Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Multivariate Analysis
ELSEVIER
Journal of Multivariate Analysis 90 (2004) 44–66
http://www.elsevier.com/locate/jmva

Problems in gene clustering based on gene expression data

Jenny Bryan[1]

Department of Statistics and Biotechnology Laboratory, University of British Columbia, 333-6356 Agricultural Road, Vancouver, BC, Canada V6M 1L2

Received 31 March 2003

Abstract

In this work, we assess the suitability of cluster analysis for the gene grouping problem confronted with microarray data. Gene clustering is the exercise of grouping genes based on attributes, which are generally the expression levels over a number of conditions or subpopulations. The hope is that similarity with respect to expression is often indicative of similarity with respect to much more fundamental and elusive qualities, such as function. By formally defining the true gene-specific attributes as parameters, such as expected expression across the conditions, we obtain a well-defined gene clustering parameter of interest, which greatly facilitates the statistical treatment of gene clustering. We point out that genome-wide collections of expression trajectories often lack natural clustering structure, prior to ad hoc gene filtering. The gene filters in common use induce a certain circularity to most gene cluster analyses: genes are points in the attribute space, a filter is applied to depopulate certain areas of the space, and then clusters are sought (and often found!) in the "cleaned" attribute space. As a result, statistical investigations of cluster number and clustering strength are just as much a study of the stringency and nature of the filter as they are of any biological gene clusters. In the absence of natural clusters, gene clustering may still be a worthwhile exercise in data segmentation. In this context, partitions can be fruitfully encoded in adjacency matrices and the sampling distribution of such matrices can be studied with a variety of bootstrapping techniques.

AMS 2000 subject classifications: 62H30; 92D10

Keywords: Cluster analysis; Microarrays; Confidence; Bootstrap

E-mail address: jenny@stat.ubc.ca
URLs: . http://hajek.stat.ubc.ca/~jenny.
[1] Supported by a grant from the National Sciences and Engineering Research Council of Canada and by a Career Award from the Micheal Smith Foundation for Health Research.

0047-259X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2004.02.011

The need to sort genes into groups based on some notion of similarity is pervasive in current genome-wide biological investigations. The hope is that similarity with respect to a measurable quantity, such as gene expression, is often indicative of similarity with respect to more fundamental and elusive qualities, such as function. Thus, these gene groups enable researchers to predict, for example, the functional role or regulatory control of a novel gene, based on knowledge of other, better-characterized members of its group. In this article, we focus on the use of high-throughput phenotypic data, primarily gene expression as detected by DNA microarrays, for the formation of gene groups.

The body of techniques known generally as "cluster analysis" or "unsupervised learning" is extremely useful for grouping genes. In the 5 years since Eisen et al. published a landmark paper [9] describing cluster analyses of gene expression data in the yeast *S. cerevisiae* and in primary human fibroblasts, this particular analytical approach has become a de facto standard for analysis of DNA microarray data, particularly when the investigation goes beyond more straightforward questions, relatively speaking, regarding differential expression between two populations or conditions. In that paper, the objects to be clustered are genes and the attributes measured on each object are gene expression levels, obtained with DNA microarrays, over a number of conditions. We will refer to such an analysis as *gene clustering*.

In this article, we critically examine prevalent approaches to gene clustering. At the risk of being abrupt, the main points of this article are given here, without the justifications and qualifications that are found in the more detailed treatment of Sections 1–4:

- *The attributes used for gene clustering should be defined in terms of parameters of the gene expression distribution*: This implies that the true gene clustering is also a well-defined parameter, for it is simply the image of the genome in attribute space under a particular clustering procedure (which often relies crucially on the choice of proximity measure).
- *Gene clustering is an exercise in data dissection or data segmentation,* i.e. there is generally an absence of any natural clusters. Evidence for natural clustering is often an artifact of preliminary gene filtering. Therefore, methods for determining the true number of clusters or for describing the strength of the clustering structure often have no *biological* interpretation.
- *The attributes available are a direct consequence of the experiment that was conducted* and the true gene clustering based on a time-course experiment will differ from that based on, say, a factorial experiment. Therefore, the experimental design should be chosen to produce (estimates of) the attributes most likely to reflect biological clusters of interest.
- *The choice of clustering procedure, including the proximity measure, has a tremendous impact on the true gene clustering.* The superiority of one algorithm over another must be established on subject matter grounds, not on statistical performance, since the two algorithms will likely identify two different clustering parameters of interest.

- Given the lack of natural gene clusters, many datasets currently subjected to cluster analysis would yield more informative results if approached with methods for supervised learning or seeded clustering.
- When gene clustering is performed on real-world datasets, the resulting clustering should be acknowledged as an estimate and appropriate measures of uncertainty should be provided.

1. Defining gene attributes and the gene clustering parameter

Here we construct a definition of a gene cluster, by first focusing on the gene-specific attributes used for clustering. Each gene g, $g \in \{1, \ldots, G\}$, has a vector of attributes $\mathbf{a}_g = (a_{g1}, \ldots, a_{gc}, \ldots, a_{gC})$, which reflect gene g's expected expression over the C conditions under study. To reinforce the notion that the attributes are parameters, we use the term "expression trajectory", as in [28], to refer to a gene-specific attribute vector and we rewrite this as a vector of (unknown) gene- and condition-specific expectations: $\boldsymbol{\mu}_{g\cdot} = (\mu_{g1}, \ldots, \mu_{gc}, \ldots, \mu_{jC})$. We can collect these expression trajectories across the genome, by stacking row-wise, into the G by C attribute matrix $\boldsymbol{\mu}$; rows of $\boldsymbol{\mu}$ are gene-specific expression trajectories and columns are condition-specific expected expression profiles. Since the true attribute matrix will be unavailable to us, we turn to the real-world datasets available for analysis. A typical gene expression dataset $\mathbf{X}$ might also be a G by C matrix in which we record one observed gene expression profile $\mathbf{X}_{\cdot c}$ for each condition c. We can regard this as an estimator $\hat{\boldsymbol{\mu}}$ of the attribute matrix $\boldsymbol{\mu}$ based on a sample of size 1. One can imagine how datasets containing more replication will provide similar, albeit higher-precision estimates of the attribute matrix. We have now defined the true gene attributes and have shown how observed data allow us to obtain estimated attributes (real-world examples follow).

The general goal in gene clustering is to group genes based on similarity with respect to their expression trajectories. Therefore, as a first step for a wide variety of clustering algorithms, we must choose a dissimilarity measure $d(\cdot, \cdot)$[2] that quantifies the proximity of genes g and b based on the expression trajectories $\boldsymbol{\mu}_{g\cdot}$ and $\boldsymbol{\mu}_{b\cdot}$. Let $D_{gb} = d(\boldsymbol{\mu}_{g\cdot}, \boldsymbol{\mu}_{b\cdot})$. We will use $\mathbf{D}$ to denote the collection of all such dissimilarities, which can be conveniently represented as a G-dimensional symmetric matrix. The estimate $\hat{\mathbf{D}}$ is easily obtained by applying our chosen distance metric to the estimated attributes.

Finally, we implement the gene grouping, typically through a combinatorial clustering algorithm. Here we employ the vocabulary of [13], in which clustering algorithms are divided into three types: mode-seeking, mixture modeling, and combinatorial algorithms, which include the most popular partitioning and hierarchical methods that make "no direct reference to an underlying probability

[2] Since the distinction between distance and dissimilarity is not crucial to the main issues addressed in this article, we will use the two terms interchangeably.

model''. Using $S(\cdot)$ to denote the clustering algorithm and $\mathbf{C}$ to denote the true clustering parameter, we have that $S(\mathbf{D}) = \mathbf{C}$. A useful, graph-theoretic encoding of the clustering parameter $\mathbf{C}$, specifically motivated by the need to study the distribution of estimated clusterings $\hat{\mathbf{C}}$, is provided in Section 3. For now, it is sufficient to note that $\mathbf{C}$ is either a partition of the G genes or a sequence of such partitions, indexed by the number of clusters. An estimated clustering $\hat{\mathbf{C}}$ is obtained by applying the chosen algorithm to the estimated distance matrix, i.e. $S(\hat{\mathbf{D}}) = \hat{\mathbf{C}}$.

