Available online at www.sciencedirect.com

ScienceDirect

www.sciencedirect.com/science/journal/16720229

GENOMICS
PROTEOMICS &
BIOINFORMATICS

Method

Mining Gene Expression Profiles: An Integrated Implementation of Kernel Principal Component Analysis and Singular Value Decomposition

Ferran Reverter*, Esteban Vegas, and Pedro Sánchez

Department of Statistics, Faculty of Biology, University of Barcelona, 08028 Barcelona, Spain.

Genomics Proteomics Bioinformatics 2010 Sep; 8(3): 200-210 DOI: 10.1016/S1672-0229(10)60022-8

Abstract

The detection of genes that show similar profiles under different experimental conditions is often an initial step in inferring the biological significance of such genes. Visualization tools are used to identify genes with similar profiles in microarray studies. Given the large number of genes recorded in microarray experiments, gene expression data are generally displayed on a low dimensional plot, based on linear methods. However, microarray data show nonlinearity, due to high-order terms of interaction between genes, so alternative approaches, such as kernel methods, may be more appropriate. We introduce a technique that combines kernel principal component analysis (KPCA) and Biplot to visualize gene expression profiles. Our approach relies on the singular value decomposition of the input matrix and incorporates an additional step that involves KPCA. The main properties of our method are the extraction of nonlinear features and the preservation of the input variables (genes) in the output display. We apply this algorithm to colon tumor, leukemia and lymphoma datasets. Our approach reveals the underlying structure of the gene expression profiles and provides a more intuitive understanding of the gene and sample association.

Key words: kernel method, biplot, gene expression profile, dimension reduction

Introduction

Microarray technology has been advanced to the point at which the simultaneous monitoring of gene expression on a genome scale is now possible. Microarray experiments often aim to identify individual genes that are differentially expressed under distinct conditions, such as between two or more phenotypes, cell lines, under different treatment types or diseased and healthy subjects. Such experiments may be the first step towards inferring gene function and constructing gene networks in systems biology.

The term "gene expression profile" refers to the gene expression values on all arrays for a given gene in different groups of arrays. Frequently, a summary statistic of the gene expression values, such as the mean or the median, is also reported. Dot plots of the gene expression measurements in subsets of arrays, and line plots of the summaries of gene expression measurements, are the most common plots used to display gene expression data (*1*).

An ever increasing number of techniques are being

*Corresponding author.
E-mail: freverter@ub.edu

applied to detect genes that have similar expression profiles from microarray experiments. Techniques like clustering (*2*) and self organization map (*3*) have been applied to the analysis of gene expression data. We can also find several applications on microarray analysis based on distinct machine learning methods such as Gaussian processes (*4*, *5*), Boosting (*6*) and Random Forest (*7*). It is useful to find gene/sample clusters with similar gene expression patterns for interpreting the microarray data.

However, due to the large number of genes involved, it might be more effective to display these data on a low dimensional plot. Recently, several authors have explored dimension reduction techniques: Alter *et al* (*8*) analyzed microarray data using singular value decomposition (SVD), Fellenberg *et al* (*9*) applied correspondence analysis to visualize genes and tissues, Pittelkow and Wilson (*10*) and Park *et al* (*11*) used several variants of biplot methods as a visualization tool for the analysis of microarray data. Visualizing gene expression may facilitate the identification of genes with similar expression patterns.

In this paper we describe a method to visualize gene expression profiles. Our procedure relies on SVD; however, unlike other methods, it incorporates an additional step that involves kernel principal component analysis (KPCA) (*12*). Kernel representation offers an alternative to nonlinear functions by projecting the data into a high-dimensional feature space, which increases the computational power of linear learning machines (*13*, *14*).

Kernel methods enable us to construct different nonlinear versions of any algorithm that can be expressed solely in terms of dot products, known as the kernel trick. Thus, kernel algorithms avoid the explicit usage of the input variables in the statistical learning task. Kernel machines can be used to implement several learning algorithms but they usually act as a black-box with respect to the input variables. This could be a drawback in biplot displays in which we pursue the simultaneous representation of samples and input variables. As we describe in greater detail below, the integrated implementation of SVD-Biplot joint with KPCA enables us to extract the nonlinear features without discarding the simultaneous display of input variables (genes) and samples (microarrays).

Method

We describe an effective procedure that allows the simultaneous display of input variables (genes) and samples (microarrays). The standard SVD-Biplot allows us to represent samples and variables jointly, but our procedure incorporates an additional extraction of nonlinear features that could improve our understanding of the relationships between samples (microarrays) and variables (genes), and enhance the detection of gene expression profiles. Since microarray data are generally nonlinear, finding methods that can handle such data is of great importance if as much information as possible is to be gleaned.

Let $\mathbf{X}$ be the preprocessed gene expression data matrix, with n samples (microarrays) in the rows and p variables (genes) in the columns. Preprocessing of the gene expression measurements needs to be considered with caution because preprocessing, such as scaling, normalization and transformation, can have a strong effect on the output visualization (*15*).

Our procedure represents the gene expression data in a feature space using a kernel method. Initially, we decompose the input data matrix $\mathbf{X}=\mathbf{GH}^T$ into singular values, for an $n\times r$ matrix $\mathbf{G}$ and a $p\times r$ matrix $\mathbf{H}$, representing sample (microarray) and variable (gene) effects, respectively.

Next, we map the variables (genes), the rows of $\mathbf{H}$, into a feature space, and we extract the main nonlinear features of variables (genes) by performing KPCA. Finally, to obtain a simultaneous plot, we project the samples (microarrays) into the subspace spanned by the leading eigenvectors from KPCA.

