PermutMatrix: A Graphical Environment to Arrange Gene Expression Profiles in Optimal Linear Order

Gilles Caraux, Sylvie Pinloche

BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 7 2005, pages 1280–1281
doi:10.1093/bioinformatics/bti141

Gene expression

PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order

Gilles Caraux[1,2,3,*] and Sylvie Pinloche[1]

[1]Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier, 161 rue Ada, 34392 Montpellier cedex 5, France, [2]Ecole Nationale Supérieure Agronomique de Montpellier, 2, Place P.Viala, 34060 Montpellier Cedex 02, France and [3]Research School of Biological Sciences, Australian National University, Canberra, Australia

Received on June 10, 2004; revised on September 3, 2004; accepted on October 4, 2004
Advance Access publication November 16, 2004

ABSTRACT

Summary: PermutMatrix is a work space designed to graphically explore gene expression data. It relies on the graphical approach introduced by Eisen and also offers several methods for the optimal reorganization of rows and columns of a numerical dataset. For example, several methods are proposed for optimal reorganization of the leaves of a hierarchical clustering tree, along with several seriation or unidimensional scaling methods that do not require any preliminary hierarchical clustering. This program, developed for MS Windows, with MS-Visual C++, has a clear and efficient graphical interface. Large datasets can be thoroughly and quickly analyzed.

Availability: http://www.lirmm.fr/~caraux/PermutMatrix/

Contact: caraux@lirmm.fr

INTRODUCTION

To analyze DNA microarray data, it is very useful to organize and cluster genes according to similarities in their expression profiles. Standard hierarchical clustering methods are appropriate for this operation, as they provide a tree that symbolizes the structure of similarities and in which clusters can be defined. Simultaneous display of the clustering tree and the colored representation of the data matrix, as proposed by Eisen *et al*. (1998), is a very useful feature, as it can be readily interpreted by biologists. The simplicity and efficacy of this representation has made it very successful. In PermutMatrix, this approach is supplemented with several optimal linear reordering methods, such as reorganization of the leaves of a clustering tree, unidimensional scaling and seriation.

Optimal reorganization of the leaves. In the standard hierarchical clustering approach, the gene order is the order in which the leaves of the clustering tree are enumerated. However, this enumeration is not unique, as the inversion of any subtree leaves does not change the general topology of the clustering tree. The number of possible enumerations is 2^{n-1}, where n is the total number of leaves in the tree. Then the question arises as to whether it is possible to choose the best possible organization of the leaves of the tree, in order to obtain the best graphical display of the data with the Eisen approach. Several methods have been proposed (Bar-Joseph *et al*., 2001; Degerman, 1982; Gruvaeus and Wainer, 1972) and are available in PermutMatrix (Fig. 1c). They differ with respect to the criterion to be optimized and the optimization algorithm.

Unidimensional scaling and seriation. Hierarchical clustering is not aimed at reordering rows and columns of a dataset, as clustering is not the same operation as ordering. Other methods are specifically designed for ordering objects, such as unidimensional scaling (Hubert and Arabie, 1986) or seriation (Kendall, 1982). Unidimensional scaling methods involve placing a set of objects along a row so that the distances between points best reflect the dissimilarities between objects. Seriation methods (Fig. 1b) assume that there is an unknown order between the objects, and they attempt to infer this order. These methods, some of which are old, have been widely developed and implemented in archaeology and psychology. For example, they were successfully used to establish the chronology of appearance of ancient objects (Marquardt, 1978). However, these methods are relatively unknown in biology. Five of them are available in PermutMatrix and are described in detail on the program website. The criteria to be optimized in the seriation methods are the same as those used in optimal reordering of the leaves of a tree. However, the algorithms implemented here are heuristics, because the research space is too wide and unstructured. There are $n!$ ways to reorganize a set of n objects.

Identification and reorganization of classes. In PermutMatrix, the clustering tree methods and seriation or unidimensional scaling methods can be combined. It is also possible to define a class by aggregating, in the clustering tree, the leaves derived from the same node. Subsequently, the tree is no longer completely ramified (Fig. 1d). This takes into account that some terminal ramifications are not significant, and associated leaves can be reordered without tree constraint. The tree reorganization occurs on two levels: the classes are reordered as the leaves of a tree, and the leaves within each class are linearly reordered by seriation or unidimensional scaling.

Manual operations. The PermutMatrix graphical interface allows several manual operations: inversion, permutation, sorting, etc. These operations can be used to refine or locally explore the optimal solutions obtained by the methods given above.

CONCLUSION

PermutMatrix is a user-friendly and exploratory work space in which the graphical Eisen approach can be easily used and extended to optimal reorganization methods, which are less utilized than

*To whom correspondence should be addressed.

