BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 21 2006, pages 2667–2673
doi:10.1093/bioinformatics/btl463

Gene expression

Targeted projection pursuit for visualizing gene expression data classifications

Joe Faith[1,*], Robert Mintram[2] and Maia Angelova[1]

[1]Northumbria University, Newcastle, UK and [2]Bournemouth University, Bournemouth, UK

Received on May 15, 2006; revised on August 24, 2006; accepted on August 25, 2006
Advance Access publication September 5, 2006
Associate Editor: Chris Stoeckert

ABSTRACT

We present a novel method for finding low-dimensional views of high-dimensional data: *Targeted Projection Pursuit*. The method proceeds by finding projections of the data that best approximate a target view. Two versions of the method are introduced; one version based on Procrustes analysis and one based on an artificial neural network. These versions are capable of finding orthogonal or non-orthogonal projections, respectively. The method is quantitatively and qualitatively compared with other dimension reduction techniques. It is shown to find 2D views that display the classification of cancers from gene expression data with a visual separation equal to, or better than, existing dimension reduction techniques.

Availability: source code, additional diagrams, and original data are available from http://computing.unn.ac.uk/staff/CGJF1/tpp/bioinf.html

Contact: joe.faith@unn.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

This article considers the problem of visualizing classifications of samples based on high-dimensional gene expression data. There are many powerful automatic techniques for analysing such data, but visualization represents an essential part of the analysis as it facilitates the discovery of structures, features, patterns and relationships, which enables human exploration and communication of the data and enhances the generation of hypotheses, diagnoses and decision making.

Visualizing gene expression data requires representing the data in two (or occasionally one or three) dimensions. Therefore, techniques are required to accurately and informatively show these very high-dimensional data structures in low dimensional representations. In the particular case considered here, showing classified gene expression data taken from cancer samples, the most useful view will be one that clearly shows the separation between classes, allowing the analyst to easily identify outliers and cases of possible misdiagnosis, and to visually compare particular samples.

There are many established techniques for viewing high-dimensional data in lower dimensional spaces. Among these, multi-dimensional scaling (MDS), including Sammon mapping, finds an arrangement of the data that best preserves the distances between points (Ewing and Cherry, 2001); VizStruct is a technique based on radial coordinates (Zhang *et al*., 2004); dendrograms may be used to linearly arrange and display clustered gene expression data (Eisen *et al*., 1998); and projection pursuit (Lee *et al*., 2005) finds linear projections that optimize some measure of their quality (the 'projection pursuit index').

Each of these techniques has limitations and advantages. MDS is able to scale to very high-dimensional data spaces but is a map-based, rather than projection-based, technique in which adding single datum requires creating a new view of the entire set; thus, it is not possible to visualize the relationships of new or unclassified samples to existing ones. VizStruct is not optimized for viewing classifications of the data, and is also only able to accurately visualize data across relatively small number of genes (e.g. 12)—hence is reliant on reducing the dimensionality of the original data through some form of feature selection. And dendrograms arrange samples in just a single dimension for display.

A fundamental advantage of using linear projections for visualization compared to, for example, MDS, is that they define a transform that can be applied to any point in gene-space. In particular, the projection contains information about the respective significance of each gene, and how they can be best combined to perform functions such as classification and genetic feature selection, or to identify gene expression signatures (Misra *et al*., 2002). Projection pursuit is a standard technique for finding linear projections optimized for particular purposes, such as classification, and has recently been applied to gene expression data (Lee *et al*., 2005).

Here we present an alternative to conventional projection pursuit for finding orthogonal and non-orthogonal 2D linear projections, which yield views of the data that are closest to a hypothesized optimal target. The method is compared both quantitatively and subjectively with existing techniques and is found to perform similarly to the best of alternatives. When combined with other techniques it can find views that are better than alternatives.

2 TARGETED PROJECTION PURSUIT

Conventional projection pursuit proceeds by searching the space of all possible projections to find that which maximizes an index that measures the quality of each resulting view. In the case considered here, a suitable index would measure the degree of clustering within, and separation between, classes of points (Lee *et al*., 2005). Targeted projection pursuit, on the other hand, proceeds by hypothesizing an ideal view of the data, and then finding a projection that best approximates that view. The intuition motivating this approach is that the space of all possible views of a high-dimensional dataset is extremely large, so search-based

*To whom correspondence should be addressed.

methods of finding particular views may not be effective. Hence, the alternative technique is pursued for suggesting an ideal view and then finding a nearest match.

