J. R. Statist. Soc. B (2002)
64, *Part* 4, *pp.* 717–736

A statistical framework for expression-based molecular classification in cancer

Giovanni Parmigiani, Elizabeth S. Garrett, Ramaswamy Anbazhagan and Edward Gabrielson

Johns Hopkins University, Baltimore, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section on '*Statistical modelling and analysis of genetic data*' *on Wednesday*, *May 22nd, 2002*, Professor D. Firth *and* Professor R. A. Bailey *in the Chair*]

Summary. Genome-wide measurement of gene expression is a promising approach to the identification of subclasses of cancer that are currently not differentiable, but potentially biologically heterogeneous. This type of molecular classification gives hope for highly individualized and more effective prognosis and treatment of cancer. Statistically, the analysis of gene expression data from unclassified tumours is a complex hypothesis-generating activity, involving data exploration, modelling and expert elicitation. We propose a modelling framework that can be used to inform and organize the development of exploratory tools for classification. Our framework uses latent categories to provide both a statistical definition of differential expression and a precise, experiment-independent, definition of a molecular profile. It also generates natural similarity measures for traditional clustering and gives probabilistic statements about the assignment of tumours to molecular profiles.

Keywords: Microarray data analysis; Mixture distributions; Molecular classification of cancer

1. Introduction

1.1. Background

Gene expression microarrays are assays for studying gene expression on a large portion of the genes on a genome of interest. See Schulze and Downward (2001) for a recent review. Cancer research is among the most important application areas for gene expression investigations. Currently, cancer classes used in prognosis and therapy decisions are defined on the basis of morphological features, sometimes complemented by single-gene or single-protein assays. Existing classes often include a broad range of malignancies with widely different prognosis, or with different responses to available therapies. For example, about 75% of infiltrating breast cancers are classified as ductal carcinomas, a category also known as 'carcinoma of no specific type'. Ductal breast cancers are highly variable in their clinical aggressiveness and response to treatment, probably as a result of different progenitor cell types and different molecular pathways that give rise to these cancers. The diversity of ductal breast cancers is reflected in their highly variable histologic appearances, but distinctive subtypes cannot be objectively recognized by morphologic criteria. See Tavassoli (1992) and Perou *et al.* (1999). More broadly, cancer is a heterogeneous disease from a genetic standpoint, so molecular classification based

Address for correspondence: Giovanni Parmigiani, Division of Oncology and Biostatistics, Johns Hopkins University School of Medicine, Suite 1103, 550 North Broadway, Baltimore, MD 21205-2011, USA.
E-mail: gp@jhu.edu

1369–7412/02/64717

on gene expression information gives hope for individualized and more effective prognosis and treatment.

The biological literature on the molecular classification of cancer by using microarrays is growing and we shall not review it here. To motivate our approach, however, it is useful to consider briefly two successful examples. Alizadeh *et al.* (2000) considered the classification of the most common class of non-Hodgkin's lymphoma, the so-called diffuse large B-cell lymphoma (DLBCL). They developed a custom-made complementary DNA array with about 20 000 genes potentially related to carcinogenesis. They measured gene expression in malignant lymphocytes from 40 patients and, to provide a frame of interpretation for the results, they also measured 56 samples of normal lymphocytes, of eight different types. They visualized data by using a colour map of gene expression for all genes after rearranging the expression matrix's rows and column using a hierarchical clustering approach (Eisen *et al.*, 1998). Using this visualization, they identified subgroups of genes with both similar expression patterns and similar biological functions. In parallel, they distinguished two subclasses of DLBCL patients, one of which they characterized by its similarity in expression to one of the normal subtypes. Finally, they correlated, informally but effectively, two of the gene groups to the two subclasses. They could also perform a preliminary validation of their classification by using prognostic data, showing differential survival in the two DLBCL subclasses.

Bittner *et al.* (2000) investigated molecular classification of cutaneous melanomas, another malignancy which is difficult to classify on the basis of morphology. They considered gene expression on a set of 31 tumours and investigated potential subtypes by using three complementary approaches: a hierarchical clustering dendrogram, based on the matrix of Pearson correlation coefficients derived from the expression measurements of all genes meeting a minimum level of expression in each hybridization, a three-dimensional multidimensional scaling (Kruskal, 1964) representation of the same matrix and the non-hierarchical clustering algorithm termed the cluster affinity search technique, developed by Ben-Dor *et al.* (1999). A visual examination of the results led to focusing on a set of 19 tumours that appear to be tightly clustered in all three approaches. They then searched for genes that are strongly associated with the subgroup of 19 by an analysis of variance in the multidimensional scaling space. The genes so identified are then grouped by expression, interpreted and used successfully in confirmatory work.

1.2. Statistical issues

As illustrated by these two applications, the main goal of microarray analyses of unclassified cancer is to identify, or hypothesize, novel cancer subtypes for later validation and prediction, with a view to individualized prognosis and therapy. At this early stage of genomic investigations, a successful classification analysis does not necessarily need to assign all available tumours to a subtype, but only to identify interesting subgroups for further analysis. Also, it is critical that molecular classes are interpretable, and amenable to further biological analysis. Eventually, classes need to be recognizable in clinical settings using less expensive and more accurate assays than expression microarrays. For this reason, it is not necessary, nor generally useful, to use all the genes to define subtypes. A critical advantage of subtypes based on a small number of genes is that they lend themselves to easier validation. Simple hypotheses about subtypes can be efficiently assessed and either discarded or pursued for clinical implementations, whereas hypotheses involving genome-wide measurements are constrained to the context of high throughput analysis.

Because of difficulties in acquiring tissue and the large costs of microarray experiments, classification is often attempted by using a relatively small number of tumours, compared with the number of genes on the array. Gene-to-tumour ratios of 100 and noisy gene expression measurements are common. These constraints and the goals outlined above suggest that initial progress and translational results are more likely to occur by aiming at classifications that involve a small number of genes and their interactions, rather than trying to exploit all gene expression information at once. Currently, the utility of a genomic classification analysis lies primarily in the comprehensive nature of the list of candidate genes. In this context, gene expression experiments are best interpreted as screening tools, whose goal is to identify candidate directions for more pointed and accurate investigations, using a variety of existing assays. Molecular profiles that are based on a small subset of genes marking the key steps in a complex carcinogenic pathway would be ideal targets.

Statistically, a prototypical molecular classification experiment is a random sample of tumours that are at present unclassifiable, each with an array of expression measurements. Formally, molecular classification could be viewed as the problem of finding subgroups of points each of which represents a tumour in a high dimensional gene space. There are numerous methodologies to approach this general type of problem. These include a variety of algorithms for finding clusters in data, reviewed by Hartigan (1975) and Quackenbush (2001) in the context of microarrays, as well as other unsupervised algorithms in pattern classification reviewed, for example, by Ripley (1996) and Duda *et al.* (2001).

There are, however, aspects that are specific to molecular classification. First, vast and growing information about human genes functions and interaction between genes is becoming available. This information is relevant and needs to be brought to bear, especially in view of the unfavourable gene-to-tumour ratios. Second, the ‘not-all-genes’ and ‘not-all-tumours’ properties of molecular classification of cancer discussed above set molecular classification aside from a large portion of general purpose clustering and classification approaches. Although traditional unsupervised classification technologies can provide useful insight and important building-blocks in molecular classification strategies, specific approaches are necessary.

1.3. Data analysis tools in molecular classification

To meet these challenges, methodologies for unsupervised classifications of microarray data are growing rapidly. See Herrero *et al.* (2001) or Segal *et al.* (2001) and references therein for recent developments. The ‘gene shaving’ algorithm, proposed by Hastie *et al.* (2000), influenced some of the ideas behind the present paper. Gene shaving searches for clusters of genes showing high variation across tumours, high correlation across the genes within a cluster and high diversity of gene expression from cluster to cluster. In applications, clusters are often selected for further analyses on the basis of expert elicitation. Cluster averages by tumour are used for classification. The same gene can belong to more than one cluster.

More broadly, most approaches pursue dimension reduction, recognizing that the gene-to-tumour ratios make it difficult to develop empirical classifications that make full use of the genomic dimensionality. We can identify two broad and somewhat overlapping tendencies: the first is to generate low dimensional summaries of the gene expression information, such as distance matrices or lower dimensional projections, like principal components. The second is to identify, via visualization, expert opinion or more formal tools, a manageable number of genes. The first approach has been prevalent in the statistical literature, whereas the second seems to be more common in successful contributions to the cancer literature. One problem with the first approach is that molecular profiles are defined in a way that is specific to the overall set of genes measured on the array. An arraywise molecular profile from a custom array can be different in meaning from a molecular profile from a generic commercial array.

