SPECTRAL MAP ANALYSIS - A METHOD TO ANALYZE GENE EXPRESSION DATA

Luc Bijnens[1], Paul J. Lewi[2], Hinrich W. Göhlmann[3], Geert Molenberghs[4] and Luc Wouters[5]

[1,2,3] Biometrics and Reporting[1], Functional Genomics[2], and Center for Molecular Design[3], Johnson & Johnson Pharmaceutical Research & Development, a division of Janssen Pharmaceutica NV, Beerse, Belgium
[4] Center for Statistics, Limburgs Universitair Centrum, transnationale Universiteit Limburg, Hasselt, Belgium
[5] Department of Biostatistics, Barrier Therapeutics NV, Geel, Belgium

KEYWORDS:
Bioinformatics; Biplot; Correspondence factor analysis; Data mining; Data visualization; Gene expression data; Microarray data; Multivariate exploratory data analysis; Principal component analysis; Spectral map analysis.

Abstract
The simultaneous measurement of the expression level of thousands of genes presents a real challenge to the information processing capability of the present computer systems and statistical software tools because of the complexity of the problems at hand and size of the data sets. These days research is concentrating on projects to find clusters in the biological samples and to identify genes related to these clusters because of the availability of the new microarray laboratory techniques.
In this study, three multivariate data analysis methods: principal component analysis (PCA), correspondence factor analysis (CFA), and spectral map analysis (SMA) are compared exactly for their ability to identify clusters of biological samples and genes using data on gene expression levels of leukemia patients (Golub, 1999). PCA has the disadvantage that the resulting principal factors are not very informative regarding differential gene expression, while CFA is sensitive to single large values and has difficulties regarding interpretation of the distances between objects. We present spectral map analysis (SMA) as an alternative method developed by Lewi (1976) and compare it with the other two methods. The importance of weighting for the level of gene expression is demonstrated. Proper weighting allows less reliable data to be down-weighted and more reliable information to be emphasized. It is shown that weighted SMA outperforms PCA and CFA in finding clusters in the biological samples and identifying genes related to these clusters. SMA addresses the data in a more appropriate manner than CFA with respect to the scale of measurement. It allows for applying a more flexible weighting to the genes and biological samples.

Introduction

Microarray technology makes use of the production of messenger RNA (mRNA). The m-RNA is produced in the internal of the cell when there is an activity in the cell that expresses a need for particular proteins. The mRNA is a sequence of bases matching to the sequence of a gene in the chromosomes of the biological cell (see figure 1). The m-RNA is then translated into chains of amino acids by the ribosomes. Those chains are then connected to form proteins. Those proteins then constitute the particular activity that the cell requested. In the microarray technology the amount of m-RNA is measured via chips that contain thousands of very small test-tubes (probes) of which each contains an over abundance of strings of bases of one type of gene (see figure 2). Each gene sequence can then hybridise with the m-RNA coming form the biological sample. Before the m-RNA is applied to the chip it is connected to a fluorescence molecule. The m-RNA that is not hybridised with the DNA sequences of the chip is washed away so that only the matching and thus hybridising (binding) m-RNA is measured. Since there is an abundance of DNA gene material in each minuscule reaction chamber the technology is able to quantify the relative amount of m-RNA in the biological cell. The more m-RNA was present in the biological cell the more fluorescence will be measured. This way the intensity of the fluorescence is directly linked to the amount of m-RNA and indirectly linked to a particular protein and activity.