Suppose X is an $n \times p$ matrix that describes the expression of p genes in n samples and T is a $n \times 2$ matrix that describes a 2D target view of those samples. We require the $p \times 2$ projection matrix, P, that minimizes the size of the difference between the view resulting from this projection of the data and our target:

$$\min\|T - XP\|, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm.

Two methods are considered for solving Equation (1), depending on whether the projection matrix is required to be orthogonal or not.

2.1 Orthogonal projections

If we make the restriction that projection P is an orthogonal-column matrix, then Equation (1) is an example of a Procrustes problem (Gower and Dijksterhuis, 2004), and a solution may be found using the following version of the singular values decomposition (SVD) method presented by Golub and Loan (1996) [see Cox and Cox (2001) for a discussion of earlier treatments].

Golub and Loan's method finds the $p \times p$ projection matrix, Q, that best maps an $n \times p$ set of data, X, onto an $n \times p$ target view, S, as follows:

$$Q = UV^{\mathrm{T}}, \tag{2}$$

where the superscript T in Equation (2) denotes the transpose operator, and where U and V are the $p \times p$ square matrices with orthogonal columns derived from the SVD of $S^{\mathrm{T}}X$:

$$S^{\mathrm{T}}X = UDV^{\mathrm{T}} \quad \text{where } D \text{ is diagonal.} \tag{3}$$

However if the target view, T, is $n \times 2$ then it can be expanded to an $n \times p$ matrix, S by padding with columns of zeroes. And the required $p \times 2$ projection, P, can be derived from Q by taking just the first two columns.

Efficient methods for SVD are available in most common mathematical and statistical packages such as MATLAB and R. Moreover the complexity of calculating a SVD is dependent on the rank of the matrix, i.e the number of linearly independent rows or columns, rather than its absolute size. Thus, where the number of samples is much less than the number of genes ($n \ll p$), then the complexity of solving a Procrustes equation will end to be dependent on the former rather than the latter. Hence, this technique scales extremely efficiently to high gene numbers.

2.2 Non-orthogonal projections

If the projection P is not required to be orthogonal then a solution to Equation (1) may be found by training a single layer perceptron with p input units and two linear output units (Fig. 1). Each of the n data rows in X are presented in turn, and standard back-propagation is used to train the network to produce the corresponding row of T in response. Once converged, the network can be used to transform data from the original gene-space to a 2D view, with the weight of the connection from the i-th input neuron to the j-th output neuron corresponding to the value of the projection matrix P_{ij}.

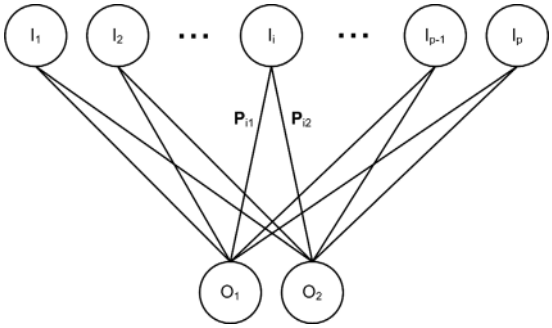

Fig. 1. Schematic diagram of a single layer perceptron for projecting p-dimensional data presented to the top input layer (I_i), to a 2D view output at the bottom layer (O_1,O_2). The connection weight (P_{ij}) describes the weight given to each gene in the projection.

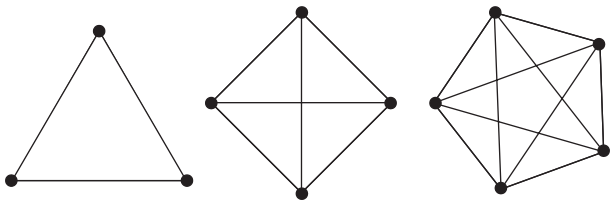

Fig. 2. Graphs of the three-, four- and five-simplex used as target views onto which the gene expression data are projected.

3 TARGETED PROJECTION PURSUIT FOR CLASSIFICATION VISUALIZATION

Given a dataset X and a target view T, then the methods described in Sections 2.1 and 2.2 will find views of X that approximate T. But what is the appropriate target view when considering the classification of gene expression data? If the samples are partitioned into k known classes then the ideal view would be that in which the classes are most clearly separated; that is, where all members of the same class are projected onto single points and where those points are evenly spaced. Thus, the ideal view is one in which all the members of each class are projected onto a single vertex of a geometric simplex.