This type of framework, in which gene groups are formed by applying a deterministic rule to parameters of the data-generating gene expression distribution, was first laid out in [27]. In [2,23,27], particular attention is given to gene groups that correspond to the presence/absence/degree of differential expression and to gene groups formed through seeded, or fixed-medoid, clustering. Also, in these articles, gene clustering is generally performed after some sort of filtering, a procedure which we show in Section 2 to have marked effect on the interpretation of gene clusters. In the context of the above work, the current article develops the framework more fully in the case of gene clustering based on higher-dimensional expression trajectories, unfiltered genome-wide data, and using unsupervised clustering methods.

We now introduce and explore two real datasets that we revisit throughout this article. Here, we specify the gene attributes we will use later for gene clustering and depict the relevant genome in attribute space, based on the estimated attributes obtained from the observed data.

1.1. Project normal mouse data from CAMDA 2002

Many readers will be familiar with the dataset from a paper by Pritchard et al. [24], which also served as one of the competition datasets for CAMDA 2002 [3]. The data consist of expression profiles for about $m = 5800$ genes, obtained from cDNA microarrays, from three specific tissues (liver, kidney, testis) from six exchangeable mice. Therefore, the gene-specific attribute is the collection of tissue-specific expected expression values (which are relative to a common reference mRNA pool derived from equal parts of all mRNA samples): $\boldsymbol{\mu}_{g\cdot} = (\mu_{g,\text{liver}}, \mu_{g,\text{kidney}}, \mu_{g,\text{testis}})$. Each tissue sample, for each mouse, was studied with four arrays; each was a comparative hybridization of the tissue sample and the common reference, with balanced dye-flipping within the four replicates. We normalized the data, within array, with the function `vsn` [15] from Bioconductor [14,16], to obtain log-ratio-like relative expression values, using the generalized log transformation proposed in [6,15]. For each mouse, for each tissue, we average across the four technical replicates. Therefore, we can use the observed data to form estimated the estimated attributes $\hat{\boldsymbol{\mu}}_{g\cdot} = (\hat{\mu}_{g,\text{liver}}, \hat{\mu}_{g,\text{kidney}}, \hat{\mu}_{g,\text{testis}})$, by simply taking the tissue-specific averages based on our $n = 6$ sample.

In Fig. 1 we depict the studied portion of the mouse genome in attribute space, through 3-D scatterplots of the estimated attributes. We note that the point cloud is concentrated around the origin, i.e. $(0, 0, 0)$, which corresponds to genes with roughly equal expression in all three tissues. Three dominant ''arms'' radiate out

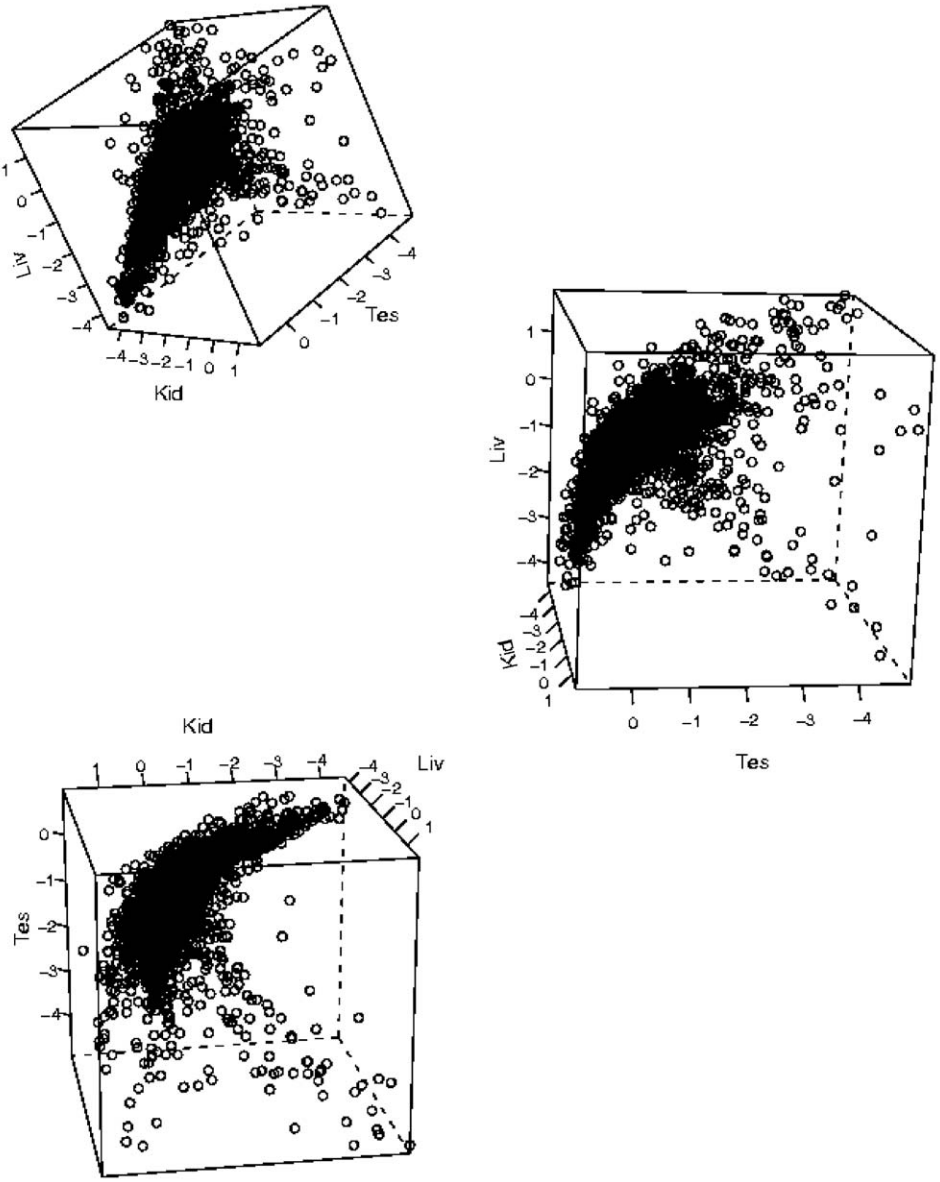

Fig. 1. Estimated attributes for CAMDA data.

from the origin, one for each tissue pair; these arms are populated by genes with approximately equal expression in those two tissues and differential expression in the third. Finally, genes do appear sporadically in other areas, which is indicative of genes with differential expression across all three tissue types. Informally, we observe

that, among these three tissues, liver and kidney present the most similar genome-wide expression profiles, as evidenced by the density of genes in the corresponding arm.

1.2. Yeast time-course data

We present a dataset generated by the van Vuuren lab at the Wine Research Institute at the University of British Columbia. Since the main biological findings have not yet been published, we only use this dataset to highlight broad features of gene expression data vis-à-vis the gene clustering problem. Without loss of statistical content, we will not provide detailed gene-specific findings or biological descriptions of gene clusters. These researchers have conducted a time-course study of yeast. The unit of study is a flask, containing a yeast culture; 15 flasks were grown under the same conditions and expression analysis was performed using Affymetrix GeneChips. The five study times are 24, 48, 60, 120, and 340 h. At each time point, samples are extracted and analyzed from 3 flasks and these flasks are then discarded; therefore $n = 3$ and there are no repeated measures. We pre-process the data using the function `rma` in the `affy` package [17] of Bioconductor [14,16]. In Panel (a) of Fig. 2, we present the expression data for three genes (to oversimplify somewhat, a probe set can be regarded as a probe for one gene) from the $m = 6871$ we analyze.[3]

We now must select the gene-specific attributes. The five study times comprise the $C = 5$ conditions of interest. We could describe an expression trajectory as the collection of time-specific expectations $\boldsymbol{\mu}_{g\cdot} = (\mu_{g1,\ldots,\mu_{g5}})$. The estimated attributes would simply be the time-specific averages, such as those highlighted with dashed lines in Panel (a) of Fig. 2. However, after considerable exploration of the raw data and in light of the research focus on broad temporal trends, we prefer to use a simple quadratic model to describe the expression trajectory for each gene g:

$$Y_g(t) = \beta_{0,g} + \beta_{1,g}t + \beta_{2,g}t^2 + \varepsilon_g(t), \tag{1}$$

where $Y_g(t)$ is the expression for gene g at time t and the gene-specific parameter $\boldsymbol{\beta}_g$ summarizes the true temporal trend. The regression parameter $\boldsymbol{\beta}_g$ captures the expression trajectory and is the basis of our chosen gene-specific attribute. The expected expression trends over time, based on the estimates $\hat{\boldsymbol{\beta}}_g$, are depicted with solid lines in the figure.