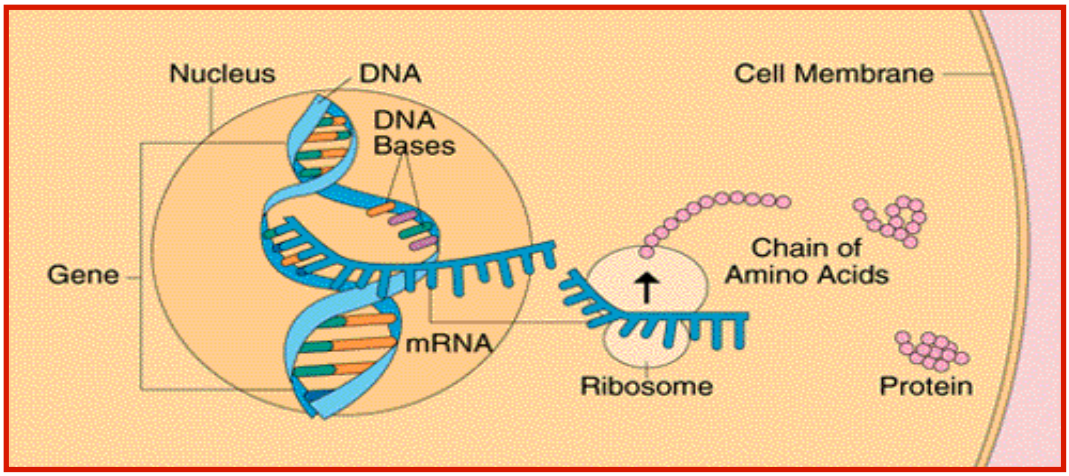

Figure 1. Schema of the translation of DNA via RNA into amino acids and proteins.

Those proteins then constitute the particular activity that the cell requested. A good summary of the biology and the technology is given by Nguyen et al 2002.

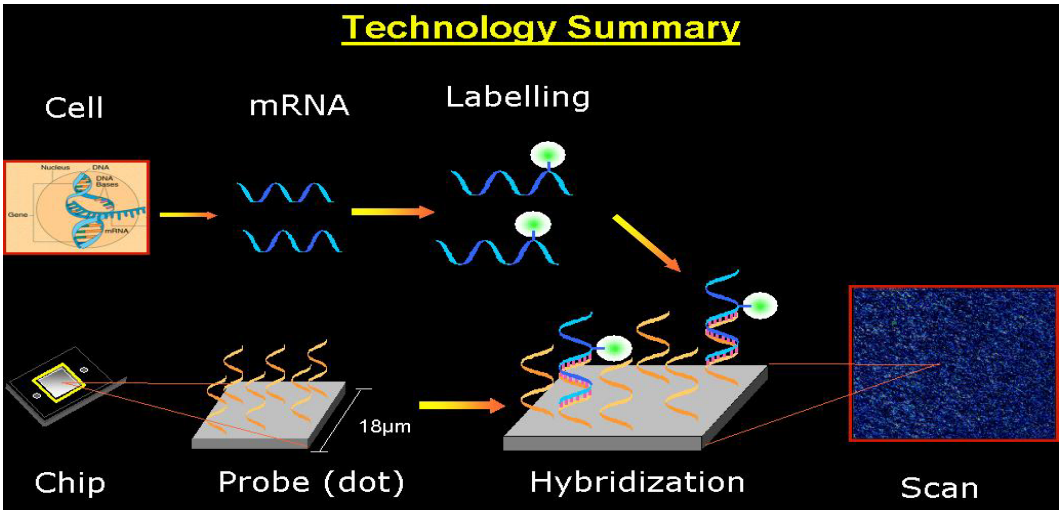

Figure 2. The microarray technology

In the present study gene expression data coming from patients with leukaemia are used (Golub et al. 1999). After data cleaning the dataset contained intensities of 5327 genes for each of 38 samples. Twenty-seven samples came from patients with ALL and eleven samples came from patients with AML. The group of samples can be split into two groups: T-lineage and B-lineage.

Statistical methods

Three different multivariate statistical methods have been used to analyse the data. Principal component analysis (PCA, Pearson 1901, Hotelling 1933), the first method, transforms the intensities into n=5327 principal axis in such a way that they are ordered in terms of the variability explained. Each of those axes is a linear combination of all of the 5327 genes. Most multivariate projection methods are based on a derived space with n orthogonal axes. Those axes are linear combinations of the original measurements (intensities) of all the genes. The axes are constructed in a way that the first axis lies in the direction (in the multivariate data space) with the largest variability and the last axis has the smallest variability. Correspondence factor analysis (CFA, Benzécri 1973 and Greenacre 1984) and spectral map analysis (SMA, Lewi 1976) are special cases of multivariate projection methods that help to reduce the complexity (dimensions) of highly dimensional data (n genes versus p samples). A classical principal component analysis will create a first axis (principle component) that maximizes the variability due to size of the intensities.

