BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 6 2001
Pages 569–570

AMADA: analysis of microarray data

Xuhua Xia and Zheng Xie

Department of Ecology & Biodiversity, University of Hong Kong, Hong Kong, Peoples Republic of China

Received on November 10, 2000; revised on January 6, 2001; accepted on February 9, 2001

ABSTRACT

Summary: AMADA is a Windows program for identifying co-expressed genes from microarray data. It performs data transformation, principal component analysis, a variety of cluster analyses and extensive graphic functions for visualizing expression profiles.

Availability: http://web.hku.hk/~xxia/software/AMADA.htm, free.

Contact: xxia@hkusua.hku.hk

INTRODUCTION

With the completion of a number of genomic sequences, the next step in understanding the operation of the genetic program is to know how the subroutines (genes) interact with each other. Take human development for example. If we designate the time of zygote formation as t_0, what genes are activated at $t_1, t_2, \ldots, t_n$? How do the products of these activated genes activate other genes and lead to the developmental cascade? If we know that Gene A activates Gene B which in turn activates Gene C, then we would be at a good position to understand what input Gene B takes and what output it generates. If Genes A, B, and C are all activated by Gene D, then we know that Genes A, B, and C share the same input, although they may produce different output.

DNA microarray technology is exactly such a method for exploring gene interactions at the genomic scale (Diehn *et al.*, 2000; Epstein and Butow, 2000; Gaasterland and Bekiranov, 2000). Centralized databases are currently being established to make the rapidly increasing data available to the public (Brazma *et al.*, 2000). However, the development of analytical methods is lagging much behind (Bittner *et al.*, 1999).

Publicly available microarray data are typically in the form of matrices, each summarizing the expression profiles of thousands of gene loci over a period of time. Such data can provide two kinds of information that is related to transcription pathways and gene interactions. The first is the co-expressed genes whose expression may be controlled by the same gene product, e.g. they may share the same transcription factor. The second is the regulator–regulatee relationship, in which one group of genes (regulatees) increase or decrease their expression consistently with the increase or decrease of the expression of another group of genes (regulators).

The co-expressed genes can be identified by calculating pair-wise similarity or dissimilarity indices among genes, and then clustering into gene clusters by using one of the many available clustering techniques (e.g. Bittner *et al.*, 1999; Chen *et al.*, 1999; Heyer *et al.*, 1999). Pearson correlation and the jack-knife correlation (Heyer *et al.*, 1999) have been proposed. The former is not robust against outliers and the latter is too time consuming to compute. An alternative is the non-parametric Spearman's r_S that is easy to compute and robust against one or multiple outliers.

For dissimilarity measures, the Euclidean distance has been suggested (Chen *et al.*, 1999; Heyer *et al.*, 1999). Other distances used in clustering algorithms include Manhattan metric, percent remoteness, chord distance, and geodesic distance (Pielou, 1984). All these distances are metric, and satisfy triangular inequality. AMADA implements these distances as well as the Pearson and Spearman correlations.

For clustering algorithms, both hierarchical and non-hierarchical ones have been proposed and used in research (Tamayo *et al.*, 1999; Tavazoie *et al.*, 1999; Eisen *et al.*, 1998; Tavazoie and Church, 1998; Wen *et al.*, 1998). The former include single-linkage, complete-linkage and average-linkage clustering (Pielou, 1984), and the latter include the k-mean clustering, the self-organization map, and the QT-clustering (Heyer *et al.*, 1999). The k-mean clustering requires the specification of the number of clusters (k) at the beginning, but k is unknown to the researcher. If the guessed k value is too large, then co-expressed genes may be split into different clusters. If the guessed value is too small, then unrelated genes may be forced into the same cluster. This problem is shared by the method of the self-organization map. The QT-clustering algorithm has three disadvantages. First, it is time-consuming. Second, the criterion of choosing the cluster with the largest number of member as the best cluster is dubious. A more sensible criterion would be to choose a cluster as the best if it has the least overlap with others. Third, the 'quality guarantee' is just

a guessed value. I note that all clusters recovered by the QT-clustering (Heyer *et al.*, 1999) can also be recovered by the average-linkage method. So the former has no obvious advantage to offset the disadvantages. AMADA implements the single-linkage, complete-linkage and average-linkage algorithms.

