TECHNICAL

REPORT

0237

Graphical exploration of gene expression data: a comparative study of three multivariate methods

L. Wouters, H.W. Göhlmann, L. Bijnens, S.U. Kass, G. Molenberghs, and P. Lewi

IAP STATISTICS

NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Graphical Exploration of Gene Expression Data:

A Comparative Study of Three Multivariate Methods

Luc Wouters[1], Hinrich W. Göhlmann[2], Luc Bijnens[3], Stefan U. Kass[2], Geert Molenberghs[1], Paul J. Lewi[4]

[1]Center for Statistics, Limburgs Universitair Centrum, transnationale Universiteit Limburg, Hasselt, Belgium

[2,3,4] Departments of Genomic Technologies[2], Global Biostatistics and Reporting[3], and Center for Molecular Design[4], Johnson & Johnson Pharmaceutical Research & Development, a division of Janssen Pharmaceutica NV, Beerse, Belgium

email: lwouters@janbe.jnj.com

Running title: Graphical Exploration of Gene Expression Data.

SUMMARY. Three methods of multivariate data analysis are compared for their ability to identify clusters of hybridizations and genes in microarray experiments. Principal component analysis (PCA) and correspondence factor analysis (CFA) have been applied already to microarray data. It is shown that PCA has the disadvantage that the resulting principal factors are not very informative, while CFA has difficulties regarding interpretation of the distances between objects. We present an alternative method, spectral map analysis (SMA) and compare it with the other methods using gene expression data in leukemia patients. It is shown that weighted SMA outperforms PCA and is at least as powerful as CFA in finding clusters in the hybridizations and identifying genes related to these clusters. SMA addresses the problem in a more appropriate manner than CFA and allows to apply a more flexible weighting to the genes and hybridizations. Proper weighting is important since it allows to downweight less reliable data and to emphasize more reliable information.

KEY WORDS: Bioinformatics; Biplot; Correspondence factor analysis; Data mining; Data visualization; Gene expression data; Microarray data; Multivariate exploratory data analysis; Principal Component Analysis; Spectral map analysis.

1. Introduction

The advent of DNA microarray technology enabling global gene expression analysis has been a fundamental breakthrough in the life sciences. The possibility to simultaneously measure the expression profile of thousands of genes allows for a better characterization of different types of a disease and for better insight in the underlying pathology, thus creating the possibility for identifying new therapeutic targets. In principle, DNA microarrays consist of some solid material upon which an array of spots of known DNA sequences are immobilized. Labeled RNA extracted from biological samples, referred to as hybridizations, is hybridized to the array. The array is scanned and the fluorescent intensity at each position in the array is considered as a measure for the expression level of the corresponding gene. At present, a typical DNA microarray contains thousands of DNA-spots. In the near future, however, improvements to the technology will probably provide information on tens of thousands of genes, eventually encompassing entire genomes.

On the other hand, the simultaneous measurement of the expression level of thousands of genes poses an enormous task to the information processing capability of present systems. Much research is still being done in the area of statistics and data mining to provide the scientific community with better tools for pattern recognition and visualisation of gene expression data. Statistical science has made significant contributions to the enhancement of the quality of raw measurements by introducing several normalisation techniques. In addition, new techniques for statistical inference have been introduced and data mining techniques for supervised and unsupervised learning have found applications. Methods of unsupervised learning clustering techniques,such as k-means clustering (Tavazoie, 1999), hierarchical clustering (Eisen, et al. 1998), and self-organizing maps (Tamayo, et al., 1999, Törönen, et al. 1999) have found widespread application in analysing and visualising gene expression data. These methods however, produce results that are highly dependent on the distance-measure and clustering technique that is used and the number of clusters in a cluster analysis can is often an issue of controversy. Furthermore, conventional clustering methods only allow for classification of either genes or hybridizations alone, but do not allow interpretations of the association between genes and hybridizations.

Another set of exploratory techniques is based upon projections of high-dimensional data in a lower dimensional space and plotting both genes and hybridizations in this lower dimensional space using the biplot (Gabriel, 1971). Principal component analysis (PCA) (Pearson, 1901, Hotelling 1933) is a well-established technique in multivariate statistics and has been applied to gene expression data (Chapman, et al., 2001, Hilsenbeck et al., 1999, Landgrebe, et al., 2002, Lefkovits, et al., 1988). Related to PCA are techniques like correspondence factor analysis (Benzécri, 1973) and spectral mapping (Lewi, 1976). Correspondence factor analysis (CFA) has recently been applied to microarray data by Fellenberg, et al. (2001). In this paper, we propose the use of a less well known technique, spectral map analysis (SMA) (Lewi, 1976) for the analysis of gene expression data. In the past, SMA has been successfully applied to a wide variety of problems ranging from pharmacology (Lewi, 1976), virology (Andries, et al. 1990), to management and marketing research (Faes and Lewi, 1987). Thielemans, Lewi, and Massart (1988) have compared SMA with PCA and CFA, using a relatively small data set from the field of epidemiology. They concluded that the appropriate technique depends upon the data to be analyzed and the features one is interested in. Up to now, the applications of SMA have always been limited to small or moderate sized data sets. The present paper illustrates the applicability of the method to large data sets and the importance of appropriate weighting in the analysis of microarray gene expression data. We will show that SMA provides the researcher with a visual data representation , useful as a tool for distinguishing patterns in the gene expression data that could be related to important biological questions and as a technique for quality control of microarray experiments.

