BIOINFORMATICS

Vol. 18 no. 12 2002
Pages 1600–1608

Between-group analysis of microarray data

Aedín C. Culhane[1,*], Guy Perrière[2], Elizabeth C. Considine[1], Thomas G. Cotter[1] and Desmond G. Higgins[1]

[1]Department of Biochemistry, University College Cork, Cork, Ireland and
[2]Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS No. 5558, Université Claude Bernard—Lyon 1, 43, blvd. du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

Received on March 12, 2002; revised on May 20, 2002; accepted on May 24, 2002

ABSTRACT

Motivation: Most supervised classification methods are limited by the requirement for more cases than variables. In microarray data the number of variables (genes) far exceeds the number of cases (arrays), and thus filtering and pre-selection of genes is required. We describe the application of Between Group Analysis (BGA) to the analysis of microarray data. A feature of BGA is that it can be used when the number of variables (genes) exceeds the number of cases (arrays). BGA is based on carrying out an ordination of groups of samples, using a standard method such as Correspondence Analysis (COA), rather than an ordination of the individual microarray samples. As such, it can be viewed as a method of carrying out COA with grouped data.

Results: We illustrate the power of the method using two cancer data sets. In both cases, we can quickly and accurately classify test samples from any number of specified *a priori* groups and identify the genes which characterize these groups. We obtained very high rates of correct classification, as determined by jack-knife or validation experiments with training and test sets. The results are comparable to those from other methods in terms of accuracy but the power and flexibility of BGA make it an especially attractive method for the analysis of microarray cancer data.

Availability: The methods described are implemented in ADE-4 which runs under MacOS and Windows, and is freely available at http://pbil.univ-lyon1.fr/ADE-4/. All scripts are available on request.

Contact: A.Culhane@ucc.ie

Supplementary information: Supplementary figures and tables are available at http://bioinfo.ucc.ie/BGA/.

INTRODUCTION

Several class prediction approaches have been applied to the analysis of microarray data (reviewed by Sherlock, 2000). These rely on the researcher specifying groupings in advance where the aim is to discover combinations of genes which can be used to classify new unknown samples. Discriminant functions (Dudoit *et al.*, 2000), artificial neural networks (Khan *et al.*, 2001), Bayesian classifiers and support vector machines (SVM; Furey *et al.*, 2000) have been applied to the problem of classifying tumour samples. These methods have been shown to be successful when applied to the problem of classification but suffer from a problem of dimensionality. For example, to properly use conventional discriminant function analysis, one must have more cases than variables, ideally by a factor of 10 or more. With microarray data sets, we usually have exactly the reverse with data sets having many thousands of variables (genes) and only a few tens of cases (microarray samples). This problem is most commonly circumvented by selecting subsets of genes in advance or iteratively during training. Typically these subsets use about 50 genes although, it has been suggested that just a few genes might be statistically preferable (Li and Yang, 2001). Such gene selections may be cumbersome to produce, possibly involving arbitrary selection criterion or may miss highly informative combinations of genes.

In this paper, we wish to describe the application of a powerful yet simple and flexible method called between-group analysis (BGA) a multiple discriminant approach that can be safely used with any combinations of numbers of genes and samples (Dolédec and Chessel, 1987). It is used in the framework of a conventional ordination technique such as COA or principal component analysis (PCA) and as such, allows for great flexibility with regard to the assumptions made in carrying out the analysis. When combined with COA it is especially powerful as it allows us to examine the correspondences between the grouped samples and those genes which most facilitate the discrimination of these groupings (Fellenberg *et al.*, 2001).

The basis of BGA is to ordinate the groups rather than

[*]To whom correspondence should be addressed

the individual samples. For N groups we find $N - 1$ eigenvectors or axes that arrange the groups so as to maximise the between group variances. The individual samples are then plotted along them. Each eigenvector can be used as a discriminator to separate one of the groups from the rest. New samples are then placed on the same axes and can be classified on an axis-by-axis basis or by proximity to the group centroids.

When applied to tumour data, BGA is fast and simple to use yet produces accurate discrimination as judged by the performance on test data or by a jack-knife analysis. Despite this simplicity, the results allow a detailed and simultaneous analysis of the entire set of genes. We are able to quickly and simply identify potential marker genes for all tumour types, including clinically important genes that were missed in other analyses or which were only found using combinations of other techniques. We can specify any groupings we wish, ask for the axes (and genes) that most discriminate these, and test the reliability and accuracy of these discriminations. Thus BGA can also be used as an exploratory technique to examine potential heterogeneity with groups.

SYSTEMS AND METHODS

Gene expression datasets

The data set from Golub *et al.* (1999) contained 72 samples from two types of acute leukaemia: 47 acute lymphoblastic leukaemia (ALL) and 25 acute myeloid leukaemia (AML). Samples were obtained from patient bone marrow or peripheral blood and the gene expression patterns analyzed on Affymetrix oligonucleotide arrays containing 6817 genes (7159 probe sets). The data set is available from http://www.genome.wi.mit.edu/cancer.