In both the melanoma and the lymphoma analyses, molecular classification results in one or two novel classes, defined by differential expression of a group of genes with similar patterns. If the sole goal is to identify subtypes, a small subset of these genes may be sufficient for assigning tumours to classes. Clusters of functionally related genes are useful in providing a framework for the interpretation of the predictor(s) and supporting the belief that the subtype may reflect a biological mechanism. Critical to the analyses of both the lymphoma and melanoma is a combination of visualization, formal quantitative analysis and informal *a priori* information on gene function. It is clear that each of these aspects contribute to the success of the analysis, and that the use of each alone is unlikely to result in similar progress.

In this paper, our aim is to provide a framework to support this tree-faceted enterprise. We propose a probabilistic definition of differential expression in the context of unsupervised classification, and we use it to define molecular profiles, and to assess quantities of potential use in classification, such as the probability that a tumour belongs to a given profile and the probability that two tumours have the same profile. Our long-term goals are

(a) to provide tools that will facilitate the use of prior knowledge about gene function in the screening process, in an interactive way, to improve the interpretation and clinical validation of the classification that will ultimately emerge from the analysis, and
(b) to capture the potentially categorical nature of differential gene expression, by using latent categorical data that can be interpreted as a gene being turned 'on' or 'off' compared with normal expression.

These categories may offer a venue for a synthesis of information across studies which is currently made difficult by low correlation across technologies and laboratories.

2. Modelling

2.1. A statistical definition of abnormal expression

Data are a $G \times T$ expression matrix A with generic element a_{gt}, representing the measured transcript abundance of gene g in tumour t, or a transformation of interest, such as the logarithm. $\mathcal{G} = \{1, \ldots, G\}$ is the set of all genes in the experiment. We assume that raw intensity values have been normalized and purified from experimental artefacts, and that the noise is stabilized with respect to abundance.

There is no obvious notion of differential expression in the context of unsupervised molecular classification of disease. Initial progress can be made by defining differential expression using empirical evidence of the presence of subgroups in expression measures. Our approach is based on defining three possible categories of expression for each entry in the matrix A, as follows:

$$e_{gt} = \begin{cases} -1 & \text{gene } g \text{ has abnormally low expression in tumour } t, \\ 0 & \text{gene } g \text{ has normal expression in tumour } t, \\ 1 & \text{gene } g \text{ has abnormally high expression in tumour } t. \end{cases}$$

This model can be used to support

(a) a statistical definition of differential expression, via a mixture approach,
(b) a precise definition of a molecular profile, independent of the set of genes measured on the array, and
(c) probabilistic statements about the relationship of tumours to profiles and tumours to each other.

To build a model we specify, for each g,

$$a_{gt}|(e_{gt}=e) \sim f_{e,g}(\cdot), \qquad e \in \{-1,0,1\}, \quad t=1,\ldots,T.$$

For gene g, the overall proportions of differentially expressed tumours in the population of unclassified tumours are $\pi_g^- = P(e_{gt}=-1)$ and $\pi_g^+ = P(e_{gt}=1)$. These are also unknown parameters. We shall use the notation $\pi_g = \pi_g^- + \pi_g^+$. Our model specification is completed by assuming that, for fixed π_g^+s, π_g^-s and fs, the e_{gt}s are independent across genes and tumours, and that, conditionally on e_{gt}s, the a_{gt}s are independent.

This model describes variation of expression across tumours. In the unsupervised case, the density $f_{0,g}(\cdot)$ describes the variation of expression for gene g in tumours that represent the modal expression for the cancer population of interest. Variation is attributable to both small biological differences between tumours of the same subtype, and imperfect measurement of abundance in the hybridization experiment. The term 'normal' in the definition of $e_{gt}=0$ refers to expression levels within the particular cancer population that is considered for molecular classification. Expression within this group may differ from typical levels in normal tissue. In some experiments, data on normal tissue or on known subtypes of cancer may be available. In these cases, expression levels in the external group could be used to identify the $f_{0,g}(\cdot)$ component. Densities $f_{-1,g}(\cdot)$ and $f_{1,g}(\cdot)$ capture the variation of expression for gene g in tumours that display underexpression or overexpression compared with the norm defined by $f_{0,g}(\cdot)$; the supports of $f_{-1,g}$ and $f_{1,g}$ are assumed to be mutually exclusive.

A key aspect of our approach is the conversion of abundance measurements into probabilities of differential expression categories. These probabilities can offer an effective way to stabilize the measurements, by eliminating a large portion of the noise and of the hard-to-cluster variation at the extremes. At the same time, they provide an interpretable scale for classification of tumours to patterns. For each data point, the probabilities of differential expression are known functions of the mixture model parameters, determined by using Bayes's rule. Specifically, we have

$$p_{gt}^+ = P(e_{gt}=1|a_{gt},\pi_g^+,\pi_g^-,f_{1,g},f_{0,g}) = \frac{\pi_g^+ f_{1,g}(a_{gt})}{\pi_g^+ f_{1,g}(a_{gt}) + (1-\pi_g^+-\pi_g^-)\, f_{0,g}(a_{gt})} \tag{1}$$

for a_{gt} and in the support of $f_{1,g}$ and $p_{gt}^+=0$ otherwise. Similarly

$$p_{gt}^- = P(e_{gt}=-1|a_{gt},\pi_g^+,\pi_g^-,f_{-1,g},f_{0,g}) = \frac{\pi_g^- f_{-1,g}(a_{gt})}{\pi_g^- f_{-1,g}\,(a_{gt}) + (1-\pi_g^+-\pi_g^-)\, f_{0,g}(a_{gt})} \tag{2}$$

for a_{gt} in the support of $f_{-1,g}$ and $p_{gt}^-=0$ otherwise. Zero terms have been omitted from both denominators above. We shall use the notation $p_{gt}=p_{gt}^+ + p_{gt}^-$.

Mixture modelling is not new to microarray data but has been mostly confined to describing the variation across genes, as in Lee *et al.* (2000). Multivariate clustering via mixtures is discussed by McLachlan *et al.* (2002) and Yeung *et al.* (2001).

2.2. Distributional assumptions

Approaches such as that above for defining categorical expression patterns in tumour populations can be implemented in a variety of ways, depending on parameterizations and distributions. In general, distributional assumptions can be specific to the application. In analysing cancer data, we find it useful to use the following specification:

$$\begin{aligned} f_{-1,g}(\cdot) &= \mathcal{U}(-\kappa_g^- + \alpha_t + \mu_g, \alpha_t + \mu_g), \\ f_{0,g}(\cdot) &= \mathcal{N}(\alpha_t + \mu_g, \sigma_g), \\ f_{1,g}(\cdot) &= \mathcal{U}(\alpha_t + \mu_g, \alpha_t + \mu_g + \kappa_g^+), \end{aligned}$$

where $\mathcal{U}$ is the uniform distribution and $\mathcal{N}$ is the Gaussian (normal) distribution. We shall use the shorthand ω for the full set of unknown parameters. Examples of normal and uniform mixtures for finding outliers and sparse clusters are discussed by Fraley and Raftery (1998).

The model can be thought of as having a systematic component $\alpha_t + \mu_g$, and a three-component mixture for the residuals. $\alpha_t + \mu_g$ is both the centre of the distribution of the normal abundance levels for gene g in tumour t and the dividing point between overexpression and underexpression. μ_g is the effect of gene g on mean normal abundance, whereas α_t is the effect of tumour t. α_t is a tumour-specific adjustment that is determined only by the normal expression levels. For example, in radiolabelled arrays, it is common to normalize data a_{gt} for each tumour by dividing by the total of abundance measurements for tumour t, as the overall signal varies with the level of activity of the isotope batch used. If the data are indeed a mixture of a core distribution and a set of dispersed points, and, if the frequency of dispersed points varies with the tumour, then the core distribution will be normalized differently from tumour to tumour, even though it is biologically the same. The α_ts provide a normalization that considers only the common normal component and not the differentially expressed component. See also Colantuoni *et al.* (2003) and Tseng *et al.* (2001).