The outline of this paper is as follows: In Section 2 a general framework for multivariate projection methods will be set up and the similarities and specifics of PCA, CFA, and SMA will be indicated. In Section 3, the different techniques will be compared using the gene expression profiles of leukemia patients (Golub, 1999). In Section 4, the advantages of weighted SMA for gene expression data will be highlighted and possible applications and limitations of the technique will be discussed.

2. Multivariate Projection Methods

The similarities and characteristics of the three multivariate projection methods, PCA, CFA, and SMA will be presented, following Lewi (1995) and Thielemans, et al. (1988).

2.1 Notation

Let $\mathbf{X}_{n\times p}$ denote the matrix containing the original expression levels x_{ij} for the expression level of *n* genes (rows) in each of *p* different hybridizations (columns). We also define two diagonal matrices with row weights $\mathbf{W}_n$ and column weights $\mathbf{W}_p$. The diagonal elements of $\mathbf{W}_n$ and $\mathbf{W}_p$ are the weight coefficients associated with the rows and columns of the matrix $\mathbf{X}$. The weight coefficients are non-negative and normalized to unit sum. An unweighted analysis is obtained by $\mathbf{W}_n = \mathrm{diag}\left(1/n\right)$ and $\mathbf{W}_p = \mathrm{diag}\left(1/p\right)$. Alternatively, the diagonal elements of $\mathbf{W}_n$ and $\mathbf{W}_p$ can be set to appropriate weighting schemes such as the row and column totals, normalized to unity, i.e. $\mathbf{W}_n = \mathrm{diag}\left(\mathbf{X1}_p \big/ \mathbf{1}_n^{\mathrm{T}}\mathbf{X1}_p\right)$ and $\mathbf{W}_p = \mathrm{diag}\left(\mathbf{1}_n^{\mathrm{T}}\mathbf{X} \big/ \mathbf{1}_n^{\mathrm{T}}\mathbf{X1}_p\right)$. There seems to be a consensus among scientists that microarray data at lower levels of expression are less reliable, so weighing for row means seems appropriate in this context. An additional advantage of defining weights is the possibility of positioning rows and columns by setting their corresponding weights to zero. Positioning is the operation where some columns or rows of the data matrix are excluded from the actual analysis, but still are represented on the map constructed on the basis of the remaining data.

2.2 General algorithm for multivariate projection methods

In the algorithms of the three multivariate projection methods the following building blocks can be distinguished: re-expression, closure, centering, normalization, factorization, and finally projection. Differences between the methods are obtained by variations in these building blocks.

a. Re-expression

It is often advantageous to re-express (i.e., transform) the data as logarithms, i.e. a new matrix $\mathbf{A}$ is obtained whose elements $a_{ij} = \log(x_{ij})$. For this operation to be valid, measurements must be made on a ratio scale and the values must be positive. Logarithmic re-expression corrects for positive skewness and reduces the effect of large influential values. A further justification of a logarithmic re-expression is the fact that in many natural systems changes occur on a multiplicative rather than an additive scale. Alternatively, one could also consider other types of re-expressions, such as reciprocals or arc sine re-expression. However, these do not possess the nice properties of logarithms, namely that differences in logarithms are related to ratios of the original data. There is also the trivial case in which the original data are left unchanged and the elements of the re-expressed matrix $\mathbf{A}$ are equal to the elements of the data matrix $\mathbf{X}$.