Since the wine researchers are primarily interested in the shape of these curves, as opposed to absolute expression levels, we focus on the linear and quadratic terms $(\beta_{1,g}, \beta_{2,g})$. Given the above model and focus, we can plot the temporal expression trends for the yeast genome in the plane. We do this in Panel (b) of Fig. 2, where each gene is represented by a point at the observed estimate $(\hat{\beta}_{1,g}, \hat{\beta}_{2,g})$. We note that the

[3] These 6871 probe sets were selected from the 9335 on this particular Affymetrix chip based on whether they derive from annotated chromosomal ORFs; this was performed independently and prior to our exploration of the expression data for gene clustering purposes.

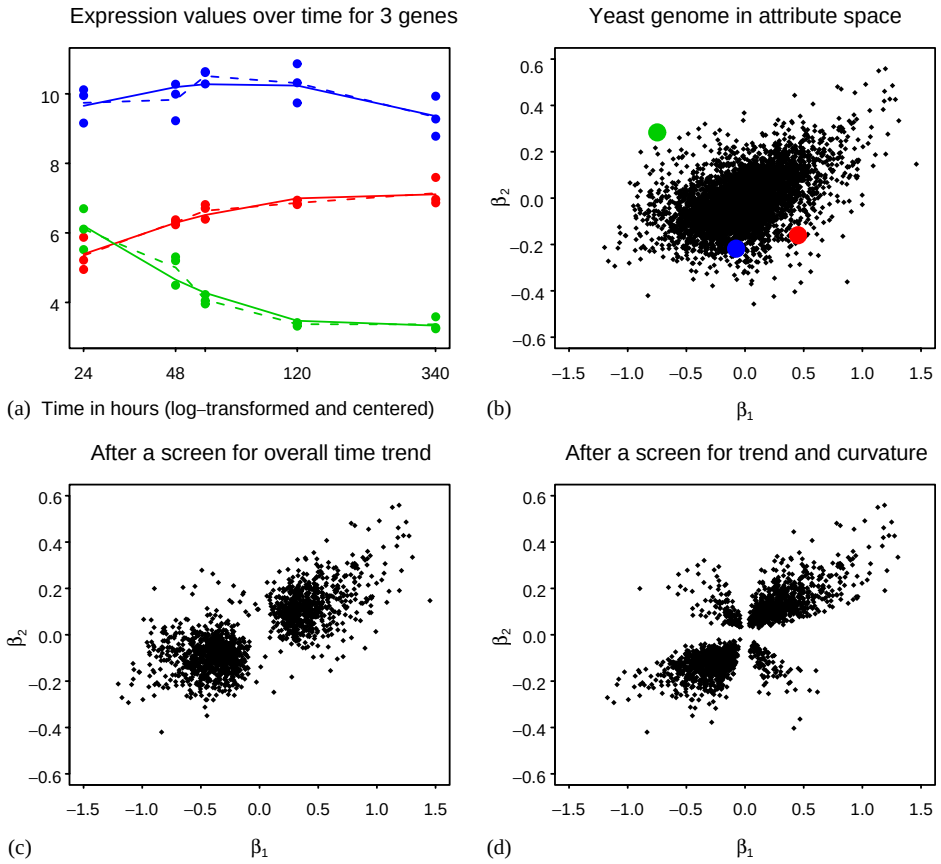

Fig. 2. Yeast time-course data. Panel (a): Expression values for three genes over time (solid dots, color-coded), time-specific averages for each gene (dashed lines), fitted quadratic model (solid lines). Panel (b): Yeast genome in attribute space, using estimated regression coefficients. Genes from panel (a) highlighted. Panel (c): Yeast genome after a screen for statistically significant temporal trend. Panel (d): Yeast genome after a screen for statistically significant linear and quadratic terms.

center of this point cloud appears to be the origin, i.e. no systematic expression change over time.

1.3. Parameter-based attributes critical for statistical formulation

From both a biological and statistical point of view, there is value in assuming the existence of a true, underlying gene clustering (that is undoubtedly context-specific, i.e. there are many biologically coherent clusterings of any given genome). The notion of a true gene clustering arises naturally, if we define the gene-specific attributes in terms of parameters. Such a framework allows us to exploit powerful statistical concepts, when we define and evaluate the success of an observed

clustering based on real gene expression data. As discussed later, it becomes possible to quantify the confidence one should have in an observed clustering and to select the experimental design or sample size that will yield an observed clustering of sufficient quality, with high probability.

The difficulty presented by the measurement error and biological variation present in gene expression profiles can be highlighted by comparing gene clustering with related, but simpler problems. We mention a couple of relevant data examples, drawn from well-known books in multivariate statistical methods. In these examples, as in gene clustering, the goal is to cluster a fixed population of objects, such as a genome, based on chosen attributes; however, in these examples, the attributes can be directly measured with certainty. In [18], they cluster 11 different modern languages based on the words used for the numbers 1–10 and cluster 22 public utilities based on economic data reported in 1975. In [19], they cluster 18 garden flowers based on objective horticultural characteristics. While there are certainly issues regarding the definition and encoding of the above attributes, in general the observed data can be regarded as fixed features of the objects of study. For example, if horticulturists agree that the Begonia *Bertinii boliveiensis* is a tuberous plant and that the Pink Carnation *Dianthus* is not, then we, as analysts, can safely regard the recorded values for the "tuberous" binary variable as biological truth. The equivalence of the observed data and the attributes implies that the observed clustering *is* the true clustering, under the choice of attributes, distance metric, and algorithm. Given the lack of stochastic behavior, a sample of size one will provide a perfect estimate of the true attributes and, therefore, will lead to perfect recovery of the clustering structure.

The mRNA transcript abundances provided by DNA microarrays represent an entirely different sort of data and, therefore, present a dramatically different sort of input for clustering procedures. This is due to unavoidable biological liability (within-unit variability), biological diversity (between-unit variability), and to the measurement noise inherent in the microarray experimental platform. We must regard the expression profile obtained from one person or one petri dish as one observation of a particular, high-dimensional random variable. When we acknowledge the use of estimated expression trajectories as estimated attributes, we can use many conventional approaches to define, asses, and control the error in observed clusterings.

1.4. Reasonableness of behavior "in the limit"

A final related point is that, to invoke statistical reasoning for gene clusters, the attributes must be defined such that larger datasets imply greater precision in estimated attributes, not different attributes altogether. This particular issue is what makes the clustering approach and data collection strategy advocated in [9] impossible to formalize statistically. In this analysis, the underlying data used for clustering yeast genes include temporal expression measured during the diauxic shift (7 timepoints), the mitotic cell division cycle (18 timepoints), sporulation (7 timepoints), low-temperature conditions (4 timepoints), high-temperature conditions

(6 timepoints) and under reducing shocks (4 time points). In total, expression profiles were collected under no less than $C \approx 75$ conditions (the above list is not exhaustive), falling into eight separate experiments.[4] Each condition is studied with one microarray, so the estimated attributes $\hat{\boldsymbol{\mu}}_{g\cdot}$ are based on a sample size of one. If one is willing to expend more arrays, the authors suggest that "… when designing experiments, it may be more valuable to sample a wide variety of conditions than to make repeat observations on identical conditions". But the implication of such a strategy is to change the gene-specific attributes themselves. Even if the true gene trajectories were available and the distance metric were fixed, there is no reason to believe that the true gene-to-gene distances would be stable as the conditions under study expand. A sequence of estimated clusterings obtained with $n = 1$ as $C \to \infty$ is a sequence of extremely imprecise estimates of ever-changing parameters, not a sequence of estimates converging in a probabilistic sense to one well-defined gene clustering.