The second data set reported gene expression profiles of four types of small round blue cell tumours of childhood (SRBCT) published by Khan *et al.* (2001). They used cDNA microarrays containing 6567 clones of which 3789 were known genes and 2778 were ESTs to study the expression of genes in neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt lymphoma (BL, a subset of non-Hodgkin lymphoma), and the Ewing family of tumours (EWS). The data set contained both tumour biopsy and cell line samples. A filtered data set containing gene expression profiles of 2308 genes in these samples is available from http://research.nhgri.nih.gov/microarray/Supplement/.

Software

BGA with COA or PCA of microarray data were computed using ADE-4 (Thioulouse *et al.*, 1997), a general purpose package for multivariate analysis, which has been used widely in the analysis of environmental and ecological data. It runs under MacOS 7 and Windows operating systems and can be downloaded from http://pbil.univ-lyon1.fr/ADE-4/. The ADE-4 modules required to perform BGA using COA or PCA are ADE-trans, FilesUtil (Transpose), CategVar (Read Categ File), PCA (Correlation Matrix PCA), COA (Correspondence Analysis), Discrimin (Initialize: Link Prep, Between Analysis: Test, Between Analysis: Run). ADE-4 can be run interactively or in batch mode. We wrote scripts in Python v2.1 to automate some analyses by calling the ADE-4 modules directly e.g. for the jack-knife analysis. Graphs were plotted using the ADE-4 modules Graph1D, Scatters and Scatterclass.

Mathematical basis of between-group analysis

BGA is carried out by ordinating groups (sets of grouped microarray samples) and then projecting the individual sample locations on the resulting axes. This is most easily done using PCA or COA. In this description, we will first describe COA and then show how we carry out the BGA on microarray data. We follow Fellenberg *et al.* (2001) and Perriere *et al.* (1996) in our notation and also the extensive documentation in the ADE-4 package (Thioulouse *et al.*, 1997). Figures to aid interpreration of this description are available on the supplementary web page.

Consider a raw data table (**N**) of gene expression data for I genes (rows) and J microarray samples (columns) with elements n_{ij}. With microarray data, the rows and columns of **N** are usually normalized in various ways. For COA we must ensure that all elements of **N** are non-negative (usually integers), by adding a constant to all values if required. We denote the row sums and column sums of **N** as n_{i+} and n_{+j} respectively. The grand total of all the elements of **N** is denoted n_{++}. The relative contribution or weight of gene i to the total variation in the data set is then denoted r_i and is calculated as

$$r_i = n_{i+}/n_{++} \tag{1}$$

while the relative contribution of sample j is denoted as c_j and is calculated as

$$c_j = n_{+j}/n_{++} \tag{2}$$

Similarly, the contribution of each individual element of **N** to the total variation in the data set is denoted as p_{ij} and is calculated as

$$p_{ij} = n_{ij}/n_{++} \tag{3}$$

This produces two vectors **R** and **C** of length I and J and one $I \times J$ matrix. We convert these into an $I \times J$ table of χ^2 values **X** using

$$x_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j} \tag{4}$$

It is this table X that is analyzed to produce the correspondence analysis. This table shows the associations between genes and samples. The total association between all genes and all samples is given by the total χ^2 value for the data set (x_{++}) which is the grand total of all the elements of **X**. COA then consists of decomposing this total χ^2 into components for each gene and each sample along each of K eigenvectors where K is $\min(I-1, J-1)$. These eigenvectors are ranked according to their eigenvalues. The total of all the eigenvalues equals the total χ^2 value for the data set. The actual method used in ADE-4 to derive the eigenvectors is general singular value decomposition (Dolédec and Chessel, 1987) where we calculate matrix **B** below as:

$$\mathbf{B} = \mathbf{D}_c^{1/2}\mathbf{X}\mathbf{D}_r\mathbf{X}^t\mathbf{D}_c^{1/2} \tag{5}$$

Here, $\mathbf{D}_c^{1/2}$ is a $J \times J$ matrix with the square roots of the elements of C along the diagonal and zeros elsewhere. Similarly D_r is an $I \times I$ matrix derived from vector R with the elements of R along the diagonal and zeros elsewhere. Finally **B** is a $J \times J$ matrix which is diagonalized to produce J eigenvalues (at least one of which will be zero) and eigenvectors.

The results of a COA are viewed by plotting the co-ordinates of all genes and samples along the top 2 or 3 eigenvectors. Groupings of samples or trends in the data set can be seen and interpreted using the proximity of genes and samples in plots as a guide. Samples and genes which are strongly associated, as measured by their χ^2 values, will lie in a similar direction from the origin.

