zero if the distance consists of a weighted combination of contributions from the variables. There is a large and growing literature devoted to this problem of selecting the variables. See, for example, De Soete et al [28], De Soete [29, 30], Van Buuren and Heiser [31], and Brusco and Cradit [32]. More generally, of course, one might suspect that different cluster structures occurred in different subsets of variables. This would be the case, for example, if people could suffer from sets of nonmutually exclusive diseases (eg, different pulmonary diseases on the one hand, and psychiatric syndromes on the other). In this case one would ideally like to search over cluster structures and over subsets of variables. In a recent paper, Friedman and Meulman [33] describe a related problem in which, although a single partitioning is sought, different subsets of variables may be dominant in each of the different clusters.

Alternatively, even when clustering genes on a relatively small number of samples, we may wish to cluster on only a subset of the samples if those samples correspond, say, to a particular group of experimental conditions. Thus we would want many "layers" of clustering based on different (and possibly overlapping) subsets of the tissue samples, with genes which are clustered together in one layer not necessarily together in another. Additive two-way analysis of variance (ANOVA) models for this purpose, termed plaid models for the rectangular blocking they suggest on the gene expression data matrix, were introduced by Lazzeroni and Owen [34].

Broadly speaking, there are two classes of clustering methods: hierarchical methods and optimisation methods. The former sequentially aggregates objects into clusters or sequentially split a large cluster into smaller ones, while the latter seeks those clusters which optimise some overall measure of clustering quality. We briefly summarise the methods below. Different algorithms are based on different measures of dissimilarity, and on different criteria determining how good a proposed cluster structure is. These differences naturally lead to different cluster structures. Put another way, such differences lead to different definitions of *what* a cluster is. A consequence of this is that one should decide what one means by a cluster before one chooses a method. The k-means algorithm described below will be good at finding compact spherical clusters of similar sizes, while the single-link algorithm is able to identify elongated sausage-shaped clusters. Which is appropriate depends on what sort of structure one is seeking. Merely because cluster analysis is an exploratory tool does not mean that one can apply it without thinking.

Hierarchical methods

Hierarchical clustering methods give rise to a sequence of nested partitions, meaning the intersection of a set in the partition at one level of the hierarchy with a set of the partition at a higher level of the hierarchy will always be equal to either the set from the lower level or the empty set. The hierarchy can thus be graphically represented by a tree. Typically this sequence will be as long as the number of observations (genes), so that level k of the hierarchy has exactly k clusters and the partition at level $k-1$ can be recovered by merging two of the sets in level k. In this case, the hierarchy can be represented by a binary tree, known as a "dendrogram." Usually the vertical scale of a dendrogram represents the distance between the two merged clusters at each level.

Methods to obtain cluster hierarchies are either top-down approaches, known as divisive algorithms, where one begins with a large cluster containing all the observations and successively divides it into finer partitions, or more commonly bottom-up, agglomerative algorithms, where one begins with each observation in its own cluster and successively merge the closest clusters until one large cluster remains. Agglomerative algorithms dominate the clustering literature because of the greatly reduced search space compared to divisive algorithms, the former usually requiring only $O(n^2)$ or at worst $O(n^3)$ calculations, whilst without reformulation performing the first stage of the latter alone requires $2^{n-1}-1$ calculations. This is reflected by the appearance of early versions of agglomerative hierarchical algorithms in the ecological and taxonomic literature as much as 50 years ago. To make divisive schemes feasible, monothetic approaches can be adopted, in which the possible splits are restricted to thresholds on single variables—in the same manner as the standard CART tree algorithm (Breiman et al [35]). Alternatively, at each stage the cluster with largest diameter can be "splintered" through allocating its largest outlier to a new cluster and relocating the remaining cluster members to whichever of the old and new clusters is closest, as in the *Diana* algorithm of Kaufman and Rousseeuw [36] which has been

implemented in the statistical programming language R. It has been suggested that an advantage of divisive methods is that they begin with the large structure in the data, again as in CART with its root split, but we have seen no examples to convince us that agglomerative methods are not equally enlightening.

