MANUAL

Roman Jaksik[1,*], Michal Marczyk[2], Andrzej Polanski[3] and Joanna Polanska[2]

[1] Systems Engineering Group, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology

[2] Data Mining Group, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology

[3]Institute of Informatics, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology

GAMRED

GaMred is a light-weight fully stand-alone application implemented in Delphi XE5 programming environment, supported in both Windows and Mac OSX operating systems and including automatic switching to parallel computing mode if provided by the hardware environment. The data analysis proceeds in four steps, each realized by one program module, 'Input data' for loading dataset for analysis, 'Configuration' for setting filter type and model parameters, 'Mixture model', where modeling results are presented and 'Filtered data' for showing final results of the filtering procedure. The four program modules and their possible options are described below.

Input data panel

The first step of the data processing involves loading the data (Figure 1). High-throughput measurements should be stored as a tab-delimited text file. It is possible to analyze any kind of data organized as a two-dimensional table, where rows represent features and columns individual samples from the experiment. Loaded data are presented on the right side of the panel. On the left side of the panel the user can select all or a specific subset of samples to be further analyzed and add the name to the study, which is used to label the result files. 'Paste' button allows for copying data from the clipboard (for example taken from an Excel spreadsheet). By clicking the 'Clear' button one may clean out the table. The 'Next' button leads to the 'Configuration' panel.

