BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 3 2002
Pages 490–491

# GeneANOVA—gene expression analysis of variance

G. Didier, P. Brézellec, E. Remy and A. Hénaut

*Laboratoire Génome et Informatique, Tour Evry 2, 523 place des terrasses de l'Agora, 91034 Evry, France*

Received on July 13, 2001; revised on September 14, 2001; accepted on September 24, 2001

**ABSTRACT**

**Summary:** GeneANOVA is an ANOVA-based software devoted to the analysis of gene expression data.
**Availability:** GeneANOVA is freely available on request for non-commercial use.
**Contact:** didier@genopole.cnrs.fr

## 1 INTRODUCTION

This Applications Note presents a software written in JAVA, especially designed to explore gene expression data. It includes an interactive bidimensional visualization module and offers several classical features such as transformations of data, classification facilities, standard visualization (plotting gene against gene and array against array) and ANOVA tools. Here, we focus our presentation on this last point.

## 2 MOTIVATIONS—INTEREST OF ANOVA

Most microarray or SAGE experiments aim at comparing variations of gene expression with biological conditions or different cell types. However there are other sources of variation in these experiments (spotting conditions, arrays…). One way of taking this fact into account is to design more complex experiments than those with only two measures per gene. For instance, Sekowska *et al.* (2001), studied in *Bacillus subtilis* the variation of expressions across two sulfur nutrients (methionine or methylthioribose). The experiments were repeated on two differents days, with two different amounts of RNA (1 or 10 $\mu$g) and one replicate for each spot. So, there are 16 measures for each of the 4107 genes, one per combination of these four factors: sulfur, day, amount and spot.

ANOVA is a statistical tool (Fisher, 1954) well suited to the study of such a problem. More precisely, it allows an estimation of the contributions of each of these factors, and possibly of the interactions between factors, in the total variation of the whole set of measures. Moreover, under statistical assumptions, see Kerr *et al.* (2000) or Fisher (1954), it gives the significance of these contributions ($P$-values).

## 3 FUNCTIONALITIES

In our software, the use of ANOVA tools requires first the definition of the *design* (i.e. the set of factors). For instance, for data from Sekowska *et al.* (2001), we have to enter the four above factors of experiments, and the factor *Gene*. The definition of the set of factors can be saved and reloaded for future utilizations.

One of the modules of the software performs ANOVA over the whole set of data and displays the classical table (see window at the top of Figure 1). This can be done for interactions of any order. In the table, the user can select or unselect factors or interactions (the unselected ones are not taken into account in the computation). This last point is sometimes necessary for some *designs*, like Latin Square for instance, where some factors and/or interactions effects are confounded (Kerr *et al.*, 2000). The table can be saved in standard text format.

In analyzing gene expression experiments, the preceding step shows how much a factor contributes to the variation, but it does not allow to detect the genes that vary significantly with a given factor. To achieve this task, another module of the software performs ANOVA for each set of measures corresponding to a gene and displays the results in a new table. However the relevance of the expression of a gene depends on two characteristics: first, the part of its variation due to the considered factor, and second, the significance of this part. We use the $P$-value for this last characteristic. However, since the software does not assess whether the modeling assumptions underlying the ANOVA hold, the validity of the precise $P$-value cannot be verified. Instead, the $P$-value should be used as an approximate indicator of the strength of the evidence for differential expression.

In order to visualize these two characteristics simultaneously, we have developed an interactive visualization module. It displays a graphic in which each gene is plotted as a point, the abscissa being the variation due to the factor normalized by the total variation of the gene, and the ordinate the logarithm of the $P$-value (see window at the bottom of Figure 1). The interface allows the user to easily locate what point(s) is

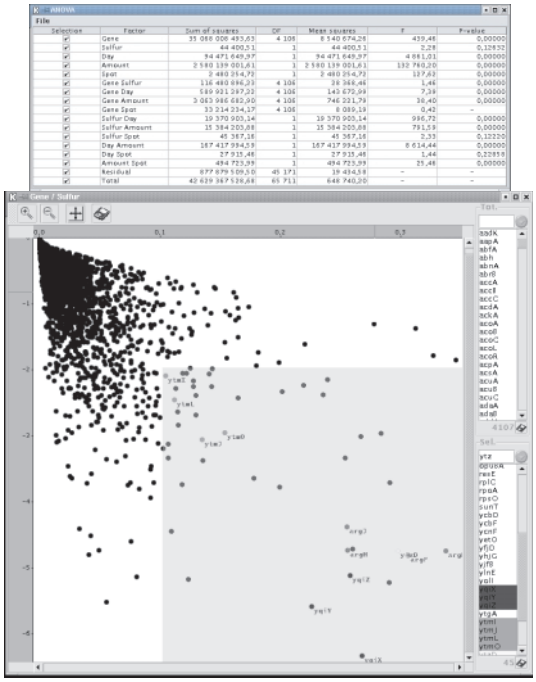

**Fig. 1.** Top: table displaying ANOVA results. Bottom: graphics module window with relevant genes (i.e. genes with large parts of variation due to the factor *Sulfur* and a high statistical significance) located in the selected area.

related to a given gene (or a set of genes) and *vice versa*.

On the right side of the image, two lists report the names of the genes. The first one (list *Tot.*) contains all the genes, and the second one (list *Sel.*) only the genes contained in the rectangular subarea defined by mouse selection (the violet zone in Figure 1). Genes in the lists can be selected by clicking on their names or by typing their prefix in the text field above each list. The color of the points on the graphic changes with the selection of the corresponding genes. A menu enables the user to display the name of the selected genes on the graphics or to attribute a new color to the corresponding points. Clicking directly on a point of the graphic with the right button shows a similar contextual menu.

In the display zone, putting the pointer on a point will display a tool tip containing the name of the gene and its coordinates.

Double clicking on a point or on the name of a gene in one of the lists opens a window containing a page resulting from an HTML request on the considered gene. This request can be set up to any HTML-interfaced database, like SWISSPROT for instance, *via Preferences* menu.

The window of the visualization module contains a toolbar which makes it possible to zoom and to save the graphics in JPEG format.

The software is able to handle datasets of usual size in microarrays experiments (several thousands of genes and a number of conditions). By using it, we easily retrieved the set of genes reported in Sekowska *et al.* (2001).

## REFERENCES

Fisher,R.A. (1954) *Statistical Methods for Research Workers*. Oliver & Boyd, London.

Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Sekowska,A., Robin,S., Daudin,J.J., Hénaut,A. and Danchin,A. (2001) Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in *Bacillus subtilis*. *Genome Biol.*, **2** 6, 1–19.