The normal component has gene-specific standard deviation σ_g. The uniform distribution is motivated here by the notion that abnormally altered expression of a gene in cancer reflects the failure of a regulatory mechanism that is present in healthy tissue, as would result, for example, from a deleterious mutation of a tumour suppressor gene. Therefore, abnormally high or low expression levels could vary over a broad range without necessarily informing on a biological subtype. In addition, the uniform distribution generally provides stable estimates of expression status probabilities, because no expression values are assigned a very low density in the differentially expressed cases. The parameters κ_g^+ and κ_g^- provide the limits of the uniform components of the mixture. Mixtures of normal and uniform distributions can lead to heavier or lighter tails than the normal, depending on the relationship between the κs and σ. Here we are interested only in inflating the tails. For this, we impose the constraint $\kappa > \kappa_0\sigma$, with $\kappa_0 > 5$ in applications.

The choice of a centre distribution can also be important. Our choice of a normal distribution works satisfactorily in the motivating application of Section 4, but it needs to be checked case by case. A gene-dependent choice of the functional form in problems with a relatively large number of tumours could result in improved classification ability.

This model can be equivalently expressed by introducing unknown quantitative expression values η_{gt} and defining $a_{gt} \sim N(\eta_{gt}, \sigma_g)$. The normal class is then defined by $\eta_{gt} = \mu_g + \alpha_t$, and σ_g unknown, the overexpressed class is $\eta_{gt} - \mu_g - \alpha_t \sim \mathcal{U}(0, \kappa_g^+)$, $\sigma_g = 0$, and the underexpressed class is $\eta_{gt} - \mu_g - \alpha_t \sim \mathcal{U}(\kappa_g^-, 0)$, $\sigma_g = 0$. The posterior means of the η_{gt} can be used to obtain multiple-shrinkage estimates of expression values. In particular we have

$$E(\eta_{gt}|a_{gt}, \omega) \approx \mu_g + \alpha_t + p_{gt}(a_{gt} - \mu_g - \alpha_t);$$

when the variability in the normal component is predominantly driven by noise, point estimates of posterior expectations of η_{gt} can provide denoised expression measurements. See also George (1986).

An alternative modelling strategy is to replace the uniform distributions with highly dispersed normal distributions, by setting $f_{-1,g}(\cdot) = \mathcal{N}^-(\alpha_t+\mu_g, \kappa_g)$ and $f_{1,g}(\cdot) = \mathcal{N}^+(\alpha_t+\mu_g, \kappa_g)$ where $\mathcal{N}^+$ and $\mathcal{N}^-$ are half-normal distributions on the positive and negative real line respectively. Similar ideas have been used in Bayesian variable selection by George and McCulloch (1993), who introduced two-element scale mixtures of normal distributions as a way of modelling latent variables that represent 'practical significance' of coefficients in a regression model.

2.3. Bayesian hierarchical analysis

Parameter estimation can be carried out by using a variety of approaches. If fast computing is critical, heuristic quantization rules could be devised to approximate the expression classes directly on the basis of the data, independently for each gene. A more systematic alternative is to use a maximum likelihood approach. Estimates can be obtained by using a Newton–Raphson algorithm. At each step, the κ_gs can be profiled out in closed form, which leads to a faster and more stable maximum likelihood algorithm than do common alternative outer distributions.

Here we have implemented a hierarchical model assuming that the tumour-to-tumour variation is described by the mixture distribution of Section 2.2 whereas, at a second level, the variation of gene-specific parameters is described by further probability distributions. This allows for an estimation of gene-specific parameters that borrows strength from the entire genomic distribution and reflects the fact that the components of variation that are driven by limitations in the technology are likely to affect the majority of genes in a similar way.

Bayesian hierarchical models lead to shrinkage estimates with good properties in the estimation of large vectors of related parameters. See Berger (1985). Newton *et al.* (2001) have considered hierarchical models for genes within a single two-dye hybridization. These have desirable denoising properties in the estimation of relative expression levels. Hierarchical models across multiple hybridizations are considered by Tseng *et al.* (2001). In our context, even greater gains are to be expected from borrowing strength across both genes and tumour types, as most genes will behave similarly in all tumours. Newton *et al.* (2001) also introduced Bayesian probabilities of differential expression in microarray analyses, again in the context of single-slide comparisons. Bayesian hierarchical models share structural assumptions with empirical Bayes analyses, reviewed by Carlin and Louis (2000), and illustrated in the context of microarrays by Efron *et al.* (2001). Related ideas underlie mixed effects models as discussed by Wolfinger *et al.* (2001).

The implementation that is considered here is based on the following second-stage distributions:

$$\begin{aligned}
\mu_g|\theta_\mu, \tau_\mu &\sim \mathcal{N}(\theta_\mu, \tau_\mu),\\
\sigma_g^{-2}|\gamma, \lambda &\sim \mathcal{G}(\gamma, \lambda),\\
\kappa_g^+|\theta_\kappa^+ &\sim \mathcal{E}(\theta_\kappa^+),\\
\kappa_g^-|\theta_\kappa^- &\sim \mathcal{E}(\theta_\kappa^-),\\
\text{logit}(\pi_g^+)|\theta_\pi^+, \tau_\pi^+ &\sim \mathcal{N}(\theta_\pi^+, \tau_\pi^+),\\
\text{logit}(\pi_g^-)|\theta_\pi^-, \tau_\pi^- &\sim \mathcal{N}(\theta_\pi^-, \tau_\pi^-),
\end{aligned}$$

where $\mathcal{G}$ is the gamma distribution and $\mathcal{E}$ is the exponential distribution. Gene-specific parameters are assumed to be independent conditionally on the hyperparameters on the right-hand sides above. The hyperparameters can be assigned dispersed proper priors, as the large number of genes allows for a data-driven estimation. One of the advantages of a hierarchical specification regards genes that show no evidence of a departure from normality. For those genes, data provide minimal information about κ_g^+ and κ_g^-. Improper priors can lead to identifiability problems. Maximum likelihood estimation of gene-specific parameters in those cases requires care and additional *ad hoc* constraints that are avoided by borrowing strength from the genomic distribution.

The α_ts are assumed independent $\mathcal{N}(0, 100)$ and constrained to sum to 0. Hierarchical modelling of the α_t is also possible and in our application leads to similar results. A proper prior on the α_ts and θ_μ is necessary, as the likelihood is uninformative about combinations in which the same constant is added to all α_ts and μ_gs.

We used a Metropolis–Hastings Markov chain Monte Carlo (MCMC) approach to obtain samples from the posterior distribution of the parameters. We augmented the data set with an unknown class indicator e_{gs} for each observation, as in Diebolt and Robert (1994) and West and Turner (1994). We use the sampling sequence $[\kappa|\omega^*]$, $[e|\kappa, \omega^*]$, $[\omega^*|\kappa, e]$, to facilitate mixing in the sampling of κs. In the expression above, the symbols refer to parameter vectors or matrices, brackets refer to posterior distributions and ω^* is ω with κ removed. The first two terms on the left-hand side combine to form $[\kappa, e|\omega^*]$. Given model parameters, the full conditional distribution of the class indicators e is given by expressions (1) and (2). Given the class indicators, the full conditional distribution of $(\pi_g^+, \pi_g^-, 1 - \pi_g^+ - \pi_g^-)$ is a Dirichlet distribution, and the full conditional of the parameters of the normal component is conjugate, with the additional constraint $\sigma < \min(\kappa_g^+, \kappa_g^-)/\kappa_0$.

3. Molecular profiles

The expression variables e can be used to construct a useful definition of a molecular profile. For a given set of G genes, we define a molecular profile to be a vector of es, i.e. a point in $\{-1, 0, 1\}^G$. A given tumour can belong to only one of these profiles. In practice, the number of profiles defined in this way on the set of all genes on a microarray is far too large to allow reliable inference based on the data sets that are typically available. Our interest is therefore in what we term marginal profiles, i.e. profiles based on a subset of genes. For example, for $G = 3$ we obtain 27 possible marginal profiles, as follows:

	gene A	gene B	gene C
profile 1	−1	−1	−1
profile 2	−1	−1	0
⋮	⋮	⋮	⋮
profile 27	1	1	1

Marginal profiles are coarse classifications, susceptible to further refinements at later times, when additional evidence becomes available. There is no implication that other genes may not display differential expression. This definition has a direct biological interpretation, is independent of the array used (as long as all G genes are measured) and is independent of the classification algorithm that is used to assign tumours to profiles.

