Software report

CLUSFAVOR 5.0: hierarchical cluster and principal-component analysis of microarray-based transcriptional profiles

Leif E Peterson

Address: Departments of Medicine, Molecular and Human Genetics, and Scott Department of Urology, Baylor College of Medicine, One Baylor Plaza, ST-924, Houston, Texas 77030, USA. E-mail: peterson@bcm.tmc.edu

Published: 24 June 2002

Genome ***Biology*** 2002, **3(7)**:software0002.1–0002.8

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/7/software/0002

Abstract

CLUSFAVOR (CLUSter and Factor Analysis with Varimax Orthogonal Rotation) 5.0 is a Windows-based computer program for hierarchical cluster and principal-component analysis of microarray-based transcriptional profiles. CLUSFAVOR 5.0 standardizes input data; sorts data according to gene-specific coefficient of variation, standard deviation, average and total expression, and Shannon entropy; performs hierarchical cluster analysis using nearest-neighbor, unweighted pair-group method using arithmetic averages (UPGMA), or furthest-neighbor joining methods, and Euclidean, correlation, or jack-knife distances; and performs principal-component analysis.

Rationale

DNA microarrays are useful for identifying genes that are co-expressed in different phenotypes, experiments, or both. Genomic regions containing *cis*-regulatory motifs can be identified by exon mapping cDNAs of co-expressed genes using BLAST [1-4]. Because *cis*-regulatory motifs act as binding sites for transcription factors that control expression, co-expressed genes sharing the same regulatory motifs are likely to be networked and under the same regulatory control [5-8].

To identify co-expressed genes using DNA microarray data, one typically uses classification and data-reduction methods such as cluster analysis and principal-component analysis (PCA). The CLUSFAVOR 5.0 computer program was developed for hierarchical cluster analysis (HCA) and PCA of DNA microarray expression data. CLUSFAVOR 5.0 was developed under the Windows operating system using Microsoft Visual Basic [9], and can therefore be installed and run on any of the 32-bit versions of Windows (95, 98, NT, 2000, or XP). CLUSFAVOR can standardize expression data, sort, and perform HCA and PCA of arrays and genes. The program accommodates missing data, can calculate replicate averages, and determine replicate outliers and drop them from the analysis. CLUSFAVOR 5.0 has primarily been used for DNA microarray data, but can be used for numerical taxonomy to identify natural groupings of variables for non-genetic data. This report reviews user specifications, input file formats, numerical methods, and output formats for the CLUSFAVOR 5.0 computer program.

CLUSFAVOR 5.0

The CLUSFAVOR 5.0 program can open either a tab-delimited text file for a new run or a pre-formatted file containing output parameters and data from a previous run. The option to view results of previous runs quickly eliminates the long wait time needed when obtaining results of large processing jobs. When opening tab-delimited text files for a new run, CLUSFAVOR 5.0 will recognize any text file whose filename ends with .txt. File formats are described in the User's Guide, available for download (see Downloading files section).

Standardization of input data

Results of multivariate statistical methods such as HCA and PCA depend strongly on the scale (the range) of data used. A common method for standardizing input data during HCA involves treating the variables (or records) being clustered as the 'objects' and the records (or variables) as the 'attributes'. When HCA is being performed on the arrays as objects,

standardized expression is calculated by subtracting the gene-specific average expression and dividing by the gene-specific standard deviation, as genes are the attributes. When cluster-analyzing the genes as objects, standardization is based on the array-specific average and standard deviation of expression. In this fashion, zero means are obtained for the attributes rather than the objects being clustered. The distance functions used for HCA in CLUSFAVOR 5.0 can be based on either Euclidean distance or correlation. When Euclidean distance is specified as the distance function and input data are standardized, a single round of standardization is performed which removes additive and multiplicative size displacements among expression profiles so that residual differences are detected. However, when correlation is specified as the distance function, the double round of standardization is more effective at removing size displacements between expression profiles. Standardization is also used by default for all PCA runs on arrays and genes. Summary statistics for array- and gene-specific average and standard deviation in expression can be saved in a tab-delimited output file.

Sorting

When working with DNA microarray data, biologists often want to know which genes have the greatest or least expression or standard deviation across the arrays. When sorting is specified at run-time, CLUSFAVOR 5.0 calculates the gene-specific coefficient of variation, average and total expression, standard deviation, and performs an ascending quicksort of each parameter being considered. Results are saved in JPG image files using a color gradient of expression specified by the user, and are also written in tabular form to tab-delimited text files. All sorting results (JPG images and text) are linked with HTML (hypertext markup language) files for viewing.

