BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 4 2001
Pages 369–370

J-Express: exploring gene expression data using Java

*B. Dysvik and I. Jonassen**

Department of Informatics, University of Bergen, HIB, N5020 Bergen, Norway

Received on September 4, 2000; revised on October 30, 2000; accepted on November 23, 2000

ABSTRACT

Summary: J-Express is a Java application that allows the user to analyze gene expression (microarray) data in a flexible way giving access to multidimensional scaling, clustering, and visualization methods in an integrated manner. Specifically, J-Express includes implementations of hierarchical clustering, *k*-means, principal component analysis, and self-organizing maps. At present, it does not include methods for comparing two or more experiments for differentially expressed genes. The application is completely portable and requires only that a Java runtime environment 1.2 is installed on the system. Its efficiency allows interactive clustering of thousands of expression profiles on standard personal computers.

Availability: http://www.ii.uib.no/~bjarted/jexpress/

Contact: bjarted@ii.uib.no

Microarray experiments produce large amounts of numerical data quantifying the (relative or absolute) expression level of each gene in a number of conditions. This data needs to be analyzed and visualized in order for it to help uncover new biological knowledge (see e.g. Brazma and Vilo, 2000). To help investigators explore their data, tools integrating powerful analysis methods together with visualization and interactive use are of paramount importance. We developed the tool J-Express that fulfills these requirements and additionally it requires relatively modest hardware resources and is completely portable.

J-Express presents a graphical user interface that integrates a number of viewing capabilities (implemented as different viewer windows), analysis methods that can be applied to the full data set or a subset selected through one of the viewers, and ways for importing and exporting data and analysis results (Figure 1). Input data are loaded as text files allowing different data delimiters (tab and space). Once a data file has been loaded, its contents is presented in a spreadsheet-like window that lets the user easily select the rows and columns containing gene identifiers and expression values. Internally its identifier and a vector of expression values represent each gene. The vectors can be subjected to one of the analysis methods and the results visualized in different viewers where the identifiers (when appropriate) are used to identify each gene.

The analysis tools include implementations of the standard hierarchical and *k*-mean clustering methods. The hierarchical clustering includes options for single, complete or average linkage and its output is a tree having the genes as leaves and clusters of similar genes (as measured by Euclidean distance between their expression vectors) as sub-trees. Substantial effort has been invested to make the clustering procedure fast and memory economical. A tree-viewer similar to that used by Eisen *et al.* (1998) allows the user to inspect the results. The viewer is interactive and when the user clicks on an internal node, a gene graph viewer shows the expression profiles of the genes in the sub-tree below the clicked node. The *k*-means method produces *k* non-overlapping clusters (*k* is chosen by the user) together containing the full gene set. The output is visualized by an array of gene graphs and clicking one of these will open a corresponding graph viewer.

An alternative to applying a clustering method to a numerical multi-dimensional data set is to apply a multi-dimensional scaling (MDS) method. This will make it possible to visualize an overview of the data in two or three dimensions. By transforming the data to a lower dimension, most often some information will be lost. An MDS method designed to keep as much of the distance information as possible is the principal component analysis (PCA) method. J-Express contains an implementation of PCA and two viewers allowing the user to view the data set in two or three dimensions. The tool shows how much of the variability is captured by each principal component and allows the user to choose which components to use for viewing the data. Regions of the 2D plots can be selected and the genes contained viewed in a gene graph or subjected to one of the other analysis methods in J-Express.

An intermediate between clustering and multi-dimensional scaling is provided by an implementation of a self-organizing map (SOM). The SOM can be seen as an intermediate in that it provides a grouping of the genes and at the same time the groups are organized in

*To whom correspondence should be addressed.

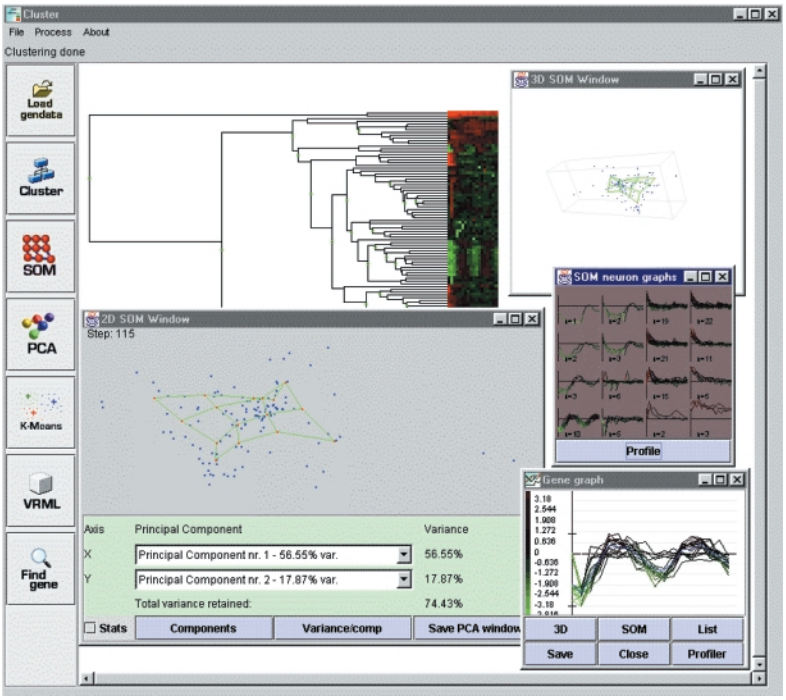

Fig. 1. A screenshot of J-Express. The main window contains pull-down menus as well as shortcut buttons. From top left corner, clockwise direction around the screenshot are shown: a tree viewer shows the result of a hierarchical clustering; a 3D view of a SOM displayed using a PCA transformation; a summary result of the SOM showed as a panel of gene graph; a gene graph window; a 2D view of a SOM. For more details, see the text and the web site.

the form of a grid (in our implementation). A special feature of the SOM implementation in J-Express is that it allows both the net and the data points to be visualized during the training of the net using the PCA method. In this way the user can get an impression of how well the SOM captures the variability of the underlying data and see how changing the parameters for the SOM algorithm affects its behavior. The output of the SOM algorithm is visualized in a similar way to that of *k*-means except that for SOM output the topology of the SOM is maintained in the layout of the viewer.

J-Express allows all plots and graphs to be saved as .gif format image files, it also allows export of cluster and tree descriptions. Furthermore, the gene graph viewer allows the user to interactively define a profile (describing allowed values for each component of an expression vector) and scan this against the complete data set. The profiles can also be saved to file and re-loaded.

In summary J-Express provides an integrated environment for exploration of gene expression data incorporating powerful and complementary analysis methods and visualization functionality. The methods implemented have been successfully applied to the analysis of gene expression data (Eisen *et al.*, 1998; Tavazoie *et al.*, 1999; Raychaudhuri *et al.*, 2000; Tamayo *et al.*, 1999). Future work will include incorporating additional methods into J-Express and investigating alternative visualization methods capable of incorporating other gene specific (and not necessarily quantitative) data.

ACKNOWLEDGEMENTS

I.J. was supported by grants from the Norwegian Research Council. The authors wish to thank Jaak Vilo and Alvis Brazma for helpful discussions.

REFERENCES

Brazma,A. and Vilo,J. (2000) Gene expression data analysis. *FEBS Lett.*, **480**, 17–24.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Raychaudhuri,S., Stuart,J.M. and Altman,R.B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*, **5**, 452–463.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovski,E., Lander,E. and Golub,T. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tavazoie,S., Hughes,J., Campbell,M., Cho,R. and Churh,G. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.