Our procedure is composed of the following steps:

1. SVD of preprocessed gene expression input matrix $\mathbf{X}=\mathbf{GH}^T$.

2. Take the rows of $\mathbf{H}$ as a set of observations and compute the corresponding kernel matrix $\mathbf{K}$.

3. Extract nonlinear gene expression features by computing KPCA on the kernel matrix $\mathbf{K}$.

4. Project the rows of $\mathbf{G}$ onto the subspace expanded by the leading eigenvectors of $\mathbf{K}$.

Step 1 allows us to represent the microarrays (samples) and the gene expressions (input variables) as points on the same r-dimensional space. We obtain the matrices:

$$\mathbf{G} = \begin{pmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{pmatrix} = \begin{pmatrix} g_{11} & \cdots & g_{1r} \\ \vdots & \vdots & \vdots \\ g_{n1} & \cdots & g_{nr} \end{pmatrix} \qquad (1)$$

and

$$\mathbf{H} = \begin{pmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_p \end{pmatrix} = \begin{pmatrix} h_{11} & \cdots & h_{1r} \\ \vdots & \vdots & \vdots \\ h_{p1} & \cdots & h_{pr} \end{pmatrix} \qquad (2)$$

where vectors $\mathbf{g}_1,\ldots,\mathbf{g}_n$ represent microarrays and vectors $\mathbf{h}_1,\ldots,\mathbf{h}_p$ represent gene expressions.

Step 2 serves to build the kernel matrix **K**. To compute **K** on the sample observations, the choice of the kernel function $K(\mathbf{x}, \mathbf{y})$ is crucial because the kernel methods use it to measure the similarity of sample observations. Gene expression data are a set of numerical vectors, so two natural candidates for kernel function $K(\mathbf{x}, \mathbf{y})$ are the polynomial kernel:

$$K(\mathbf{x}, \mathbf{y})=(\mathbf{xy}+c)^t$$

where t, $c>0$ are free parameters, and the radial basis kernel is:

$$K(\mathbf{x}, \mathbf{y})=\exp(-c\|\mathbf{x}-\mathbf{y}\|^2)$$

where $c>0$ is a free parameter. Once we select a suitable kernel function, we compute the $p\times p$ kernel matrix $\mathbf{K}=(K(\mathbf{h}_i,\mathbf{h}_j))$.

Step 3 allows us to extract nonlinear features of gene expressions by solving the dual eigenvalue problem, such as Formula 8 showed below. Notice that in this step we have considered the vectors $\mathbf{h}_1,\ldots,\mathbf{h}_p$, representing gene expressions (Formula 2), as input observations. At the end of this step we find: $\boldsymbol{\alpha}^1,\ldots,\boldsymbol{\alpha}^r$, the set of normalized eigenvectors, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$, the corresponding set of nonzero eigenvalues of **K**. Formula 9 (see below) allows us to obtain the coordinates of the gene expression $\mathbf{h}_l$ (l=1,…,p) by projecting onto the leading kernel principal components

$$\left\langle \mathbf{v}_j, \phi(\mathbf{h}_l) \right\rangle = \sum_{i=1}^{p} \alpha_i^j K(\mathbf{h}_i, \mathbf{h}_l) \qquad (3)$$

where j=1,...,r. As is customary, we use a small number of eigenvectors in order to obtain a graphical output in low dimension.

The stability of KPCA, that is, the extent to which the projection captures new data sampled according to the same distribution as the training data, increases when the eigenvalues become small at an initial stage in the spectrum of **K**. Provided we project into a space whose dimension exceeds the index of this stage, we expect to capture most of the variance of the unseen data.

Step 4 is to project the sample (microarrays) points $\mathbf{g}_l$, l=1,...,n (see Formula 1), onto the subspace spanned by the eigenvectors of **K**. The new coordinates are given using Formula 9, which is

$$\left\langle \mathbf{v}^j, \phi(\mathbf{g}_l) \right\rangle = \sum_{i=1}^{p} \alpha_i^j K(\mathbf{h}_i, \mathbf{g}_l) \qquad (4)$$

where j=1,...,r.

We can analyze and better understand the association between genes and samples by superimposing the coordinate Formulas 3 and 4 in the same plot. Furthermore, this integrated implementation of SVD-Biplot and KPCA provides a supplementary tool to detect gene expression profiles in microarray experiments. For brevity we refer to our method as the KPCA-Biplot.

Kernel selection and automatic tuning

We could consider the choice of the kernel and the tuning procedure for the kernel parameters as a model selection problem.

As is well known in the context of supervised learning (*e.g.*, classification tasks), we can handle this selection by including a cross-validation step in the procedure. We therefore guide the kernel selection and the tuning of the parameters automatically using a cross-validation score function based on the minimization of the misclassification rate.

In the context of unsupervised learning (*e.g.*, dimension reduction tasks), an approach to include a cross-validation step for data-driven kernel selection, when the samples are grouped (*e.g.*, different treatment types, cell lines and phenotypes), is to adopt a strategy similar to that followed in the supervised case (*16*). In summary, assuming a fixed number of dimensions, usually 2 or 3, we perform the dimension reduction for each kernel. Then, using the sample coordinates in the reduced dimension space, a classifier is trained, and assessed with respect to a test dataset. Finally, we select the kernel that minimizes the misclassification rate.