The mixture analysis provides a way to evaluate membership probabilities for each tumour. Using conditional independence of the genes, we can determine the probability that each of the tumours belongs to any given profile, as a function of the ps. Specifically,

$$p(e_{1t}, \ldots, e_{Gt}|\omega) = \prod_g (p_{gt}^-)^{I\{e_{gt}=-1\}} (p_{gt}^+)^{I\{e_{gt}=1\}} (1 - p_{gt}^+ - p_{gt}^-)^{I\{e_{gt}=0\}},$$

where I is the binary indicator of the condition in parentheses. Some tumours may be easily classifiable, whereas for others the probabilities may spread over several competing profiles. It is also straightforward to compute the probability that any two tumours have the same profile over a set of genes $\mathcal{G}_0$. Because e_{gt} and $e_{gt'}$ are conditionally independent given the model parameters, the conditional probability that two tumours t and t' have the same expression status on all genes in $\mathcal{G}_0$ is

$$q(t, t', \mathcal{G}_0) = \prod_{g \in \mathcal{G}_0} \{p_{gt}^- p_{gt'}^- + p_{gt}^+ p_{gt'}^+ + (1 - p_{gt}^+ - p_{gt}^-)(1 - p_{gt'}^+ - p_{gt'}^-)\}. \tag{3}$$

Analogous computations apply to gene patterns across tumours. A special role is played by the profile of all zeros, indicating that a gene is expressed at a normal level in all tumours, or is varying by a degree that is not differentiable from noise. The probability of this pattern for gene g is $p_g^0 = \Pi_t\,(1 - p_{gt})$. The expected number of tumours in which gene g is at a normal level is $n_g = \Sigma_t\,(1 - p_{gt})$. Estimates of both these quantities can be used to exclude from analysis genes with low discriminatory ability; n_g is less sensitive than p_g^0 to the presence of very small probabilities, which may be unstable.

Gene patterns across tumours can be used to mine for genes that show promise as subtype predictors. Measures of variability of the overexpression and underexpression probabilities are examples of useful gene-specific summaries. In general, summaries will have to strike a balance between the fraction of tumours that show evidence of differential expression and the reliability with which each tumour can be attributed to each class. In Section 4 we shall use a class of summaries that allows us to control for the first component, by mining for genes matching patterns of the type 'n^- tumours are underexpressed and n^+ are overexpressed', irrespective of the order of the tumours. We define $E(n^-, n^+)$ to be the set of all such patterns, i.e. the set of all T-dimensional vectors taking values in $\{-1, 0, 1\}$ and such that n^- entries are -1 and n^+ entries are 1. The corresponding probability is

$$r_{n^-,n^+}(g) = p\{e_{g1}, \ldots, e_{gT} \in E(n^-, n^+)|\omega\}; \tag{4}$$

r depends on the target frequencies n^- and n^+ and also reflects uncertainty in the expression status of genes: as values of p_{gt}^+ and p_{gt}^- move away from the extremes, r decreases irrespective of the target pattern.

Lastly, we can express the probability that two genes g and g' have the same pattern over all tumours as

$$q(g, g') = \prod_t \{p_{gt}^- p_{g't}^- + p_{gt}^+ p_{g't}^+ + (1 - p_{gt}^+ - p_{gt}^-)(1 - p_{g't}^+ - p_{g't}^-)\}. \tag{5}$$

Although $q(g, g')$ is not a distance, it does provide an interpretable measure of closeness for describing the gene space. The closeness $q(g, g)$ of a gene to itself (i.e. the probability that it would have the same true profile as another gene with an identical observed expression) defines a measure of internal consistency.

4. Molecular analysis of ductal breast cancer

4.1. Data and preprocessing

The data set that is analysed in this section includes measurements of gene expression in 80 ductal carcinomas of the breast, also known as carcinoma of no specific type. We use a custom breast cancer array developed in the Gabrielson Laboratory at Johns Hopkins University, to classify these tumours according to gene expression profiles. We have selected frozen samples of ductal breast cancers and isolated tumour cells from the samples by using a rapid mechanical microdissection technique. Total ribonucleic acid (RNA) was labelled and hybridized to the array membranes, and hybridization to specific spots was quantitatively measured by using a phosphoimager. The reproducibility of these measurements is high, as shown by Johnston *et al.* (2002). Our collection of tumours includes primary breast tumours (including tumours of low and high pathological grades) and metastatic breast cancers, chosen to represent highly aggressive cancers. In the selection of specimens for these studies, the histology was reviewed to attempt to represent ductal cancers with different histologic features. Some bias against small cancers with tubular features may be inevitable because these samples may not yield sufficient RNA for analysis.

Radiolabelling arrays produce a single intensity measurement per spot. For initial graphical diagnostics we constructed a reference by computing the median intensities by gene over all tumours. Scatterplots of intensity *versus* this reference indicated a subset of genes with a pronounced, roughly linear, loss of signal on a subset of the tumours. One of these tumours is shown in Fig. 1(a). To address this issue, we used a latent class mixture model postulating two classes of genes: class 0 for genes with no loss of signal and class 1 for genes with loss of signal. After class assignment, we carried out separate normalizations. The resulting data could be analysed jointly over the two classes but, because the reasons leading to the loss of signal have yet to be established, here we only present analyses using genes that are likely to be in class 0, a total of 2897 genes.

In general, a linear loss of signal on a relatively large portion of the genes is not uncommon in microarray experiments with radiolabelling filters. It may result from tissue admixture or contamination, or from array manufacturing problems. When the size of the class 1 subset is large, standard normalization procedures based on robust regression, as in Yang *et al.* (2002) and Irizarry *et al.* (2001), would not be appropriate. In view of these difficulties, the resulting arrays are sometimes discarded. A latent class model provides a systematic way of treating these cases.

More specifically, let y_{gt} be the natural logarithm of intensity measures for gene g in tumour t, m_g be the median of log-intensity for gene g across all tumours and c_g the binary class indicator. We specified a model based on two separate regression equations for the two classes, i.e.

$$y_{gt} = \beta_{t1} + \beta_{t2}c_g + \beta_{t3}m_g + \beta_{t4}c_g m_g + \varepsilon_{ts},$$

with ε_{gt} normally distributed with mean 0 and variance $\zeta_{gt}^2 = 1/\rho_t^2(1+\delta c_g)$, specified so that it is tumour specific and that $\zeta_{gt}^2/\zeta_{g't}^2 = 1$ if genes g and g' are in the same group and $\zeta_{gt}^2/\zeta_{g't}^2 = 1+\delta$ otherwise. We define $p(c_g = 1) = \xi$ and set dispersed priors $\rho_t^{-2} \sim \mathcal{G}(0.001, 0.001)$, $\xi \sim \mathcal{U}(0, 1)$ and $1+\delta \sim \mathcal{U}(0.2, 5)$, so the ratio of variances of the two groups ranges between $1/5$ and 5, and values less than 1 are favoured. The regression equation above is used for class assignment, but not for subsequent abundance correction.

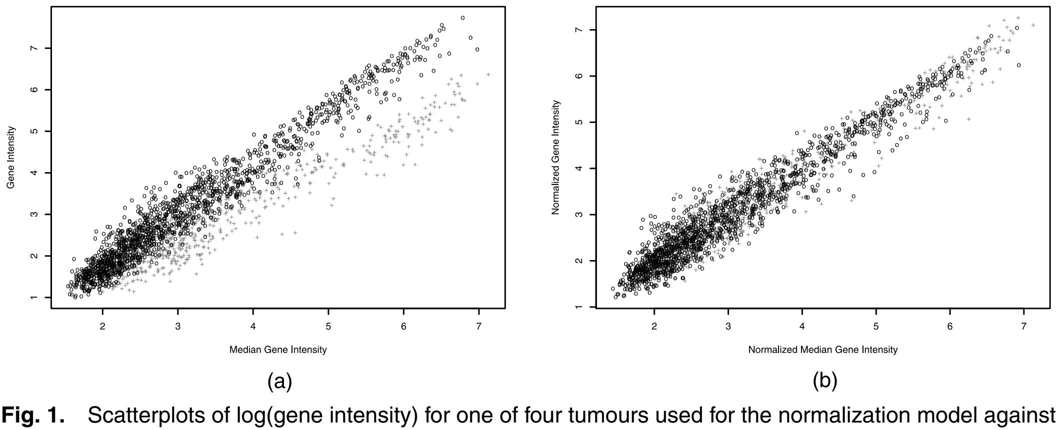

Fig. 1. Scatterplots of log(gene intensity) for one of four tumours used for the normalization model against the corresponding median log(gene intensity) (medians are taken over all the samples; ○, genes assigned to class 0; +, genes assigned to class 1): (a) raw y_{gt} *versus* m_g; (b) normalized $a_{gt} + m_g$ *versus* m_g