b. Closure

Closure is defined as the operation of transforming the data into relative values such that they sum to unity. Closure requires the data to be non-negative and measured in the same units. From the matrix $\mathbf{A}$ with the re-expressed data, a new table $\mathbf{B}$ is obtained by either column, row, or global closure. In column-closure each element a_{ij} of $\mathbf{A}$ is divided by the corresponding column marginal total of $\mathbf{A}$, i.e. $b_{ij} = {}^{a_{ij}}\!/_{a_{+j}}$, where $a_{+j} = \sum_{i=1}^{n} a_{ij}$. Column-closure imposes a linear constrained on the rows of the matrix. As a consequence, when $n \leq p$, it reduces the rank of the data matrix by one. In row-closure the elements of the matrix $\mathbf{B}$ are obtained from $\mathbf{A}$ by dividing each element by the corresponding row marginal total, i.e. $b_{ij} = {a_{ij}}\!\big/\!{a_{i+}}$, where $a_{i+} = \sum_{j=1}^{p} a_{ij}$. A linear constraint is imposed on the columns of the matrix, resulting in a rank-reduction by one when $p \leq n$. Double closure consists of the combined operation of dividing each element a_{ij} of the data matrix $\mathbf{A}$ by its corresponding row and column marginal total. The result is then multiplied by the total sum of $\mathbf{A}$ to yield a dimensionless number. We thus have: $b_{ij} = \frac{a_{ij} a_{++}}{a_{i+} a_{+j}}$, where $a_{++} = \sum_{j=1}^{p} \sum_{i=1}^{n} a_{ij}$. Double closure always involves a reduction of the rank of the original data matrix by one. The operation of double closure combined with weighting of rows and columns by their corresponding marginal totals forms the core of CFA. Of course, in an algorithmic approach one should also consider the trivial case of no closure in which $b_{ij} = a_{ij}$

c. Centering

Centering is defined as a correction of $\mathbf{B}$ for a mean value to yield the centered matrix $\mathbf{Y}$. There are different ways to derive mean values from a matrix, each resulting in a different way to center the data. Geometrically, centering involves a translation to the origin of the data in the column space, the rowspace, or in both. In column centering the matrix $\mathbf{Y} = \mathbf{B} - \mathbf{1}_n \mathbf{m}_p^{\mathrm{T}}$ contains deviations from the weighted column means $\mathbf{m}_p^{\mathrm{T}} = \mathbf{1}_n^{\mathrm{T}} \mathbf{W}_n^{\mathrm{T}} \mathbf{B}$. In row centering $\mathbf{Y} = \mathbf{B} - \mathbf{m}_n \mathbf{1}_p^{\mathrm{T}}$ is the matrix with deviations from the weighted row means $\mathbf{m}_n^{\mathrm{T}} = \mathbf{B} \mathbf{W}_p \mathbf{1}_p$. In global centering the matrix $\mathbf{Y} = \mathbf{B} - m$ contains the deviations of the elements $\mathbf{B}$ from the global weighted mean $m = \mathbf{1}_n^{\mathrm{T}} \mathbf{W}_n^{\mathrm{T}} \mathbf{B} \mathbf{W}_p \mathbf{1}_p$. Simultaneous centering by rows and columns yields the double-centered matrix of deviations from row and column means $\mathbf{Y} = \mathbf{B} - \mathbf{1}_n \mathbf{m}_p^{\mathrm{T}} - \mathbf{m}_n \mathbf{1}_p^{\mathrm{T}} + m \mathbf{1}_n \mathbf{1}_p^{\mathrm{T}}$. The operation of double-centering involves a projection of the data matrix on a hyperplane that runs through the origin and is orthogonal to the line of identity. The result is a reduction by one of the rank of the original matrix. The dimension that is lost is related to a component of “size” that is present in the data and is oriented along the identity line. This “size” component is common to all elements of the data table and often obscures the important information that is present in the data. Applying double-centering after logarithmic re-expression is the very essence of SMA. It is interesting to note the close analogy between double closure and double centering after logarithmic re-expression. For the centering part of the algorithm, we also define the trivial case of no centering with $\mathbf{Y} = \mathbf{B}$.

d. Normalization

Normalization or standardization is the operation of dividing $\mathbf{Y}$ by the square root of the mean sums of squares or norm, yielding a normalized matrix $\mathbf{Z}$. There are several ways to compute the norm of a matrix each resulting in a different method of normalization. In column-normalization the normalized results is obtained as $\mathbf{Z} = \mathbf{Y}\mathbf{D}_p^{-1}$, with the weighted column-norm $\mathbf{D}_p$ defined as $\mathbf{D}_p = \mathrm{diag}\left(\left(\mathbf{Y}^{\mathrm{T}}\right)^2 \mathbf{W}_n \mathbf{1}_n\right)^{1/2}$.