Computational complexity

The speed and the time of calculations have always been important problems in the implementation of algorithm in multivariate data analysis. Microarray datasets typically consist of thousands of variables and less than 100 samples. For such huge datasets, the computation time needed for matrix decomposition using classical SVD algorithm and/or the eigenvalue decomposition for the kernel matrix may be excessive.

Alter *et al* (*8*) propose an efficient algorithm for SVD for genome-wide expression data. In summary, the algorithm is as follows. For the dimension of data matrix $\mathbf{X}$ ($n\times p$), where n denotes the number of samples and p denotes the number of genes, we assume that the data are centered. For matrices $\mathbf{U}_{n\times r}$, $\mathbf{D}_{r\times r}$ and $\mathbf{V}_{r\times p}$, the $p\times p$ sample correlation matrix admits the following eigenvalue decomposition $\mathbf{A}=\mathbf{X}^T\mathbf{X}=\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T=\mathbf{V}\mathbf{D}^2\mathbf{V}^T$, and the $n\times n$ gene correlation matrix admits the following eigenvalue decomposition $\mathbf{B}=\mathbf{X}\mathbf{X}^T=\mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}\mathbf{D}\mathbf{U}^T=\mathbf{U}\mathbf{D}^2\mathbf{U}^T$. We therefore calculate the SVD of $\mathbf{X}$ (step 1 of the proposed method) by diagonalizing $\mathbf{B}$ and computing $\mathbf{V}$ as $\mathbf{V}=\mathbf{X}^T\mathbf{U}\mathbf{D}^{-1}$.

Furthermore, since the kernel matrix $\mathbf{K}$ is $p\times p$ (where p is assumed to be from thousands to tens of thousands in real application), reducing the cost of computing might also be appropriate in the kernel matrix eigenvalue decomposition (step 3 of the proposed method). When we need to analyze a large number of genes, we may want to work with an algorithm for computing only the largest eigenvalues, as for instance the power method with deflation (*17*).

SVD

Let $\mathbf{X}$ be the preprocessed gene expression data matrix, with n samples (microarrays) in the rows and p variables (genes) in the columns. Underlying the biplot techniques (*18, 19*) is the SVD:

$$\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{V}^T \tag{5}$$

where $\mathbf{U}$ and $\mathbf{V}$ are matrices of size $n\times r$ and $p\times r$, respectively, with orthonormal columns so that $\mathbf{U}^T\mathbf{U}=\mathbf{V}\mathbf{V}^T=\mathrm{I}_r$, $\mathbf{D}$ is an $r\times r$ diagonal matrix with elements $\lambda_1\geq\ldots\geq\lambda_r$ in the diagonals, and r is the rank of $\mathbf{X}$, so usually $r = \min(n, p)$. Let us define $\mathbf{D}^{\alpha}=\mathrm{diag}(\lambda_1^{\alpha},\cdots,\lambda_r^{\alpha})$ and let $\mathbf{G}=\mathbf{U}\mathbf{D}^{\alpha}$ and $\mathbf{H}=\mathbf{V}\mathbf{D}^{1-\alpha}$, where $0\leq\alpha\leq 1$, thus $\mathbf{X}$ can be factorized as

$$\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{V}^T=\mathbf{U}\mathbf{D}^{\alpha}\mathbf{D}^{1-\alpha}\mathbf{V}^T=\mathbf{G}\mathbf{H}^T$$

for an $n\times r$ matrix $\mathbf{G}$ and $p\times r$ matrix $\mathbf{H}$. Thus $\mathbf{X}$ can be decomposed into two sets of matrices $\mathbf{G}$ and $\mathbf{H}$, representing row and column effects, respectively.

KPCA

Given a set of observations $\mathbf{x}_i\in\Re^p$, $i=1,\ldots,n$, let us consider a feature space F related to the input space by a map ϕ: $\Re^p\rightarrow F$, which may be nonlinear. We assume that we are dealing with centred data $\sum_{i=1}^{n}\phi(\mathbf{x}_i)=0$. In F the covariance matrix takes the form

$$\mathbf{C}=\frac{1}{n}\sum_{j=1}^{n}\phi(\mathbf{x}_j)\phi^T(\mathbf{x}_j)$$

We seek eigenvalues $\lambda\geq 0$ and nonzero eigenvectors $\mathbf{v}\in F\backslash\{0\}$ satisfying $\mathbf{C}\mathbf{v}=\lambda\mathbf{v}$. All solutions $\mathbf{v}$ with $\lambda\neq 0$ lie in the span of $\{\phi(\mathbf{x}_i)\}_{i=1}^{n}$ as shown in the literature (*14*). This has two consequences: first we may instead consider the set of equations

$$\left\langle\phi(\mathbf{x}_j),\mathbf{C}\mathbf{v}\right\rangle=\lambda\left\langle\phi(\mathbf{x}_j),\mathbf{v}\right\rangle \tag{6}$$

for all $j=1,\ldots,n$, where $\left\langle\cdot,\cdot\right\rangle$ denotes the dot product defined in F. Second, there exist coefficients α_i, $i=1,\ldots,n$, such that

$$\mathbf{v}=\sum_{i=1}^{n}\alpha_i\phi(\mathbf{x}_i) \tag{7}$$

Combining Formulas 6 and 7 we obtain the dual representation of the eigenvalue problem for nonzero eigenvalues:

$$\mathbf{K}\alpha=n\lambda\alpha \tag{8}$$

where $\mathbf{K}=(K(\mathbf{x}_i,\mathbf{x}_j))$, i, $j=1,\ldots,n$, is the kernel matrix and K is a kernel function such that the dot product in F satisfies $\left\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\right\rangle=K(\mathbf{x}_i,\mathbf{x}_j)$. Let $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_n$ be the eigenvalues of $\mathbf{K}$ and $\alpha^1,\ldots,\alpha^n$ be the corresponding set of normalized eigenvectors, with λ_r being the last nonzero eigenvalue. For the purpose of principal component extraction, we need to compute the projections onto the eigenvectors $\mathbf{v}^j$ in F, $j=1,\ldots,r$. Let $\mathbf{x}$ be a test point, with an image $\phi(\mathbf{x})$ in F. Then

$$\left\langle\mathbf{v}^j,\phi(\mathbf{x})\right\rangle=\sum_{i=1}^{n}\alpha_i^jK(\mathbf{x}_i,\mathbf{x}) \tag{9}$$

which is the j-th nonlinear principal component corresponding to ϕ.