In fact, it is unclear that the type of attribute chosen in [9] is particularly desirable for forming biologically coherent gene clusters. The relevant aspects of this attribute are its dimension ($C > 75$) and its coverage of many loosely related conditions (recall the eight separate experiments). Conventional biological wisdom dictates that most genes are not expressed under most conditions much of the time, at least for many higher-level organisms. Conversely, at any given time and under fixed conditions, we only expect to see measurable expression for some relatively modest fraction of the genome. By extension, we may conclude that most gene networks or other functionally linked gene groups are only activated under certain conditions and/or for limited time periods. Therefore, it seems to follow that for most such gene groups, as we gradually study them under an increasingly large number of diverse conditions, some of which do not occur in nature, the proportion of conditions under which there is measurable expression and evidence for coexpression will get arbitrarily small. Thus, the true gene distances, based on such high-dimensional and eclectic trajectories, will tend to approach some degenerate value for most gene pairs as C increases. If the above intuition holds, it suggests that there is more value in clustering genes based on well-replicated experiments spanning a relatively small set of related conditions rather than on meta-datasets built by combining data from across different experiments, labs, and array platforms. If there is interest in investigating "consensus" gene clusters for an organism, in large and diverse meta-datasets, it may be advisable to first form estimated clusterings *within* experiments and then combine the results *across* experiments using established techniques for meta-analysis. This would, in particular, provide a statistically sound blueprint for the integrated analysis of disparate high-throughput experiments, i.e. for combining expression data and molecular interaction data.

[4] All expression measurements are relative to an experiment-specific reference sample, which is typically chosen to be an mRNA pool taken from time 0. Furthermore, these data are analyzed after taking base 2 logarithms, i.e., the authors work with log ratios.

2. Gene clustering differs from subject or condition clustering

We make a conceptual distinction between the motivations and methods for grouping genes versus the grouping of expression profiles derived from different populations or conditions. The grouping of expression profiles is a usually a classic application of cluster analysis, in the sense of seeking *natural clusters*. In a review of cluster analysis over 30 years ago [4], Cormack laments the fact that clusters are rarely formally defined, but highlights two intuitive qualities that most analysts value: internal cohesion and external isolation. The first edition of Everitt's text on cluster analysis offers more concrete details about natural clusters:

> A description of what constitutes a cluster which probably agrees closely with our intuitive understanding of the term is given by considering entities as points in a p-dimensional space, with each of the p variables being represented by one of the axis [sic] of this space. The variable values for each entity now define a p-dimensional co-ordinate in this space. Clusters may now be described as continuous regions of this space containing a relatively high density of points, separated from other such regions by regions containing a relatively low density of points. Clusters described in this way are sometime referred to as *natural clusters*. [10, p. 44]

When approaching a dataset consisting of expression profiles generated under different experimental conditions or from different populations,[5] it is reasonable to assume that each population generates expression profiles from a particular distribution. Thus the data can be regarded as observations from a mixture, with each component identifying one population (this is discussed on [4, p. 324] in the context of taxonomy, in which the objects to be grouped are species, based on randomly sampled individuals from these populations). If there is sufficient distinction between these components, they may also be identified with modes of the mixture distribution and observed datasets may very well exhibit natural clusters. As an aside, we note that variable (or gene) selection for the purposes of cluster analysis is an attempt to find a subspace of the attribute space in which there exists natural clustering structure. In this context, three important points aid the statistical formulation of cluster analysis:

- A cluster is well-defined; it corresponds to a component (perhaps even a mode) and, therefore, to a population. It follows that the true number of clusters is well-defined. Furthermore, the equivalence of clusters and populations implies that each cluster has external meaning.
- Well-defined measures of clustering strength can be constructed. By this we refer to quantities that can be computed for the true data-generating mixture distribution. Quantities like the silhouette [19] and the gap statistic [26] are empirical versions of such measures, reflecting the degree of internal cohesion and

[5] Hereafter we use the term "population" broadly, regardless of whether the source of the expression profile is a true subpopulation, a treatment group, an experimental condition, or a genotype.

external isolation in the observed data. Generally speaking, these are measures of what is often called the internal validity of a clustering.

- Well-defined measures of observed clustering validity can be constructed. By this we refer to measures of concordance between observed cluster (i.e. population or component) labels and the true labels. Quantities like prediction strength, employed in [5], or misclassification rate are available as a direct consequence of the external meaning of a cluster. Such measures are often referred to as external criteria or indices.

We note that the external isolation or separation of clusters is an aspect of the configuration of a set of objects in attribute space that is beyond the control of the analyst (more correctly, it is a property of the underlying mixture that generated those objects). In the absence of isolation, the exercise of grouping is variously referred to as dissection [4,20] or segmentation [13] and is considerably more difficult to define and motivate statistically than the natural clustering described above. To paraphrase Cormack, all groups of objects can be dissected—not all can be clustered. We claim that gene clustering is generally an exercise in segmentation. It then follows that many clustering methods and measures of strength and validity, with their implicit assumption of natural clusters, are less relevant and interpretable than they might initially seem and than we might hope.

We first justify our assertion of the lack of natural gene clusters logically and then proceed to examine some real datasets. Given the set of conditions under study and the organism of interest, the genome is a large set of points in the attribute space. Each point represents a particular gene expression trajectory, i.e. the expected (relative) expression of a gene across the conditions of interest. First, we find it implausible to make a meaningful mixture model assumption for the genome. By this we mean, a mixture in which the number of components is relatively small compared to the genome, i.e. genes fall into a small number of large classes that generate expression data according to the same distribution. It seems much more realistic that each gene has a unique trajectory. For analytical convenience, one may choose to make a mixture model assumption, but it is important to acknowledge that the mixture is a computational construct that does not directly reflect biological realities. Second, we find it implausible that the genome would naturally fall into clusters separated by empty areas of the attribute space. Why would certain trajectories be exhibited by many genes and other, similar trajectories be exhibited by essentially no genes? The latter point is especially noteworthy, as we see below, because common methods of data pre-processing—specifically, the pre-screening of genes—will create such sparsely populated areas as an artifact.

2.1. *Natural clusters not evident in real, unfiltered data*

We briefly revisit the CAMDA data introduced in Section 1.1 and depicted in Fig. 1. Although the mouse genome does not inhabit a simple spherical or ellipsoid region in the liver–kidney–testis space, it also does not fall into clear natural clusters, i.e.

disjoint regions separated by relatively empty areas in the expression trajectory space. The most obvious structure here is the presence of the three arms, described earlier, radiating out from the large collection of genes near the origin. Rather than attempting to recover this in an unsupervised fashion, one could simply define several classes of genes explicitly (e.g. same expression in tissues A and B, but down-regulated in tissue C), proceed to classify genes based on observed attributes, and provide measures of confidence in the classification. At the end of such an analysis, one would have the added benefit of knowing exactly what each gene class signified. The usual practice of gene filtering would generally have the effect of artificially removing the large gene-point cloud near the origin, leaving the three "arms" behind for possible re-discovery by unsupervised algorithms. Although there is certainly nothing inherently wrong about the latter approach, it seems to be an unnecessarily indirect strategy.

Let us now recall the yeast time-course data introduced in Section 1.2. We are studying gene expression over 5 timepoints and each gene trajectory is parameterized by $\boldsymbol{\beta}$ from model (1). Since the focus is on the shape of the temporal trend, the attributes for gene clustering are $(\beta_{1,g}, \beta_{2,g})$. The yeast genome is depicted in the plane in Panel (b) of Fig. 2, based on attributes estimated from the observed data. A striking feature of this plot is the lack of evidence of any natural clustering structure. That is, we see no regions in the plane that are densely populated and that are separated from other such regions by more sparsely populated areas.

What is interesting is the ease with which we can induce an apparent clustering structure, by applying the typical gene filters used in such cluster analyses. Generally, the first step in gene clustering is to eliminate genes that do not exhibit sufficient evidence for expression differences across the conditions being studied. The usual motivation for this is to reduce computation and to avoid further study of "uninteresting" genes. A notable side effect, however, can be the creation of a clustering structure that was not present in the original genome-wide data. Here we use the fitted model to retain genes with evidence of nontrivial temporal trend. Specifically, we retain only those genes exhibiting p-values less than 0.001 for the F-statistic that tests the intercept-only model versus the quadratic model of Eq. (1). Such genes are depicted in Panel (c) of Fig. 2. We see that the remaining 1640 genes now fall into two apparent natural clusters, roughly corresponding to genes with a positive time trend (i.e., $\beta_{1,g} > 0$) and those with a negative time trend (i.e., $\beta_{1,g} < 0$). We can construct another model-based filter that retains only those genes exhibiting p-values less than 0.15 for both the linear and quadratic coefficients. Such genes are depicted in Panel (d) of Fig. 2 and we see that the 1917 retained genes now fall into four natural clusters. These clusters correspond well to the four possible combinations of overall time trend (up vs. down regulation) and concavity (up vs. down). At this point, we make an informal conjecture that the number of gene clusters an analyst will uncover after filtering this data can be predicted quite accurately, based on the dimension of the attribute space and the number of independent constraints in the filter. Once again, if one is ultimately going to form the four gene groups suggested by Panel (d) of Fig. 2, which indeed constitute an

interesting segmentation of the yeast genome, then there are much more transparent ways to achieve this than to apply a gene filter, followed by unsupervised cluster analysis. See [2,21,23,27] for illustrations of seeded clustering, in which clusters are anchored at known attributes. See [1] for an application of direct gene classification.