Having selected an appropriate distance metric between observations, this needs to be translated into a "linkage metric" between clusters. In model-based clustering this again follows immediately, but otherwise natural choices are single link, complete link, or average link.

Single-link (or nearest neighbour) clustering defines the distance between two clusters as the distance between the two closest objects, one from each cluster (Sokal and Sneath [37] Jardine and Sibson [38]). A unique merit of the single-link method is that when one makes a choice between two equal intercluster distances for a merger, it will be followed by a merger corresponding to the other distance, which gives the method a certain type of robustness to small perturbations of the distances. Single-link clustering is susceptible to chaining: the tendency for a few points lying between two clusters to cause them to be joined. Whether this really is a weakness depends on what the aim is—on what one means by a "cluster." In general, if different objects are thought to be examples of the same kind of thing, but drawn at different stages of some developmental process, then perhaps one would want them to be assigned to the same cluster.

Complete-link (or furthest neighbour) clustering defines the distance between two clusters as the distance between the two furthest objects, one from each cluster. It is obvious that this will tend to lead to groups which have similar diameters, so that the method is especially valuable for dissection applications. Of course, if there are natural groups with very different diameters in the data, the smallest of these may well be merged before the large ones have been put together. We repeat, it all depends on one's aims and on what one means by a cluster.

Average-link (or centroid) clustering defines the distance between two clusters as the distance between the centroids of the two clusters. If the two clusters are of very different sizes, then the cluster that would result from their merger would maintain much of the characteristics of the larger cluster; if this is deemed undesirable, median cluster analysis which gives equal weighting to each cluster can be used.

Lance and Williams [39] present a simple linear system as a unifying framework for these different linkage measures.

After performing hierarchical clustering there remains the issue of choosing the number of clusters. In model-based clustering, this selection can be made using a model choice criterion such as Bayesian information criterion (Schwarz [40]) or in a Bayesian setting with prior distributions on model parameters, choosing the clustering which maximises marginal posterior probability. Otherwise, less formal procedures such as examining the dendrogram for a natural cut off or satisfying a predetermined upper bound on all within-group sums of squares are adopted.

Optimal partitioning methods

Perhaps more in tune with statistical ideas are direct partitioning techniques. These produce just a single "optimum" clustering of the observations rather than a hierarchy, meaning one must first state how many clusters there should be. In dissection analysis the number of groups is chosen by the investigator, but in cluster analysis the aim is to discover the naturally occurring groups, so some method is needed to compare solutions with different numbers of groups as discussed above at the end of hierarchical clustering.

For a fixed number of clusters k, a partitioning method seeks to optimise a clustering criterion; note, however, that the fact that no hierarchy is involved means that one may not be able to split a cluster in the solution with k clusters to produce the $k + 1$ cluster solution, and thus care must be taken in choosing a good starting point. Although, in principle, all one has to do is search over all possible allocations of objects to classes, seeking that particular allocation which optimises the clustering criterion, in practice there are normally far too many such possible allocations, so some heuristic search strategy must be adopted. Often, having selected an initial clustering, a search algorithm is used to iteratively relocate observations to different clusters until no gain can be made in the clustering criterion value.

The most commonly used partitioning method is "k-means" clustering (Lloyd [41] and MacQueen [42]). k-means clustering seeks to minimise the average squared distance between observations and their cluster centroid. This strategy can be initiated by specifying k centroids perhaps independently from the data, assigning each datum to the closest centroid, then recomputing the cluster centroids, reassigning each datum, and so on. Closely related to k-means clustering is the method of self-organising maps (SOM) (Kohonen [43]); these differ in also having prespecified geometric locations on which the clusters lie, such as points on a grid, and the clusters are iteratively updated in such a way that clusters close to each other in location tend to be relatively similar to one another. More generally, these optimisation methods usually involve minimising or maximising a criterion based on functions of the within-group ($\mathbf{W}$) and between-group ($\mathbf{B}$) (or, equivalently, the total $\mathbf{T}$) sum of squares and cross products matrix familiar from multivariate ANOVA. In fact, k-means clustering minimises trace ($\mathbf{W}$). Other common alternatives are minimising det ($\mathbf{W}$) and maximising trace ($\mathbf{BW}^{-1}$). For more details see Everitt [44].