The effect of column-normalization in the column space is to weight each column-dimension proportional to the inverse of its mean sum of squares. In row-space the effect is a sphericization, such that the points are forced to lie on a hypersphere. Column-normalization after column-centering is a standard operation in PCA. In row-normalization $\mathbf{Z} = \mathbf{D}_n^{-1}\mathbf{Y}$ with the weighted row-norm $\mathbf{D}_n = \mathrm{diag}\left(\mathbf{Y}^2 \mathbf{W}_p \mathbf{1}_p\right)^{1/2}$. The geometric interpretation of row-normalization is similar to that of column-normalization with the row and column spaces interchanged. Normalization for the weighted global norm $d = \mathbf{1}_n \mathbf{W}_n \mathbf{Y}^2 \mathbf{W}_p \mathbf{1}_p$ yields the global-normalized matrix

$\mathbf{Z} = \frac{1}{d}\mathbf{Y}$. Finally, for the sake of completeness, we have the case of no normalization where $\mathbf{Z} = \mathbf{Y}$.

e. Factorization

Factorization of $\mathbf{Z}$ yields factors that are orthogonal to one another and account for a maximum of the variance of the data. For a weighted analysis, the multivariate projection methods under consideration rely on the generalized singular value decomposition as factorization method. The generalized singular value decomposition of $\mathbf{Z}$ is defined as:

$$\mathbf{W}_n^{1/2}\mathbf{Z}\mathbf{W}_p^{1/2} = \mathbf{U\Lambda V}^{\mathrm{T}} \quad [1]$$

where $\mathbf{\Lambda}$ is an $r \times r$ matrix of singular values, r being the rank of $\mathbf{W}_n^{1/2}\mathbf{Z}\mathbf{W}_p^{1/2}$. In addition, we have $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_r$ and $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_r$. Consequently, we have $\left(\mathbf{W}_n^{-1/2}\mathbf{U}\right)^{\mathrm{T}} \mathbf{W}_n \left(\mathbf{W}_n^{-1/2}\mathbf{U}\right) = \mathbf{I}_r$ and $\left(\mathbf{W}_p^{-1/2}\mathbf{V}\right)^{\mathrm{T}} \mathbf{W}_p \left(\mathbf{W}_p^{-1/2}\mathbf{V}\right) = \mathbf{I}_r$.

f. Projection

Projection of the results of the generalized singular value decomposition along the first few common factors yields the biplot (Gabriel, 1971). Different biplots, with characteristic geometric properties, can be constructed using combinations of two factor-scaling coefficients α and β, set to, for example, 0, 0.5, or 1, where the weighted factor scores $\mathbf{S}$ and factor loadings $\mathbf{L}$ are obtained from [1] by $\mathbf{S} = \mathbf{W}_n^{-1/2}\mathbf{U\Lambda}^{\alpha}$ and $\mathbf{L} = \mathbf{W}_p^{-1/2}\mathbf{V\Lambda}^{\beta}$. It is easy to show that the above expressions for $\mathbf{S}$ and $\mathbf{L}$ can also be written as $\mathbf{S} = \mathbf{Z}\mathbf{W}_p^{1/2}\mathbf{V\Lambda}^{\alpha-1}$ and $\mathbf{L} = \mathbf{Z}\mathbf{W}_n^{1/2}\mathbf{U\Lambda}^{\beta-1}$. The latter form, though more complex, is required for positioning supplementary rows or columns by setting their respective weights to zero. The following cases of factor scaling can be distinguished:

- $\alpha = 1$, $\beta = 1$ referred to as eigenvector scaling. This type of symmetric scaling is customary in CFA. Distances of points in row-space as well as in column-space are reproduced in the plot, as well as the correlation structure of the column-variables.
- $\alpha = 1$, $\beta = 0$ referred to as unit column-variance scaling, is customary in PCA. Only distances between row-points are preserved in this asymmetric type of scaling. In full factor-space the distances of the column-items from the origin are constant and the correlation structure between column-variables is not reproduced.
- $\alpha = 0$, $\beta = 1$ referred to as unit row-variance scaling, is also customary in PCA. Only distances between the column-items and the correlation structure between column-variables are preserved. In full factor-space the distances of the row-points from the origin are constant.
- $\alpha = 0.5$, $\beta = 0.5$ referred to as singular value scaling, is customary in SMA. This type of factor-scaling is a compromise between the versions given above. Distances between row-points and the correlation structure of the column-variables are not fully reproduced. The distortion is most pronounced when the ratios between the eigenvalues $\left(\mathbf{\Lambda}^2\right)$ associated with the axes of the biplot are very large or very small.

Having defined a general framework encompassing the three projection methods, we will now discuss their different characteristics.

2.1 Principal component analysis

Historically, PCA dates back to Pearson (1901) and Hotelling (1933). In the algorithm described above it is defined as: constant weighting of rows and columns, optional re-expression, column-centering, column-normalization, and factor scaling with either symmetric eigenvector scaling with $\alpha = 1$, $\beta = 1$, asymmetric unit column-variances with $\alpha = 1$, $\beta = 0$, or asymmetric unit row-variances with $\alpha = 0$, $\beta = 1$.