Clusters that can be identified based on the first axis will simply differ in absolute size of the intensities. CFA was originally developed for contingency tables and decomposes the chi-square statistic. Distances therefore have a chi-square distribution. In a SMA of log transformed data the distances are proportional to ratios of genes or samples. In microarray data we are mainly looking for contrasts and not simply high or low intensities. For that reason both CFA and SMA have the appropriate properties (double closure for CFA and double centering for SMA) that remove the size component from the data. CFA and SMA will look for contrasts in intensities between genes without the nuisance effect of the absolute values of the intensities. However, microarray data tend to be more reliable with increasing intensity. In order to deal with that, re-introduction of the size component via variable weighting proportional to the mean intensities of genes and samples is required. Since there is a high difference in dimension between rows and columns in SMA it is possible to introduce scaling. This operation pulls the genes away from the center of a biplot while it leaves the samples at their original places (Wouters et al 2003). A biplot (Gabriel 1971 and Chapman 2001) created by the first two axes displays the maximal separation of both the genes and the samples. Coinciding clusters of samples and genes on the biplot indicate the genes (signatures) that are responsible for the separation of the samples. Genes that are located in the general direction of a sample on the biplot should be looked at as potential signatures for the separation of that sample versus the others.

Principal component analysis

PCA was carried out after logarithmic re-expression of the gene expression profiles. 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 3), an asymmetric scaling with unit column-variance was used to allow for better visual discrimination between the different samples. 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 samples (average variance = 2). A consequence of unit column-variance factor scaling is that correlations and distances between samples are not represented in the biplot. However, in exploring gene expression data only patterns in the distribution of the biological samples are of direct interest. In Figure 3, 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 only 3% of the total variance. The remaining principal components were considered to reflect random disturbances.

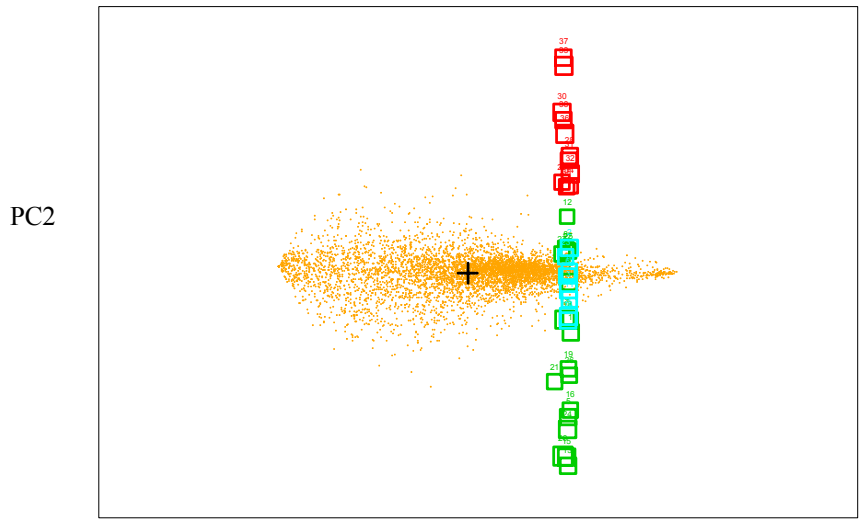

Figure 3: Principal component analysis (PCA) The yellow dots represent genes and the boxes represent the diseases: AML (red), ALL-T (blue) and ALL-B (green).

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 samples. Differences between biological samples 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.

Correspondence factor analysis

The biplot obtained from CFA on the Golub data is depicted in Figure 4. The same asymmetric unit column-variance scaling was used as in PCA, to allow for optimal visual discrimination of the different samples. While distances between samples 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 4 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 samples, an approach that is much more relevant to the problem. In Figure 4, genes are distributed in a funnel-like pattern and there is a clear separation between ALL and AML patients with only two 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 tempting to identify a few genes that could be used in characterizing the three pathological classes. Gene probes located at the poles of the triangular-like shape should be characteristic for a given class of leukemia. However, for the two gene probes identified as the top left and right pole only a few valid measurements were made and the results depended largely on the expression level obtained in a single patient. This underscores the, in this case, less desirable sensitivity of CFA to single large values. There is also 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 question the applicability of CFA in microarray data analysis.