Model-based partitioning methods are essentially mixture decomposition methods. Most commonly, mixtures of normal distributions are assumed, so that each cluster is characterised by an unknown mean and covariance matrix pair. Notable works in this area include Wolfe [21], Richardson and Green [45], and Fraley and Raftery

[20, 46, 47]. The authors of the latter provide the accompanying free software MCLUST, which uses the EM algorithm for parameter estimation to avoid the Markov chain Monte Carlo (MCMC) techniques required in Bayesian method of Richardson and Green [48] and Bensmail et al [19]. It should be remarked that when it comes to parameter estimation in mixture modelling, one has to be careful of the nonidentifiability of the mixture component labels; to get around this problem order constraints are placed on the parameters, often artificial, so that only a unique permutation of the component labels is supported.

Gene expression clustering

There are many instances of reportedly successful applications of both hierarchical clustering and partitioning techniques in gene expression analyses. This section illustrates the diversity of techniques which have been used.

Eisen et al [24] used agglomerative hierarchical clustering with their uncentred correlation-based dissimilarity metric as described above for growth time-course microarray data from budding yeast. This approach has since been followed in similar studies by Chu et al [49], Spellman et al [50], Iyer et al [51], Perou et al [52] and Nielsen et al [4]. Alternatively, Wen et al [53] used Euclidean hierarchical clustering on vectors with the time series of expression levels for each concatenated with the slopes between them to take into account offset but parallel patterns.

Turning to nonmodel-based partitioning methods, SOMs have been favoured; Tamayo et al [54] used SOMs for clustering of different time series of gene expression data. Similar approaches have also been used by Golub et al [1] for cancer tissue class discovery and prediction and Kasturi et al [55] for gene expression time series, where the latter first normalises the data to allow the use of Kullback-Leibler divergence as the distance metric. Tavazoie et al [56] represented expression time series in T-dimensional space and used the k-means clustering algorithm.

To find more subtle cluster structures, many model-based variations have been developed beyond these generic methods. This has been especially beneficial in the context of time series of gene expression samples. Ramoni et al [57] modelled gene expression time series with autoregressive processes, providing the accompanying free software CAGED. Luan and Li [58] clustered gene expression time series with mixed effects with B-splines; Bar-Joseph et al [59] used cubic splines for each gene with spline coefficients constrained to be similar for genes in the same cluster. They also used a time warping algorithm to align time series with similar expression profiles in different phases. Wakefield et al [60] performed clustering using a full MCMC Bayesian approach, with a basis function representation for the expression time series incorporating random effects. Yeung et al [61] used the mixture of normal distributions software MCLUST of Fraley and Raftery [20, 46, 47] for a variety of real and synthetic gene expression data sets, some time indexed. Pan et al [62] used the same model as MCLUST but on a two-sample t-statistic of differential expression for each gene rather than the full gene expression data matrix. Medvedovic and Sivaganesan [63] used the Gibbs sampling methods of Neal [64] for Dirichlet process mixture models to give a Bayesian version. Alon et al [65] used a divisive algorithm iteratively fitting two Gaussians at each stage with self-consistent equations. Heard et al [66] used a mixture of Gaussian processes with basis function representations for clustering of gene expression time series, with a conjugate model removing the need for MCMC.

Graphical models have also been attempted. Ben-Dor et al [67] gave two alternative graphical model-based clustering algorithms, clustering genes on a similarity matrix, PCC and CAST. Zhou et al [8] connected genes with highly correlated gene expression in a graphical model and clustered genes through a shortest-path analysis identifying "transitive genes." Dobra et al [68] attempted to actually model the whole covariance structure of the genes using Gaussian graphical models.