DESCRIPTION

AMADAs user interface is similar to that of Microsoft Excel. A typical analysis identifying co-expressed genes can be done in four steps:

(1) Click **File**|**Open** to read in a data file in either Excel format or text format (tab- or comma-delimited).

(2) Optionally click **Data**|**Standardize values in rows** to transform each expression profiles to have a mean equal to 0 and a variance equal to 1.

(3) Click **Analysis**|**Clustering**. Click one of the seven similarity or distance indices (Pearson correlation, Spearman correlation, Euclidean distance, Manhattan metric, percent remoteness, chord distance, and geodesic distance) and one of the clustering algorithms (single-linkage, complete-linkage, average-linkage). Blanks are automatically taken as missing values. There are two common ways of handling missing values: pair-wise deletion and case-wise deletion. For illustration, with the data shown below:

GeneName	T0	T10	T20	T30	T40	T50
18srRnaa	22			43	23	29
18srRnab	5	9	−13	−9	−14	−13
18srRnac	3	−2	13		6	5

……

Case-wise deletion will delete the column headed by T10, T20 and T30 and the distance or similarity indices between, say, 18srRnab and 18srRnac are computed from the remaining three pairs of values. With pair-wise deletion, the distance or similarity indices between 18srRnab and 18srRnac are computed from the five pairs of values. AMADA handles missing values by pair-wise deletion that uses more information.

(4) AMADA now does a cluster analysis and presents a graphic window displaying a huge bifurcating tree. Genes with similar expression profiles (i.e. co-expressed genes) are clustered together. At the beginning, only the root node is displayed, and you need to left-click this node to expand the tree. The tree is scrollable and clickable. Right-click a node brings up a pop-up menu with three items which can also be accessed from the menu bar under the Tree menu, i.e. **Show subtree**, **Expression Plot**, and **Node properties**. Click **Show subtree** will display the subtree with the selected node as root. Click **Expression Plot** will display the expression profile of the loci clustered under the selected node, i.e. loci with similar expression profiles. Click **Node properties** will show basic properties of the node, such as how far it is from the tip and how many nodes grouped under this selected node, etc.

ACKNOWLEDGEMENTS

This study is supported by a CRCG grant (10203043/27662, 10203435/27662) and a RGC grant (HKU7265/00M) from the Research Grant Council of Hong Kong to X.X., and a Chinese Ministry of Education grant (99K68027) to Z.X.

REFERENCES

Bittner,M., Meltzer,P. and Trent,J. (1999) Data analysis and integration: of steps and arrows. *Nature Genet.*, **22**, 213–215.

Brazma,A., Robinson,A., Cameron,G. and Ashburner,M. (2000) One-stop shop for microarray data. *Nature*, **403**, 699–700.

Chen,Y., Bittner,M.L. and Rougherty,E.R. (1999) Issues associated with microarray data analysis and integration. *Nature Genet.*, **22**, 213–215.

Diehn,M., Eisen,M.B., Botstein,D. and Brown,P.O. (2000) Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nature Genet.*, **25**, 58–62.

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.

Epstein,C.B. and Butow,R.A. (2000) Microarray technology—enhanced versatility, persistent challenge. *Curr. Opin. Biotechnol.*, **11**, 36–41.

Gaasterland,T. and Bekiranov,S. (2000) Making the most of microarray data. *Nature Genet.*, **24**, 204–206.

Heyer,L.J., Kruglyak,S. and Yooseph,S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.*, **9**, 1106–1115.

Pielou,E.C. (1984) *The Interpretation of Ecological Data: a Primer on Classification and Ordination*. Wiley, New York.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S. *et al.* (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. and Church,G.M. (1998) Quantitative whole-genome analysis of DNA–protein interactions by in vivo methylase protection in *E.coli*. *Nat. Biotechnol.*, **16**, 566–571.

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

Wen,X., Fuhrman,S., Michaels,G.S., Carr,D.B., Smith,S., Barker,J.L. and Somogyi,R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA*, **95**, 334–339.