3. The clustering parameter

When unsupervised gene clustering is performed and acknowledged as an (likely) exercise in data segmentation, it can still provide considerable benefits in the interpretation of gene expression trajectories across entire genomes. In fact, Eisen et al. describe the analysis of [9] as a method of "organizing" or "illuminating order" in expression data, with no explicit claims of recovering underlying natural clusters. Nongenomic examples of worthwhile data segmentation include splitting a homogeneous population of student (grades) into the classes A through F and the division of houses in a town into postal districts [11, p. 7].

However, the lack of natural clustering structure has certain implications for the statistical formulation of gene clustering. Recall the three points outlined in Section 2, for the case of condition/subject clustering. The recurring theme here was that the identification of clusters and distinct data-generating mechanisms gave absolute meaning to each cluster and, ultimately, allowed many concepts to be well-defined, e.g., cluster number, clustering strength, observed clustering validity. In the absence of a plausible mixture model assumption, the true gene clusters have no external meaning, but are only implicitly defined by the chosen procedure to map the genome in attribute space into a clustering structure. We propose that the true gene clustering parameter is better summarized through the collection of all possible gene pair connections, rather than through cluster labels, which have no external meaning. We say there is a connection between two objects, or genes, when they belong to the same cluster.

We assume that the clustering procedure is chosen from the two broad classes of partitioning and hierarchical methods. Note that a hierarchical clustering is simply a sequence of partitions, with a special nested structure induced by recursive splitting or merging. To describe a partition, we can use the graph theoretic concept of an adjacency matrix (graph theoretic approaches to clustering are already noted in the 1971 review of [4, p. 330]). The objects to be clustered are the genes and these can be regarded as the nodes of a graph. The connections defined above comprise the edges of the graph. The adjacency matrix of a graph, or partition, is simply a symmetric matrix $\mathbf{J}$ of indicators in which J_{gb} indicates for an edge between nodes g and b or, equivalently, joint cluster membership of objects g and b. Note that such a matrix can always be transformed into block-diagonal form through rearrangements of the rows and columns; blocks then correspond to clusters and the number of blocks equals the number of clusters.

Given a partitioning method, one can obtain a length m sequence of partitions by iteratively directing the algorithm to divide the objects into K groups, where

$K \in \{1, \ldots, m\}$. We encode each partition in an adjacency matrix and denote the sequence as $\bar{\mathbf{J}} = (\mathbf{J}_0, \mathbf{J}_1, \ldots, \mathbf{J}_{m-1})$. The adjacency matrix $\mathbf{J}_K$ encodes a partition containing $m - K$ clusters. Therefore, $\mathbf{J}_0$ is the trivial partition in which each object comprises a cluster, for a total of m clusters of size 1. At the other extreme, $\mathbf{J}_{m-1}$ is another trivial partition in which all objects are collected into one cluster of size m. If the partition sequence $\bar{\mathbf{J}}$ results from a hierarchical method, then it can be collapsed into one matrix $\mathbf{L}$ that simply stores element-wise the index K of the adjacency matrix $\mathbf{J}_K$ in which the corresponding connection is first established. We say that the connection between objects g and b is of level l, when $L_{gb} = l = \arg\min\{d : J_{gb,d} = 1, d \in \{0, \ldots m-1\}\}$ and we call $\mathbf{L}$ the level matrix.

With both partitioning and hierarchical methods, one generally records information above and beyond that contained in $\bar{\mathbf{J}}$ or $\mathbf{L}$. For example, with partitioning-around-medoids (PAM) of Kaufman and Rousseuw [19], the diameter and separation of each cluster are reported, which reflect internal cohesion and isolation, respectively. One also records the silhouette of each object and its average globally and by cluster; the silhouette reflects how well-matched an object is with its cluster. With agglomerative methods, each partition is accompanied by the distance between the most recently merged clusters. These quantities, in addition to many others, are generally used in the graphical presentation of the clustering. They may also be used when deciding the "correct" number of clusters, i.e. the choice of K for partitioning or where to prune a hierarchy. We believe that, when clustering is used for dissection or segmentation, these quantities hold considerably less interest than when natural clusters are suspected. Therefore, we focus on the partition sequence $\bar{\mathbf{J}}$, a particular partition $\mathbf{J}_K$, or the level matrix $\mathbf{L}$ as the main clustering parameter of interest $\mathbf{C}$. If the analyst seeks a partition with a certain number of clusters, the choice of K can be informed by practical considerations (what number of clusters provides a sufficient reduction in the number of entities to be described?) or biological interpretability (which K induces clusters with good biological coherence?). In the absence of natural clusters, *there is no true number of clusters*.

Any clustering procedure maps a point cloud in the attribute space to the relevant clustering space. We emphasize that two different clustering procedures will often produce two different clusterings of the same genome, even when using the true trajectories as attributes and the same proximity measure. Using our notation, let $S_1(\mathbf{D}) = \mathbf{C}_1$ and $S_2(\mathbf{D}) = \mathbf{C}_2$ and we see that, for general $\mathbf{D}$ and $S_j(\cdot)$, there is no reason that C_1 should equal C_2. This is why the often-asked question "which clustering algorithm is best?" is an ill-posed inquiry, at least from a purely statistical point of view. Different choices of algorithm imply different parameters of interest, just as the competing choices of mean and median provide distinct measures of central tendency. The relative advantages of the mean and the median depend on qualities of the data-generating distribution and on the analytical goal and, similarly, the optimal choice of clustering algorithm is context specific. If one algorithm produces gene clusters that have greater overlap with the biological clusters of interest, then this is excellent proof of its superiority in that application; note that

this determination requires subject-matter information and a stated analytical purpose. From a purely computational standpoint, we can only appropriately compare algorithms with respect to bias, variance, and computational efficiency when they implicitly define the same (or, perhaps, extremely similar) clustering parameters.

4. Sampling distribution of the estimated clustering

Given the possible clustering parameters of interest laid out in Section 3, we note that the fundamental clustering object is a partition, which can be encoded in an adjacency matrix $\mathbf{J}$. Here we discuss the sampling distribution of estimated partitions $\hat{\mathbf{J}}$ and particular features of that distribution that are relatively accessible and interpretable. These provide the information on statistical certainty that can accompany estimated gene clusterings, i.e., these are the measures of clustering validity that are important to estimate and provide when clustering is largely an exercise in data dissection. In the presence of natural clustering, these measures are also of great interest and relevance, but only in conjunction with many other internal and external indices. The general framework for viewing a gene clustering as a parameter of interest that arises from the application of a clustering procedure to the data-generating gene expression distribution is developed in [2,23,27]. These papers do not provide a full treatment of unsupervised clustering, which is developed here. In [2,23,27], one can find important results regarding the consistency of estimated attributes, even as m grows faster than n.

To begin, we can study the frequency with which individual edges appear in estimated clusterings or, equivalently, the distribution of individual elements of the estimated adjacency matrix. For a given experiment and sample size, each of these is a particular Bernoulli random variable. For the g, b gene pair, there is a true edge state J_{gb} and reappearance probability $q_{gb,n} = q_{gb} = P(\hat{J}_{bg} = 1)$, which is the Bernoulli parameter mentioned above. This quantity q_{gb} reflects the true distance D_{gb}, the distribution of the estimated distance $\hat{D}_{gb}$ around the truth, and also whether the genes g and b happen to lie in an area of the attribute space that is near cluster boundaries. Therefore, q_{gb} is driven by the error contained in estimated attributes and by the joint effect of the clustering algorithm and the true configuration of the genome in attribute space. Intuitively, we expect/hope that the q_{gb} for adjacent genes, i.e., $J_{gb} = 1$, are close to 1 and are close to 0 otherwise. For finite n, if the $q_{gb,n}$ were available, we could use them to refine the more granular information contained in an estimated clustering $\hat{\mathbf{J}}$. One could present an alternative partial clustering that only includes estimated edge states, be they present or absent, for gene pairs where q_{gb} is sufficiently close to either 0 or 1. If α is a user-specified cutoff between 0.5 and 1, we could report $\hat{J}_{gb}$ for g, b pairs, such that $q_{gb} \in \{0, 1-\alpha\}$ or $q_{gb} \in \{\alpha, 1\}$. If $q_{gb} > \alpha$ for all g, b such that $J_{gb=1}$ and $q_{gb} < 1-\alpha$ for all g, b such that $J_{gb} = 0$, then this can be described as a partial clustering in which each edge state has

marginal reappearance probability greater than α. If the condition is violated, then the probabilistic description will only be approximately true.