Instead of working on the raw gene expression matrix (genes $\times$ arrays), Alter et al [69] used singular value decomposition (cf principal component analysis) to analyse microarray data in the reduced diagonalised "eigengenes"$\times$"eigenarrays" space and filter out the eigengenes or eigenarrays inferred to represent experimental noise. Clustering in this new space was performed by Holter et al [70, 71] using standard hierarchical techniques; by Hastie et al [11] using "gene shaving," which identifies subsets of the genes with coherent expression patterns and large variations across samples; and by Wall et al [72] using thresholding on the magnitude of the elements of the left singular vectors (gene coefficient vectors), this thresholding enabling genes inhibited or promoted by the same transcription regulator to be clustered together. Clustering on principal components using k-means, Euclidean hierarchical average link and CAST was tested by Yeung and Ruzzo [73] and showed no benefits over clustering on the raw data.

Heyer et al [74] devised QT-clustering. There, for robustness to outliers, the *jackknife correlation* between two gene expression vectors is taken as the minimum of the correlation between the whole of both vectors or the correlation of the two vectors with any single component deleted. Clusters are then iteratively generated, with each made to be as large as possible subject to a threshold on the diameter of the cluster.

It will be apparent that much of the above hinges on how the distance between profiles is measured. Indeed, in general, different ways of measuring distance will lead to different solutions. This leads on to the question of how to assess the performance of different methods. In general, since most of these problems are fundamentally exploratory, there is no ideal answer to this. Datta and Datta [75] compared k-means, hierarchical (raw and partial least squares regression), MCLUST,

Diana and Fanny (a fuzzy k-means algorithm, see Kaufman and Rousseeuw [36]) for real temporal and replicate microarray gene expression data, scoring each method using measures of cluster overlap and distance, and overall favoured Diana. Yeung et al [76] compared k-means clustering, CAST, single-, average- and complete-link hierarchical clustering, and totally random clustering for both simulated and real gene expression data, scoring each method predictively using a jackknife procedure to obtain an adjusted "figure of merit," and favoured k-means and CAST. Gibbons and Roth [77] compared k-means, SOMs, and hierarchical clustering of real temporal and replicate microarray gene expression data, using a figure of merit which scores against random assignment, and favoured k-means and SOMs. For single method cluster validation, Li and Wong [78] used bootstrap sampling to check cluster membership robustness after hierarchical clustering.

In general, different clustering methods may yield different clusters. This is hardly surprising, given that they define what is meant by a cluster in different ways. It is true that if there is a very strong clustering in the data, one might expect consistency among the results, but it is less true that differences in the discovered cluster structure means that there is no cluster structure.

PATTERN DISCOVERY

Cluster analysis partitions a data set and, by implication, the space in which the data are embedded. All data points, and all possible data points, are assigned to an element of the partition. Often, however, one does not wish to make such grand sweeping statements. Often one merely seeks to find *localised* subsets of objects, in the sense that a set of objects are behaving in an unexpectedly similar way, regardless of the remainder of the objects. In the context of gene expression data, this would mean that amongst the mass of genes, each with their own expression profile, a (possibly) small number had unusually similar profiles. (As mentioned earlier, this idea can be generalised—one might be interested in detecting negatively correlated expression profiles—but we will not discuss such generalisations here.) In the context of nucleotide sequencing, it would mean that interest lay in identifying sequences which were very similar, without any preconceptions about what sort of sequence one was searching for. In both of these examples, one begins, as one does in cluster analysis, with the concept of a distance between elements (expression profiles or nucleotide sequences), but here, instead of using this distance to partition the data space, one merely uses it to find locally dense regions of the data space. Note that, in these two examples, the distance measures used are very different: classic multivariate distance measures (Euclidean distance being the most familiar) can be used in the first case, but the second case requires measures of distances between sequences or strings of symbols, such as the Levenshtein distance (Levenshtein [79]). In such situations, a natural way to define distance is in terms of the number of edit operations needed to make one string identical to the other. If edit operations are defined as being one of insertion, deletion, or substitution, then the Levenshtein (or "edit" or "sequence") distance between two strings is the minimum number of such operations needed to convert from one string to the other. A distance of 0 corresponds to an exact match. Since an optimisation is involved here, to find the minimum, many distance measures for strings use ideas of dynamic programming.