BGA is carried out where we can specify G groups of samples in advance. The purpose of the analysis is then to ordinate these groups so as to separate them maximally in some space. This is achieved by grouping the J columns (samples) and calculating the vector **C** of column weights with G elements where each element is the sum of column weights for one group i.e. $\mathbf{C}_g$ is the sum of the sample weights for group g. Matrix $\mathbf{D}_c$ then has $G \times G$ elements and the COA is carried out as before using Equation (5) above. The result of this is to produce G-1 eigenvectors with the co-ordinates of all genes and of the group centroids. Finally, all individual samples are plotted on the G-1 eigenvectors as supplemental points.

BGA can also be carried out using PCA. We used the default PCA method in ADE-4 where the raw data set is normalized row by row (gene by gene) such that each row has zero mean and unit standard deviation. By default, using ADE-4 for PCA, the column weight vector **C** is set to give uniform weights that total to 1 (i.e. all c_j are set to $1/J$) and similarly the row weight vector **R** elements are set to $1/I$. BGA with PCA can then be carried out exactly as before using the normalized data matrix (X) and the row and column weight vectors **R** and **C**. These are subjected to a generalized singular value decomposition.

If there are only 2 groups, then the result of the analysis will be a single vector with the positions of all samples and all genes. We can plot new samples as supplemental points and classify them according to which group they are nearest to. We use the following formula for the weighted mean to calculate a threshold where $\bar{X}_1$ is the mean for group 1 and $\bar{X}_2$ is the mean for group 2, and SD_1 and SD_2 are the group standard deviations:

$$\frac{\bar{X}_1 SD_2 + \bar{X}_2 SD_1}{SD_1 + SD_2} \tag{6}$$

This is a simple method, which will assign all new samples as belonging to one group depending on whether the co-ordinate falls above or below the threshold.

If there are more than 2 groups, then there will be G-1 axes. We can then treat each axis as a discriminator of one group where each axis serves to separate one group from the rest. Thus an individual sample is classified to a group if its co-ordinate along any axis, falls on the correct side of the threshold for that axis. If it fails to fall on the correct side of any of the thresholds, it will belong to the remaining group. When classifying new samples we simply first assign each sample to the nearest group centroid where proximity is measured by Euclidean distance in the space described by the G-1 axes. We then reclassify each sample if it can be allocated to any one group by falling on the correct side of a threshold on any of the axes.

Validation and accuracy assessment

We tested the method by measuring the percentage of test samples that could be correctly classified. We also used a simple jack-knife procedure which involves removing a single sample from the data set and carrying out an entire BGA analysis of the remaining samples. The removed sample is then classified and the success or failure of the classification is recorded. This is repeated for all samples and the result shown as the percentage of samples that can be correctly classified.

RESULTS

Classification of acute leukaemia samples

Using sample information from Golub *et al.* (1999) training samples were categorized as AML or ALL and were subjected to BGA using PCA and COA. The result of a BGA using two groups is a single axis or discriminator separating the two groups. The axis discriminating ALL and AML samples is shown Figure 1a. The individual microarray training samples are then plotted on the same axis and this gives a visual indication of the degree to which the two groups are separable, using this discriminating axis.

We first tested the validity of the discriminator using a remove-one jack-knife test. This showed that we could correctly predict group membership for 35 of the 38