More globally, we can study the proportion of true edge states that are recovered, possibly separated according to states where $J_{gb} = 0$ and 1. For each estimated clustering $\hat{\mathbf{J}}$, one can populate a 2×2 matrix, in which the rows correspond to the true edge state and the columns correspond to the estimated edge state; the sum of the counts in the four cells is the number of gene pairs, i.e. $\tilde{m} = m(m-1)/2$:

	$\hat{J}_{gb} = 0$	$\hat{J}_{gb} = 1$	
$J_{gb} = 0$	M_{00}	M_{01}	$\tilde{m}_0$
$J_{gb} = 1$	M_{10}	M_{11}	$\tilde{m}_1$
	M_0	M_1	$\tilde{m}$

From $c_j = \tilde{m}_j/\tilde{m}$, we learn the overall connectivity of the clustering. The maximum connectivity occurs with almost trivial clusterings in which there are $K-1$ singleton clusters and one cluster of size $m-K+1$; the minimum connectivity occurs when the K clusters are as equally sized as possible. We refer to $fid = (M_{00} + M_{11})/\tilde{m}$ as the fidelity of an estimated clustering and to $sens_0 = M_{00}/\tilde{m}_0$ and $sens_1 = M_{11}/\tilde{m}_1$ as its negative and positive sensitivity, respectively. Note that $fid = c_0 sens_0 + c_1 sens_1$. The expectations of these random variables provide useful global measures of the extent to which the true adjacency is reflected in observed adjacencies. If known, we could describe an estimated clustering as having the property that the expected proportion of recovered edge states is $E(fid)$ or that the expected proportion of recovered (lack of) edges is $E(sens_1)$ ($E(sens_0)$).

We note the close connection to well-established criteria for the external validity of a clustering. The Rand statistic [25] is exactly equal to *fid* and many other criteria, such as the Jaccard and Fowlkes and Mallows statistics, can be easily expressed based on the contingency table described above [22]. While many of the known properties of these statistics are important to note, their relative weaknesses and merits have generally been studied only in the context of natural clusters. Finally, we refer the reader to [2,23,27] for results that, relying on the consistency proofs mentioned above for estimated attributes, show that the reappearance probability $q_{gb,n}$ approaches J_{gb} as $n \to \infty$, for any gene pair g, b.

4.1. Resampling to estimate clustering validity

In general, it will be impossible to determine quantities like q_{gb}, $E(fid)$, and $E(sens_j)$ through sampling theory. Therefore, we estimate these quantities by creating appropriate bootstrap datasets and applying the same dissimilarity and clustering procedure to yield bootstrap clusterings. We then use empirical proportions and averages to estimate these quantities. Since the objects to cluster, i.e. the genes are fixed, we are not resampling from the genes, but are creating new observations of the estimated gene attributes. Depending on the type of experiment, this could result in resampling individuals to form bootstrap samples or could result

in directly generating bootstrap expression trajectories from the observed data and some modeling assumptions. Based on B realizations of a bootstrap gene attribute matrix, we form bootstrap clusterings $\hat{J}^*$. We estimate the reappearance probability $q_{gb,n}$ with the average g,bth element of $\hat{J}^*$ and denote it $\hat{q}_{gb,n} = \hat{q}_{gb}$. Similarly, we estimate the expected fidelity and sensitivity with the average proportion of data-generating edge states or data-generating (lack of) edges recovered (omitted) in the $\hat{J}^*$. This is precisely the strategy introduced in [27] for assessing sampling variability in gene clusterings, although the adjacency matrix encoding of an unsupervised clustering was not used.

Below we implement the bootstrap for the yeast time-course data, but first we briefly present possible strategies for datasets like the CAMDA mouse data. Recall that the gene-specific attribute is $\boldsymbol{\mu}_{g\cdot} = (\mu_{g,\text{liver}}, \mu_{g,\text{kidney}}, \mu_{g,\text{testis}})$, which we estimate based on averages of tissue-specific (relative) expression across $n = 6$ mice. This gives us the estimated attribute $\hat{\boldsymbol{\mu}}_{g\cdot}$. To obtain bootstrap attributes, one could conduct a nonparametric bootstrap in this case, by repeatedly resampling with replacement from the 6 mice. Given the extremely small sample size, it may be more desirable to make some parametric assumptions; we could, for example, assume that the underlying relative expression measurements are normally distributed. One would then have to confront the issue of correlation across genes for a given tissue and correlation across tissues for measurements of one gene made across the tissues on the same mouse. To obtain a correlation structure from which one can generate new observations, one generally must assume independence or shrink the observed correlation matrix towards the identity, in order to work around the singularity of the observed correlation. Even when these practical concessions make it implausible to blithely regard the bootstrap reappearance proportions as good estimates of q_{gb}, one can still regard this as an extremely relevant simulation study that provides some quantitative information about statistical uncertainty in a difficult situation. We also point out, as was discussed in [4], that the normality assumption *for the estimated attributes* becomes more plausible as n increases, since most estimated attributes are asymptotically normal. Therefore, the parametric bootstrap can be extremely useful even in the absence of normality in the underlying expression data.

4.2. Analysis of the yeast time-course data

We applied three clustering algorithms to a subset of the yeast time-course data introduced earlier. Five hundred genes were selected at random to produce a computationally manageable example, but we stress that genes were not selected according the estimated attributes they exhibit. We specified the gene distance to be the squared Euclidean distance between gene-specific attributes, which we choose in Section 1.2 to be $(\beta_{1,g}, \beta_{2,g})$. The algorithms were AGNES, DIANA, and PAM, which are described fully in [19] and implemented in the `cluster` package in R. These are agglomerative hierarchical, divisive hierarchical, and partitioning algorithms, respectively, and each was used to cluster the 500 genes into $K = 2, \ldots, 5$ groups. The observed clusterings are presented in Fig. 3, using the same axis

limits as in Fig. 2. For the moment, we assess these clusterings as if they were based on the true gene attributes; that is, we evaluate them with respect to their utility as a data segmentation. We once again note the lack of natural clusters and see that the unsupervised gene clusters formed essentially produce gene groups based on the direction and magnitude of the linear trend. It is difficult, if not impossible, to declare any algorithm or choice of K as the optimal one. We do suggest that the clusterings produced by AGNES for $K = 2$ and 3 are rather uninformative, given

