

Visualization of Microarray Data ¹

Alex Gribov ²

University of Augsburg, e-mail: alex967600@freenet.de

Abstract: In this paper the challenges of graphically displaying and analysing microarray data and their accompanying datasets are described and discussed using a dataset on Leukaemia patients from a joint research project between three groups in Germany (Augsburg University, Ulm University Clinic, German Cancer Research Centre Heidelberg). The paper also describes and illustrates the software SEURAT, which has been developed to display and interactively work with microarray data and their associated datasets in combination. SEURAT provides an array of graphical features and is able to take advantage of modelling tools available in R through making use of the package Rserve.

Keywords: Microarray data, Matrix visualization, Heatmap, SEURAT

1. Introduction

The analysis of microarray data has become very important in recent years. Medical researchers use them to analyse possible genetic causes for diseases. The reactions of individual genes to testing are described as their expression and these data are displayed in matrix visualisations or heatmaps of genes by patients (Figure 1). Unusually for statistical data, there are far more variables (genes) than cases (patients) and the patients are generally plotted in the columns with the genes in the rows. It is also important to note that gene expression data have always been filtered and standardised to some extent before display. With this description, visualization for these data involves finding ways of effectively organising and displaying a matrix of several thousand rows by at most a few hundred columns. This is not an easy task and the paper will discuss some of the tools and strategies developed to deal with it. However, there is a lot more to microarray data that needs to be taken into account.

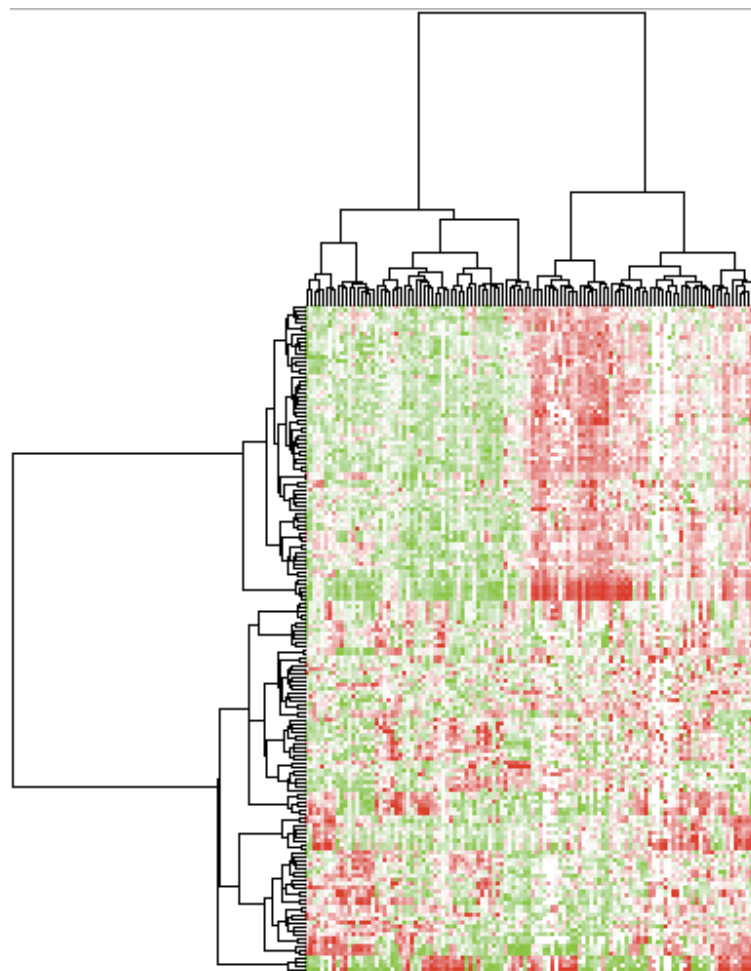
2. Four datasets

When microarray data are collected, then data about the patients are also collected in a second dataset. Any analysis of the gene expression data should reflect the different types of patient involved. This interaction can work in both directions, either to see how gene expression data varies for different patient groups or to see which patients are characterised by particular gene expression patterns. Obviously, we want to be able to visualize these interactions together with the basic expression data.