We fitted the normalization model based on four tumours showing strong evidence of a linear loss of signal. Our implementation used MCMC sampling and the software BUGS, by Thomas *et al.* (1992). We obtained posterior probabilities of class membership for each gene and used those for class assignments. To correct for a minor low intensity bias remaining in some filters, we performed a separate abundance adjustment for each of the two subclasses. We applied a LOESS fit to $y_{gt} - m_g$ *versus* m_g, to obtain residuals r_{gt}. To facilitate the fit of a hierarchical distribution of gene-specific variances we performed a variance stabilization, fitting a second LOESS curve on the squared residuals *versus* m_g, as proposed by Colantuoni *et al.* (2003). The square roots of the residuals from the second LOESS fit, s_{gt}, were then divided by the median value across tumours, to preserve the scale. The variance stabilization term from the second LOESS fit was $s^*_{gt} = s_{gt}/\text{median}(s_{g1}, \ldots, s_{gT})$. Finally we evaluated the adjusted expression measures $a_{gt} = m_g + r_{gt}/s^*_{gt}$. Scatterplots of a_{gt} *versus* m_g are shown in Fig. 2.

4.2. Estimation

As the next step we fitted the model of Section 2.3 to the a_{gt}, using an R implementation of the MCMC algorithm. We used vague priors $\theta^+_\mu \sim \mathcal{N}(0, 100)$, $\theta^-_\mu \sim \mathcal{N}(0, 100)$, $(\tau^+_\mu)^{-2} \sim \mathcal{G}(1, 0.1)$, $(\tau^-_\mu)^{-2} \sim \mathcal{G}(1, 0.1)$, $\theta^+_\kappa \sim \mathcal{N}(0, 100)$, $\theta_\pi \sim \mathcal{N}(0, 100)$, $(\tau^+_\pi)^{-2} \sim \mathcal{G}(1, 0.1)$ and

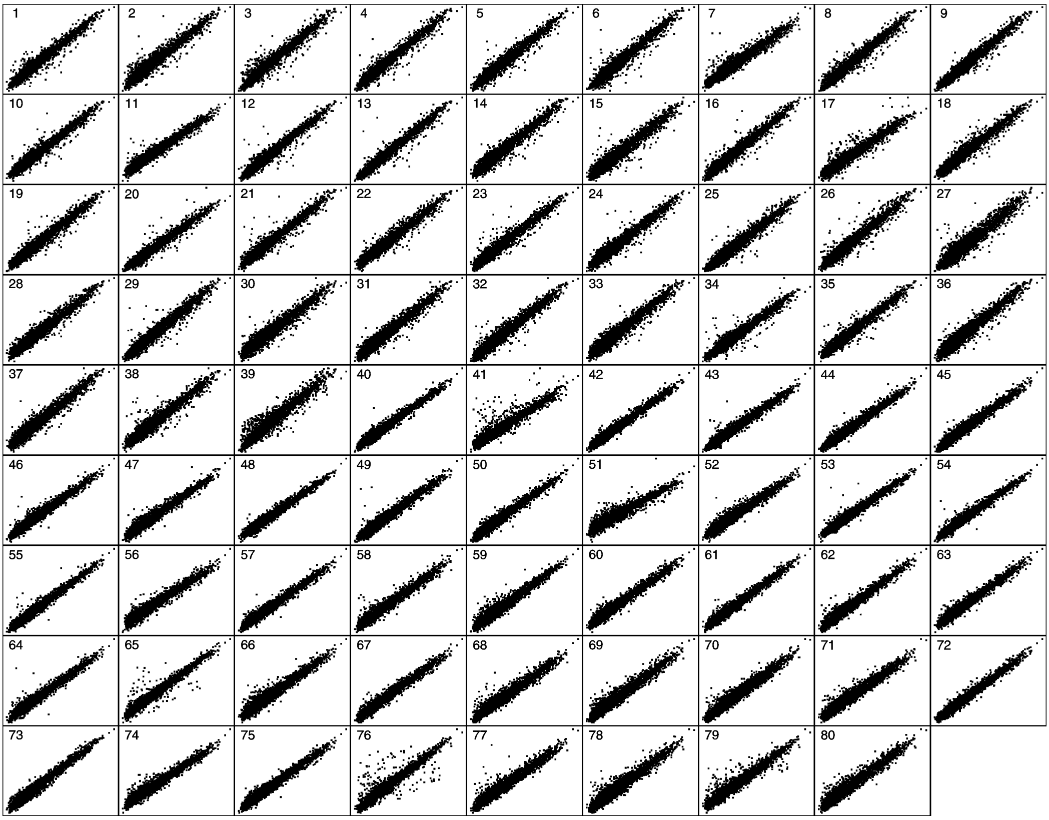

Fig. 2. Scatterplots of normalized expression measurements $a_{gt} + m_g$ *versus* medians m_g for the 80 tumours studied, and for the genes selected for further analysis, after normalization: genes correspond to points and tumours to scatterplots: points away from the line $x = y$ indicate that gene expression differs from the median of that gene over all samples

$(\tau_\pi^-)^{-2} \sim \mathcal{G}(1, 0.1)$, $\lambda \sim$ flat and $\gamma \sim$ flat. The sensitivity to the specific vague prior used is very limited in our application. The convergence of the hyperparameters is fast. The convergence of gene-specific parameters varies with the gene, with κs being the slowest mixing parameters. The MCMC output provides point estimates and assessments of uncertainty on any of the quantities of interest. Hats will denote point estimates based on the chain. The spot-specific $\hat{p}_{gt}$ generally converge more rapidly than the mixture weights and uniform limits. A careful choice of the initial values substantially improves the speed of convergence. We found it useful to perform preliminary gene-by-gene 2-means or 3-means clustering to estimate class assignments and to initialize the chain at the corresponding parameter estimates.

Fig. 3 illustrates estimated mixture components for four genes, selected to illustrate a range of situations. The fit of the normal component is evaluated by the *qq*-plots. The dark section corresponds to the estimated normal component; linearity of the dark section indicates a good fit and is common to the majority of genes in this experiment. A good fit of the inner normal component is an important element of our approach: an inner distribution that is skewed, but smoothly so, could make both estimation and interpretation problematic.

Summary measures of non-linearity of the residuals of the regression of the normal *qq*-plot, using $1 - \hat{p}_{gt}$ as the weights, can provide a mining tool for selecting genes for further visualization and diagnostics. Another useful mining tool is the estimated gene self-consistency $q(g, g)$. Although it is common for the inner distribution to fit well, it is less common that the differentially expressed values will be uniformly distributed. This lack of fit of the uniform distribution is not critical, as it typically coexists with a good estimate of the probability of differential expression. Gene 1753 in Fig. 3 highlights a potentially problematic situation for our model. There are either more than three groups or a departure from normality in the $e = 0$ group. Also, because groups have similar sizes, the choice of the class labelled $e = 0$ is sensitive to the initialization of the chain.