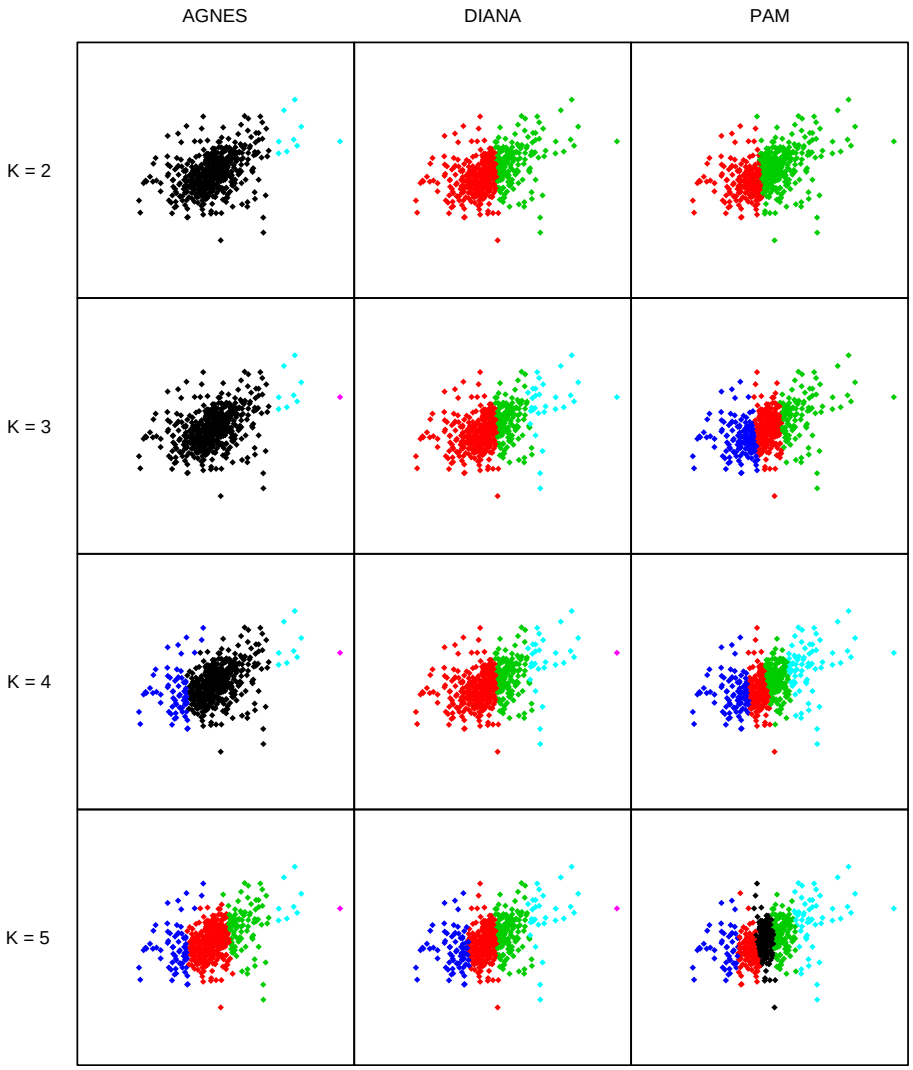

Fig. 3. Clusterings of the yeast genome, based on estimated attributes. Color schemes chosen only to aid visual comparisons; there is no objective cluster matching across methods or, for PAM across K.

that 491 of the 500 genes are placed in one group. This distinctly fails to meet the data reduction goal of dissection. In fact, for all K (AGNES) and for $K>3$ (DIANA), the hierarchical algorithms choose a partition that includes a singleton cluster. Although there is no basis for objective cluster matching across the methods, it is interesting to note substantial concordance across the AGNES $K=5$, DIANA $K=5$, and PAM $K=4$ clusterings. The Rand statistics, or proportion of concordant edge states, are 0.76 for AGNES $K=5$ and DIANA $K=5$, 0.68 for AGNES $K=5$ and PAM $K=4$, and 0.762 for DIANA $K=5$ and PAM $K=4$.

We now acknowledge the estimated nature of the clusterings and use the bootstrap to estimate certain features of the cluster sampling distribution. For each gene, we have the estimated attribute $(\hat{\beta}_{1,g}, \hat{\beta}_{2,g})$ and a corresponding estimated covariance matrix. From this, we generate bootstrap attributes by resampling from a bivariate normal at the observed attribute and estimated covariance matrix. We apply the same distance measure and clustering algorithms. This was done for $B=100$ bootstrap datasets and reappearance probabilities q_{gb}, the expected fidelity $E(fid)$, and expected sensitivities $E(sens_j)$ were estimated with relative frequencies and averages of proportions.

In Fig. 4, we provide smoothed histograms of the reappearance proportions $\hat{q}_{gb}$, for gene pairs with a data-generating edge present and absent, separately. As indicated earlier, the tendency for gene pairs that are connected in the data-generating distribution to exhibit (estimated) q_{gb} near 1 is seen in many cases, but not in all. In fact, as K grows, the distribution of $\hat{q}_{gb}$ among adjacent genes in the data-generating distribution grows flatter for both PAM and DIANA. This is implied by the fact that these algorithms are actually grouping the genes into K clusters, each of considerable size, and, therefore, there are many gene pairs that have small (estimated) distance, but that also lie in a region of the attribute space where cluster boundaries tend to occur. In Table 1 we provide the average bootstrap fidelity and sensitivities. We see that the overall recovery of data-generating edge states is quite high; the average bootstrap fidelity is generally around 80%. For $K<4$, AGNES recaptures edge states very well, albeit for clusterings that are potentially rather uninteresting. This is inevitable, as the fidelity of high-connectivity estimated clusterings will tend to be close to 1. The failure of AGNES to effectively recapture "lack of edges", i.e. $J_{gb=0}$, is an interesting illustration of the statistical instability of an agglomerative algorithm versus divisive and partitioning methods. AGNES partitions $m=500$ objects into, say, $K=3$ clusters by selecting the $m-K=497$th element in a sequence of partitions enacted on estimated attributes. Compare this to the relative stability of DIANA, which induces a very similar data-generating clustering, by selecting the 2nd element in a partition sequence. For the gene clustering problem, this provides a vivid illustration that a hierarchical algorithm "can never repair what was done in previous steps" [19, p. 44] and, when K is small relative to m, it may be advisable to employ divisive partitioning methods.

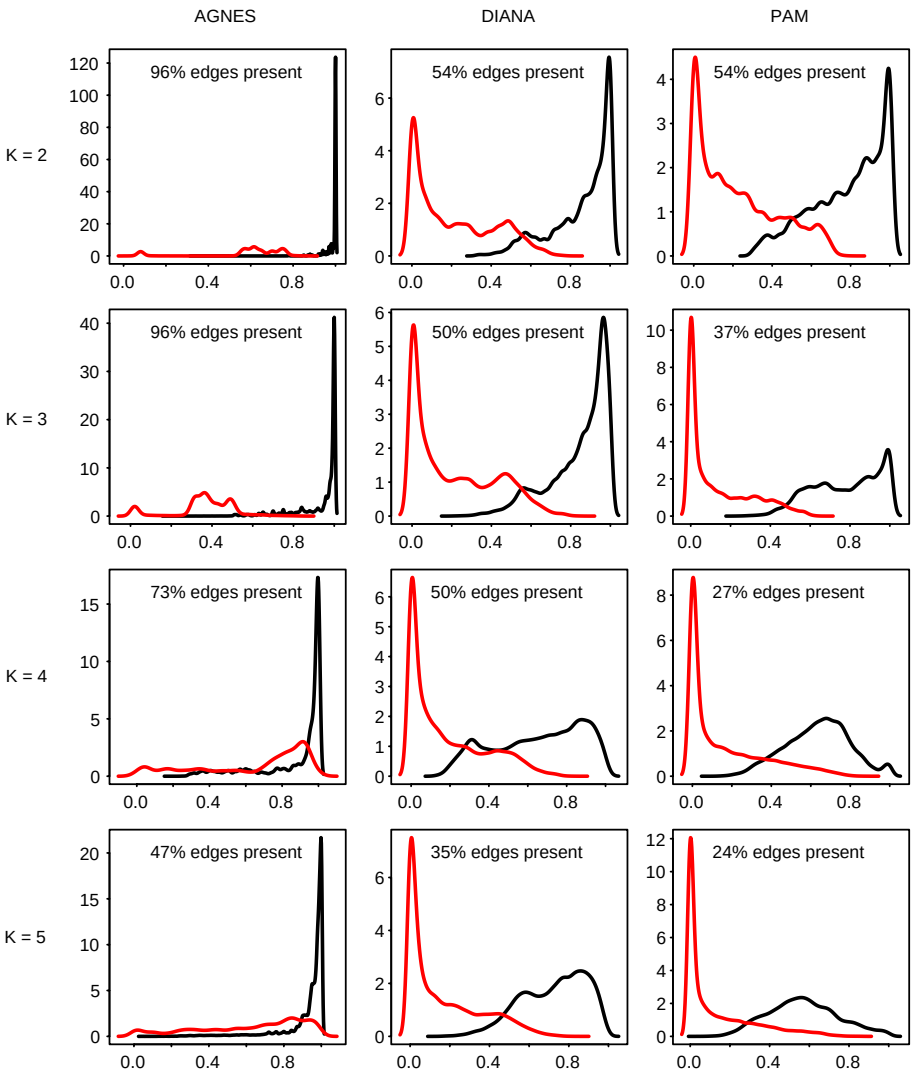

Fig. 4. Edge reappearance proportions in the bootstrap clusterings. Smoothed histograms of the $\hat{q}_{gb}$ for edges present in the data-generating, observed clustering (black) and for edges absent (red).