Validation

In this section we illustrate the application of KPCA-Biplot with data from the colon tumor (*20*), leukemia (*21*) and lymphoma (*22*) datasets. In these examples, the aim of the KPCA-Biplot is to detect genes (variables) that have a similar pattern of up/down-regulation for each sample. By simultaneously displaying both the samples and the genes on the same plot, it is possible both to visually detect genes that have similar profiles and to interpret this pattern by reference to the sample groups.

From the position of the genes relative to the samples, it can be deduced that genes, which lie relatively close to any given group, will have higher values (up-regulated) in that group than in the other groups. Genes lying on the opposite side of the origin from a given group will tend to have lower values (down-regulated) in that group. Then gene profiles are useful to reveal differential expression between sample groups.

As an example, we describe the profiles of some illustrative genes that are located away from the central gene cloud in each genomic dataset. In particular, with the aim of detecting different profiles, we explore several directions from the origin of the graphical output and describe the profiles of a set of genes that lie in those directions. By combining the KPCA-Biplot and the plot of the profiles, we can represent all the different kinds of profile on one plot.

Colon tumor dataset

This dataset is composed of 2,000 genes in 40 colon tumor samples and 22 normal colon tissue samples. Gene expression levels were analyzed with Affymetrix oligonucleotide array. The dataset is available on the web at http://www.molbio.princeton.edu/colondata. We complete the preprocessing of the gene expression data with a microarray standardization and gene centring (*10*).

We performed the KPCA-Biplot procedure on this dataset as detailed in the previous section. Initially, we factorize the input data matrix using SVD with α=0.5. Next, we compute the kernel matrix in Step 2 using the radial basis kernel with c=0.1. The resulting biplot is given in **Figure 1**. It shows the projection onto the two leading kernel principal components of genes and microarrays, which were obtained by using Formulas 3 and 4, respectively. Genes are shown as a tilde and microarrays are shown by different marks according to the class of colon tissues to which they belong.

The first axis mainly distinguishes the sample and gene clouds. It also separates those highly expressed genes from the origin. The second axis reveals normal and tumor samples. Nevertheless, normal and tumor samples are not completely separated in this 2D plot. In addition, some genes are labeled with their gene numbers and these have been chosen to illustrate different profiles. Genes that lie towards the left and bottom will have higher expression in tumor samples. In contrast, genes at the top, close to the normal samples, will tend to have higher expression in these samples.

Figure S1 shows the gene expression profiles of some genes labeled in Figure 1. The gene expression profiles are computed from the expression values on all arrays for a given gene in different groups of arrays, and they are summarized in a line plot connecting the median of each group. Each row of Figure S1 contains the gene expression profiles of genes lying in the same directions of the KPCA-Biplot. In order to explore different profiles, we have selected three directions as an example. Genes T95018, T61602, T58861, T48804 and T57633 (the top row) have higher expression in tumor samples. In contrast, genes R78934, T92451, M33680 and T42-control (the bottom row), are most strongly expressed in normal cells. Genes T51560, T49703, U05012, T51496 and R44112 (the middle row) are equally expressed in both samples.

These results are consistent with those reported by Alon *et al* (*20*), in which the analysis of the same dataset by using two-way data clustering reveals a group of genes whose expression is correlated across tissue types. In particular, it reports a list of 48 ESTs homologous to ribosomal proteins within the set of 2,000 genes used for the clustering. The intensity of the ribosomal protein genes is relatively low in normal colon tissues and high in the colon tumor tissues.