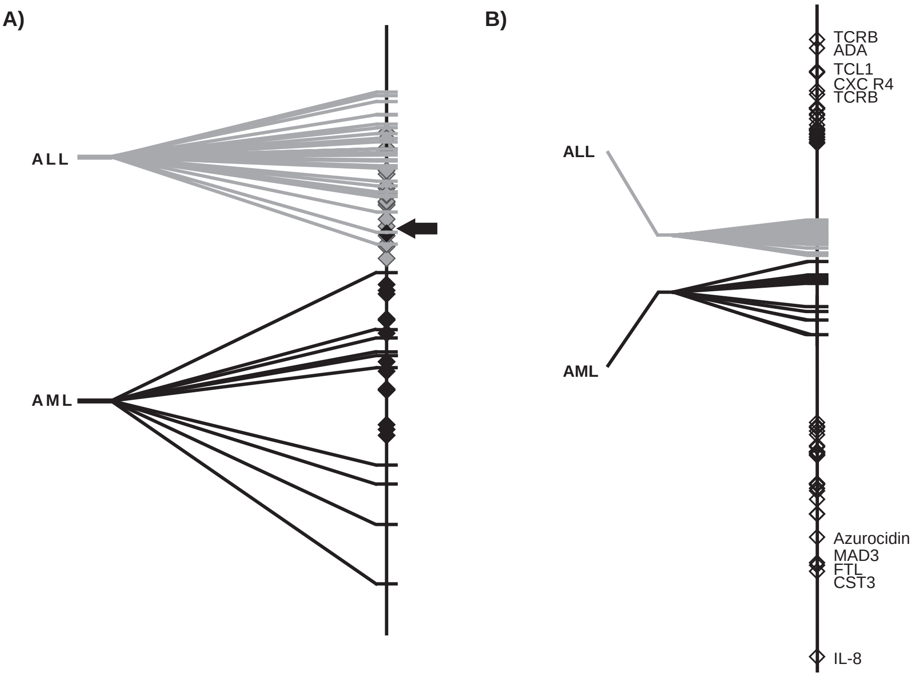

Fig. 1. Discrimination between acute myeloid and acute lymphoblastic leukaemias. The results of a BGA using COA on the leukaemia data set of Golub *et al.* (1999) are shown. (a) The single axis of the analysis with all 38 training and 34 test samples plotted. The analysis begins with the calculation of the co-ordinates of the two group centres (ALL and AML). Then the 38 training samples (indicated by grey (ALL) and black (AML) slanted lines) are plotted accordingly. Finally, the 34 test samples are projected on this axis, indicated by grey (ALL) and black (AML) filled diamonds. A threshold is calculated (represented by a black arrow), between the two training groups and this is used to distinguish and classify the test samples into the two groups. (b) The same analysis as in (a) but with the positions of the 25 most extreme genes from either end of the axis indicated by unfilled diamonds. The most extreme 10 of these are labelled.

training samples (92%) using BGA with either COA or PCA. It should be noted that this is a global analysis where we have used the entire data set with no selection of genes or adjustment.

Secondly, the 34 supplementary blind test samples from Golub *et al.* (1999) were projected onto the discriminating axis and each sample was assigned to AML or ALL groups (Figure 1a). Using BGA with PCA 82.4% of samples were correctly classified. But when BGA with COA was used all AML cases (15/15) and 16/20 ALL cases were correctly assigned (88.2%). In general BGA using COA tended to out perform PCA ordination. Since COA has the further advantage that the genes and samples can be projected along the same axes, we only show results for BGA with COA in the rest of this paper.

Identification of discriminating genes and potential molecular markers

After a BGA, the samples are separated along axes and if we use COA we also get the co-ordinate of every gene. The genes most responsible for separating the groups are located at the ends of the axes. The 25 genes with the most extreme co-ordinates for each group are displayed in Figure 1b. ALL predictor genes included lymphoid specific genes, the T-cell-specific tyrosine kinase p56 lck, and the oncogene TCL1. Genes that distinguished AML included genes that were specific to myeloid lineage cells, a number of antioxidant enzymes and the AML immunological markers myeloperoxidase, lysozyme and ferritin. A full table of discriminating genes is available in the **Supplementary information**.

Exploring subclasses in the data

Using BGA, any groupings can be specified, and the axes (and genes) that most discriminate these can be examined. Thus we sought to determine whether ALL T and B cell lineage samples could be distinguished using BGA. Training samples consisting of 8 T cell and 19 B cell ALL samples were subjected to BGA using COA. In jackknife tests 26/27 samples were correctly assigned to B or T cell ALL classes. Only one B cell ALL was incorrectly predicted. This BGA analysis was tested by projecting 20 supplementary test ALL samples (19 B cell and 1 T cell ALL) onto the discriminating axis. All supplementary