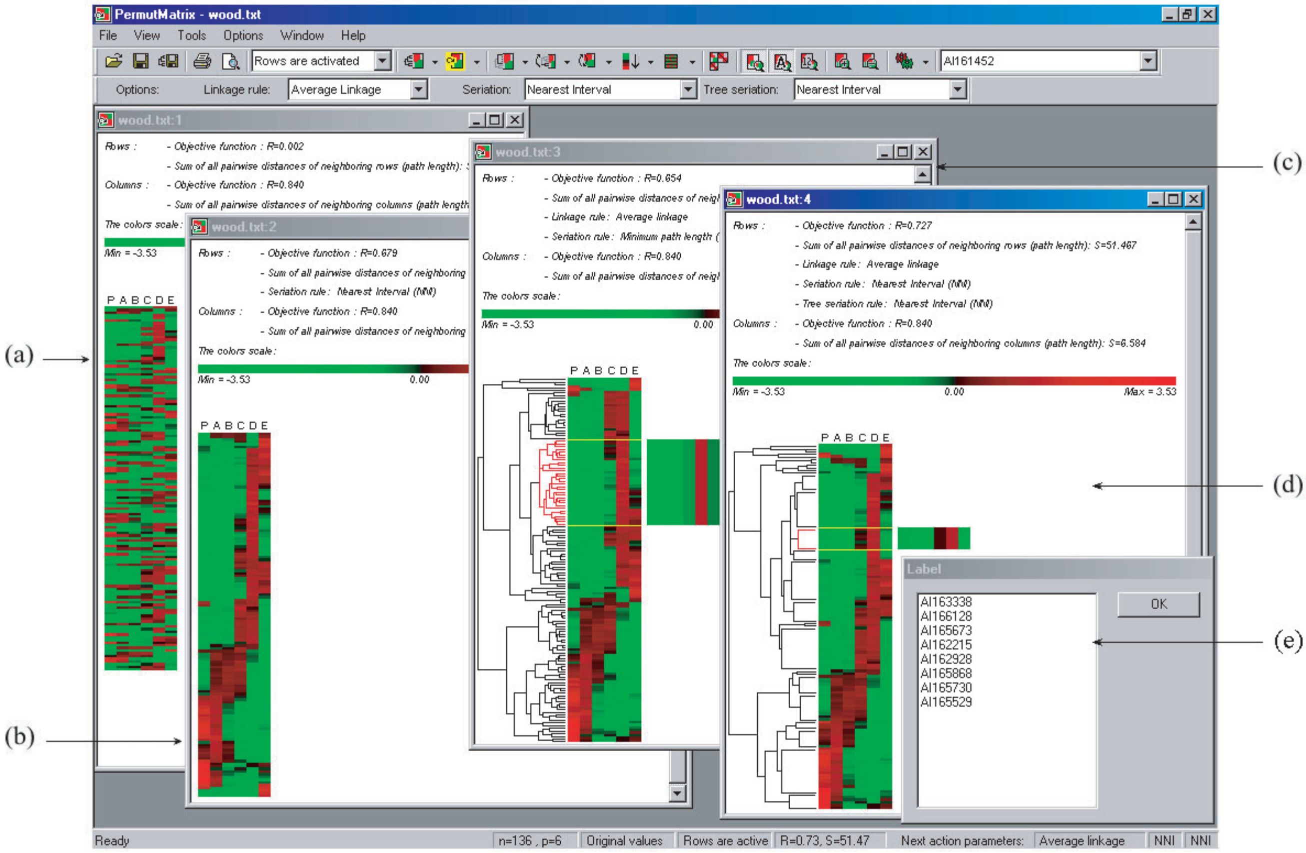

Fig. 1. PermutMatrix has a standard multiwindow operating environment. Various clustering and seriation methods can be selected via the toolbar and parameter options. Each result is presented in a separate window. The above figure, shows the results of a sample treatment. Window **(a)** shows the initial data (log2-ratio) prior to clustering, window **(b)** displays the results of a zero-constraint seriation operation (Gelfand, 1971), window **(c)** contains a hierarchical clustering result, after optimal reorganization (Bar-Joseph *et al*., 2001) and window **(d)** shows a reorganized set of classes. In this figure, there are also interpretation aids: the mean expression profile of a class of genes in given in windows (c and d) whereas a classified list of designated gene labels is shown in window **(e)**.

hierarchical clustering. These methods usually yield different and complementary results, therefore contributing to the understanding and identification of different gene expression profiles. It was designed for MS Windows and accepts any input data file in a standard text file format or in Eisen's Cluster format.

ACKNOWLEDGEMENTS

This program was partly developed during a sabbatical position at the Research School of Biological Sciences of the Australian National University. We are grateful for all support and friendship during this period. We have also been supported by Montpellier L-R Genopole.

REFERENCES

Bar-Joseph,Z., Gifford,D. and Jaakkola,T.S. (2001) Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, **17**, S22–S29.

Degerman,R. (1982) Ordered binary trees constructed through an application of Kendall's tau. *Psychometrika*, **47**, 523–527.

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.

Gelfand,A.E. (1971) Rapid seriation methods with archaeological application. In Hodson,F.R., Kendall,D.G. and Tautu (eds), *Mathematics in the Archaeological and Historical Sciences*. Edinburgh University Press, Edinburgh, pp. 186–201.

Gruvaeus,G. and Wainer,H. (1972) Two additions to hierarchical cluster analysis. *Br. J. Math. Stat. Psychol.*, **25**, 200–206.

Hubert,L.J. and Arabie,P. (1986) Unidimensional scaling and combinatorial optimization. In de Leeuw,J., Heiser,W., Meulman,J. and Critchley,F. (eds), *Multidimensional Data Analysis*. DSWO Press, Leiden, The Netherlands, pp. 181–196.

Kendall,D.G. (1982) Seriation. In Kotz,S. and Johnson,N.L. (eds), *Encyclopedia of Stastistical Sciences*. Wiley-Interscience, New York, NY, Vol. 8, pp. 417–424.

Marquardt,W.H. (1978) Advances in archaeological seriation. In Schiffer,M.B. (ed.), *Advances in Archaeological Method and Theory*. Academic Press, Orlando, FL, Vol. 1, pp. 257–314.