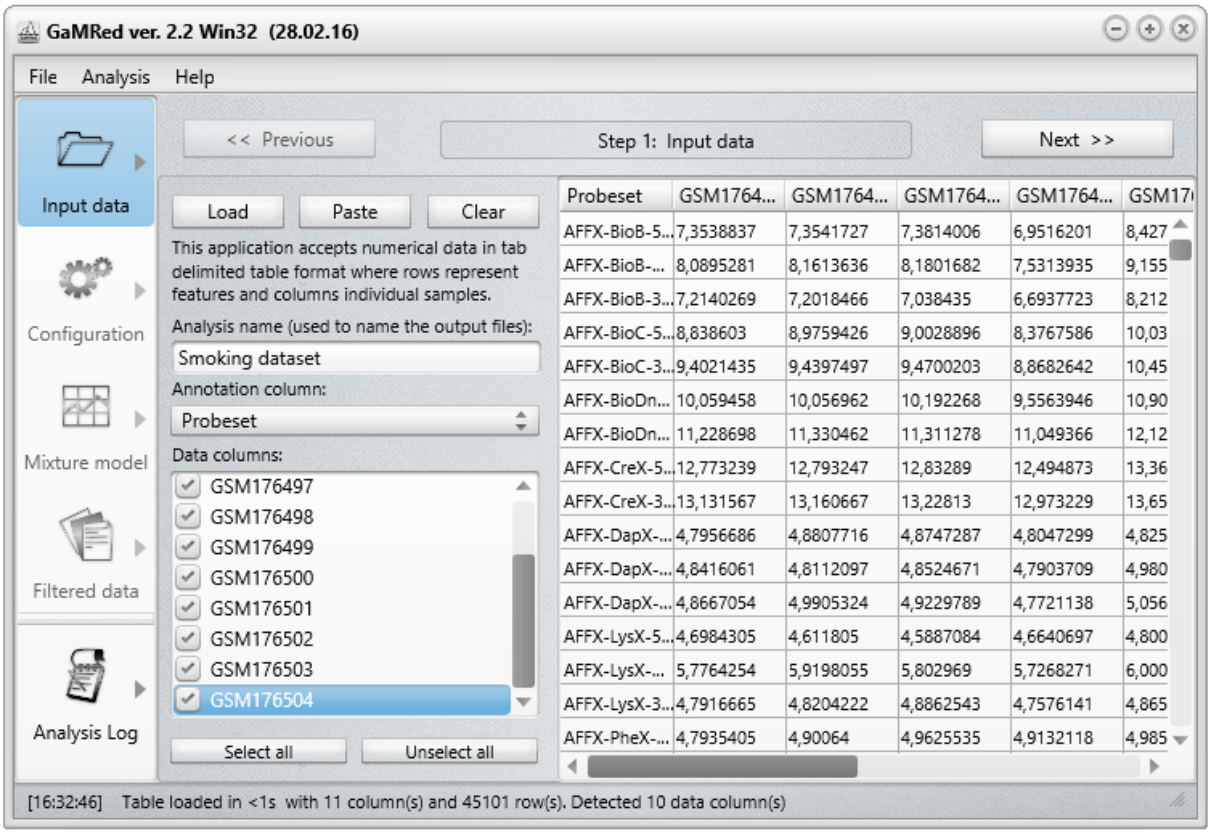

Figure 1 Input data panel

Configuration panel

The first task of the feature pre-filtering is to choose the filter type (Figure 2). The user can decide whether to analyze distributions of signals obtained by summarization of high throughput data, defined by logarithms of mean expression (S filter), variances (LV filter or V filter) or to analyze expression (in the logarithmic scale) of individual samples (IS). Below the drop-down list, there is a hint for usage of each filter type. The user can manually set a measurement scale of data present in text file (log2 or linear), however the program will also attempt to automatically recognize the scale based on the expression values. Initial conditions of the EM iterations used for Gaussian mixture model are set randomly; there is an option to choose the number of performed random trials (number of individual EM executions, each started with randomly chosen initial conditions). This reduces the risk of reaching a local maximum of the likelihood function in the EM iterations. By selecting a smaller range for the number of model components and lower EM algorithm precision level the working time can be reduced. The user can also set the maximum number of iterations, after which the EM algorithm will stop. There is also an option for placing one Gaussian component at zero, useful for datasets with fold change values. On the right hand side of the panel the user can read a description for each parameter after placing the mouse pointer on a desired term or the question mark next to its value. Again the 'Next' button leads to the panel that includes the next stage of the analysis - 'Mixture model'.