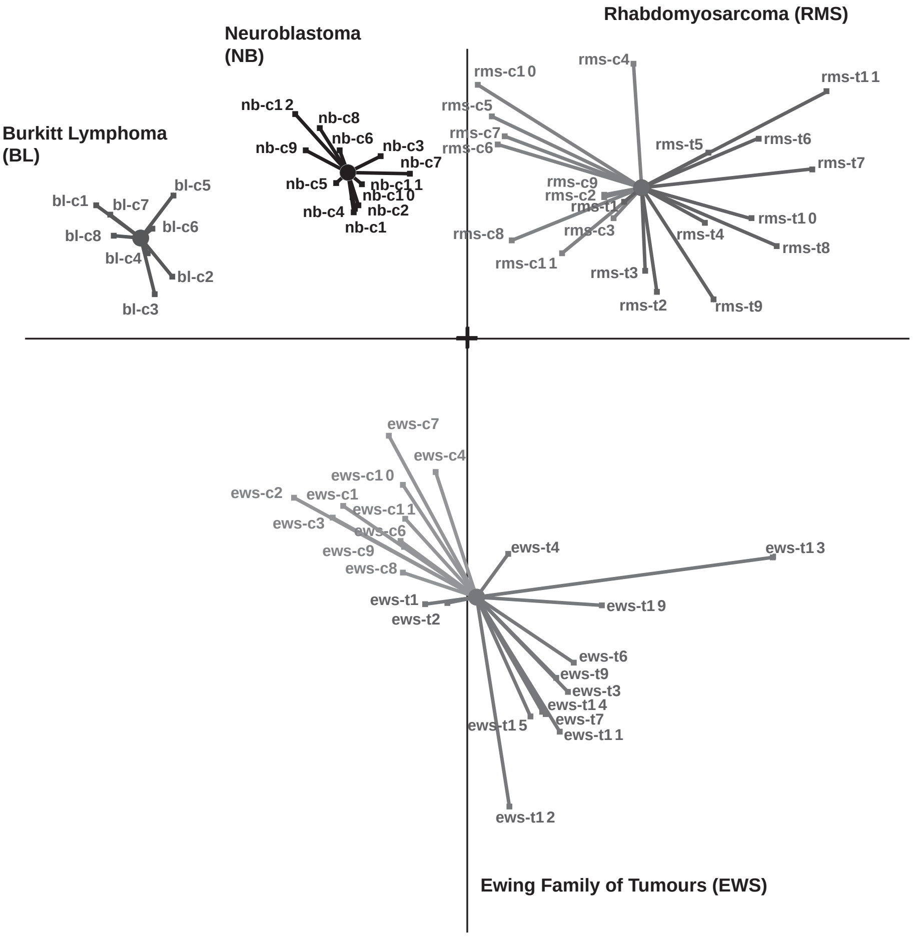

Fig. 2. Discrimination of four subtypes of SRBCTs. The first two axes of a BGA using COA of the data from Khan *et al.* (2001) are shown. The four groups are: neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt lymphoma (BL, a subset of non-Hodgkin lymphoma), and the Ewing family of tumours (EWS). The four tumours were discriminated by three axes, and this graph shows axes 1 and 2. Axis 1 (horizontal) can be used as a discriminator to distinguish BL from the other three while axis two (vertical) discriminates EWS from the rest. Filled circles and squares represent group centroids and samples respectively. A line connects each training sample to the centre of its group. Training sample labels –T (e.g. rms-t5) and –C (e.g. rms-c3) represent tissue biopsy and cell line samples respectively. Cell line samples of RMS and EWS are highlighted in paler shades of grey.

B cell ALL cases (19/19) were correctly assigned, but sample 67 the only T cell ALL sample from peripheral blood was not predicted, resulting in a test accuracy of 95%. The genes discriminating T and B cell ALL are given in the **Supplementary information**.

Discrimination of small round blue cell tumours of childhood

BGA can be used with any number of groups. We illustrate this using four groups of tumours from the data set of Khan *et al.* (2001). The training set was subjected to BGA with COA, and the results are shown in Figure 2. In this case, we have 4 groups and 3 discriminating axes. Each axis serves to separate one of the 4 groups from the rest. Axis 1 discriminates BL from the rest; axis 2 serves to separate EWS from the rest and axis 3 can be used to discriminate NB from the rest of the groups. 3D graphs of these results are available on the supplementary web site.

Class prediction of test samples. We tested the discrimination using a set of 20 test samples from EWS, RMS, and NB tumours and EWS, NB and BL cell lines. We also included two control samples from normal muscle tissue and three from unrelated cancer cell lines. Samples were projected onto the discriminating axes of the training data and the class of unknown samples were assigned as described in the methods. The co-ordinates of the projected blind samples can be seen in Figure 3.