4.3. Link to the problem of regions

We would like to point out a link between the reappearance probabilities q_{gb} described above and certain Bayesian posterior probabilities outlined in work by Efron and Tibshirani [8]. In what they term the "problem of regions", the analyst wants to ascertain which one of a discrete set of possibilities applies to a continuous

Table 1
Overall recovery of edge states in bootstrap clusters for yeast time-course data. Averages across the $B = 100$ bootstrap clusterings

	AGNES			DIANA			PAM		
K	$sens_0$	$sens_1$	fid	$sens_0$	$sens_1$	fid	$sens_0$	$sens_1$	fid
2	0.41	0.99	0.97	0.79	0.86	0.83	0.78	0.79	0.78
3	0.64	0.94	0.93	0.79	0.84	0.82	0.87	0.78	0.84
4	0.35	0.88	0.74	0.82	0.65	0.74	0.84	0.64	0.78
5	0.38	0.93	0.64	0.84	0.73	0.80	0.87	0.57	0.80

parameter. For example, we might focus on the modality of a density or the degree of a polynomial regression function. They describe a very straightforward "first-order bootstrap" approach to summarizing the evidence in observed data for different values of the discretized parameter of interest. One simply generates bootstrap datasets from the observed distribution and estimates the discretized parameter of interest. The relative frequencies associated with each element of the discrete parameter space are used to estimate a true probability they call the "confidence value". In certain cases, the confidence values are shown to be posterior probabilities for the presence of that feature, given a flat prior on the original parameter of interest.

In our case, the underlying parameter is the data-generating gene expression distribution and, for gene pair g, b, we are interested in whether this implies that an edge between genes g and b is present or absent. The reappearance probabilities q_{gb} are equivalent to the confidence values described above and our resampling strategy for estimating q_{gb} is equivalent to the first-order bootstrap of Efron and Tibshirani. However, in the clustering context, it would be difficult if not impossible to demonstrate the direct connection to Bayesian posterior probabilities that was possible in some simple settings. Nonetheless, it is interesting to note the relationship and the heuristic interpretation of $\hat{q}_{gb}$ as the evidence contained in the observed data for an edge connecting genes g and b.

We also point out that the nonparametric bootstrap was used by Felsenstein [12] and Efron et al. [7] to assess confidence in phylogenetic trees, obtained from hierarchical clustering of species using sequence information at individual loci. The resampling strategy employed actually requires resampling from the loci and, therefore, treats them as draws from a population of loci that somehow contribute exchangeable information on phylogeny. If implemented on typical multi-condition microarray data, this would be equivalent to making random selections of *attributes* for the purposes of bootstrap clustering results. Since we wish to summarize clustering uncertainty arising from the error in observed attributes, which are considered to be fixed, this is distinct from the gene clustering problem.

5. Conclusions

In this work, we assess the suitability of cluster analysis for the gene grouping problem confronted with microarray data. The formal definition of attributes as (deterministic functions) of the data-generating parameters greatly facilitates the statistical treatment of gene clustering, by providing a gene clustering parameter of interest. For simplicity, we have focused solely on condition-specific expectations, possibly linked through a model, as gene-specific parameters, but we acknowledge that covariance and/or correlation are also extremely relevant for forming gene attributes in many studies; we regret that these issues cannot be investigated here. We point out that genome-wide collections of expression trajectories often lack natural clustering structure, prior to ad hoc gene filtering. The gene filters in common use can induce a certain circularity to gene cluster analyses: genes are points in the attribute space, a filter is applied to depopulate certain areas of the space, and then clusters are sought (and often found!) in the "cleaned" attribute space. As a result, statistical investigations of cluster number and clustering strength are just as much a study of the stringency and nature of the filter as they are of any biological gene clusters. In the absence of natural clusters, gene clustering is often still a worthwhile exercise in data segmentation. In this context, partitions can be fruitfully encoded in adjacency matrices and the sampling distribution of such matrices can be studied with a variety of bootstrapping techniques. This also implies that simulation studies can be used to determine the sample size needed to create estimated attributes and, therefore, estimated gene clusterings that meet the analyst's criteria for statistical stability.

References

[1] J. Bryan, Gene classification based on deletion set studies, Refereed abstract and talk for Cold Spring Harbor Laboratory/Wellcome Trust Conference on Genome Informatics available from author's website at hajek.stat.ubc.ca/~jenny;
K. Baetz, L. McHardy, K. Gable, T. Tarling, D. Rebérioux, J. Bryan, T. Dunn, P. Hieter, M. Roberge, Yeast genome-wide drug-induced haploinsufficiency screen to determine drug mode of action, in Proc. Natl. Acad. Sci. in press.
[2] J. Bryan, K.S. Pollard, M.J. van der Laan, Paired and unpaired comparison and clustering with gene expression data (special issue on bioinformatics), Statist. Sinica 12 (1) (2002) 87–110.
[3] CAMDA, 2002, http://www.camda.duke.edu.
[4] R. Cormack, A review of classification, J. Roy. Statist. Soc. Ser. A 134 (3) (1971) 321–367.
[5] S. Dudoit, J. Fridlyand, A prediction-based resampling method for estimating the number of clusters in a dataset, Genome Biol. 3 (7) (2002) 0036.1–0036.21.
[6] B. Durbin, J. Hardin, D. Hawkins, D. Rocke, A variance-stabilizing transformation for gene-expression microarray data, Bioinformatics 18 (2002) S105–S110.
[7] B. Efron, E. Halloran, S. Holmes, Bootstrap confidence levels for phylogenetic trees, Proc. Nat. Acad. Sci. 93 (1996) 13429–13434.
[8] B. Efron, R. Tibshirani, The problem of regions, Ann. Statist. 26 (5) (1998) 1687–1718.
[9] M. Eisen, P. Spellman, P. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Nat. Acad. Sci. 95 (1998) 14863–14868.
[10] B. Everitt, Cluster Analysis, Heinemann Educational Books, London, 1974.

[11] B.S. Everitt, S. Landau, M. Leese, Cluster Analysis, Arnold, London, 2001.
[12] J. Felsenstein, Confidence limits on phylogenies: an approach using the bootstrap, Evolution 39 (4) (1985) 783–791.
[13] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, 2001.
[14] http://www.bioconductor.org.
[15] W. Huber, A. von Heydebreck, H. Sueltmann, A. Poustka, M. Vingron, Variance stabilization applied to microarray data calibration and to the quantification of differential expression, Bioinformatics 18 (2002) S96–S104.
[16] R. Ihaka, R. Gentleman, R: A language for data analysis and graphics, J. Comput. Graph. Statist. 5 (3) (1996) 299–314.
[17] R.A. Irizarry, L. Gautier, L.M. Cope, An r package for analyses of affymetrix oligonucleotide arrays, in: G. Parmigiani, E. Garett, R. Irizarry, S. Zeger (Eds.), The Analysis of Gene Expression Data: Methods and Software, Statistics for Biology and Health, Springer, New York, 2003 (Chapter 4).
[18] R.A. Johndon, D. Wichern, Applied Multivariate Statistical Analysis, Prentice-Hall, Englewood Cliffs, NJ, 2002.
[19] L. Kaufman, P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 1990.
[20] M. Kendall, Discrimination and classification, in: P. Krishnaiah (Ed.), Proceedings of Symposium Multivariate Analysis, Dayton, Ohio, Academic Press, New York, 1966, pp. 165–185.
[21] M.K. Kerr, G. Churchill, Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments, Proc. Nat. Acad. Sci. 98 (2001) 8961–8965.
[22] G.W. Milligan, S. Soon, L.M. Sokol, The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5 (1983) 40–47.
[23] K.S. Pollard, M.J. van der Laan, Statistical inference for simultaneous clustering of gene expression data, Math. Biosci. 176 (2002) 99–121.
[24] C. Pritchard, L. Hsu, J. Delrow, P. Nelson, Project normal: defining normal variance in mouse gene expression, Proc. Nat. Acad. Sci. 98 (23) (2001) 13266–13271.
[25] W. Rand, Objective criteria for the evaluation of clustering methods, J. Amer. Statist. Assoc. 66 (1971) 846–850.
[26] R. Tibshirani, W. Guenther, T. Hasite, Estimating the number of clusters in a data set via the gap statistic, J. Roy. Statist. Soc. Ser. B 63 (2001) 411–423.
[27] M.J. van der Laan, J. Bryan, Gene expression analysis with the parametric bootstrap, Biostatistics 2 (4) (2001) 445–461.
[28] J.C. Wakefield, C. Zhuo, S.G. Self, Modelling gene expression data over time: curve clustering with informative prior distributions, Bayesian Statist. 7 (2003) 711–722.