Note that PCA makes a clear distinction between row- and column-items in the centering and normalization procedure. Therefore, one distinguishes classical R-mode analysis and its complement Q-mode analysis on the transposed data matrix.

2.2 Correspondence factor analysis

CFA has been developed by Benzécri (1973) and is adequately described by Greenacre (1984). This multivariate projection method was originally developed for the analysis of contingency tables but has also been applied to other tables with non-negative values (Fellenberg, et al., 2001). CFA involves the following steps: weighting of rows and columns by marginal row and column totals, no re-expression, double-closure, double-centering, global normalization, and symmetric eigenvector factor-scaling ($\alpha = 1$, $\beta = 1$). The double-closure and double-centering transformations are symmetric with respect to the rows and columns of the data table. In the CFA-biplot distances of the row- and column-items from the centre of the biplot are interpreted as chi-square values.

2.3 Spectral map analysis

SMA was originally developed for the display of activity spectra of chemical compounds (Lewi, 1976). The algorithm for spectral mapping is characterized by: constant weighting of rows and columns or weighting by some properly chosen weighting factor, logarithmic re-expression, double-centering, global normalization, and factor scaling using either symmetric scaling with singular values ($\alpha = 0.5$, $\beta = 0.5$) or asymmetric scaling with unit-column variance ($\alpha = 1$, $\beta = 0$). A further characteristic of SMA is that in the biplot the areas of the symbols are made proportional to a selected column, or to marginal row- and column-totals.

The double-centering transformation in SMA is symmetric with respect to the rows and columns of the data table. As a result of the double-centering, all absolute aspects of the data are removed. What remains are contrasts between the different rows (genes) and contrasts between the different columns (hybridizations) of the data table. These contrasts can be expressed as ratios due to the logarithmic transformation. The contrasts can be understood as specificities of the different genes for the different hybridizations. Conversely, they refer also to the specificities or preferences of the different hybridizations for some of the genes. Therefore, one could state that SMA provides a visualisation of the interactions between genes and hybridizations. An advantage of SMA over CFA is that the scope of SMA is not limited to contingency tables and cross-tabulations. In addition, SMA offers the possibility to use other weighting factors than the marginal totals.

2.4 Implementation

The general algorithm as described above has been implemented in the open source language R. Analysis of a 5000 x 40 data matrix takes about 20 secs on a 750 MHz Intel processor with 500 Mbyte RAM. The library with functions for analysis, plotting, and printing is freely available at http:\\www.xxxxxxx.xxx.

3. Application

In a recent study, Golub et al. (1999) obtained gene expression profiles of 38 patients suffering from acute leukemia. In the following, we will refer to this data set as MIT1. Patients were diagnosed as suffering from either acute myeloid leukemia (18 patients) or acute lymphoblastic leukemia (20 patients). The latter class could further be subdivided in B-cell and T-cell classes. In addition to the initial 38 patients Golub, et al. also considered a second validation sample (MIT2) of 34 patient for which the gene expression profile was determined. Both data sets are available on http://www-genome.wi.mit.edu/mpr/data_set_ALL_AML. The original data were preprocessed as follows: genes that were called "absent" in all hybridizations were removed from the data sets, since these measurements are considered unreliable by the manufacturer of the technology. Negative measurements that were present in the data were set to 1.. The resulting data set contained 5327 genes of the 6817 originally reported by Golub and co-workers.

3.1 Principal component analysis

PCA was carried out after logarithmic re-expression of the gene expression profiles in MIT1. Since gene expression data are positively skewed and can contain large influential values, we considered a logarithmic re-expression appropriate. For the construction of the biplot (Figure 1), an asymmetric scaling with unit-column variance ($\alpha = 1$, $\beta = 0$) was used to allow better visual discrimination between the different hybridizations. This special type of factor scaling was considered optimal for extreme rectangular matrices of microarray data where variability between the genes (average variance log transformed data = 6.4) is much higher than between the different hybridizations (average variance = 2). A consequence of unit-column variance factor scaling is that correlations and distances between hybridizations are not represented in the biplot. However, in exploring gene expression data only patterns in the distribution of the hybridizations are of direct interest. In Figure 1, the horizontal axis of the biplot, represents the first principal component that accounts for 71 % of the total variance in the data. The second principal component is represented by the vertical axis of the biplot and explains another 3 % of the total variance. The horizontal axis is dominated entirely by a global component related to the size of the measurements and does not contribute any information about the differential expression of genes in the hybridizations. Differences between hybridizations are found only along the vertical axis. Only a difference between the ALL and AML groups is eminent, while data from ALL B-lineage and ALL T-lineage completely overlap one another. Furthermore, it is impossible to use the biplot for selecting genes that discriminate best between the ALL and AML classes.