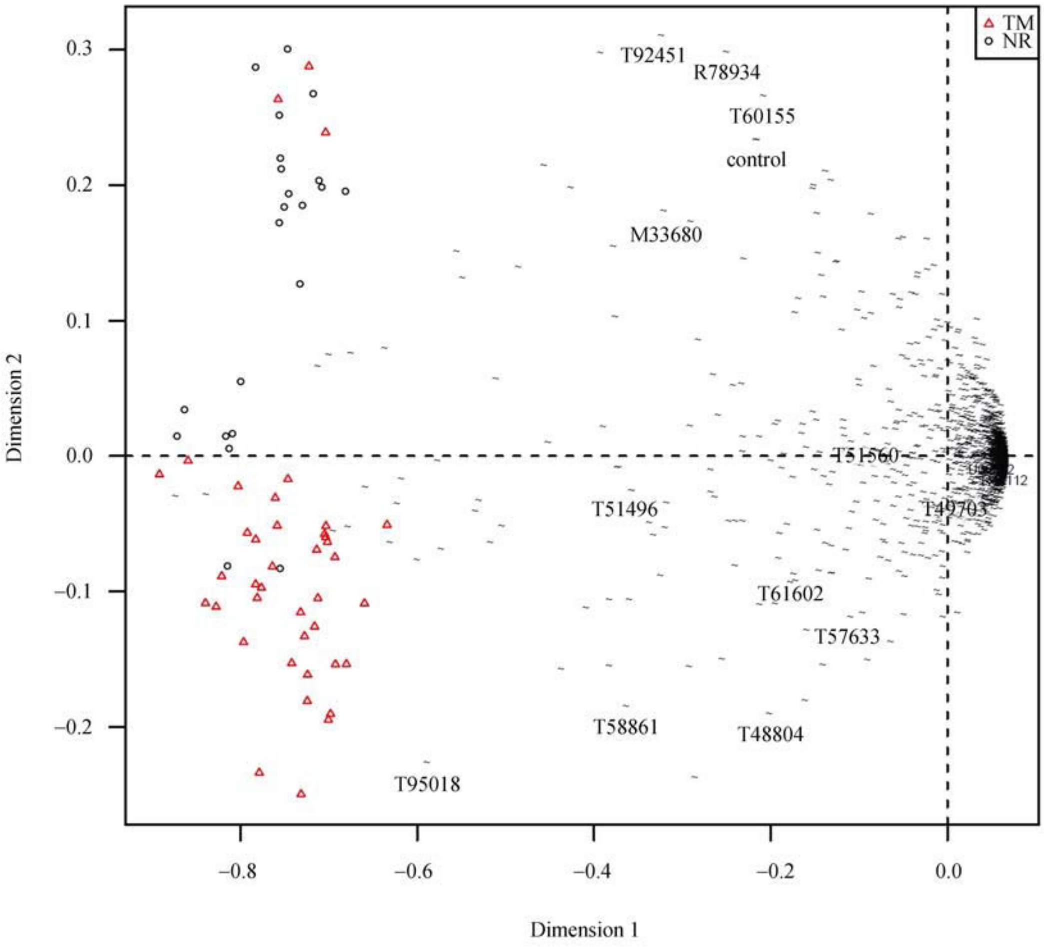

Figure 1 KPCA-Biplot of tumor colon dataset. Microarrays are marked according to the class of tissue (see plot legend) and genes are shown as a tilde. Some differentially expressed genes are labeled with their numbers.

We remark that genes T95018, T61602, T58861, T48804 and T57633, which are highlighted by our method, belong to the list of 48 ESTs homologous to ribosomal proteins reported by Alon *et al* (*20*).

Figure 2 shows the standard biplot representation obtained in analyzing the same dataset. We observed that the sample groups are separated mainly along the second axis but the separation is not completed. In this example, with respect to the sample representation, both the standard biplot and KPCA-Biplot detect the group structure in reduced dimension but the two groups are not fully separated by either method. With respect to the gene representation, both the standard biplot and our method reveal the normal and tumor profiles but we observed that with the KPCA-Biplot the genes are more clearly separated, which can be helpful for mining the microarray data.

Leukemia dataset

The leukemia dataset is composed of 3,051 gene expressions in three classes of leukemia: 19 cases of B-cell acute lymphoblastic leukemia (ALL), 8 cases of T-cell ALL and 11 cases of acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays. The data can be downloaded from http://www.genome.wi.mit.edu.

The data were preprocessed according to the protocol described by Dudoit *et al* (*23*). In addition, we complete the preprocessing of the gene expression data with a microarray standardization and gene centring.

We performed the KPCA-Biplot procedure on this dataset as detailed in the previous section. Initially, we factorize the input data matrix using SVD with