Hierarchical cluster analysis (HCA)

Hierarchical cluster analysis (HCA) is an exploratory multivariate statistical method for identifying 'natural' groupings of objects considered in an analysis. The least distance, $D(r,s)$, between two objects r and s (arrays r and s) consisting of n_r and n_s elements is first identified. The distance function, $D(r,s)$, can be based on either the Euclidean distance ($0 \le D(r,s) < +\infty$) or 1-correlation ($0 \le D(r,s) \le 2$). Objects r and s are joined to form a new 'node' u with $n_u = n_r + n_s$ elements. Next, distances between the newly formed node u (comprised of objects r and s) and all other nodes v (or objects) are calculated as

$$D(u,v) = \begin{cases} \min\{D(r,v), D(s,v)\}, \textbf{single linkage} \\ (D(r,v)n_r + D(s,v)n_s)/n_u, \textbf{UPGMA} \\ \max\{D(r,v), D(s,v)\}, \textbf{complete linkage} \end{cases} \quad (1)$$

where $D(r,v)$ is the distance between nodes r and v joined previously, $D(s,v)$ is the distance between nodes s and v joined previously. Single linkage (nearest neighbor), unweighted pair-group method using arithmetic averages (UPGMA), or complete linkage (furthest neighbor) is specified by the user, and n_u, n_r, and n_s are the number of objects in the nodes. After all the new distances between node u and other nodes have been calculated, a search for the smallest distance is conducted, followed by calculation of new distances. This is done repeatedly until all clusters have joined. The CLUSFAVOR 5.0 algorithm first clusters the arrays, and then the genes.

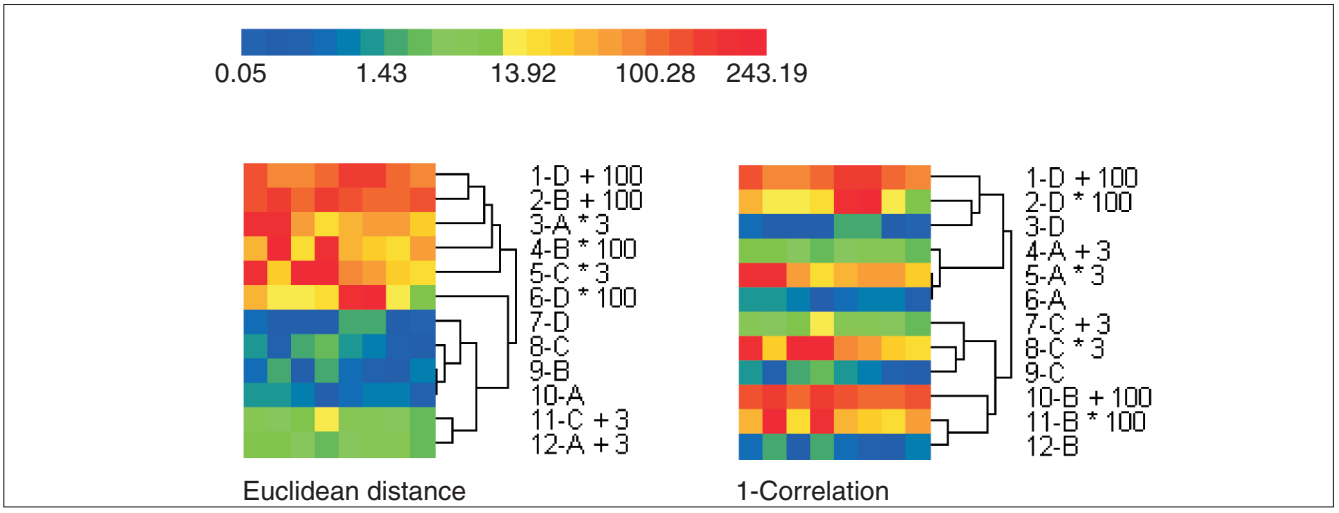

Figure 1
Differences between Euclidean distance and correlation in hierarchical cluster analysis (HCA) of expression for four genes (A, B, C and D) to which constant values of 3 and 100 were added or multiplied. Euclidean distance did not ignore additive and multiplicative translations from base values of A, B, C and D. However, correlation ignored translations because the correlation between a profile and that same profile to which a constant is added or multiplied is unity. The lengths of dendogram arms shown relate to joining step rather than joining distance.

Frequency histograms of array- and gene-specific pairwise distances and gene-specific expression values can also be generated in JPG format and linked to an HTML file for viewing. The choice of Euclidean distance or 1-correlation as a distance function is based on whether or not the user wants to ignore additive and multiplicative translations between expression values for pairs of objects. Figure 1 illustrates this for four genes (A, B, C and D) whose expression values were increased by adding and multiplying with constant values of 3 and 100. Euclidean distance did not ignore the additive and multiplicative translations from base values of A, B, C, and D and clustered the expression profiles according to their level of expression. Correlation, however, ignored the additive and multiplicative translations in expression values and clustered together genes to which the constants were added and multiplied, as the correlation between an expression profile and that same profile plus or times a constant is unity. This shows why cluster runs based on correlation can result in the clustering together of genes with low and high expression values. Figure 2 illustrates a cluster image display of gene expression among selected genes using the National Cancer Institute (NCI) 60 cancer cell line data [10] based on standardized input data, UPGMA, and Euclidean distance.