¹The present paper is financially supported by the José Carreras Foundation

²Address of correspondence: Dept. of Computer Oriented Statistics and Data Analysis, Augsburg University, 86135, Augsburg, Germany

Figure 1: A heatmap of gene expression data for leukaemia patients after clustering

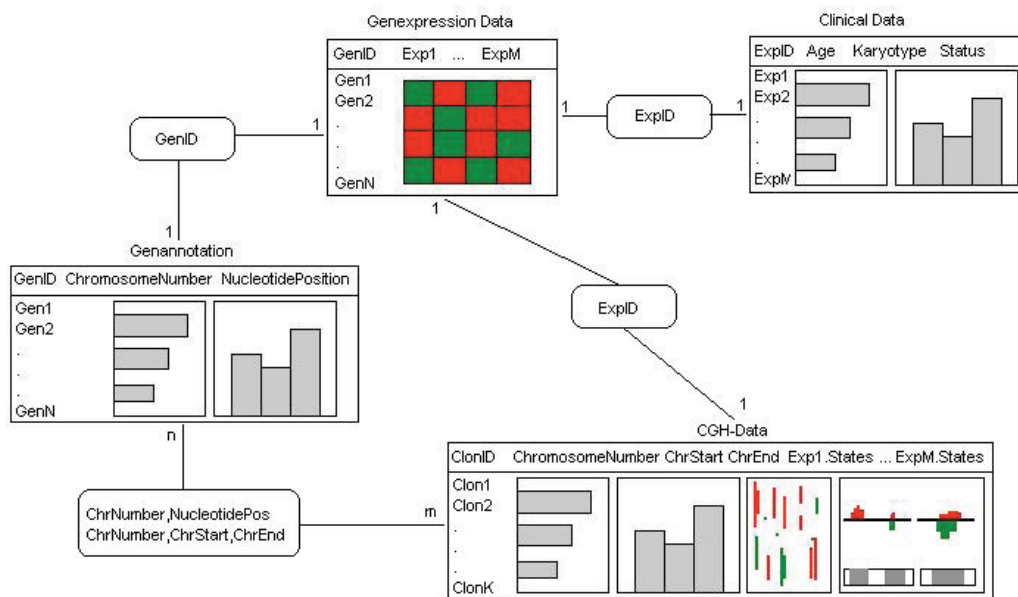


Just as information is available on the patients, there is substantial further data available on the genes in a third dataset. Given a set of genes with high values, it is important to try to interpret their functional role and this means having direct access to the corresponding database information. Gene chromosome location can also be of interest and whether genes are part of a clone (effectively a short length of a chromosome). Information on clones is available in a fourth data set. The four kinds of dataset and the connections between them are displayed in Figure 2.

3. Visualizations in SEURAT

The software SEURAT has been developed to provide interactive linked visualizations for all the various kinds of data associated with microarrays. As well as offering heatmaps for gene expression data and standard displays like histograms and barcharts for patient data, it also uses the Rserve package to provide access to the broad range of clustering and seriation tools in R. The advantage of this approach is, of course, that if any new

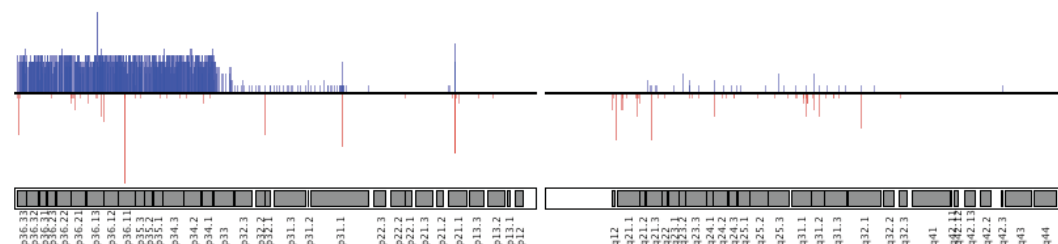
Figure 2: Four kinds of dataset associated with microarray studies



procedure becomes available in R, it can easily be incorporated in SEURAT. A good example is given by the package *amap*, which offers fast hierarchical clustering, much faster than using *hclust* directly, and is now used in SEURAT.