The k-simplex, or hypertetrahedron, is the generalization of an equilateral triangle ($k = 3$) or tetrahedron ($k = 4$) to higher dimensions. That is, the simplest possible polytope in any given space, that in which all vertices are equidistant from each other. The k-simplex itself is a polytope in $k - 1$ dimensions, but 2D graphs of the three-, four- and five-simplices are shown in Figure 2.

For example, given a set of samples taken from three classes over a large number of dimensions then the ideal view of that data would approximate an equilateral triangle, with the samples of each class clustered at the vertices, and hence would show the clustering within, and the separation between, classes. Whether or not an accurate approximation to such a view can be found depends on how well separated the original data is.

The significance of using a k-simplex rather than just a regular polyhedron as our projection target can be shown by considering the case of $k = 4$. It may be supposed that the separation of classes could be effectively shown by projecting the members of each class onto the vertices of a square. However, the vertices of a square are

not equidistant: the two diagonal pairs of vertices are further apart from the pairs of vertices on each edge. Therefore, using a square as a projection target would entail breaking symmetry; effectively assuming that the pairs of classes that are mapped onto the diagonally opposed vertices are further separated than the other pairs. And this assumption may not be justified. Mapping to the tetrahedron, on the other hand, makes no such assumption. Symmetry is not broken since each pair of vertices are equally separated.

It may be the case in fact that certain classes are more closely related than others, in which case the projection pursuit procedure will produce a view in which they are shown closer together; however, this breaking of symmetry will be due to the nature of the data rather than any assumption made on the experimenters part. This issue is explored empirically below.

The procedure of mapping data onto a target view can also be considered in two other ways, other than the geometric interpretation given above. First, as a set of binary classification problems and second as a spatial classification problem.

First, note that the coordinates of the vertices of the k-simplex can be generated by taking the rows of the k-dimensional identity matrix, i.e. the unit diagonal matrix,

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Now, if C_i is the set of members of the i-th class (with complement $\bar{C}_i$), then mapping the sample classes onto the k-simplex is equivalent to k individual binary classification problems, in which the i-th column of our projection matrix, denoted as $P_{.i}$, maps the members of C_i to 1 and the members of $\bar{C}_i$ to 0. [A similar technique for reducing a multiclass classification problem to multiple binary classifications is explored by Shen and Tan (2006).]

Alternatively, the projection onto a simplex can be thought of as mapping the original data into 'class space'—a k-dimensional space in which the j-th ordinate of the i-th point represents how closely the i-th sample is related to the members of the j-th class.

Whether considered as a mapping onto a simplex, as a combination of binary classification tasks or as a mapping into class space, the view of the data produced by targeted projection pursuit is k-dimensional. Therefore where there are two or three classes the result can be visualized directly. However, when $k > 3$ then a further dimension reduction step is required to view the data. Here we use principal components analysis (PCA) on the rows of our projection matrix, $P_{i.}$, each a k-dimensional vector, to find a lower dimensional projection that best preserves the information in P. (The first two principal components are chosen, based on the square roots of the eigenvalues of the covariance matrix.) Thus, we have a two-stage dimension reduction process, each stage of which is based on a linear projection; therefore, the combined result is itself a linear projection (Fig. 3).

Note that by taking a linear projection of the original data that approximates a k-simplex we are effectively ignoring all information about the distances between individual points *per se*, and instead utilizing information about the relationships with other points discriminated by class. Contrast this with MDS, which finds a representation of the data that best preserves distances between data points and ignores classification [though a version of MDS that uses cluster information for visualization purposes is discussed by Schwenker *et al*. (1996)].

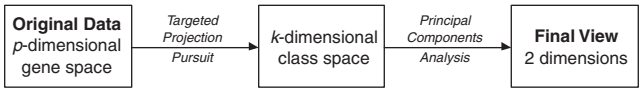

Fig. 3. Two-stage dimension reduction. Targeted projection pursuit is used to reduce the original high-dimensional dataset to k dimensions (where k is the number of classes in the original data). Principal components analysis is then used to find a 2D projection of the k-dimensional view.

4 METHODS

The targeted projection pursuit techniques outlined in Sections 2 and 3 were tested for their ability to produce 2D views of data that clearly separate sample classes. The techniques were tested on three publicly available datasets, and the views were compared with the output from standard dimension reduction techniques. The views of each dataset produced by each technique were tested both quantitatively and qualitatively. The views were quantitatively compared in two ways: first, by submitting them to a standard classification algorithm and measuring the resulting generalization performance; and, second, by using a standard statistical measure of class separability. The views were qualitatively compared by visual inspection.