[GEERT] IK ZOU OP DEZE PLAATS EEN IETS MEER EXPLICIETE CONCLUSIE TOEVOEGEN. WAT IS DE LES? HOE GAAN WE MET DEZE BOODSCHAP VERDER? DE LEZER ZOU HIER ANDERS WAT IN HET ONGEWISSE GELATEN ZIJN.

[GEERT] ALLE PLOTS ZIJN 2-DIMENSIONAAL; DE VRAAG KAN ONTSTAAN: IS HET WERKELIJK ZO DAT HET GEKOZEN AANTAL FACTOREN ALTIJD 2 IS, IS DAT GEWOON UIT GRAFISCH GEMAK, OF IS ER EEN ANDERE REDEN? DIT VERDIENT ENEIGE DISCUSSIE

3.2 Correspondence factor analysis

The biplot obtained from CFA on the original data in MIT1 is depicted in Figure 2.. The same asymmetric unit-column variance scaling was used as in PCA, to allow optimal visual discrimination of the different hybridizations. While distances between hybridizations are not represented in this type of scaling, the weighted distances of genes from the center are interpreted as chi-square values. In CFA sums of squares are expressed as chi-square values and the global weighted sum of squares is defined as the global chi-square. The horizontal axis of the biplot in Figure 2 accounts for 17 % of the global chi-square, while the vertical axis accounts for an additional 10 %. In contrast to PCA the first dominant component is not related to size. CFA highlights the differential genetic profiles of the different hybridizations, an approach that is much more relevant to the problem. In Figure 2, genes are distributed in a funnel-like pattern and there is a clear separation between ALL and AML patients with only 2 patients that overlap one another. In contrast to PCA, B-lineage and T-lineage classes within the ALL group are also separated from one another. It is possible to identify a few genes that could be used in characterizing the three pathological classes. However, there is a problem with the interpretation of the numerical value of the distances between genes. Since in CFA, distances refer to chi-square values that have a meaning only for contingency tables and not for continuous data as is the case in gene expression experiments, one could seriously question the applicability of CFA in microarray data analysis.

3.3 Spectral map analysis

In SMA, we considered both constant weighting and variable weighting proportional to the row marginal totals. The latter was motivated by the fact that differences found at lower levels of gene expression are less reliable than differences at a high level of expression.

a. Unweighted SMA

The results of SMA with constant weighting factors are depicted in Figure 3. Asymmetric unit column-variance was used as factor scaling in the construction of the biplot. Genes located near the center of the map are still displayed as dots, while the 0.5 % (27) most distal genes are displayed as circles with areas proportional to their marginal row mean. In addition, these genes were labelled with their accession number. SMA, like CFA, stresses the differential genetic profile of the different hybridizations, but in contrast to CFA relative distances can be interpreted and quantified as ratios. The three classes of hybridizations cluster around the three poles of a triangle. The horizontal axis that accounts for 10 % of the interaction variance appears to be dominated by the ratio in gene expression profiles of the AML to the ALL class. The vertical axis, accounting for an additional 7 % of the interaction, is related to the ratio of the ALL T-cell versus the ALL B-cell class. However, like in PCA there is a significant amount of overlap between these two subclasses. Genes that occupy the most extreme positions on the map are differentially expressed between the different classes of hybridizations. For instance, the gene with accession number X82240 on the left pole of the triangle is a gene that has on average a high absolute level of expression, as is indicated by the area of the associated circle, and is selective for the ALL-B class. This gene is contrasted to a cluster of genes concentrated around the right pole that are selective for AML patients and a set of genes located on the top pole associated with the ALL-T class. It is however, questionable whether the genes with relatively high values for the ALL-T class make sense in reality, since the three genes on the top of the triangle all correspond to Affymetric control genes. These genes are placed on the microarrays by the manufacturer as a means of internal control.

b. Weighted SMA

In a second SMA, we used variable weighting for the genes, with weights proportional to the mean expression levels of the genes. SMA and construction of the biplot was carried out as above. The resulting biplot is depicted in Figure 4. The pattern formed by the different hybridizations lies in between the result obtained by CFA and unweighted SMA. Also here, it is possible to identify a triangular-like shape with three poles corresponding to the three classes of leukemia. The horizontal axis of the map is dominated by the ratio in gene expression between the AML and ALL class and accounts for 13 % of the total interaction variance. The vertical axis is dominated by the contrast between the ALL T-cell and ALL B-cell group and accounts for an additional 12 % of the interaction. In contrast to the former unweighted SMA, the three classes of leukemia are completely separated from one another. Note that, compared to Figure 3, the direction of the vertical axis is reversed. All of the genes that are located distal from the center could have a physiological meaning. It is noteworthy to mention that only 3 of the 27 most distal genes were among the 50 genes selected by Golub, et al. to discriminate between the different classes of disease.