One stream of work aiming at detecting locally dense accumulations of data points goes under the name of "bump hunting" (eg, Silverman [80] and Harezlak [81]). Early work concentrated on unidimensional problems, but this has now been extended to multiple dimensions. An example is the PRIM algorithm of Friedman and Fisher [82], which embeds the ideas in a more general framework. Work which is intrinsically multivariate includes the PEAKER algorithm described in [83, 84], which identifies those data point locations which have a higher estimated data probability density than all local neighbours.

Although we have described the exercise as being one of finding localised groups of objects in the data set, in fact the aim is really typically one of inference. For example, the question is not really whether some particular expression profiles in the database are surprisingly similar, but whether these represent real underlying similarities between genes. There is thus an inferential aspect involved, which allows for measurement error and other random aspects of the process producing the data to make statements about the underlying structure. The key question implicit in this inferential aspect is whether the configuration could have arisen by chance, or whether it is real in the sense that it reflects an unusually high local density in the distribution of possible profiles. Sometimes the unusually high local probability densities are called "patterns" (eg, Hand et al [10]).

In order to make a statement about whether a configuration of a few data points is *unexpectedly* dense, one needs to have some probability model with which to compare it. In spatial epidemiology this model is based on the overall population distribution, so that, for example, one can test whether the proportion of cases of illness is unexpectedly high in a particular local region. In general, however, in bioinformatics applications such background information may not be available. DuMouchel [85] gives a particularly telling example of this in the context of adverse drug reactions, where there is no information about the overall number of times each drug has been prescribed. In such cases, one has to make reasonable assumptions about the background model. This is not necessarily problematic: the aim is, after all, an exploratory one. A basic form of background model is an inhomogeneous Poisson process, with the local intensity being based on any information one does have. Since, typically, more than one variable is involved, background models based on independence or information about known

relationships (such as time order) between the variables are often used.

From an inferential perspective, the key issue is one of multiplicity. With a large data space, perhaps with many observations, there is considerable opportunity for a large number of local maxima of an estimated density function. Deciding which of these maxima are genuine and which are attributable to chance is a nontrivial problem, and one which has been of concern in more general data mining contexts. Traditional statistical approaches to the multiplicity problem focus on controlling familywise error rate, setting a limit on the proportion of cases where there is no underlying distributional structure which are detected as significant. The consequence is that only the largest probability peaks exceed the chosen threshold. Benjamini and Hochberg [86], however (although in a rather different context), suggested controlling the (expected value of the) proportion of structures detected as significant where no real structure existed. This does not require so great a sacrifice of power for the individual tests. More generally, there is an accumulating body of statistical work in the area of *scan statistics* (eg, Glaz et al [87]). The intuitive idea here is that one scans a window over the data space, seeking positions where some function of the data within the window (eg, in our case, a count of the data points) exceeds some critical value. To date most of this work has focused on low-dimensional cases.

Compared to cluster analysis, pattern discovery is a relatively new area of investigation, but, like cluster analysis, the ideas have been developed by several different intellectual communities for different problem domains contemporaneously. These include speech recognition, text processing (which, of course, has received a dramatic boost with the web), real-time correction of keyboard entry errors, technical analysis ("chartism") in tracking stock prices, association analysis (in data mining, including its subdiscipline of market basket analysis), configural frequency analysis, and other areas. Ideas developed in one area can often be ported across to others, and, in particular, to bioinformatics—just as cluster analysis has been. There are also areas of statistics, which, although they have their own special emphases, are closely related, such as outlier detection.

Bump hunting, pattern discovery, or peak detection methods seem to have been applied relatively rarely in the analysis of microarray data to date, and yet in many problems such tools are arguably more appropriate than cluster analysis. In particular, in cases where the aim is to group the genes on the basis of their time or experimental condition expression profiles, many, perhaps most, of the genes will be doing nothing of interest. Including those in the partitioning is at best pointless and at worst may be misleading.