CLUSFAVOR 5.0 can also perform HCA using 'jack-knife' distance functions [11]. Consider n genes on p arrays. For each of the $n(n^*1)/2$ pairwise correlation coefficients for expression profiles of genes i,j (i,j = 1,2,...,n), calculate p (k = 1,2,...,p) correlation coefficients, each time dropping expression values for the kth array. This is the process of jack-knifing, where data are dropped during calculation. Thus, instead of having one correlation coefficient per pair of expression profiles (for genes i and j) over p arrays, we get p correlation coefficients based on the kth expression value (for both genes) dropped from the calculation. For notation, call the correlation coefficient with the first pair of expression values from array 1 dropped $r(i,j)^{(1)}$, call the second correlation coefficient with values for array 2 dropped $r(i,j)^{(2)}$, and so forth up to $r(i,j)^{(p)}$. For genes i and j, take the minimum jack-knife correlation, that is $min\{r(i,j)^{(1)}, r(i,j)^{(2)},..., r(i,j)^{(p)}\}$, subtract this from 1, and use this as the distance $D(i,j)$ for genes i and j in HCA. This will ensure that the greatest distances (that is, 1-correlation) between pairs of genes are used, as the smallest correlation coefficients are used. The advantage of using jack-knife distance functions is that false positives due to outlier effects are minimized; however, the disadvantage is long run-times due to calculation

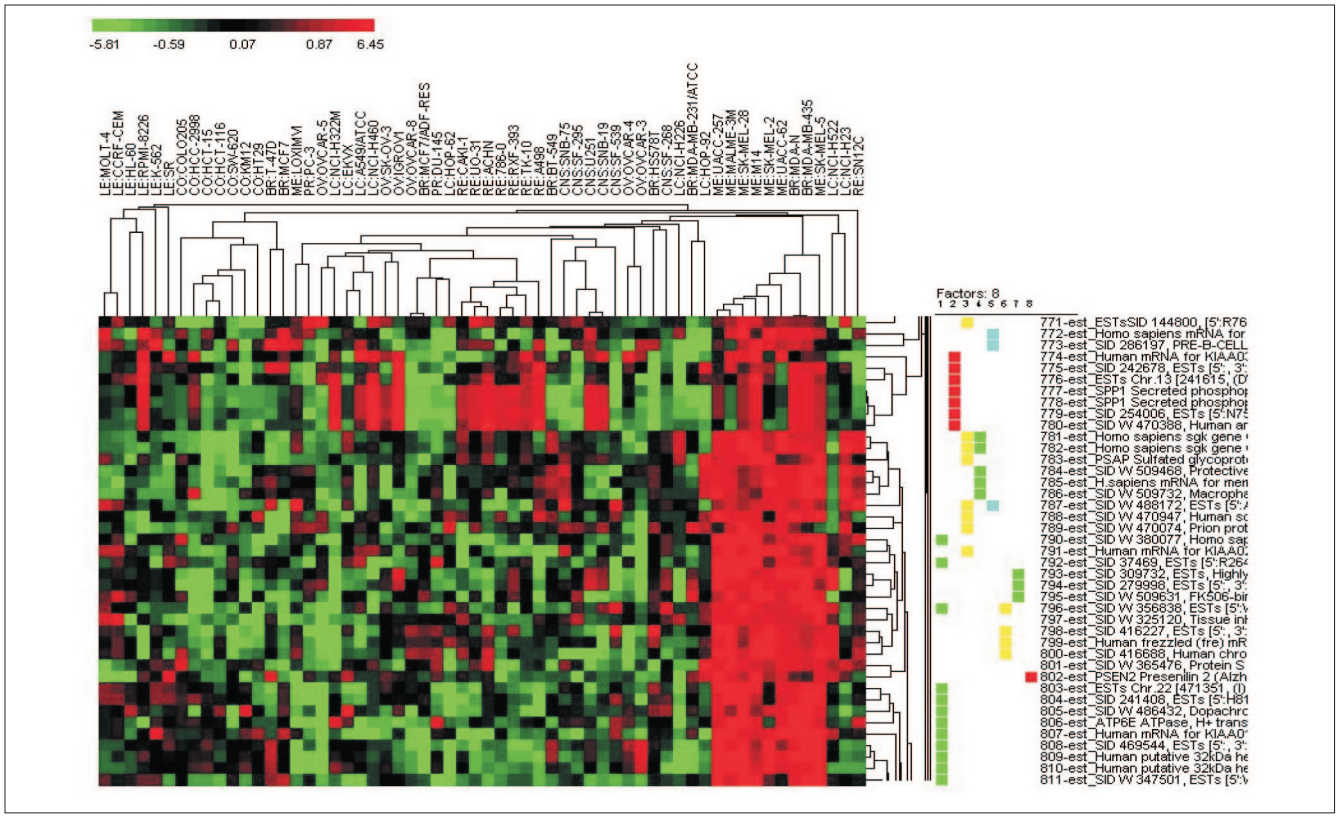

Figure 2
HCA results showing cluster image display for selected genes with color gradient for standardized gene expression and dendograms for arrays and genes. Results of principal-component analysis (PCA) on the selected genes are shown on the right. Data from Ross *et al.* [10]. Expression data were standardized (color is z-score based on array average and standard deviation). The distance function is based on Euclidean distance.