[GEERT] IS DE REVERSIE VAN DE ASSEN GEWOON EEN ARTEFACT (INVARIANTIE OP ORIENTATIE) OF IS ER MEER STRUCTUREELS AAN DE HAND. OOK DIT VERDIENT ENIGE DISCUSSIE.

In a subsequent analysis (Fig. 5) we carried out a weighted SMA using the 27 genes identified in Figure 4. Since row and column variances are now comparable, the biplot was constructed using singular values ($\alpha = 0.5$, $\beta = 0.5$) as method for factor scaling. The horizontal and vertical axis explain 43 % and 32 % of the global interaction variance. Using only this small subset of 27 genes allows complete separation of the three pathological classes. Figure 6 shows the hybridizations obtained in the second data set (MIT2) positioned on the biplot based on MIT1. AML and ALL-B class can clearly be distinguished from one another without any overlap. There is only one possible mismatch, the only hybridization in MIT2 that was identified as ALL-T.

c. SMA as a tool to quantify differential gene expression

The maps shown in Figures 4 and 5 suggest an even further reduction of the data. Indeed, the genes located at the poles of the triangle formed by the three pathological classes almost completely represent the interaction that is present in the first factorial plane. To emphasize this point we constructed the biplot in Figure 7 using only the expression profile of the genes with accession numbers X82240, X76223, and M82546. This case of spectral map analysis where only three columns are considered is also referred to as multivariate ratio analysis (MRA) and has found applications in the field of ecology (Hermy and Lewi, 1991). MRA differs from conventional SMA only by the application of asymmetric unit column-variance ($\alpha = 1,\ \beta = 0$) as method for factor scaling. All 72 hybridizations present in MIT1 and MIT2 data sets were plotted on the plane determined by these three gene-poles. In addition axes were drawn through the poles of the triangle. These axes allow quantification of the different ratios in gene expression that can be calculated from the data. The map allows to read-off the differential genetic profile of each of the hybridizations with respect to the three characteristic genes. This possibility to read off ratios is a major advantage of SMA as compared to CFA. Figure 7 shows that hybridizations whose expression profile is not specific for any of the genes and consequently are located at the center of the map, also have a low level of expression for all three genes, as is indicated by the extreme small areas of the corresponding squares. Genetic specificity as expressed by the differential ratio between any two genes can be substantial and can amount to a factor of 10,000 or more, as is illustrated by the axis M82546 versus X76223. Furthermore, it is shown that, using only three genes, the three pathological classes can be discriminated to a substantial extent.

4. Discussion

The results obtained in the previous section indicate that weighted SMA is a valuable tool for the analysis of gene expression microarray data. Weighted SMA and CFA outperform conventional PCA in visualizing the data, determining clusters of hybridizations and genes, correlating hybridizations with gene expression profiles, and reducing the data. An advantage of SMA over CFA is the possibility to interpret distances as ratios, while CFA does not allow such an intuitive approach. A limitation with regard to interpretation of the spectral map would be the abundance of groupings in the different hybridizations as is the case in some data mining applications. However, for such applications one could consider exploring subsets of the data instead of entire data sets.

Assessing the quality of experiments that are as complex as microarray experiments is a non-trivial matter and is of paramount importance to the scientist. Quality assurance already starts in the laboratory by a careful conduct of the experiment. Proper normalisation of the raw data will further enhance the quality and make results comparable for different genes. However, it happens in some occasions that mistakes happen in the laboratory processing of the samples. It has been our experience that, in addition to the above, constructing spectral maps of the data is extremely helpful for determining such outlying results.

[GEERT] WAAR DIENT DEZE LAATSTE COMMENTAAR TOE. IK DENK NIET DAT DE LEZER ER AL TE VEEL WIJZER VAN WORDT. OFWEL VERDUIDELIJKEN EN ER WAT VERDER OP IN GAAN, OFWEL SCHRAPPEN.

Apart from the data analytic aspects of this report, it is noteworthy to mention that the three genes selected in the construction of Figure 7, could be related to leukemia. Only one gene was also present in the set of 50 genes used by Golub, et al. for class determination. M84526 is known to be an AML predictive gene and was also selected by Golub, et al. The two other genes X82240 and X76223 were not selected by Golub, et al. but were reported to be involved in T-cell leukemia.

References

Andries, K., Dewindt, B., Snoeks, J., Wouters, L., Moereels, H., Lewi, P.J., Janssen, P.A.J. (1990). Two groups of rhinoviruses revealed by a panel of antiviral compounds present sequence divergence and differential pathogenicity. *Journal of Virology* **64**, 1117-1123.