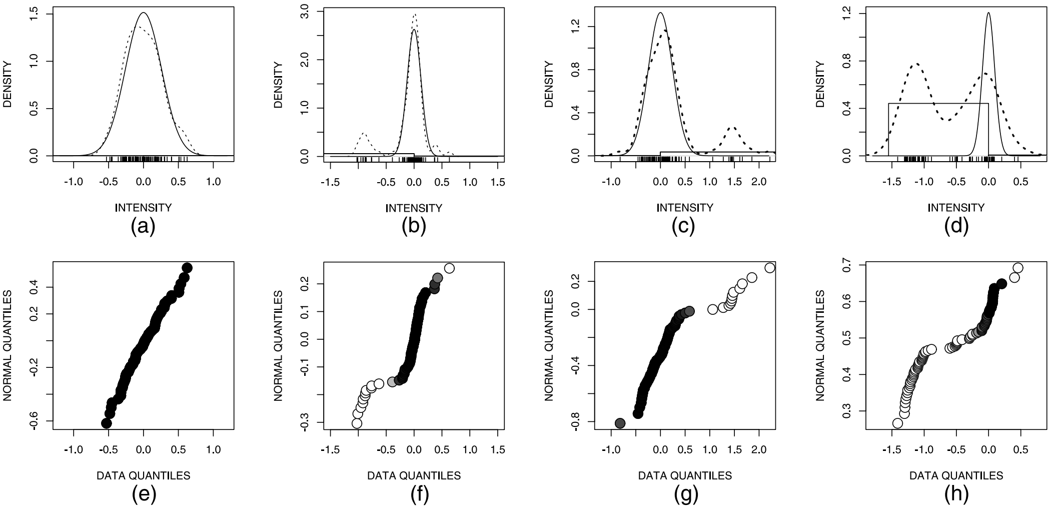

Fig. 3. Estimated mixture components for genes (a) 20, (b) 2818, (c) 2634 and (d) 1753 (the vertical marks are the estimated residuals $a_{gt} - \mu_g - \alpha_t$; $\cdots\cdots$, kernel density estimate of the distribution of the residuals; ——, best fitting uniform and normal components of the mixture, multiplied by the corresponding mixture weights (some of the uniform components are too close to the axis to appear)) and (e)–(h) corresponding normal quantile plots, with shades of grey proportional to the probability $1 - \hat{p}_{gt}$ of being from the normal distribution

4.3. Visualization of profile information: genome-wide approaches

A useful step towards molecular classification is identifying a subset of candidate genes for subsequent validation. The results of the mixture analysis can contribute to this process, by providing the basis for visualization and mining tools. For example, because of our stipulation that p_{gt}^{+} and p_{gt}^{-} cannot both be positive, we can represent expression probabilities in a single dimension, the estimated signed differential expression probabilities $\hat{p}_{gt}^{+} - \hat{p}_{gt}^{-}$. This ranges from -1 to 1 and can be used for imaging in grey scales or with different colour codes for positive and negative differential expression. To sort genes and tumours by using hierarchical clustering, all traditional similarity metrics can be applied directly to $\hat{p}_{gt}^{+} - \hat{p}_{gt}^{-}$. In addition, measures with an interpretation that is more directly related to the problem at hand are available. We use the probability that two tumours have the same pattern, calculated for the whole gene set, as defined by q of equation (3). Alternatively, we can build a less stringent criterion by computing the probability that two patterns are the same up to a given number of samples. Visualizations based on probabilities are suitable for comparisons across microarray technologies, especially when using the now prevalent oligonucleotide arrays, whose ability to measure underlying transcript abundance varies markedly with both the gene and the manufacturer.

Fig. 4 compares a visualization of probabilities with a visualization of observed expression. Only genes with $\Sigma_t \, p_{gt} > 2$ are shown. In Fig. 4(a) we show the centred data, interpretable as log-ratios of expression to the gene median. Rows and columns are sorted by using the divisive hierarchical algorithm `diana`, in R (Rousseeuw *et al.*, 1996) with Euclidean distance. In Fig. 4(b) are the signed probabilities of differential expression. Rows and columns are sorted again by using `diana`, with similarity q. This example illustrates the denoising that takes place as a result of mixture modelling and suggests that clustering using the probability of differential expression may be less sensitive to noise-driven artefacts than clustering in the original scale.

In this application, there is a large subset of genes with bimodal distributions as can be seen in Fig. 4. This can result from large subclasses of cancers, but also from systematic variation at the analytical level, without an underlying biological explanation. Confirmatory laboratory work is in progress to clarify this issue. In the presence of significant bimodality, it is possible that tumour classes that are labelled, say, 0 and -1 should have been labelled 1 and 0, as determining the modal class in the data when two classes are of similar sizes is difficult. Empirically, it is possible that the normal class is not the best represented. An incorrect identification of the normal class in pronouncedly bimodal genes is not necessarily problematic for the molecular classifications of the samples, but it makes an interpretation of covariation patterns of the genes involved less reliable.

4.4. An iterative approach examining small subsets of candidate genes

A complementary tool to the visualization of clustered genomic profiles of Fig. 4 is the visualization of a selected subset of candidate genes, in concert with other information such as functional annotations. Unlike clustering, this can be an iterative process in which several subsets are successively identified and examined, on the basis of various mining criteria. We illustrate one iteration in Fig. 5, which shows a collection of genes with high potential for discrimination. Genes may be grouped. Each group is formed by iteratively choosing a seed gene, and then identifying a set of genes that are similar to it for display. The goals of displaying groups are to reduce redundancy in the set of seed genes, to provide additional context to facilitate interpretations of subtypes, and to make it less likely to select isolated artefacts. It is not critical that the gene groups contain all the genes that are actually involved in a co-regulated pattern. Although such a result would be of great biological interest, it is often too difficult a task for the

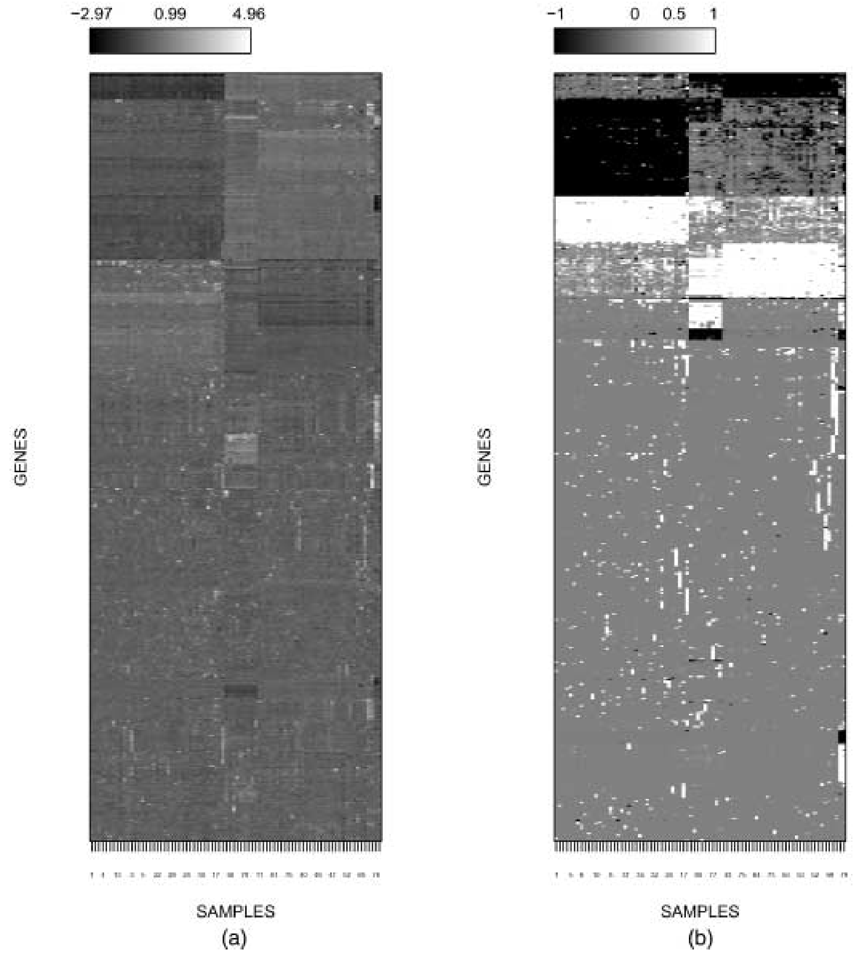

Fig. 4. Alternative visualization strategies: (a) centred log-expressions and (b) probabilities of differential expression (the row and column orderings differ)

type of data sets at hand. Likewise, it is not critical that the gene sets be disjoint. This approach requires a measure of discriminatory power for ranking seed genes, a measure of similarity of gene patterns for forming groups and preliminary filters to exclude unlikely candidates. Many alternatives are available, and the choice between them can affect the results.