CONCLUSION

Cluster analysis with gene expression data has its own aspects, perhaps notably that of high dimensionality and low number of cases for some problems. However, in other ways this domain avoids issues which are important for other applications. The question of scaling the variables has been mentioned above: typically gene expression variables are commensurate. Choice of variables is a critical problem when one is trying to classify cells or samples, on the basis of the genes (and hence with a very large number of variables) but unimportant when one is trying to classify genes themselves, perhaps on the basis of very few expression conditions. This is a crucial issue, since cluster structure is always in the context of the variables chosen to describe the objects.

The most important point to bear in mind when considering using cluster analysis is that different methods have different (often implicit) definitions of what is meant by a cluster. If one is searching for compact spherical structures in the database, for example, one should not use a method which is likely to throw up long attenuated structures—and vice versa. Of course, in a completely exploratory situation, one can argue that any kind of structure could be of interest. This is true and provides a case for using multiple different methods (with multiple different distance measures) in the hope that some interesting structure may be found. In general, however, one does better by constraining the exploration in the light of what one already knows or believes likely to be the case: as Louis Pasteur said, "chance favours only the prepared mind."

Even more generally than the question of whether researchers have always used the appropriate method of cluster analysis when analysing their microarray data is the question of whether *any* form of cluster analysis is appropriate. Cluster analysis is a partitioning tool, assigning each of the data points to a unique (in general) cluster. Often, however, much, perhaps most, of the data points are uninteresting, with concern only being with particular local regions of the data space. In this context, pattern discovery methods in particular seem relevant to the analysis of microarray data. These tools identify subgroups of objects (eg, genes) which have similar profiles, regardless of the profile shapes of the other objects.

ACKNOWLEDGMENT

The work of Nicholas Heard described in this paper was funded by the Wellcome Trust grant number 065822.

REFERENCES

[1] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 1999;286(5439):531–537.

[2] Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* 2000;403(6769):503–511.