and screening of $n(n^*1)/2^*p$ correlation coefficients rather than the $n(n^*1)/2$ typically calculated. Figure 3 shows how large values of standardized expression can strongly bias the correlation between expression profiles of two genes. After removing the outlier values of expression from array 7, correlation is reduced from 0.83 to -0.37.

Principal-component analysis (PCA)

Principal-component analysis (PCA) is useful for reproducing the total variance among a large number of variables using a much smaller number of unobservable variables or dimensions called latent factors. CLUSFAVOR 5.0 uses the principal-component solution to the factor model for extracting factors (components). This is accomplished by use of the principal-axis theorem, which says that for a gene-by-gene ($n \times n$) correlation matrix **R,** there exists a rotation matrix **E** and diagonal matrix **Λ** such that **ERE′=Λ**. The principal form of **R** is given as

$$\underset{n\times n}{\mathbf{R}} = \underset{n\times n}{\mathbf{E}\Lambda\mathbf{E}'} = \begin{bmatrix} e_{11}e_{12} & \cdots & e_{1n} \\ e_{21}e_{22} & \cdots & e_{2n} \\ \vdots & \vdots & \vdots \\ e_{n1}e_{n2} & \cdots & e_{nm} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} e_{11} & e_{21} & \cdots & e_{n1} \\ e_{12} & e_{22} & \cdots & e_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1n} & e_{2n} & \cdots & e_{nm} \end{bmatrix} \quad (2)$$

where columns of **E** and **E′** are the eigenvectors and diagonal entries of **Λ** are the eigenvalues. In CLUSFAVOR 5.0, only