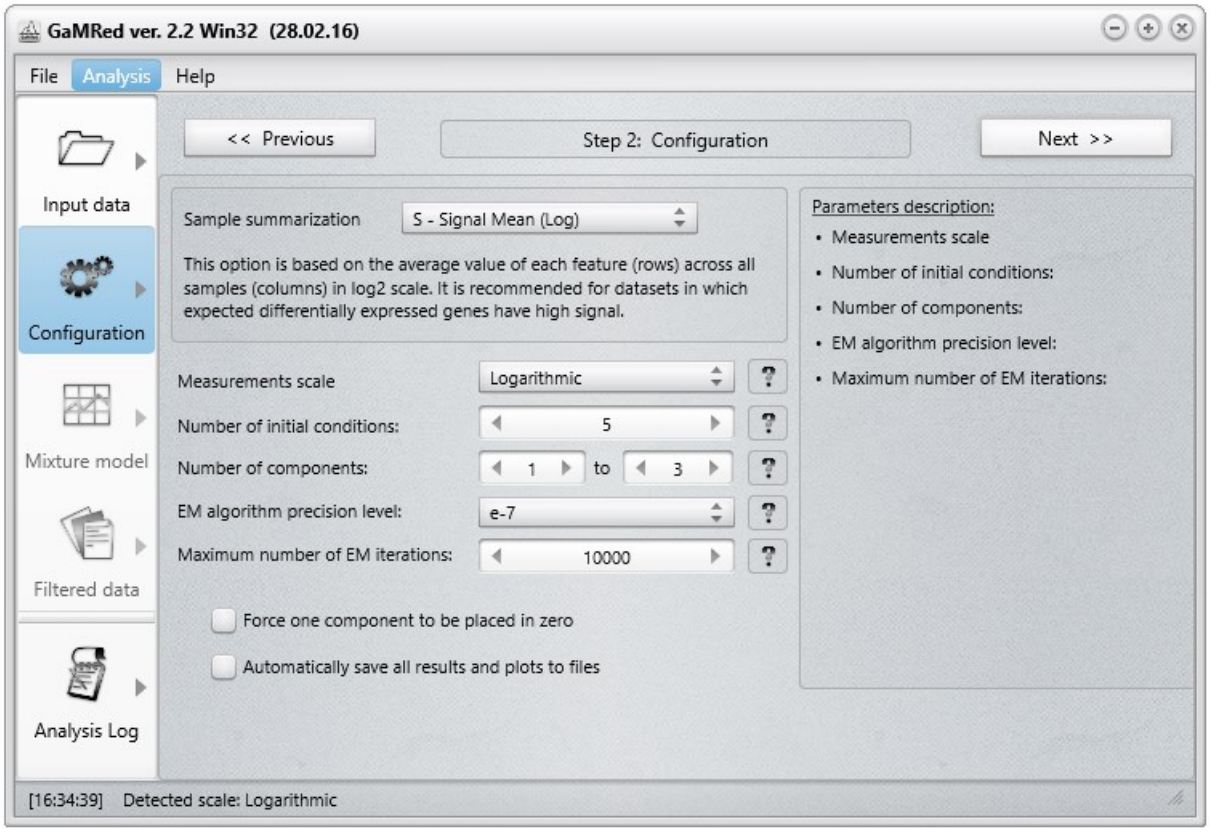

Figure 2 Configuration panel

Mixture model panel

The third panel (Figure 3) presents results of Gaussian mixture modeling using the chosen summarization method. In order to find the cut-off threshold for pre-filtering, the user can choose between k-means, top-3 and manual methods (Marczyk, et al., 2013). On the right hand side of the panel there is a plot, which illustrates the Gaussian mixture model decomposition of the analyzed signal. A histogram of the signal is drawn in grey, the GMM model distribution function is represented as a white line, components are shown in red (non-informative) and green (informative). The dashed vertical line shows the value of the estimated threshold. The user can also mark all components intersection points on the graph. There is an option to save or copy the picture to the clipboard. Again, the 'Next' button leads to the panel that includes the next stage of the analysis - 'Filtered data'.

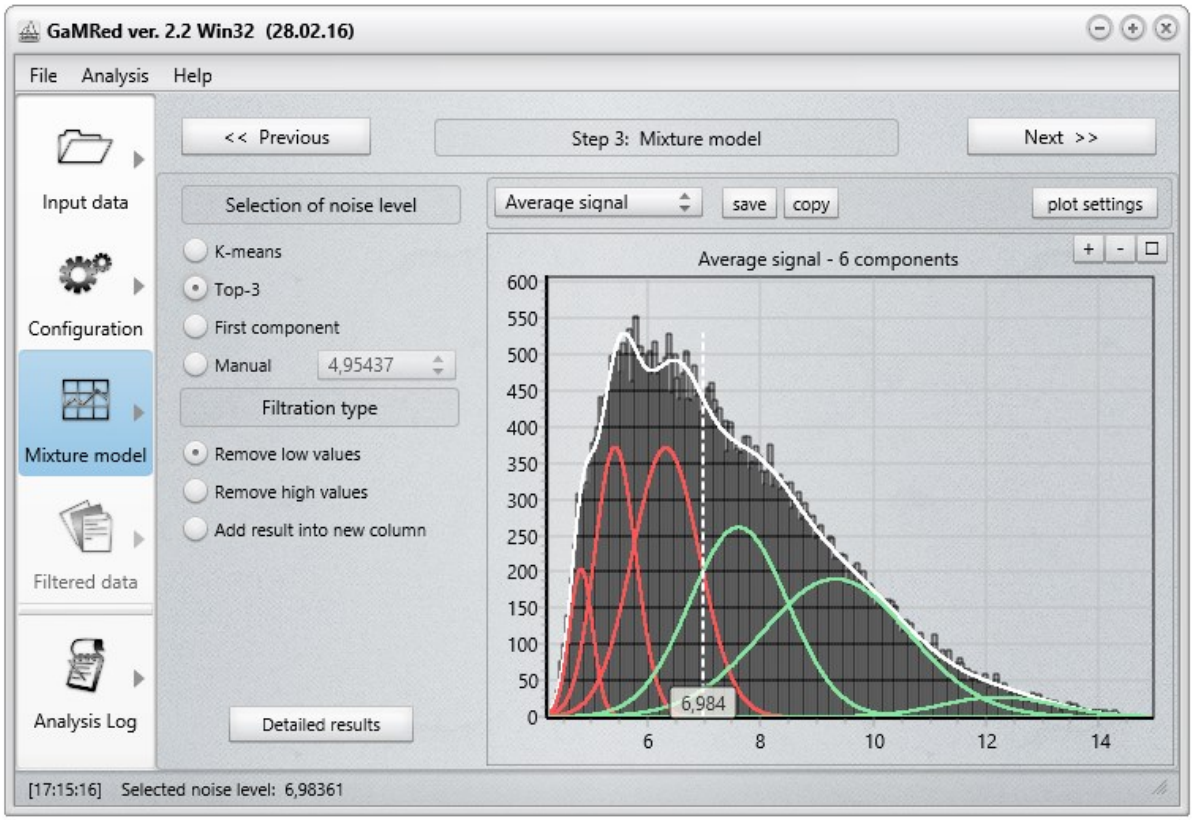

Figure 3 Mixture model panel

Filtered data panel

The fourth panel (Figure 4) includes a table with features (rows) which remain in the dataset after the filtering procedure. The user can save the table as a .txt file or copy it to the clipboard. Below the table there is an information on how many features were filtered out.