Information on genes can be tied to their location on a chromosome. Figure 3 shows an example with gains above and losses below the line. The chromosome bands are also displayed and can be selected individually.

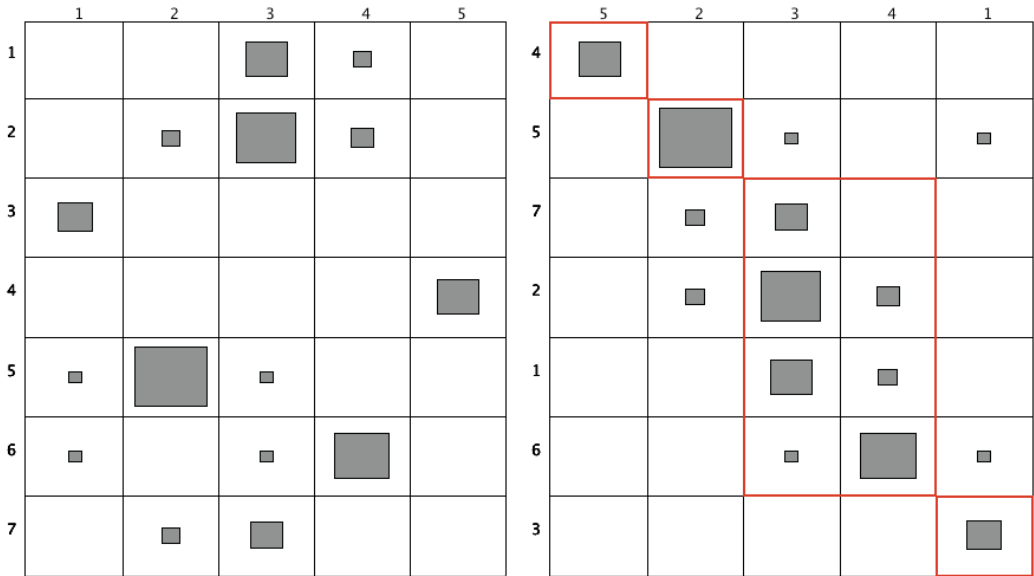
Figure 3: Gains and losses plotted for a chromosome



Heatmaps for microarray data are often clustered, as in Figure 1. There are many, many different ways of carrying out cluster analyses and it is not easy to decide which one to choose. An important step is to be able to compare two clusterings produced by different approaches. One way to do this is to draw up a confusion matrix, a table showing how many cases are in both cluster *i* of the first clustering and cluster *j* of the second. Displaying the table as a fluctuation diagram in which each table entry is represented by a rectangle whose area is proportional to its count is more effective (cf. the left-hand plot in Figure 4). Still more effective is to order the two clusterings in the rows and columns to try to maximise the agreement between them. SEURAT includes a heuristic method for

doing this and the result for the example is shown in the right-hand plot of Figure 4.

Figure 4: Comparing two clusterings, unsorted (left) and sorted (right)



4. Further developments

Linking between several datasets raises complicated issues. Interactive graphical tools have in the main been limited to analysing single datasets and innovative structures are needed for coping with these new challenges. It is not so much a matter of the visualizations themselves, how they look and how they are drawn, it is much more a matter of organising and managing them effectively. One picture may be worth a thousand words, as the old saying has it, but several pictures properly linked are even better.