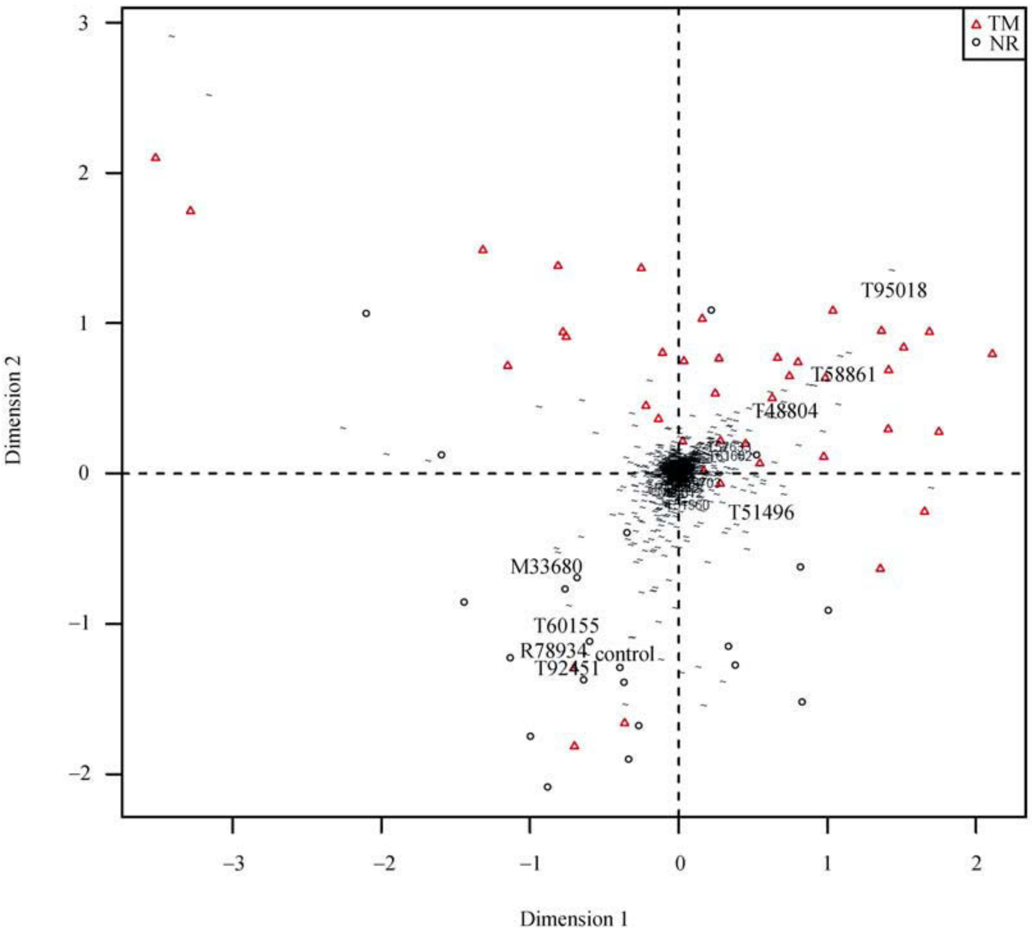

Figure 2 Standard Biplot of tumor colon dataset.

α=0.5. Next, we compute the kernel matrix in Step 2 using the radial basis kernel with c=0.001, which is set heuristically. In this step we also use the polynomial kernel but its behavior, in this case, is similar to that of the radial basis kernel. For brevity we only show results obtained with the radial basis kernel. The resulting biplot is given in **Figure 3**. It shows the projection onto the two leading kernel principal components of genes and microarrays, which were obtained by using Formulas 3 and 4, respectively. Genes are shown as a tilde and microarrays are shown by different marks according to the class of leukemia to which they belong. The first axis separates ALL (B and T-cells) and AML samples. In particular, the 2D plot detects the three classes of leukemia.

Additionally, in Figure 3 some genes are labeled with their probe IDs and these genes have been chosen to illustrate different profiles. It is expected that genes that lie towards the right will have higher expression values in the AML samples than in the ALL samples. In contrast, genes on the left will have higher expression values in the ALL samples than in the AML samples. Genes placed near the bottom will tend to have higher expression values in T-cells than in B-cells.

Figure S2 shows the gene expression profiles of some genes labeled in Figure 3. The gene expression profiles are computed from the expression values on all arrays for a given gene in different groups of arrays and they are summarized in a line plot connecting the median of each group. Each row of Figure S2 contains the gene expression profiles of genes lying in the same directions of the KPCA-Biplot. In order to explore different profiles, we have selected three directions as an example. It is observed that genes lying close to any group of leukemia have higher values in that group than in other groups, and genes towards the opposite group have lower values in that group than in other groups.

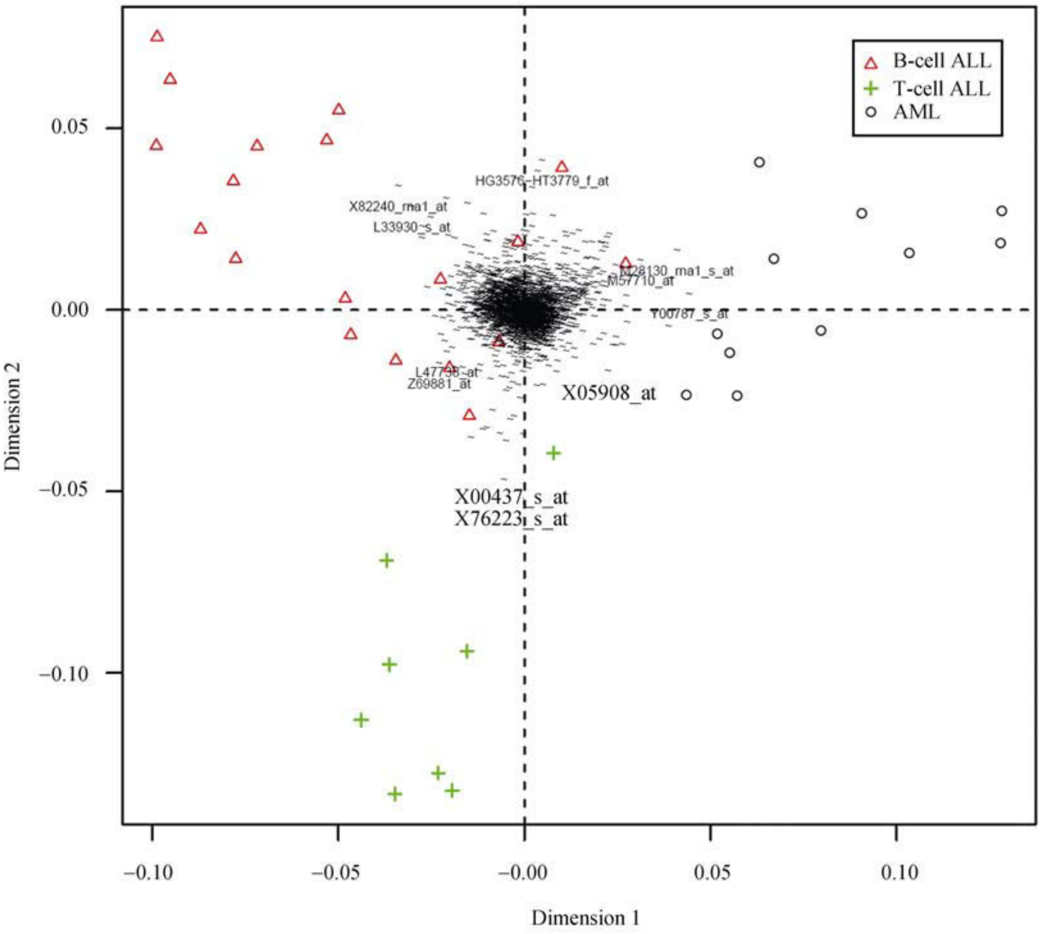

Figure 3 KPCA-Biplot of leukemia dataset.