Benzécri, J.P. (1973). *L'analyse des données. Vol II. L'Analyse des Correspondences*. Gounod, Paris.

Chapman, S., Schenk, P., Kazan, K., Manners, J. (2001). Using biplots to interpret gene expression patterns in plants. *Bioinformatics* **18**, 202-204.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA* **95**, 14863-14868.

Faes, W., Lewi, P.J. (1987). Spectramap: the story behind your numbers. *The International Management Development Review* **3**, 183-187

Fellenberg, K., Hauser, N., Brors, B., Neutzner, A., Hoheisel, J., Vingron, M. (2001) Correspondence analysis applied to microarray data. *Proceedings of the National Academy of Sciences USA* **98**, 10781-10786.

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

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press, London.

Hermy, M., Lewi, P.J. (1991). Multivariate ratio analysis, a graphic method for ecological ordination. *Ecology* **72**, 735-738.

Hilsenbeck S.G., Friedrichs, W.E., Schiff, R., O'Connell, P., Hansen, R.K., Osborne,C.K., Fuqua, S.A. (1999). Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *Journal of the National Cancer Institute* **91**, 453-459.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417-441.

Landgrebe, J., Welzl, G., Metz, T., van Gaalen, M.M., Ropers, H., Wurst, W., Holsboer, F. (2002). Molecular charisation of antidepressant effects in the mouse brain using gen expression profiling. *Journal of Psychiatric Research* **36**, 119-129.

Lefkovits, I., Kuhn, L., Valiron, O., Merle, A., Kettman,J. (1988). Toward an objective classification of cells in the immune system *Proceedings of the National Academy of Sciences USA* **85,** 3565–3569

Lewi, P.J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. *Arzneimittel Forschung (Drug Research)* **26**, 1295-1300

Lewi, P.J. (1995). Spectral-mapping of drug-test specificities. In: *QSAR: Chemometrics in Molecular Design*. H. van de Waterbeemd (Ed.), VCH, Weinheim, Germany, pp. 219-253

Pearson, K. (1901) On lines and planes of closest fit to points in space. *Philosophical Magazine*, **2**, 559-572.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation *Proceedings of the National Academy of Sciences USA* **96**, 2907-2912.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J.,Church , G.M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281-285.

Thielemans,A., Lewi,P.J., Massart, D.L. (1988). Similarities and differences among multivariate display techniques illustrated by Belgian cancer mortality distribution data. *Chemometrics and Intelligent Laboratory Systems* **3**, 277-300.

Törönen, P., Kolehmainen, M., Wong, G., Castrén, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Letters* **451**, 142-146.

Legend to Figures

Figure 1. Biplot of the results of PCA. First principal component (horizontal axis) and second principal component (vertical axis) account for 71 % and 4 % respectively of the global variance. Small dots represent genes. The three classes of leukemia are identified by squares (AML), triangles with top up (ALL B-Cell), or triangles with top down (ALL T-cell).

Figure 2. Biplot of the results of CFA. Horizontal axis represents 17 % and vertical axis 10 % of the global chi-square. Codings of the different hybridizations are identical to Figure 1.

Figure 3. Biplot of the results of an unweighted SMA. Horizontal axis represents 10 % and vertical axis 7 % of the global interaction variance. Codings of the different hybridizations are identical to Figure 1. The 0.5 % most distal genes are labeled and represented as circles with areas proportional to the marginal mean.

Figure 4. Biplot of the results of SMA weighted for marginal row and column means. Horizontal axis represents 14 % and vertical axis 12 % of the total interaction variance. Note the reversal of the vertical axis as compared to Figure 3. Codings of the different hybridizations are identical to Figure 1. The 0.5 % most distal genes are labelled and represented as circles with areas proportional to the marginal mean.

Figure 5. Biplot of the results of SMA weighted for marginal row and column means using only the 0.5 % genes most distant from the center. Horizontal axis represents 43 % and vertical axis 32 % of the total interaction variance. Codings of the different hybridizations are identical to Figure 1. Genes are represented as circles with areas proportional to the marginal mean.

Figure 6. Positioning of the 34 additional patients (MIT2) on the biplot of Figure 4. Codings of the different hybridizations are identical to Figure 1. Genes are represented as circles with areas proportional to the marginal mean.

Figure 7. All 72 hybridizations plotted on the plane determined by the ratios between the three most extreme genes X82240, X76223, and M82546, of Figure 5. Differential gene expressions of individual patients can be read from the calibrated axes. Hybridizatons are coded as: M for AML, B for ALL B-cell, and T for ALL T-cell class patients.

[GEERT] DE FIGUREN ZELF VERDIENEN DUIDELIJKE ASSEN MET LABELS ENZ.