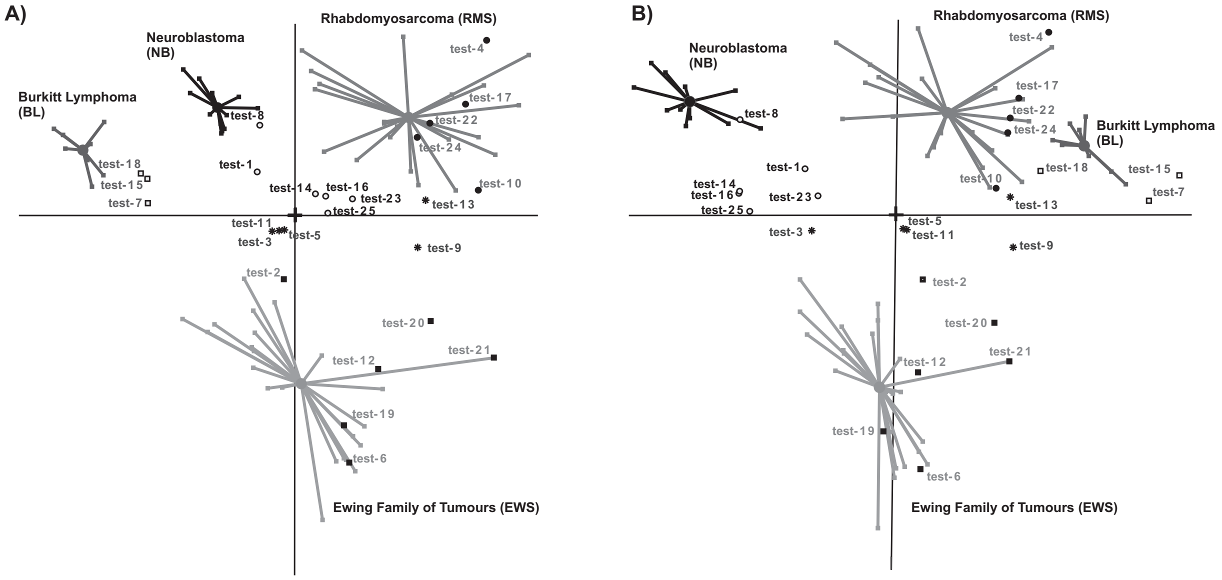

Fig. 3. Projection of supplementary blind samples of SRBCTs. The same analysis as in Figure 2 but with the projected locations of the 25 test samples, labelled test-1 to test-25 where (a) shows the first two axes (1,2) and (b) shows axes 3 and 2. Training sample labels are omitted for clarity. The test samples are assigned to one of the four groups, based on proximity to the group centroids and position in relation to each discriminating axis. The 25 test samples are labelled and the symbols represent the group to which they should be assigned: BL (open squares), NB (open circles), RMS (closed circles) and EWS (closed squares). The control samples, which do not belong to one of the 4 groups, are represented by asterisk symbols.

Almost all EWS, BL, NB and RMS test samples were projected closest to their respective clusters and 19/20 samples were correctly predicted. All NB training samples were cell line samples but NB test data contained both cell line and biopsy samples. Figure 3b shows that the test cell line samples were projected closest to the NB training cluster and the NB biopsy samples were further from the cluster. One NB test biopsy sample was not classified. The two normal skeletal muscle samples clustered closest to the RMS cluster and the three unrelated cancer cell lines clustered in the centre of the figures as they could not be assigned clearly to any of the training classes.

Heterogeneity in the training samples. In Figures 2 and 3, a clear separation of cell line versus biopsy samples can be seen within EWS and RMS classes. Continuous passaging, potential contamination or the matrix upon which cells are cultured may influence gene expression *in vitro*. Equally biopsy samples frequently contain other cell populations due to components such as surrounding normal tissue, stroma, vasculature or immune elements. Thus we investigated whether these sample populations could be differentiated.

EWS and RMS samples were analyzed independently. EWS and RMS training samples were each subjected to BGA with COA, and the supplementary test samples were projected onto the discriminating axes. All EWS cell and tissue samples, and all RMS tissue samples in the test data set were correctly predicted (100% accuracy). Discriminating genes included collagens, matrix metalloproteinases and other cell matrix associated proteins, which were expressed in tissue samples, but were expressed at a low level or not detected in cell line samples.

Identification of discriminating genes and potential molecular markers. After a BGA, the samples are separated along axes. The genes that are most responsible for separating the groups are those with the highest (or lowest) co-ordinates along these axes. In the Khan data set, we have 4 groups and 3 axes; each axis serves to distinguish one group from the other three (Figures 2 and 3). Low scores (co-ordinates) on axes one, two and three were associated with genes that were relatively highly expressed in BL, EWS and NB respectively. High scores on both axes one and two were associated with genes upregulated in RMS (Figure 4).

The list of discriminating genes identified by BGA was very similar to those reported by Khan *et al.* (2001). On axes one, two and three, 17/20, 15/20 and 17/20 of the most discriminating genes were among the top 96 genes previously reported by Khan *et al.* (2001). Moreover the rank of the top 10 genes distinguishing EWS predicted by Khan *et al.* (2001) exactly matched those identified by BGA. A full table of predictor genes is given in the