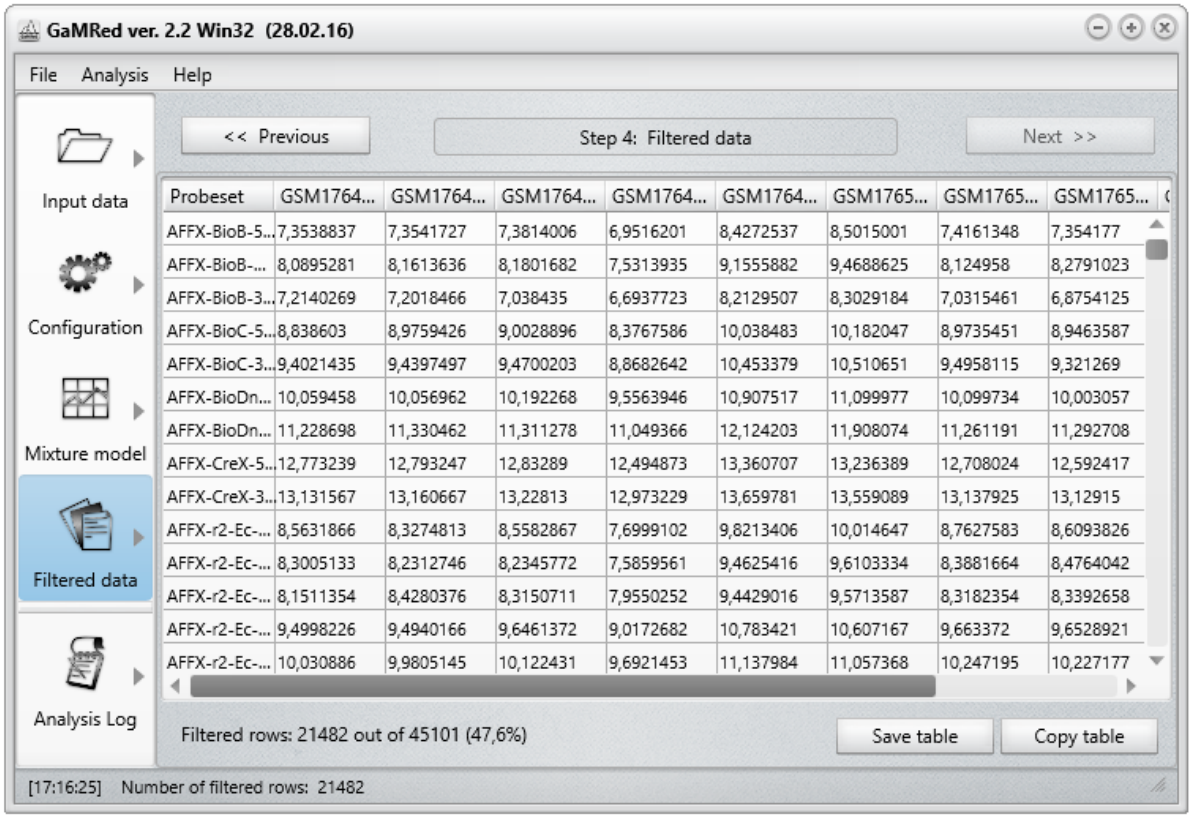

Figure 4 Filtered data panel

An additional panel, named 'Analysis log', contains information about the number of available processor cores that will be used in the data processing, time consumption for each step and all methods chosen by the user. There are also other results of the analysis, like the estimated filtering threshold or the number of features, that were filtered out.

References

Marczyk M, Jaksik R, Polanski A, Polanska J: *Adaptive filtering of microarray gene expression data based on Gaussian mixture decomposition*. BMC Bioinformatics 2013, 14(1):101.