The following dimension reduction techniques were compared:

- SLP: The result of targeted projection pursuit using a single layer linear perceptron network, followed by PCA.
- PRO: The result of targeted orthogonal projection pursuit using the solution to a Procrustes equation, followed by PCA.
- PP: The linear projection produced by search-based projection pursuit (Lee *et al*., 2005).
- SAM: The result of a Sammon MDA (Ewing and Cherry, 2001).
- VS: The result of a VizStruct projection onto radial coordinates (Zhang *et al*., 2004).

Further details of the techniques used, including original source code where appropriate, is available from the associated website.

The following datasets were used:

- LEUK: This dataset is the result of a study of gene expression in two types of acute leukaemia: acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML) (Golub *et al*., 1999). The samples consist of 38 cases of B-cell ALL, 9 cases of T-cell ALL and 25 cases of AML with the expression levels of 7219 genes measured. Note that, following Lee *et al*. (2005), the B-cell and T-cell ALL samples are considered as separate classes.
- SRBCT: This dataset comprises cDNA microarray analysis of small, round blue cell childhood tumors (SRBCT), including neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt Lymphoma (BL; a subset of non-Hodgkin lymphoma) and members of Ewing's family of tumors (EWS). Expression levels from 6567 genes for 83 samples were taken (Khan *et al*., 2001).
- NCI: This dataset records the variation in gene expression among the 60 cell lines from the National Cancer Institute's anticancer drug screen (Scherf *et al*., 2000). It consists of eight different tissue types where cancer was found: nine breast, five central

nervous system (CNS), seven colon, six leukaemia, eight melanoma, nine non-small-cell lung carcinoma (NSCLC), six ovarian, two prostate and eight renal. A total of 9703 cDNA sequences were used.

No further normalization was applied to any dataset, beyond that described in the original references, though the top 50 most discriminatory genes were chosen on the basis of the ratio of their between-group to within-group sums of squares (Dudoit *et al*., 2002).

The classification algorithm used for the quantitative evaluation was *K*-nearest neighbours (KNNs). This choice of algorithm was motivated by two considerations. The first is that it is known to be effective at discriminating classes of tumour using gene expression data (Dudoit *et al*., 2002). The second consideration is KNN is an instance- and distance-based measure in which the classification of an instance is dependent on the classes of its nearest neighbours. It is assumed that this measure would accord better with human judgement than a probabilistic attribute-based measure such as Naïve Bayes—even though the latter may have superior classification performance in some cases. The Weka implementation of this algorithm was used (Witten and Frank, 2005), tested using 10-fold cross-validation and a simple percentage accuracy score found $k = 5$ nearest neighbours were used, since this minimized mean cross-validation error.

Note that the accuracy of classification using KNN for each view tested is not equivalent to a true generalization performance since the views were produced using the full datasets, rather than a training subset. This is because it is the class separation within each view that is being tested, rather than the performance of the classifier. Given a view, we would like to know how visually separated the classes in the data are—operationalized as classifier generalization—not which technique produces the best generalization performance as a classifier.

The statistical measure of class separability used to compare views was a version of Fishers' linear discriminant analysis index (I_{LDA}) introduced by Lee *et al*. (2005), based on the ratio of between-groups to within-groups sum of squares. If V_{ij} is the view of the j-th member of the i-th class then let

$$B = \sum_{i=1}^{k} n_i(\bar{V}_{i.} - \bar{V}_{..})(\bar{V}_{i.} - \bar{V}_{..})^{\mathrm{T}}\text{: between-group sum of squares}$$

$$W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{V}_{ij} - \bar{V}_{i.})(\bar{V}_{ij} - \bar{V}_{i.})^{\mathrm{T}}\text{: within-group sum of squares}$$

Thus B is a measure of the variance of the centroids of the classes, and W is a measure of the variance of the instances within each class. In order to get a projection pursuit index in the range [0,1], with increasing values corresponding to increasing class separation then, I_{LDA}, a version of Wilks Lamda, a standard test statistic used in multivariate analysis of variance, is used:

$$I_{\mathrm{LDA}} = 1 - \frac{|W|}{|W + B|}.$$

The R-code implementation of I_{LDA} distributed by Lee *et al*. (2005) was used to measure the class separability of the resulting views.