[3] Bittner M, Meltzer P, Chen Y, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature.* 2000;406(6795):536–540.
[4] Nielsen TO, West RB, Linn SC, et al. Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet.* 2002;359(9314):1301–1307.
[5] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA.* 2002;99(10):6567–6572.
[6] Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E. A statistical framework for expression-based molecular classification in cancer. *J Roy Statist Soc Ser B.* 2002;64(4):717–736.
[7] Dhillon IS, Marcotte EM, Roshan U. Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*. 2003;19:1612–1619.
[8] Zhou X, Kao MC, Wong WH. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci USA.* 2002;99(20):12783–12788.
[9] Kendall MG. *Multivariate Analysis.* 2nd ed. New York, NY: Macmillan; 1980.
[10] Hand DJ, Adams NM, Bolton RJ, eds. *Pattern Detection and Discovery.* New York, NY: Springer; 2002.
[11] Hastie T, Tibshirani R, Eisen MB, et al. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 2000;1(2):Research0003.
[12] Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist Sinica.* 2002;12(1):111–139.
[13] Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 2002;30(4):e15.
[14] Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003;4(2):249–264.
[15] Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 2003;19(2):185–193.
[16] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics.* 2001;17(6):520–525.
[17] Turkheimer FE, Duke DC, Moran LB, Graeber MB. Wavelet analysis of gene expression (WAGE). In: *Proceedings of the IEEE International Symposium on Biomedical Imaging: Macro to Nano. Vol* 2. Arlington, Va; 2004:1183–1186.
[18] Bock H. Probabilistic models in cluster analysis. *Comput Stat Data An.* 1996;23:5–28.
[19] Bensmail H, Celeux G, Raftery AE, Robert CP. Inference in model-based cluster analysis. *Stat Comput.* 1997;7:1–10.
[20] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc.* 2002;97(458):611–631.
[21] Wolfe JH. *A Computer Program for Maximum-Likelihood Analysis of Types.* San Diego, Calif: US Naval Personnel Research Activity; 1965. USNPRA Technical Bulletin 65-15.
[22] Gower JC. Measures of similarity, dissimilarity, and distance. In: Kotz S, Johnson NL, Read CB, eds. *Encyclopedia of Statistical Sciences. Vol* 5. New York, NY: John Wiley & Sons; 1985:397–405.
[23] Gordon AD. *Classification*. 2nd ed. Boca Raton, Fla: Chapman and Hall/CRC; 1999.
[24] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression *Proc Natl Acad Sci USA.* 1998;95(25):14863–14868.
[25] Milligan GW. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika.* 1980;45:325–342.
[26] DeSarbo WS, Mahajan V. Constrained classification: the use of a priori information in cluster analysis. *Psychometrika*. 1984;49:187–215.
[27] Fowlkes EB, Gnanadesikan R, Kettenring JR. Variable selection in clustering. *J Classification.* 1988;5(2):205–228.
[28] De Soete G, DeSarbo WS, Carroll JD. Optimal variable weighting for hierarchical clustering: an alternating least squares algorithm. *J Classification.* 1985;2:173–192.
[29] De Soete G. Optimal variable weighting for ultrametric and additive tree fitting. *Qual Quant.* 1986;20:169–180.
[30] De Soete G. OVWTRE: a program for optimal variable weighting for ultrametric and additive tree fitting. *J Classification.* 1988;5:101–104.
[31] Van Buuren S, Heiser W. Clustering n objects into k groups under optimal scaling of variables. *Psychometrika.* 1989;54:699–706.
[32] Brusco MJ, Cradit JD. A variable selection heuristic for k-means clustering. *Psychometrika*. 2001;66:249–270.
[33] Friedman JH, Meulman JJ. Clustering objects on subsets of attributes (with discussion). *J Roy Statist Soc Ser B.* 2004;66(4):815–849.
[34] Lazzeroni LC, Owen A. Plaid models for gene expression data. *Statist Sinica.* 2002;12(1):61–86.
[35] Breiman L, Friedman JH, Olshen R, Stone CJ. *Classification and Regression Trees.* Belmont, Calif: Wadsworth Advanced Books and Software; 1984.
[36] Kaufman L, Rousseeuw PJ. *Finding Groups in Data. An Introduction to Cluster Analysis.* New York, NY: John Wiley & Sons; 1990.
[37] Sokal RR, Sneath PHA. *Principles of Numerical Taxonomy*. San Francisco, Calif: WH Freeman; 1963.
[38] Jardine N, Sibson R. *Mathematical Taxonomy*. London, UK: Wiley; 1971.