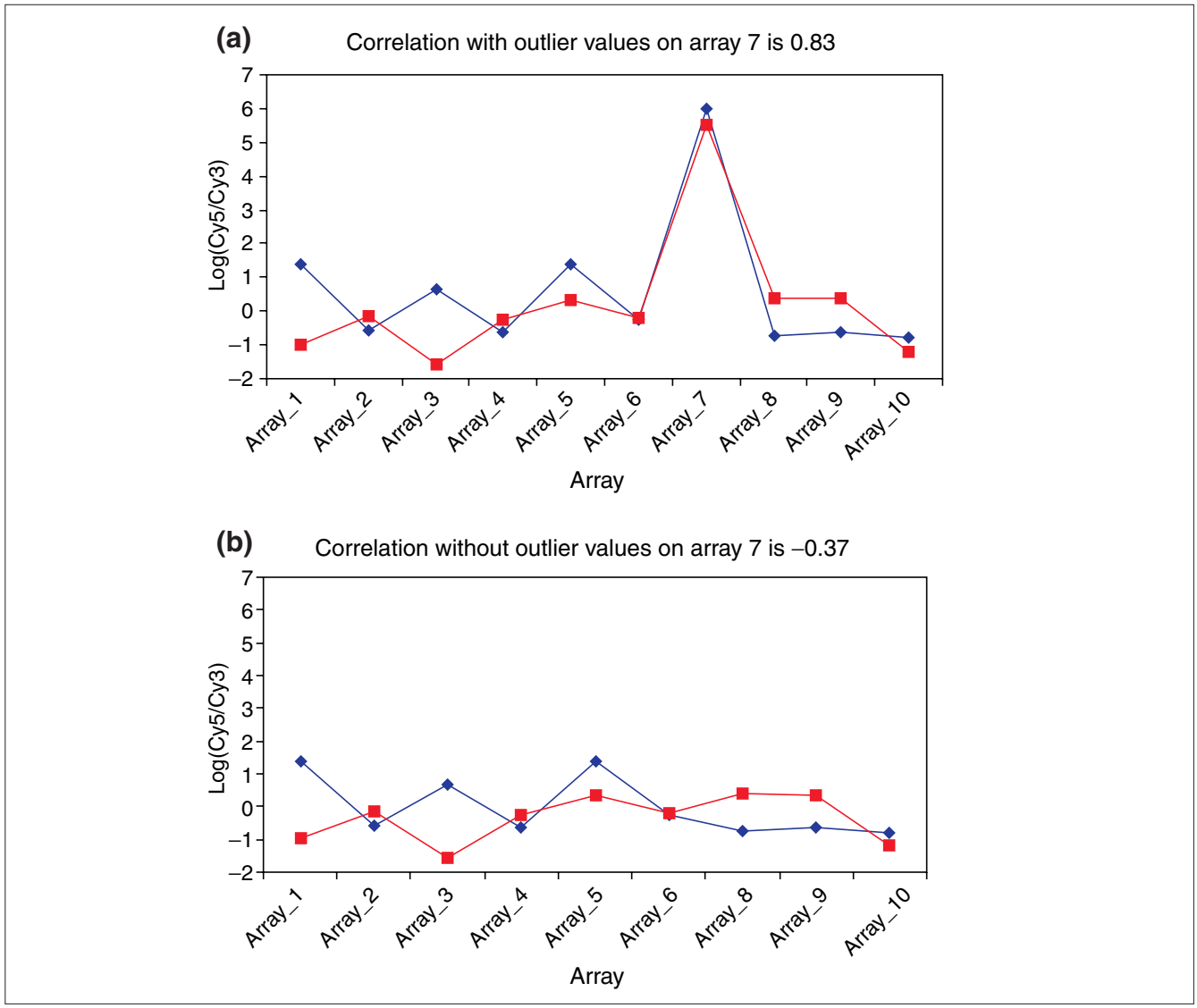

Figure 3
Jack-knife distance functions for HCA minimize bias due to outlier effects. Correlation between expression profiles for genes 1 and 2 drops from **(a)** 0.83 to **(b)** -0.37 when expression values for array 7 (outlier) are dropped from the calculation of correlation.

components whose eigenvalues exceed unity, $\lambda_j > 1$, are extracted from $\mathbf{\Lambda}$ and sorted such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 1$. The 'loading' or correlation between genes and extracted components is represented by a matrix in the form

$$\underset{n\times m}{\mathbf{L}} = \begin{bmatrix} \sqrt{\lambda_1}e_{11} & \sqrt{\lambda_2}e_{12} & \ldots & \sqrt{\lambda_m}e_{1m} \\ \sqrt{\lambda_1}e_{21} & \sqrt{\lambda_2}e_{22} & \ldots & \sqrt{\lambda_m}e_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\lambda_1}e_{n1} & \sqrt{\lambda_2}e_{n2} & \ldots & \sqrt{\lambda_m}e_{nm} \end{bmatrix} \quad (3)$$

where rows represent genes and columns represent components, and, for example, $\sqrt{\lambda_1}e_{11}$ is the loading (correlation) between gene 1 and component 1. After component extraction and loading calculations are completed, the CLUSFAVOR 5.0 program performs a varimax orthogonal rotation of components so that each gene mostly loads on a single component [12]. The cluster image display in Figure 2 includes results of PCA performed on a group of genes selected interactively following the HCA run. Genes that load strongly negative (< -0.45) or strongly positive (> 0.45) on a single component are indicated by use of two arbitrary colors in the column for each component. Genes with identical color patterns in one or more columns can be considered as having similar expression profiles within the selected group of genes. PCA can also be run on an entire data set to produce groups of genes with similar loading patterns, and results for this run option are provided in Figures 4 and 5. Figure 4 illustrates a group of $N = 29$ genes with strong positive loading (> 0.45), with component 3 of 59 components extracted from the correlation matrix of 1,416 genes in the NCI 60 cancer cell line data [10]. Note that these 29 genes were mostly upregulated in the leukemia cell lines. Figure 5 illustrates the average and standard deviation of standardized expression of the same 29 genes, also generated by CLUSFAVOR 5.0.

Viewing results in HTML

Cluster image displays (such as that in Figure 1) for each group of 100 genes are saved separately in JPG format and are linked to an HTML file for viewing with a web browser. This enables the user to view all cluster images for thousands of genes and also to export results quickly to either public or password-protected directories for web publication or collaborative data analysis review. PCA does not depend on the number of genes in a run, and always generates JPG image files (such as Figures 4 and 5) that are linked to an HTML file for browser viewing. During HTML viewing of PCA output, a user can click on a command button in the HTML file to retrieve cDNA sequences (in FASTA format from the National Center for Biotechnology Information