The effect of the symmetry assumption mentioned in Section 3 was tested by varying the order in which the classes of data were taken, and finding whether this had an effect on the classification performance and class separability of the resulting views produced by SLP.

Table 1. Comparison of class separability following dimension reduction for visualization.

Dataset	LEUK		SRBCT		NCI	
Genes	7129		2308		9712	
Samples	72		83		61	
Classes	3		4		8	
Class separation measure	I_{LDA}	5NN	I_{LDA}	5NN	I_{LDA}	5NN
SLP	0.999	100.0	0.999	100.0	0.999	83.6
PRO	0.966	97.2	0.966	89.2	0.894	49.2
Dimension reduction technique						
PP	0.972	98.6	0.988	100.0	0.981	62.3
SAM	0.959	97.2	0.911	95.2	0.927	54.1
VS	0.952	95.8	0.637	56.6	0.838	32.8
SLP-PP	0.999	100.0	1.000	100.0	1.000	95.1

Each dimension reduction technique (SLP, PRO, PP, SAM, VS and SLP-PP) is evaluated on each dataset (LEUK, SRBCT, NCI), and the separability of the resulting view tested using both 5-nearest neighbours classification (5NN, generalization error in %) and a version of Wilks Lamda ($0 < I_{\mathrm{LDA}} < 1$).

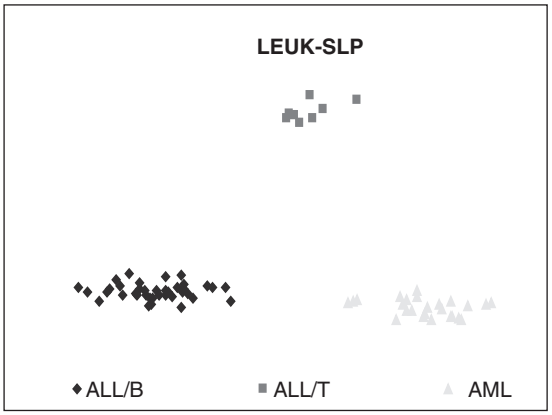

Fig. 4. View of LEUK dataset generated by SLP method: this method finds a projection in which all three classes are very clearly separated. A colour version of this figure appears in the Supplementary data.

5 RESULTS

The quantitative comparison of the four projections on the three datasets is shown in Table 1, and a sample of the resulting views are given in Figures 4–11. A complete set of views are available on the accompanying website.

The first aspect of the results to note is that the choice of dimension-reduction technique can alter radically the resulting view of the data, judged both quantitatively and qualitatively. The structure and relationship between clusters appears very differently in each view, resulting in very different performances of classification algorithms. The choice of dimension reduction technique clearly matters in visualizing high-dimensional data such as gene expression data.

The second aspect to note is that quantitative measures such as I_{LDA} or classification performance are not a reliable indicator of visual class separation. For example, the view of the NCI dataset

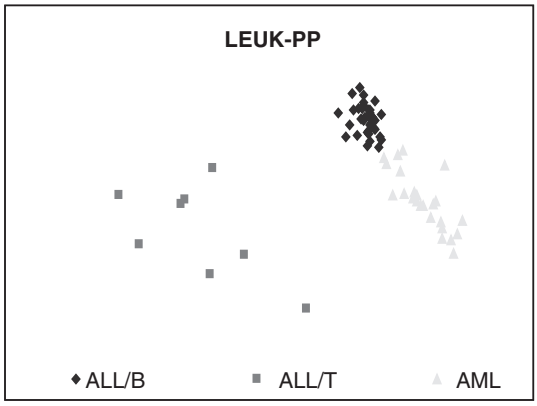

Fig. 5. View of LEUK dataset generated by PP method: conventional search-based projection pursuit finds a view in which there is little clear separation between some samples of ALL/B and AML. A colour version of this figure appears in the Supplementary data.

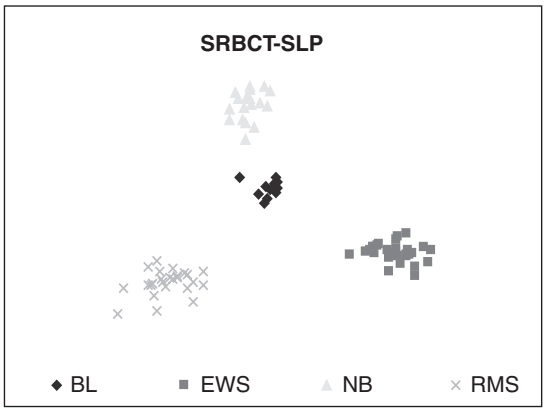

Fig. 6. View of SRBCT dataset generated by SLP method, showing a clear separation between four classes. A colour version of this figure appears in the Supplementary data.