Genes M57710_at, M28130_rna1_s_at and Y00787_s_at are highly expressed on AML samples and they have a similar profile as shown in Figure S2. In the opposite direction we can find genes L47738_at and Z69881_at, which are down-regulated in AML samples. As expected, these genes have the opposite profile of expression.

The expression of the gene HG3576-HT3779_f_at is highest in AML and ALL-B cell samples but lowest in ALL T-cell samples. In contrast, genes X76223_s_at and X00437_s_at are most strongly expressed in ALL T-cell samples. The shape of this set of genes is opposite to that of HG3576-HT3779_f_at.

Genes X82240_rna1_at and L33930_s_at are highly expressed on ALL-B cell samples. In the opposite direction we can find the gene X05908_at, which is highly expressed in AML and ALL T-cells but expressed most weakly in ALL-B cells.

These results indicate that our method provides a useful tool to find genes with different expression. Our method is complementary to other current methods to detect genes with different expression. For example, genes Z69881_at and M28130_rna1_s_at are 2 of the 50 genes selected in Golub's study (*21*) to differentiate AML from ALL cells. Genes HG3576-HT3779_f_at, X76223_s_at and X82240_rna1_at are also selected in Pittelkow and Wilson's study (*10*).

Lymphoma dataset

The lymphoma dataset comes from a study of gene expression of three prevalent lymphoid malignancies: B-cell chronic lymphocytic leukemia (B-CLL), follicular lymphoma (FL) and diffuse large B-cell lymphoma (DLCL). Among 96 samples we took 62 samples containging 4,026 genes in three classes: 11 cases of B-CLL, 9 cases of FL and 42 cases of DLCL. Gene expression levels were measured using two-channel cDNA microarrays. The data can be obtained from http://genome-www.stanford.edu/lymphoma.

After preprocessing, all gene expression profiles were base-10 log-transformed and standardized to zero mean and unit variance in order to prevent single arrays from dominating the analysis. Finally, we complete the preprocessing of the gene expression data with gene centring.

To perform the KPCA-Biplot procedure, we factorize the input data matrix using SVD with α=0.5. Next, we compute the kernel matrix using the radial basis kernel with c=0.01. In this step we use alternatively the polynomial kernel, but it yields the poorest representation of the dataset.

The resulting biplot is given in **Figure 4**. It shows the projection onto the two leading kernel principal components of genes and microarrays, which is obtained by using the steps of kernel selection and automatic tuning. Genes are shown as a tilde and microarrays are shown by different marks according to the type of lymphoma to which they belong. The three classes are clearly separated. The first axis separates DLCL cells and the others, and the second axis separates FL and B-CLL cells.

As we can see in Figure 4, some genes labeled with their clone IDs have been selected to visualize different profiles in this example. It is expected that genes lying towards the right will have higher values for DLCL samples than for FL and B-CLL samples. Genes placed near the top will tend to have higher expression values in FL samples than in B-CLL samples.

Figure S3 shows the profiles of some genes that have been labeled in Figure 4. Each row of Figure S3