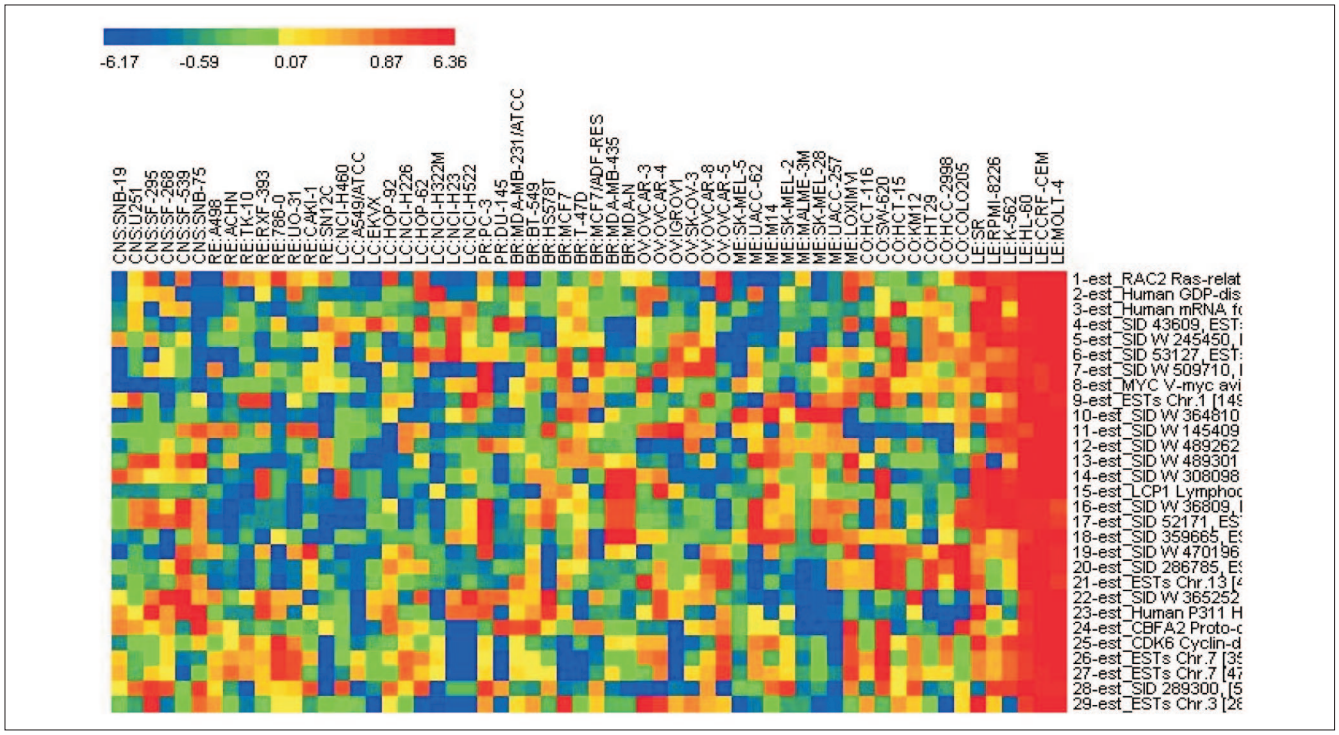

Figure 4
PCA results showing a group of $N = 29$ genes with strong positive loading (> 0.45) on component 3 of 59 components. Genes were extracted from the correlation matrix of 1,416 genes in the NCI 60 cancer cell line data [10]. Note that these 29 genes were upregulated in leukemia cell lines (with annotation prefix 'LE'). Expression data were standardized (color is z-score based on array average and standard deviation).