In our illustration we apply preliminary filters based on a minimum $\Sigma_t p_{gt}$ of 10 and a minimum internal consistency score $q(g, g)$ of 60/80. We measure discriminatory power with the probability r of matching a target pattern, given in equation (4). The pattern used in Fig. 5 is $n^- = n^+ = 25$, i.e. genes are ranked on the basis of how similar they are to a hypothetical gene that is overexpressed in 25 tumours and underexpressed in 25 other tumours, in any order. Very low values of n^- and n^+ may lead to mining genes whose pattern is the result of noise or other artefacts and are unlikely to be useful. Very large values, with no normal expression, may result from an improper fit of the mixture model. Within these extremes, a wide range of target frequencies could lead to useful genes. For example, if it were known that a fraction of about 20% of tumours had early local recurrence, we may mine using $n^- + n^+ = 0.2T$, even in the absence of matched phenotype information on the tumours. The display of Fig. 5 is somewhat

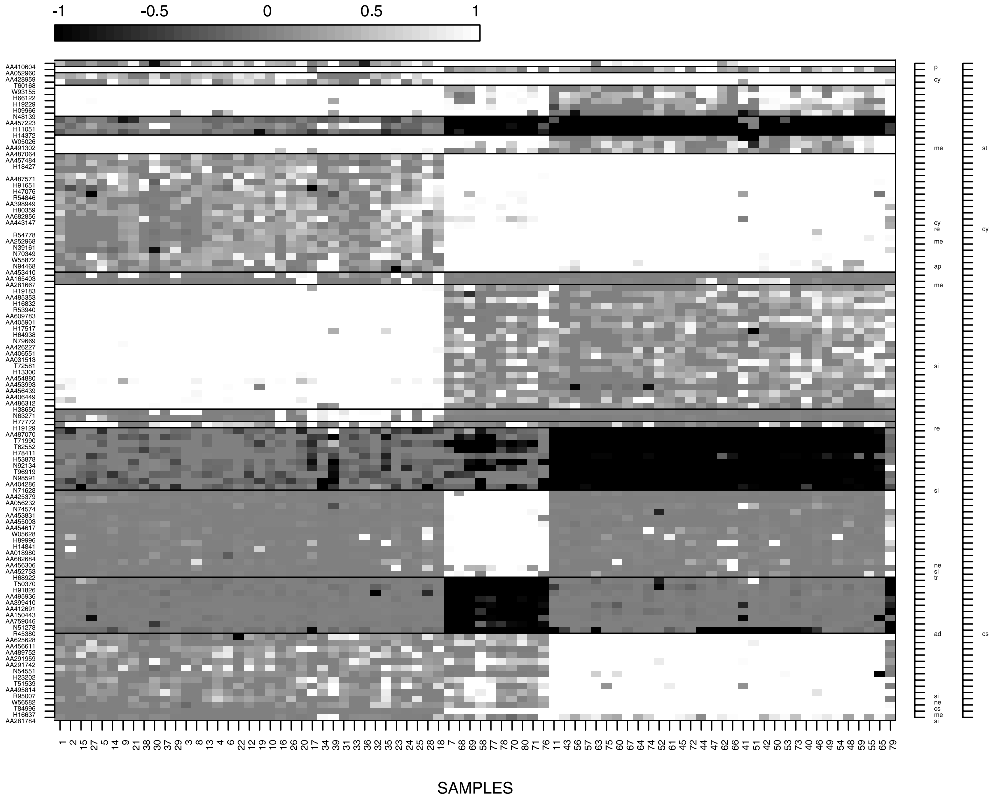

Fig. 5. Visualization of gene expression for the elicitation of classification genes: each column corresponds to a tumour, with all 80 tumours represented, and ordered via the same hierarchical clustering as in Fig. 4; each row corresponds to a gene, with genes identified by genbank accession numbers; grey scales represent differential expression probabilities; horizontal sections corresponding to gene groups are separated by black lines; at the right-hand side, for each gene, is listed information about the family it belongs to, with each gene potentially belonging to more than one family

sensitive to the choice of n^- and n^+. In this application, as n^- and n^+ are decreased below 10, genes with large overexpressed or underexpressed subsets of samples are no longer selected. The order in which seed genes are ranked is sensitive to the choice of n^- and n^+ but the overall order of groups is less so. As the computation of r is time consuming, we approximate it by the probability that the vector $e_{g1}, \ldots, e_{gT}$ is equal to the pattern in the set $E(n^-, n^+)$ with the highest probability of being correct. This can be a poor approximation to the probability r, as it is only the largest term in a sum of many terms, but it is likely to preserve the ranking of potential seed genes.

We measure similarity by the probability q, given in equation (5), that a gene has the same pattern as the seed. In Fig. 5, we set the threshold for adding genes to a group at 60/80, imposing that genes in a group differ from the seed in no more than 20 tumours in total. The sensitivity to this threshold is naturally high, as one can modulate it to create groups that go from being empty to having the size of the entire genome. Here we tuned the threshold starting from high values, i.e. few genes. The goal here is not so much to identify clusters of co-regulated genes but to select enough co-regulated genes to provide a context for candidate predictors, and alternatives to expressed sequence tags in defining marginal molecular profiles.

Table 1. Abbreviations for gene families

ad	adhesion	in	invasion
an	angiogenesis	me	metabolism
ap	apoptosis	ne	neuroendocrine
cc	cell cycle	re	receptors
ch	checkpoint, mitosis, G2-modifiers	rn	RNA processing and ribosomes
cy	cytokines	si	signal transduction
cs	cytoskeletal and secretory	st	stress
da	DNA damage and repair	tr	transcription factors
dn	DNA binding and modification		

At the right-hand side of Fig. 5, for each gene, we list information about the family that it belongs to, where families are defined on the basis of the biological function of the gene (the gene family keys are given in Table 1). Each gene may belong to more than one family, with up to two listed here. More elaborate versions of this graph may use additional gene-specific information to provide a richer context, e.g. by using genomic search tools such as DRAGON (Bouton and Pevsner, 2000). The list of genbank accession numbers on the left-hand side of Fig. 5 can be uploaded to DRAGON to generate a Web page with links to database and literature information about each of the genes. No obvious functional similarity within groups seems to emerge in this case. Some of the rows correspond to expressed sequence tags of yet unknown function, whence the lack of functional class information.

4.5. *Visualization of profile probabilities*

The analysis of Sections 4.3 and 4.4, as well as additional validation with confirmatory assays, may lead to the identification of genes for molecular classification. The results of the mixture model can then be useful in representing the assignment of tumours to subtypes and the associated uncertainty. To illustrate, we select one gene from three of the groups in Fig. 5, focusing on N94468 (jun B proto-oncogene) from the ninth group from the bottom, AA486312 (cyclin-dependent kinase 4) from the seventh group from the bottom and AA453831 (hepatoma-

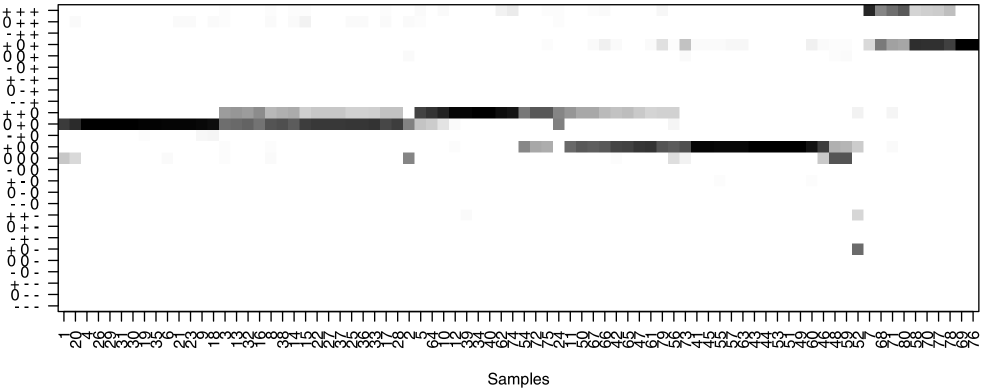

Fig. 6. Molecular profiles probabilities: each row corresponds to one of the 27 molecular profiles defined by the expression status of genes N94468, AA486312 and AA453831; each column corresponds to a tumour—for example, the point for row (1,−1,0) for tumour 79 is the probability that the true expression indicators for tumour 79 are (1,−1,0) with regard to the genes in question

derived growth factor) from the third group from the bottom. All three have a probability near 0 of being affected by a loss of signal in the analysis of Section 4.1.

Fig. 6 shows the molecular profile probabilities, with tumours sorted by using hierarchical clustering. Three large subclasses, (1,1,0), (0,1,0) and (1,0,0), emerge, as well as two smaller ones: (1,0,1) and (1,1,1). There is uncertainty about the classifications, especially for tumours that are likely to be in class (0,1,0), many of which also have a significant probability of belonging to class (1,1,0). This display is complementary to genome-wide colour maps. An alternative approach to that presented here is to use genome-wide maps to select genes and then to use our mixture analysis and this graph to represent the ensuing assignment probabilities.