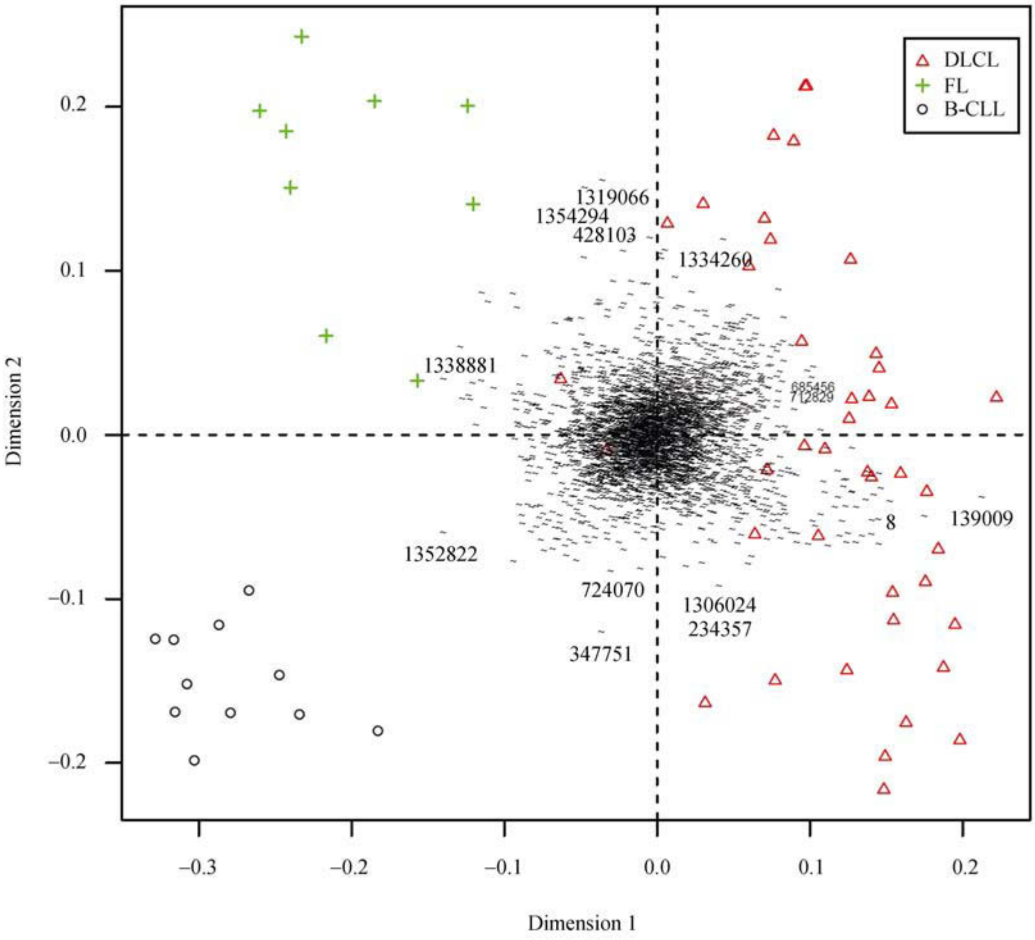

Figure 4 KPCA-Biplot of lymphoma dataset.

contains the gene expression profiles of genes lying in the same directions of the KPCA-Biplot. In order to explore different profiles, we have selected four directions as an example. It is observed that genes lying close to any group of lymphoma are up-regulated in that group, and genes lying opposite a group are down-regulated.

Genes 1319066, 1354294 and 428103 show an expression profile with high values in FL samples and low values in DLCL and B-CLL sample cells. In contrast, genes 234357 and 1306024 have opposite profile of expression.

Genes 347751 and 724070 are highly expressed in B-CLL samples and weakly expressed in FL and DLCL samples. In the opposite direction we can find gene 1334260, which is down-regulated in B-CLL samples and up-regulated in the others.

Genes 139009 and 8 are highly expressed in DLCL samples. In contrast, gene 1338881 has the reverse shape. Finally, we observed gene 1352822, which has a linear profile, and genes 685456 and 712829, lying in the opposite direction, also with linear profile but with the opposing slopes.

Conclusion

In this paper we propose an exploratory method, called KPCA-Biplot, that combines KPCA and SVD-Biplot for elucidating relationships between samples (microarrays) and variables (genes). The main properties of KPCA-Biplot are the extraction of nonlinear features together with the preservation of the input variables (genes) in the output display. The method described here is easy to implement and facilitates the identification of genes that have similar or reversed profiles. Our results indicate that KPCA-Biplot is complementary to other tools currently used for finding gene expression profiles, with the advantage that it can capture the usual nonlinear nature of microarray data.

Acknowledgements

We thank the anonymous reviewers for many suggestions. This research was funded in part by grant MEC-MTM2008-00642.

Authors' contributions

FR was responsible for the development of the procedure, suggested the combination of KPCA and SVD-Biplot, and was involved in critical revision of the manuscript. EV implemented the approach, coded the procedures and drafted the manuscript. PS prepared analysis and results. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Chambers, J.M., *et al*. 1983. *Graphical Methods for Data Analysis*. Duxbury Press, Belmont, USA.

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

3 Tamayo, P., *et al*. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96: 2907-2912.

4 Chu, W., *et al*. 2005. Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatcis* 21: 3385-3393.

5 Zhao, X. and Cheung, L.W. 2007. Kernel-imbedded Gaussian processes for disease classification using microarray gene expression data. *BMC Bioinformatcis* 8: 67.

6 Dettling, M. 2004. BagBoosting for tumor classification with gene expression data. *Bioinformatcis* 20: 3583-3593.

7 Diaz-Uriarte, R. and Alvarez de Andres, S. 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatcis* 7: 3.

8 Alter, O., *et al*. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* 97: 10101-10106.

9 Fellenberg, K., *et al*. 2001. Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. USA* 98: 10781-10786.

10 Pittelkow, Y.E. and Wilson, S.R. 2003. Visualisation of gene expression data—the GE-biplot, the Chip-plot and the Gene-plot. *Stat. Appl. Genet. Mol. Biol.* 2: Article 6.

11 Park, M., *et al*. 2008. Several biplot methods applied to gene expression data. *J. Stat. Plan. Inference* 138: 500-515.

12 Schölkopf, B., *et al*. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* 10: 1299-1319.

13 Shawe-Taylor, J. and Cristianini, N. 2004. *Kernel Meth-*

ods for Pattern Analysis. Cambridge University Press, UK.

14 Schölkopf, B. and Smola, A.J. 2002. *Learning with Kernels—Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, USA.

15 Quackenbush, J. 2002. Microarray data normalization and transformation. *Nat. Genet.* 32: S496-501.

16 Elizondo, D.A., *et al*. 2008. Dimensionality reduction and microarray data. In *Principal Manifold for Data Visualization and Dimension Reduction* (eds. Gorban, A.N., *et al*.), pp. 293-308. Springer, Berlin, Germany.

17 Diamantaras, K.I. and Kung, S.Y. 1996. *Principal Component Neural Networks*. Wiley, New York, USA.

18 Gabriel, K.R. 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58: 453-467.

19 Gower, J.C. and Hand, D.J. 1996. *Biplots*. Chapman and Hall, London, UK.

20 Alon, U., *et al*. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96: 6745-6750.

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

22 Alizadeh, A.A., *et al*. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403: 503-511.

23 Dudoit, S., *et al*. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Assoc.* 97: 77-87.

Supplementary Material

Figures S1-S3

DOI: 10.1016/S1672-0229(10)60022-8