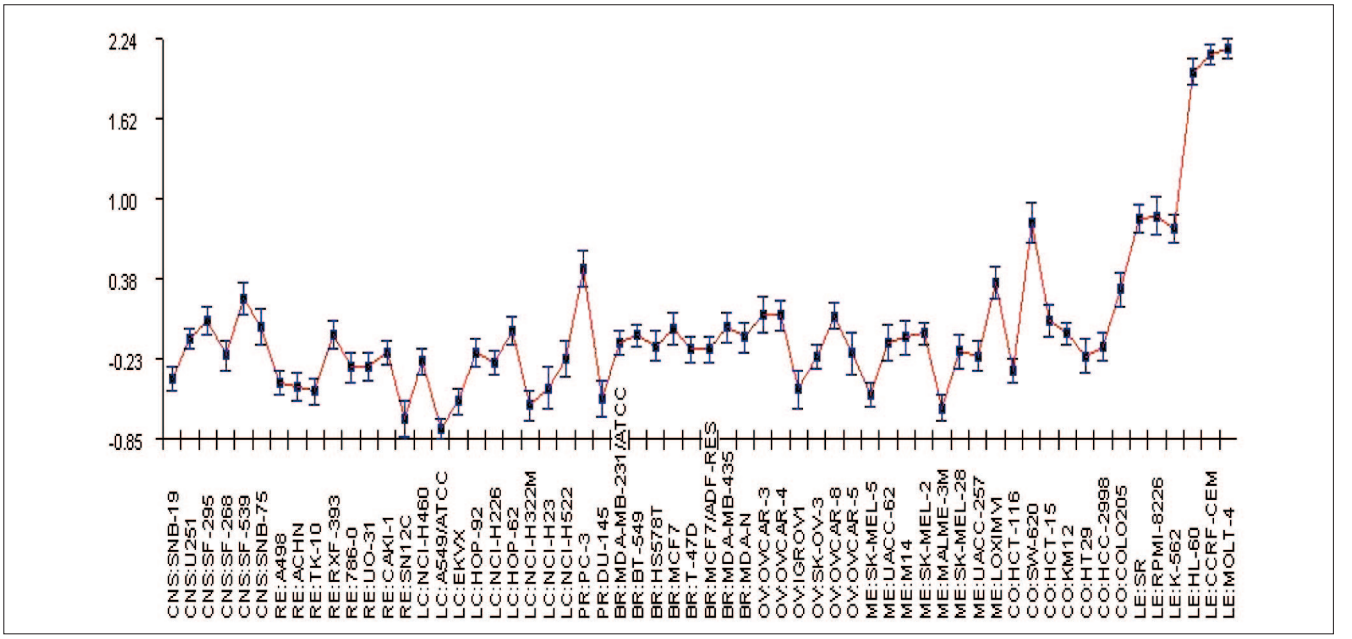

Figure 5
PCA results showing average and standard deviation of standardized expression values for the same *N* = 29 genes shown in Figure 4. Note that these genes were mostly upregulated in leukemia cell lines.

(NCBI)) for each group of genes identified. The cDNA sequences can then be used to search for upstream regions using BLAST; these can be used for *cis*-regulatory motif searching [2-8].

Benchmarking CLUSFAVOR with the statistical analysis package SPSS

Simulated gene expression data for 60 arrays and 120 genes were generated for benchmarking. Data in arrays (columns) 1-10 were based on 120 pseudo-random uniform variates, U(0,1). Arrays 11-20 were filled with 120 standard normal variates, N(0,1). Arrays 21-30 had elements distributed N(3,1), arrays 31-40 N(-3,1), 41-50 with N(0,3), and 51-60 with 120 variates distributed N(0,0.3). Various additive and multiplicative translations were applied in the rows so that expression also varied over the genes. Expression values in rows 1-40 (genes 1-40) were not transformed. However, in rows 41-60 a constant of 3 was added to all array elements, in rows 61-80 a constant of 3 was subtracted from all array values, in rows 81-100 a constant of 3 was multiplied with all array entries, and in rows 101-120 a constant of 0.3 was multiplied with all array elements.

Results for Euclidean distance, correlation, eigenvalues, unrotated component loadings, and rotated component loadings were compared with results from SPSS Version 10 [13]. HCA and PCA run results from CLUSFAVOR were virtually identical (within the machine precision used) when compared with SPSS results. HCA with CLUSFAVOR resulted in identical distance-function matrices and agglomeration schedules. PCA comparisons based on a data set containing 60 arrays and 300 genes (see [14]) resulted in identical loadings of the 300 genes on 24 unrotated components, and no more than a 0.001 bias between SPSS and CLUSFAVOR for 300 loadings on the 24 rotated components (Figure 6). These results indicate good agreement between CLUSFAVOR and SPSS. Certainly, there are other algorithms with which CLUSFAVOR can be benchmarked, such as Eisen's cluster program [15], SAS® [16], Statistica® [17], S-Plus® [18], and R [19]; however, results described above and available at [14] are considered as a first-pass comparison against results obtained from the long-standing commercial statistical software package SPSS.

Downloading files

CLUSFAVOR 5.0 can be downloaded from [14]. Users must first install Version 2.0 in order to obtain the base set-up, and then can download the executable file (clusfavor.exe) for Version 5.0, and save this file into the directory where Version 2.0 was installed. CLUSFAVOR 5.0 is copyright protected against commercial gain, and has a 90-day non-exclusive license that can be extended free of charge for non-profit institutions.

Acknowledgements

Algorithm development for CLUSFAVOR 5.0 was supported by NCI grant CA-78199-04.