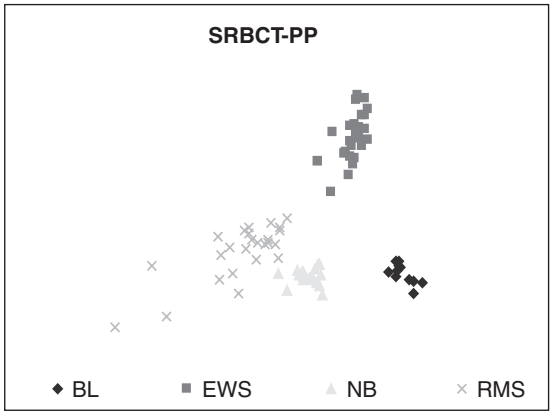

Fig. 7. View of SRBCT dataset generated by PP method: search-based projection pursuit distinguishes all classes, though with little clear separation between samples of NB and RMS. A colour version of this figure appears in the Supplementary data.

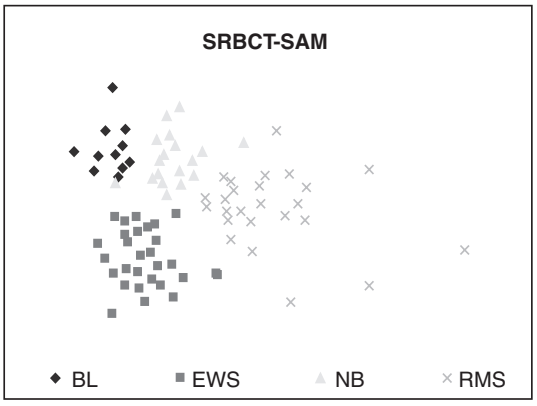

Fig. 8. View of SRBCT dataset generated by SAM method, showing one effect of the 'curse of dimensionality': Sammon mapping, like other MDS methods, tries to find a representation that preserves distances between points. Where the original data dimensionality is large there is less variance in intra- and inter-class distances, and hence there is little 'bunching' or class separation in the lower dimensional representation. A colour version of this figure appears in the Supplementary data.

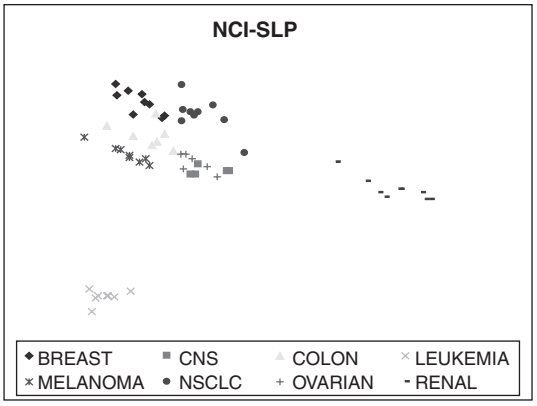

Fig. 9. View of NCI dataset generated by SLP method: as the number of classes increases to eight, so does the amount of visual overlap as the higher-dimensional simplex is viewed in two dimensions using PCA. Only the leukaemia and renal cancer cases are clearly separated. A colour version of this figure appears in the Supplementary data.

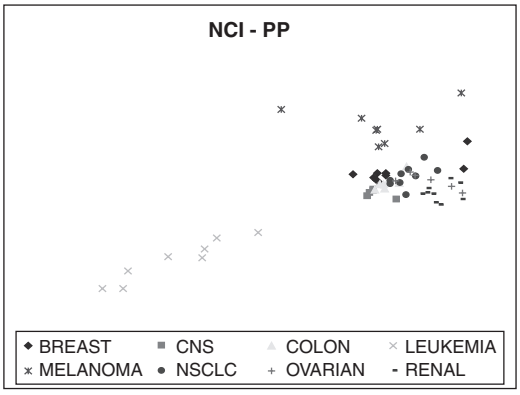

Fig. 10. View of NCI dataset generated by PP method: search-based projection pursuit is only able to clearly distinguish leukaemia and melanoma cases. A colour version of this figure appears in the Supplementary data.