[39] Lance GN, Williams WT. A general theory of classification sorting strategies. I. Hierarchical systems. *Comput J*. 1967;9:373–380.
[40] Schwarz G. Estimating the dimension of a model. *Ann Statist.* 1978;6(2):461–464.
[41] Lloyd SP. *Least Squares Quantization in PCM.* Murray Hill, NJ: Bell Laboratories; 1957. Internal Technical Report. Published in IEEE Transactions on Information Theory.
[42] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, eds. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol 1 of Statistics*. Berkeley, Calif: University of California Press; 1967:281–297.
[43] Kohonen T. *Self-organizing Maps.* Berlin, Germany: Springer; 1995.
[44] Everitt BS. *Cluster Analysis.* 3rd ed. London, UK: Edward Arnold; 1993.
[45] Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components. *J Roy Statist Soc Ser B.* 1997;59:731–792.
[46] Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput J.* 1998;41(8):578–588.
[47] Fraley C, Raftery AE. MCLUST: software for model-based cluster analysis. *J Classification.* 1999;16(2):297–306.
[48] Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J Roy Statist Soc Ser B.* 1997;59(4):731–792.
[49] Chu S, DeRisi J, Eisen MB, et al. The transcriptional program of sporulation in budding yeast. *Science.* 1998;282(5389):699–705.
[50] Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell.* 1998;9(12):3273–3297.
[51] Iyer VR, Eisen MB, Ross DT, et al. The transcriptional program in the response of human fibroblasts to serum. *Science.* 1999;283(5398):83–87.
[52] Perou CM, Jeffrey SS, van de Rijn M, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci USA.* 1999;96(16):9212–9217.
[53] Wen X, Fuhrman S, Michaels GS, et al. Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci USA.* 1998;95(1):334–339.
[54] Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA.* 1999;96(6):2907–2912.
[55] Kasturi J, Acharya R, Ramanathan M. An information theoretic approach for analyzing temporal patterns of gene expression. *Bioinformatics.* 2003;19(4):449–458.
[56] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet.* 1999;22(3):281–285.
[57] Ramoni MF, Sebastiani P, Kohane IS. Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci USA.* 2002;99(14):9121–9126.
[58] Luan Y, Li H. Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics.* 2003;19(4):474–482.
[59] Bar-Joseph Z, Gerber G, Gifford D, Jaakkola T, Simon I. A new approach to analyzing gene expression time series data. In: *Proceedings of the 6th Annual International Conference on Research in Computational Molecular Biology (RECOMB '02).* Washington, DC; 2002:39–48.
[60] Wakefield J, Zhou C, Self S. Modelling gene expression over time: curve clustering with informative prior distributions. In: Bernardo JM, Bayarri MJ, Berger JO, Heckerman D, Smith AFM, West M, eds. *Proceedings of the 7th Valencia International Meeting. Vol 7 of Bayesian Statistics*. New York, NY: The Clarendon Press, Oxford University Press; 2003:721–732.
[61] Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. *Bioinformatics.* 2001;17(10):977–987.
[62] Pan W, Lin J, Le CT. Model-based cluster analysis of microarray gene-expression data. *Genome Biol.* 2002;3(2):Research0009.
[63] Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics.* 2002;18(9):1194–1206.
[64] Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat.* 2000;9(2):249–265.
[65] Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA.* 1999;96(12):6745–6750.
[66] Heard NA, Holmes CC, Stephens DA. *A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves*. London, UK: Imperial College; 2004. Technical Report.
[67] Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol.* 1999;6(3–4):281–297.
[68] Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. *J Multivariate Anal*. 2004;90(1):196–212.
[69] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA.* 2000;97(18):10101–10106.

[70] Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci USA.* 2000;97(15):8409–8414.
[71] Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR. Dynamic modeling of gene expression data. *Proc Natl Acad Sci USA.* 2001;98(4):1693–1698.
[72] Wall ME, Dyck PA, Brettin TS. SVDMAN—singular value decomposition analysis of microarray data. *Bioinformatics.* 2001;17(6):566–568.
[73] Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. *Bioinformatics.* 2001;17(9):763–774.
[74] Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* 1999;9(11):1106–1115.
[75] Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics.* 2003;19(4):459–466.
[76] Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. *Bioinformatics.* 2001;17(4):309–318.
[77] Gibbons FD, Roth FP. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.* 2002;12(10):1574–1581.
[78] Li C, Wong WH. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.* 2001;2(8):Research0032.
[79] Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory.* 1966;10(8):707–710.
[80] Silverman BW. Using kernel density estimates to investigate multimodality. *J Roy Statist Soc Ser B.* 1981;43(1):97–99.
[81] Harezlak J. *Bump Hunting Revisited* [master's thesis]. Vancouver, BC, Canada: Department of Statistics, University of British Columbia; 1998.
[82] Friedman JH, Fisher NI. Bump hunting in high-dimensional data (with discussion). *Stat Comput.* 1999;9:123–162.
[83] Adams NM, Hand DJ, Till RJ. Mining for classes and patterns in behavioural data. *J Opl Res Soc.* 2001;52:1017–1024.
[84] Bolton RJ, Hand DJ, Crowder MJ. Significance tests for unsupervised pattern discovery in large continuous multivariate data sets. *Comput Statist Data Anal.* 2004;46(1):57–79.
[85] DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system (with discussion). *Am Stat.* 1999;53:177–202.
[86] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B.* 1995;57(1):289–300.
[87] Glaz J, Naus J, Wallenstein S. *Scan Statistics.* New York, NY: Springer; 2001.