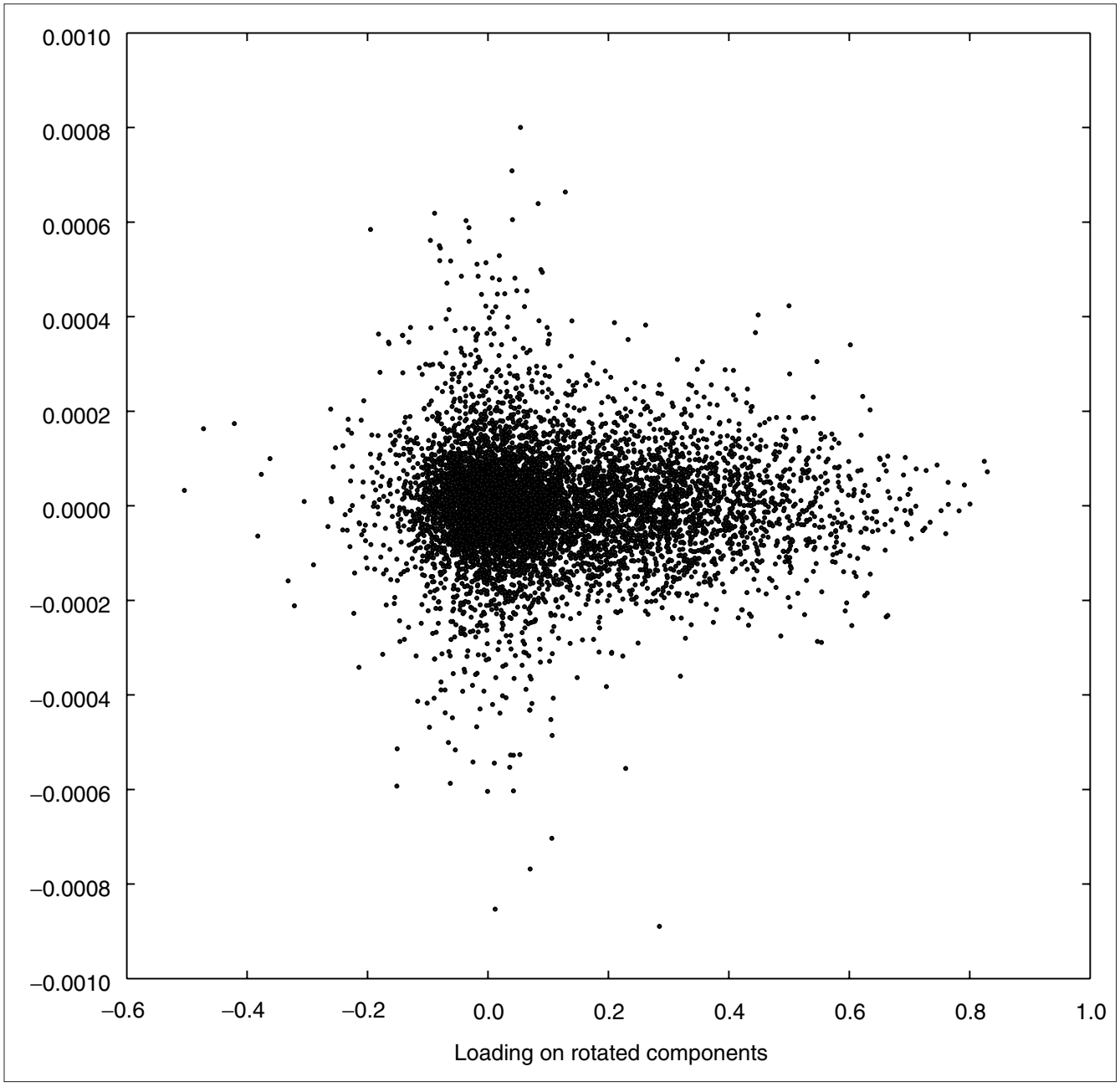

Figure 6
PCA bias showing difference between SPSS and CLUSFAVOR results for loadings of 300 gene-expression profiles on 24 rotated components (Varimax). Plot indicates that less than a 0.001 bias in loadings was obtained from a run containing 300 genes with 60 arrays.

References

1. **BLAST** [http://www.ncbi.nlm.nih.go v/BLAST/]
2. Suzuki Y, Ishihara D, Sasaki M, Nakagawa H, Hata H, Tsunoda T, Watanabe M, Komatsu T, Ota T, Isogai T, *et al.*: **Statistical analysis of the 5′ untranslated region of human mRNA using "Oligo-Capped" cDNA libraries.** *Genomics* 2000, **64:**286-297.
3. Arnone MI, Davidson EH: **The hardwiring of development: organization and function of genomic regulatory systems.** *Development* 1997, **124:**1851-1864.
4. Manson McGuire A, Church GM: **Predicting regulons and their cis-regulatory motifs by comparative genomics.** *Nucleic Acids Res* 2000, **28:**4523-4530.
5. Wuensche A: **Genomic regulation modeled as a network with basins of attraction.** *Pac Symp Biocomput* 1998:89-102.
6. D'haeseleer P, Liang S, Somogyi R: **Genetic network inference: from co-expression clustering to reverse engineering.** *Bioinformatics* 2000, **16:**707-726.
7. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.
8. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27:**167-171.
9. **Microsoft Visual Basic.NET** [http://msdn.microsoft.com/vstudio/]

10. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24:**227-235.
11. Heyer LJ, Kruglyak S, Yooseph S: **Exploring expression data: identification and analysis of coexpressed genes.** *Genome Res* 1999, **9:**1106-1115.
12. Kaiser HF: **The varimax criterion for analytic rotation in factor analysis.** *Psychometrika* 1958, **23:**187-200.
13. **SPSS** [http://www.spss.com]
14. **CLUSFAVOR** [http://mbcr.bcm.tmc.edu/genepi/]
15. Eisen, MB, Spellman, PT, Brown, PO, Botstein, D: **Cluster analysis and display of genome-wide expression patterns**, *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
16. **SAS** [http://www.sas.com]
17. **Statistica** [http://www.statsoft.com]
18. **Insightful** [http://www.insightful.com]
19. **The R project for statistical computing** [http://www.r-project.org]