The uncertainty represented here concerns the assignment of tumours to marginal profiles conditionally on the model parameters. Additionally, there is uncertainty about the model parameters themselves, and uncertainty in the selection of the marginal profile. The latter uncertainty is complex to quantify because the selection process involves several steps, not all of which are easily quantified. Parameter uncertainty, however, can be quantified by using the results from the MCMC output. Each of the spots in Fig. 6 is a function of unknown parameters and for each we can derive a posterior probability distribution.

5. Gene interactions

Complex interactions between expression levels of several genes are likely to be present in cancer data. This can be the result of carcinogenic pathways, some of which are well understood, but many of which may still be unknown. Naturally, an exploration of these interactions is of interest biologically and eventually clinically. The approach described here can be extended in a direct way to the exploration of gene interactions. Of special concern is the case in which a cancer subtype is not distinguishable on the basis of a marginal inspection of each gene separately but may become apparent when multiple genes are considered simultaneously.

One way of thinking about this is to work with subsets of genes, rather than individual genes. Let s be a set of genes, i.e. a subset of $\mathcal{G}$, and let $\mathcal{S}$ be the set of all subsets under consideration. Typically, data sets will only permit a consideration of sets of moderate size, even though genetic pathways may involve a large number of genes. For illustration, we consider subsets of two genes. We can again work with a statistical definition of differential expression, constructed via a mixture approach, by specifying, for each set s comprising two genes g and g', a joint distribution of a_{gt} and $a_{g't}$

$$a_{gt}, a_{g't} | (e_{gt} = e, e_{g't} = e') \sim f_{e,e'}(\cdot)$$

with (e, e') taking values in the set $\{(1,1), (-1,1), (1,-1), (-1,-1), (0,0)\}$, with probabilities $\pi_g^{++}, \pi_g^{-+}, \pi_g^{+-}, \pi_g^{--}, 1 - \pi_g^{++} - \pi_g^{-+} - \pi_g^{+-} - \pi_g^{--}$. The cases $\{(0,1), (0,-1), (1,0), (-1,0)\}$, have been omitted to reflect the fact that the pathway is either activated, in which case both genes are altered, or not, in which case both genes are normal.

A natural extension of the model of Section 2 to this case is to choose a mixture of a bivariate normal distribution for the (0, 0) component, and four mutually exclusive bivariate uniform distributions for the differentially expressed components. Specifically,

$$f_{0,0}(\cdot) = \mathcal{N}_2 \left\{ \begin{pmatrix} \alpha_t + \mu_g \\ \alpha_t + \mu_{g'} \end{pmatrix}, \begin{pmatrix} \sigma_g^2 & 0 \\ 0 & \sigma_{g'}^2 \end{pmatrix} \right\},$$

$$f_{-1,1}(\cdot) = \mathcal{U}_2 \left\{ \begin{pmatrix} -\kappa_g^- + \alpha_t + \mu_g \\ \alpha_t + \mu_{g'} \end{pmatrix}, \begin{pmatrix} \alpha_t + \mu_g \\ \kappa_{g'}^+ + \alpha_t + \mu_{g'} \end{pmatrix} \right\},$$

where $\mathcal{N}_2$ is a bivariate normal distribution and $\mathcal{U}_2$ is a bivariate uniform distribution. Densities for other differentially expressed cases can be defined similarly. The one-dimensional marginal densities for g and g' correspond to the sets including only g and only g', as desired. Gene selection and visualization can now proceed from the ss rather than the gs.

6. Discussion

We have proposed a strategy for the analysis of gene expression in unclassified tumours, a setting in which there is no natural reference, and often no reliable phenotypic information. We introduced a statistical definition of differential expression based on latent classes, developed a probabilistic definition of the molecular profile in this context and proposed a statistical implementation based on mixture modelling. We discussed generating model summaries that have simple probabilistic interpretations and using them for visualization, gene clustering and tumour classification.

A critical assumption is that the important aspect of the variation of gene expression across tumours can be captured sufficiently well by a three-way categorical variable. Taken literally, this is unlikely to be true. Yet, at this preliminary phase of genomic research, it may be more efficient to focus on gross features, rather than attempting to detect the results of subtle changes in expression that may be obscured by natural biological variation and noise in measurements. Our categorization may play a constructive role in enabling a combination of results across microarray technologies that may have different gene-specific sensitivities and/or non-linearities, and contribute to the critical step of synthesizing information on genomic information.

The strategy outlined here is likely to work well when the normal class is relatively large. When there is evidence of multiple large subgroups, the identification of which class is the normal one is both more arbitrary and more sensitive to the initialization of the MCMC estimation algorithm. In some experiments, mixture modelling can be informed by additional normal tissue or by existing known subclasses, both of which could be used to help to identify which of the latent classes is defined to be normal.

We considered molecular classification based on gene expression information only. Additional support for the presence of subclasses can be provided by a variety of additional molecular and more traditional measurements, including information on the outcomes for patients or their response to therapy. The presence of this additional information adds complexity but also increases the likelihood of a successful classification.

Acknowledgements

Supplementary information and software are available at `http://astor.som.jhmi.edu/poe`. The work was supported by the National Institutes of Health under the Johns Hopkins ‘Specialized program of research excellence in breast cancer’ grant P50CA88843 and grant 5P30CA06973-39. We thank Peter Müller for a useful suggestion.

References

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

Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999) Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–297.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N. and Trent, J. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Bouton, C. M. and Pevsner, J. (2000) DRAGON: database referencing of array genes online. *Bioinformatics*, **16**, 1038–1039.

Carlin, B. P. and Louis, T. A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton: Chapman and Hall.

Colantuoni, C., Henry, G., Bouton, C. M. L. S., Zeger, S. L. and Pevsner, J. (2003) SNOMAD: biologist-friendly web tools for the Standardization and Normalization of MicroArray Data. In *The Analysis of Gene Expression Data: Methods and Software* (eds G. Parmigiani, E. S. Garrett, R. A. Irizarry and S. L. Zeger). New York: Springer. To be published.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc.* B, **56**, 363–375.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001) *Pattern Classification*. New York: Wiley.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Ass.*, **96**, 1151–1160.

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

Fraley, C. and Raftery, A. E. (1998) How many clusters?: which clustering method?—answers via model-based cluster analysis. *Comput. J.*, **41**, 578–588.

George, E. I. (1986) Minimax multiple shrinkage estimation. *Ann. Statist.*, **14**, 188–205.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

Hartigan, J. A. (1975) *Clustering Algorithms*. New York: Wiley.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D. and Brown, P. (2000) "Gene Shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, research0003.1–research0003.21.

Herrero, J., Valencia, A. and Dopazo, J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.

Irizarry, R. A., Parmigiani, G., Guo, M., Dracheva, T. and Jen, J. (2001) A statistical analysis of radiolabeled gene expression data. In *Computing Science and Statistics: Proc. 33rd Symp. Interface*. Fairfax: Interface Foundation of North America.

Johnston, J., Parmigiani, G., Gabrielson, E. and Anbazhagan, R. (2001) Gene expression profiling of formalin-fixed paraffin embedded tissues using cDNA microarrays. *Technical Report 02-01*. Johns Hopkins Oncology Center, Baltimore.

Kruskal, J. B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrica*, **29**, 1–27.

Lee, M. L., Kuo, F. C., Whitmore, G. A. and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natn. Acad. Sci. USA*, **97**, 9834–9839.

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

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. and Tsui, K. W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., Lashkari, D., Shalon, D., Brown, P. O. and Botstein, D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natn. Acad. Sci. USA*, **96**, 9212–9217.

Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.*, **2**, 418–427.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rousseeuw, P., Struyf, A. and Hubert, M. (1996) Clustering in an object-oriented environment. *J. Statist. Sftwr.*, **1**, 1–30.

Schulze, A. and Downward, J. (2001) Navigating gene expression using microarrays—a technology review. *Nat. Cell Biol.*, **3**, E190–E195.

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

Tavassoli, F. A. (1992) *Pathology of the Breast*. Norwalk: Appleton and Lange.

Thomas, A., Spiegelhalter, D. J. and Gilks, W. R. (1992) BUGS: a program to perform Bayesian inference using Gibbs sampling. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 837–842. Oxford: Clarendon.

Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. and Wong, W. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucl. Acids Res.*, **29**, 2549–2557.

West, M. and Turner, D. A. (1994) Deconvolution of mixtures in analysis of neural synaptic transmission. *Statistician*, **43**, 31–43.

Wolfinger, R. D., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**, 625–637.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucl. Acids Res.*, **30**, e15.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.