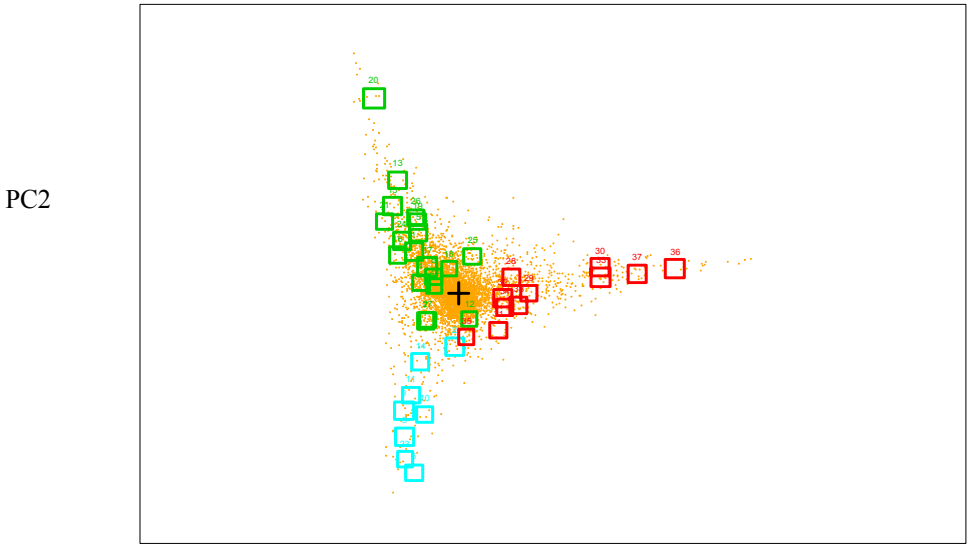

Figure 4: Correspondence Factor Analysis (CFA). The yellow dots represent genes and the boxes represent the diseases: AML (red), ALL-T (blue) and ALL-B (green).

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 higher levels.

In a weighted SMA, we used variable weighting for the genes and samples, with weights proportional to the mean expression levels of genes and samples, respectively. SMA and construction of the biplot was carried out as above. The resulting biplot is depicted in Figure 5. The pattern formed by the different samples lies in between the result obtained by CFA and unweighted SMA.

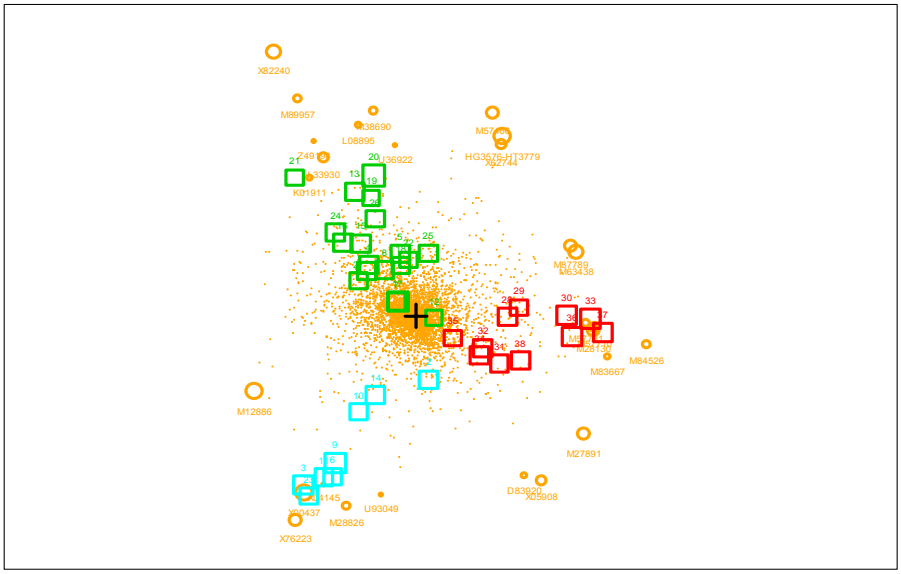

Figure 5: Spectral map of the Golub training dataset. The yellow dots represent genes and the boxes represent the diseases: AML (red), ALL-T (blue) and ALL-B (green). The sizes of the symbols are proportional to the absolute intensity of the genes for the dots and of the diseases for the boxes.

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 14% 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. All of the genes that are located distal from the center could have a physiological meaning. It is noteworthy to mention that only 4 of the 27 most distal genes were among the 50 genes selected by Golub et al. (1999) to discriminate between the different classes of disease.