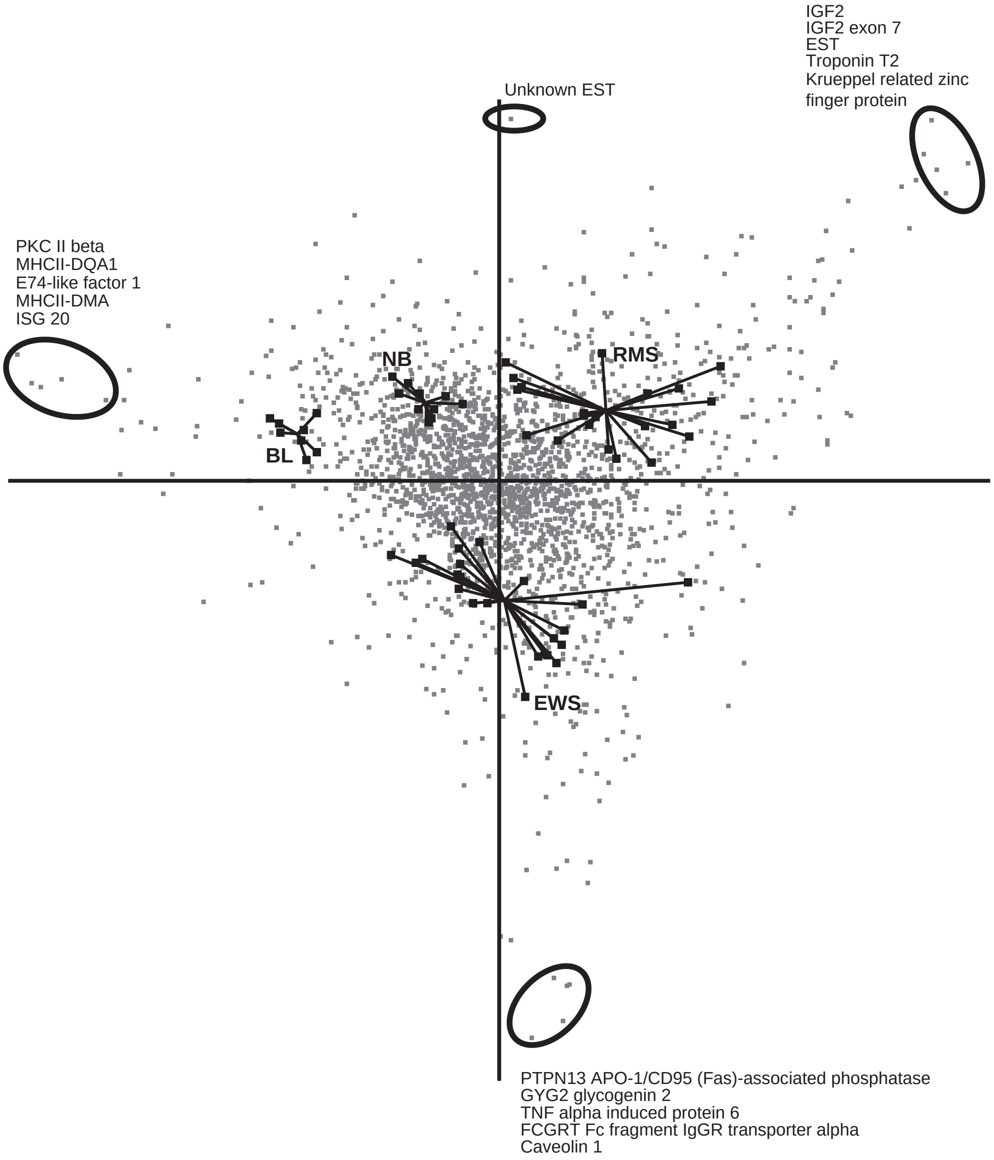

Fig. 4. Gene predictors of 4 subtypes of SRBCTs. The same analysis as in Figures 2 and 3 but with the locations of all genes indicated by small grey squares. The training data microarray samples are shown as black lines connecting the position of each sample (black square) to its group centre. Only the names of the most extreme genes from the ends of the axes (axes 1 and 2) are indicated and highlighted with rings. Highly negative scores on axis 1 and 2 correlated with highly expressed genes in BL and EWS respectively. Positive scores on both axis 1 and 2, were associated with genes that were highly expressed in RMS.

Supplementary information.

DISCUSSION

BGA was used to discriminate between AML and ALL classes in the data set from Golub *et al.* (1999). Golub *et al.* (1999) reported 36/38 samples were correctly called in jack-knife tests. BGA is comparable as 35/38 samples were correctly assigned using the full data set, and between 34 and 38 samples were predicted if genes were ranked and filtered to remove noisy data (data available on supplementary web page). Thus BGA is robust and performs comparably to other studies without the absolute need for gene filtering.

Although BGA with PCA ordination tended to out perform COA in jack-knife tests, BGA assignment of supplementary test samples was more accurate using the latter. Between 30/34 and 33/34 of supplementary test samples were correctly assigned depending on whether the genes were filtered to reduce background noise. BGA outperforms or performs with similar effectiveness to other approaches. For example, using a weighted voting scheme, 29/34 blind test samples were predicted correctly (Golub *et al.*, 1999) and SVM methods assigned between 30 and 32 test samples correctly (Furey *et al.*, 2000).

We observed that where samples were consistently predicted in all BGA analyses, the samples also had high prediction strength (PS) scores in the Golub study (Golub *et al.*, 1999). For example, Golub *et al.* (1999) failed to classify training sample 12 (an ALL B-cell sample from bone marrow) and test sample 67 (the only peripheral blood ALL T cell sample in the data set) as these samples had the lowest PS scores in tests. A further study by Furey *et al.* (2000) was unable to classify sample 67. Equally these samples were misclassified in almost all BGA/COA analyses.