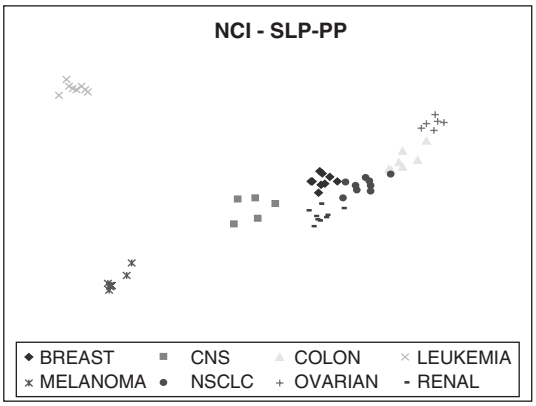

Fig. 11. View of NCI dataset generated by hybrid SLP-PP method: combining projection pursuit methods separates classes more clearly than either method alone. Leukaemia, CNS and melanoma cases are clearly distinguished, and some separation between all other classes. A colour version of this figure appears in the Supplementary data.

generation using SLP has an extremely high I_{LDA} index of 0.999, but there is visual confusion between most of the classes (Fig. 9). In another example, SAM produced a view of the SRBCT with a 5NN classification performance of 95.2%, but with many outliers between classes.

Overall, VizStruct performed least well in separating classes. Although the difference between VizStruct and the other techniques was least for the low-k case (LEUK), the difference became more marked as the number of classes increased. This poor performance is unsurprising, since this technique is not explicitly designed to accentuate classifications [though see Zhang *et al*. (2004)].

The Sammon mapping performed well in separating classes, but its output was marked by the 'curse of dimensionality': in high-dimensional spaces, the variance in distances between randomly distributed points decreases. Sammon mapping attempts to preserve the distance between data points, and hence the resulting views tend to be evenly distributed, with little bunching of points belonging to a single class (Fig. 8). Classification algorithms may succeed in ascribing points to classes—and hence the classification scores for SAM are similar to those for the linear mappings—but this may not be an accurate reflection of the perceived class separation.

The projection pursuit methods SLP and PP performed best in general, finding linear projections that clearly separated all classes where the number of cancer types was small (LEUK, SRBCT). However, as the number of classes increases the performance of all methods degrades, rendering them ineffective with little consistent distinction between classes.

Conventional search-based projection pursuit also suffered from unreliability. Since it is partly a stochastic technique, the results could differ. Over a sequence of 100 trials, the values for I_{LDA} for PP applied to the NCI set ranged from 0.935 to 0.992 (mean = 0.978, SD 0.00924). The values for I_{LDA} and 5NN shown in Table 1, and the view show in Figure 10 are for a projection of near-mean I_{LDA} value.

Varying class order was found to have no effect on the classification performance or class separability of the views produced by SLP (though the orientation of each view may be altered). Thus, this technique is not affected by the symmetry assumption embodied in targetting simplex-views of the data.

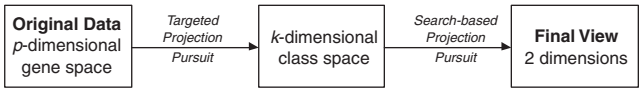

Fig. 12. Hybrid targeted and search-based projection pursuit. Targeted projection pursuit is used to reduce the dataset to k dimensions as before (Figure 3), but now search is used to find the optimal two-dimensional projection of this view.

5.1 Hybrid projection pursuit

The targeted methods (SLP and PRO) performed relatively poor in the higher-k case (NCI), compared with their success on the lower-k cases (LEUK, SRBCT). This suggests that the drop in performance is due to the second stage of the two-stage reduction process, where PCA is used to reduce the dimensionality from k-dimensional class space to the 2D visualization, rather than the reduction from the original gene-space to k-dimensional class-space (Fig. 3). This is presumably because classes that are separated in k-space may overlap when viewed in two dimensions.

This hypothesis was tested by testing a hybrid dimension reduction technique, in which SLP was used to reduce the dimensionality to k and then search-based projection pursuit was used to find a 2D view of the result (Fig. 12). Note that the combined effect of this hybrid technique is still a linear projection of the original data. This technique (SLP-PP) was found to be highly effective with a clear visual separation between classes (Fig. 11). It thus seems that a limiting factor on search-based projection pursuit is the problem of searching a very large space using a stochastic technique, such as simulated annealing. Combining search-based projection pursuit with SLP reduces the size of the space for the former task from 50×2 dimensions to 8×2 in this case, and the increase in performance is marked.