In a subsequent analysis (Fig. 6), we carried out a weighted SMA using the 27 genes identified in Figure 5. Since row and column variances are now comparable, the biplot was constructed using singular values as the method for factor scaling. The horizontal and vertical axes 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.

Discussion and conclusions

The results obtained in the previous section illustrate the impact of the different methods. The characteristic difference between conventional PCA on the one hand and CFA and SMA on the other hand are the operations of double-closure and double-centering. The

double-closure operation in CFA eliminates the size factor that is related to the first

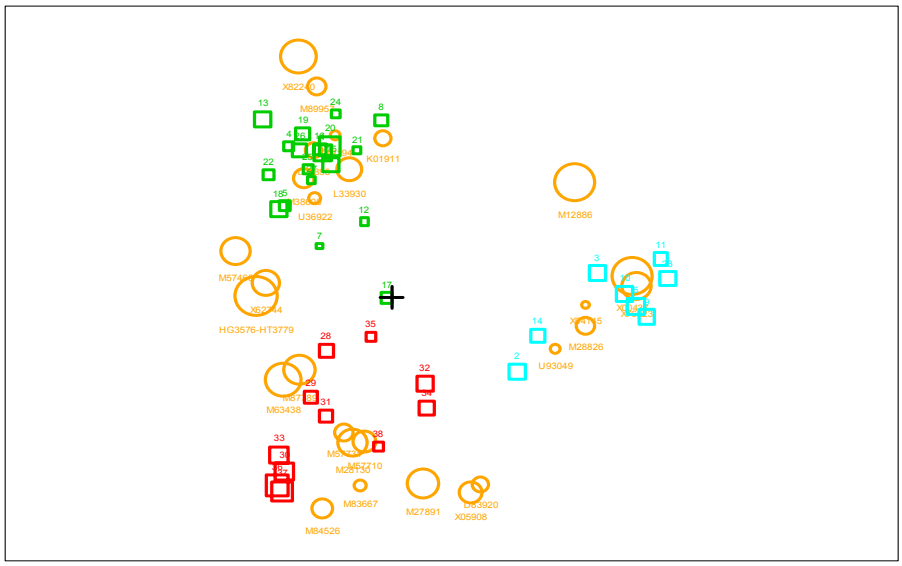

Figure 6: Spectral map of the 27 most extreme genes of the Golub training dataset. The yellow dots represent genes and the boxes represent the diseases: AML (red), ALL-T (blue) and ALL-B (green). The sizes of the symbols are proportional to the absolute intensity of the genes for the dots and of the diseases for the boxes.

dominant component in PCA and stresses differences among the genes and among the samples. The same effect is obtained by double-centering after logarithmic re-expression in SMA. Although, mathematically, these two operations are related, the results can differ substantially as is illustrated by the differences in the biplots obtained from CFA and SMA, respectively. Re-expressing the data to logarithms downplays very large contrasts that result from extreme outcomes. This is a desirable property for the analysis of gene expression data that typically suffer from the presence of severely outlying measurements.

A drawback of the logarithmic re-expression is that contrasts at a less reliable level of gene expression are considered of equal importance, as are contrasts at a more reliable level. Incorporating weights proportional to the marginal totals in the centering, normalization can counteract this phenomenon, and factorization building blocks leading to weighted SMA.

Our results 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 samples and genes, correlating samples with gene expression profiles, and reducing the data. An advantage of SMA over CFA is the possibility of interpreting distances as ratios, while CFA does not allow such an intuitive approach.

Apart from the data analytic aspects of this report, it is noteworthy to mention that the three most influential genes were identified in the literature to be related to leukemia (Wouters et al. 2003).

Literature

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.

Gabriel, K.R. (1971). The biplot graphical display of matrices with applications to principal component analysis. *Biometrika* **58**, 453-467.

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

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.

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

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

Nguyen, D., Bulak, A.A., Wang, N. and Caroll, R.J. (2002). DNA-Microarray experiments: biological and technological aspects. Biometrics 58, 701-717

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

Wouters, L., Göhlmann, H.W., Bijnens, L., Kass S.U., Molenberghs, G. and Lewi P.(2003). Graphical exploration of gene expression data: a comparative study of three multivariate methods (Biometrics, accepted).