One major advantage of BGA is the ease with which we can identify discriminating genes. Genes that discriminated ALL and AML included IL-8, lysozyme and adenosine deaminase, MB-1, topisomerase II, which is consistent with other studies (Golub *et al.*, 1999). However of the top 25 genes which most differentiated each class, only 6/25 ALL and 7/25 AML discriminating genes were also reported by Golub *et al.* (1999). Of these genes not reported by other studies, many were clinically significant. For example the oncogene TCL1 that has been tightly linked to the pathogenesis of mature T-cell leukaemia (Pekarsky *et al.*, 2001) and the chemokine receptor CXCR4 which is highly expressed in ALL, and is reported to play an essential role in liver, spleen, lymph nodes and brain invasion in ALL (Crazzolara *et al.*, 2001) were both identified as ALL predictor genes.

Any sub-grouping can be examined easily and rapidly using BGA. B and T cell lineage ALL samples were discriminated using BGA with COA. In jack-knife analysis, 96% of samples were correctly assigned as B or T cell. Using an unsupervised approach, Golub *et al.* (1999) were able to distinguish these subclasses in four self organizing map clusters. Furey *et al.* (2000) were able to assign all B and T cell cases using SVM if a larger training set of all training and test samples were utilized. BGA easily differentiated these two groups and MHC and T cell receptor (TCR) genes were found to discriminate B and T cell ALL samples respectively. Interestingly, the top six genes that discriminated ALL T cell samples also matched those reported by Grant *et al.* (2001).

We were also able to use BGA to discriminate the four SRBCT subclasses and 19/20 supplementary test samples were correctly classified. Khan *et al.* (2001) were able to assign all 20 test samples, of which 17 were confidently predicted using an artificial neural network (ANN) trained on 96 genes. Genes which effectively discriminated SRBCTs, as identified by BGA, correlated well with those reported by Khan *et al.* (2001). BGA was trained on the complete data set of 2308 genes and the results are comparable to the more complex ANN approach. Furthermore cell line and tissue samples could be differentiated. Thus genes specific to *in vitro* cell passaging or to heterogeneity in tissue samples, which are not significant to disease prognosis can be identified and flagged.

The main limitation of our method is how it deals with samples when they do not belong to any of the groups that were used to train the analysis. Currently, each test sample must be assigned to one of the groups. This is due to the crude class assignment method we use which allocates each sample to the nearest group centroid. This can be overcome by using a more probabilistic class assignment method that would indicate the chances of a sample belonging to each class.

BGA is clearly a useful and general purpose technique for dealing with grouped microarray samples. It can be considered as a variation of COA or PCA for dealing with grouped samples or as a multiple discriminant method. Either way, it is simple to use, fast and flexible and produces useful visual summaries of data sets as well as accurate assignments to classes.

ACKNOWLEDGEMENTS

This work was supported by grants from Enterprise Ireland and Retinitis Pigmentosa Ireland. The authors thank Sharon McKenna for useful discussions.

REFERENCES

Crazzolara,R., Kreczy,A., Mann,G., Heitger,A., Eibl,G., Fink,F.M., Mohle,R. and Meister,B. (2001) High expression of the chemokine receptor CXCR4 predicts extramedullary organ infiltration in childhood acute lymphoblastic leukaemia. *Br. J. Haematol.*, **115**, 545–553.

Dolédec,S. and Chessel,D. (1987) Rhythmes saisonniers et composantes stationelles en milieu aquatique I—Description d'un plan d'observations complet par projection de variables. *Acta Oecologica Oecologica Generalis*, **8**, 403–426.

Dudoit,S., Fridlyand,J. and Speed,T.P. (2000) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, **97**, 77–87.

Fellenberg,K., Hauser,N.C., Brors,B., Neutzner,A., Hoheisel,J.D. and Vingron,M. (2001) Correspondence analysis applied to microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 10781–10786.

Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R. and Caligiuri,M.A. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Grant,G.R., Manduchi,E. and Stoeckert,C. (2001) Using non-parametric methods in the context of multiple testing to determine differentially expressed genes. In Lin,S.M. and Johnson,K.F. (eds), *Methods of Microarray Data Analysis*. Kluwer, Boston, pp. 37–56.

Khan,J., Wei,J.S., Ringner,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. *et al.* (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Li,W. and Yang,Y. (2001) How many genes are needed for a discriminant microarray data analysis? In Lin,S.M. and Johnson,K.F. (eds), *Methods of Microarray Data Analysis*. Kluwer, Boston, pp. 137–150.

Pekarsky,Y., Hallas,C. and Croce,C.M. (2001) The role of TCL1 in human T-cell leukemia. *Oncogene*, **20**, 5638–5643.

Perriere,G., Lobry,J.R. and Thioulouse,J. (1996) Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. *Comput. Appl. Biosci.*, **12**, 519–524.

Sherlock,G. (2000) Analysis of large-scale gene expression data. *Curr. Opin. Immunol.*, **12**, 201–205.

Thioulouse,J., Chessel,D., Dolédec,S. and Olivier,J.M. (1997) ADE-4: a multivariate analysis and graphical display software. *Stat. Comput.*, **7**, 75–83.