This hybrid method thus combines the strengths of targeted- and search-based projection pursuit. Targeted projection pursuit is able to find effective projections from very high dimensions, but only to k-dimensional subspaces. Whereas search-based projection pursuit produces better projections to two-dimensions than PCA, but looses effectiveness and reliability as the dimensionality of the original space increases.

6 DISCUSSION

The high dimensionality of microarray data introduces the need for visualization techniques that can 'translate' these data into lower dimensions without losing significant information, and hence assist with data interpretation. Many dimension reduction techniques are available, but in this paper we introduce the novel concept of targeted projection pursuit—that is, finding views of data that most closely approximate a given target view—and demonstrate the use of solutions of Procrustes equations and trained perceptron networks to achieve this end. In this particular case, we explore the possibility of using targeted projection pursuit to find views that most clearly separate classified datasets.

Targeted projection pursuit was evaluated in comparison with three very different established dimension reduction techniques, on three publicly available datasets. When discriminating a small number of cancer classes the performance of the technique matched or bettered that of established methods. When presented with a large

number of classes (eight) the technique combined effectively with other existing techniques to produce views of the data that showed the separation between sample classes more effectively than the alternatives evaluated.

The technique is also able to scale to large numbers of genes: the version involving the targeted pursuit of orthogonal projections (PRO) is able to handle an input dimensionality of tens of thousands of genes without feature selection.

Note that the use of a target view does not constitute a limitation of the technique. The target plays the role of a hypothesis—in this case that the samples can be classified based on gene expression levels—and the resulting views illustrates how well the data meets that hypothesis. (And by using a fully symmetrical simplex as the target view, no assumptions about the relationships between classes are made.) Other hypothesis-targets could be used in other cases, such as using a circular target to explore cyclical process in samples from a time-series, or a rectilinear target to explore the existence of simple linear relations. The same classification visualization technique employed here to classify samples in gene-space could also be applied to the transpose problem; that of visualizing the classification of genes on the basis of their expression profiles in varying conditions, and so explore relationships between gene function rather than between samples.

Targeted projection pursuit is a general purpose technique for finding views of data that approximate optimal targets. This paper discussed just one specific application to the problem of visualizing classified microarray data. The authors are currently exploring other applications in visualizing high-dimensional biological data, including constructing a tool that would allow an user to interactively explore the space of possible views of high-dimensional datasets.

As mentioned in Section 1, one of the principal reasons for choosing a visualization technique based on a linear projection rather than, say, MDS, is that the resulting projection can yield useful information about the relative significance of particular genes, including their respective weights for classification (Misra *et al*., 2002). This paper has discussed the derivation of such projections and the future works will explore the significance of the resulting information.

Conflict of Interest: none declared.

REFERENCES

Cox,M.F. and Cox,M.A.A. (2001) *Multidimensional scaling*. Chapman and Hall, London.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc*., **97**, 77–87.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Ewing,R.M. and Cherry,J.M. (2001) Visualization of expression clusters using Sammon's non-linear mapping. *Bioinformatics*, **17**, 658–659.

Golub,J. and Loan,C.F. (1996) *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore MD, US.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gower,J.C. and Dijksterhuis,G.B. (2004) Procrustes Problems. Oxford University Press, Oxford, UK.

Khan,J. *et al*. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med*., **7**, 673–679.

Lee,E.K. *et al*. (2005) Projection pursuit for exploratory supervised classification. *J. Comput. Graph. Stat*., **14**, 831–846.

Misra,J. *et al*. (2002) Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res*., **12**, 1112–1120.

Scherf,U. *et al*. (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet*., **24**, 236–244.

Shen,L. and Tan,E.C. (2006) Reducing multiclass cancer classification to binary by output-coding and SVM. *Comput. Biol. Chem*., **30**, 63–71.

Schwenker,F. *et al*. (1996) Visualization and analysis of signal averaged high resolution electrocardiagrams employing cluster analysis and multidimensional scaling. In: Proceedings of the Computers in Cardiology 1996. IEEE Press, Indianapolis, IN, US, pp. 453–456.

Witten,I.H. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd edn. Morgan Kaufmann, San Francisco, CA.

Zhang,L. *et al*. (2004) VizStruct: exploratory visualization for gene expression profiling. *Bioinformatics*, **20**, 85–92.