GaMRed

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

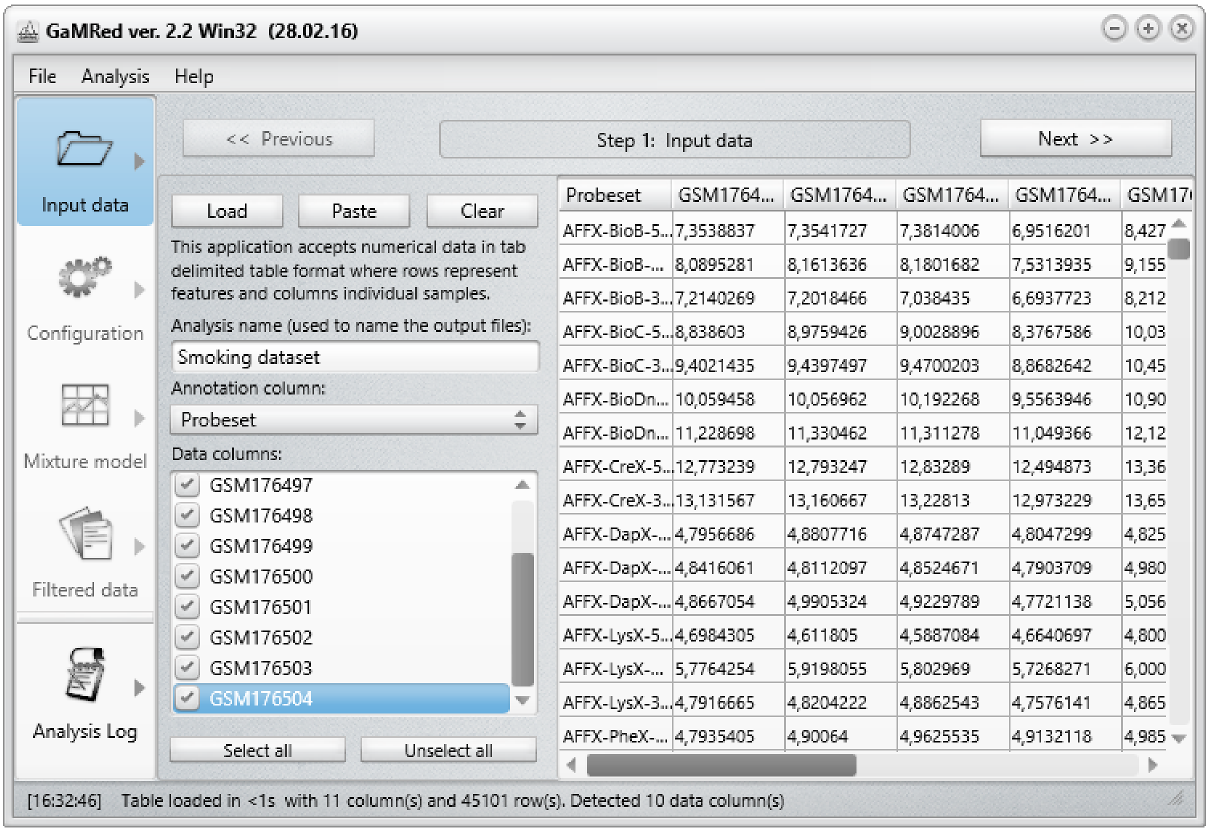
[3]Institute of Informatics, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology

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.



This application accepts numerical data in tab delimited table format where rows represent features and columns individual samples.

Analysis name (used to name the output files): Smoking dataset

Annotation column: Probeset

Data columns:

- ☑ GSM176497
- ☑ GSM176498
- ☑ GSM176499
- ☑ GSM176500
- ☑ GSM176501
- ☑ GSM176502
- ☑ GSM176503
- ☑ GSM176504

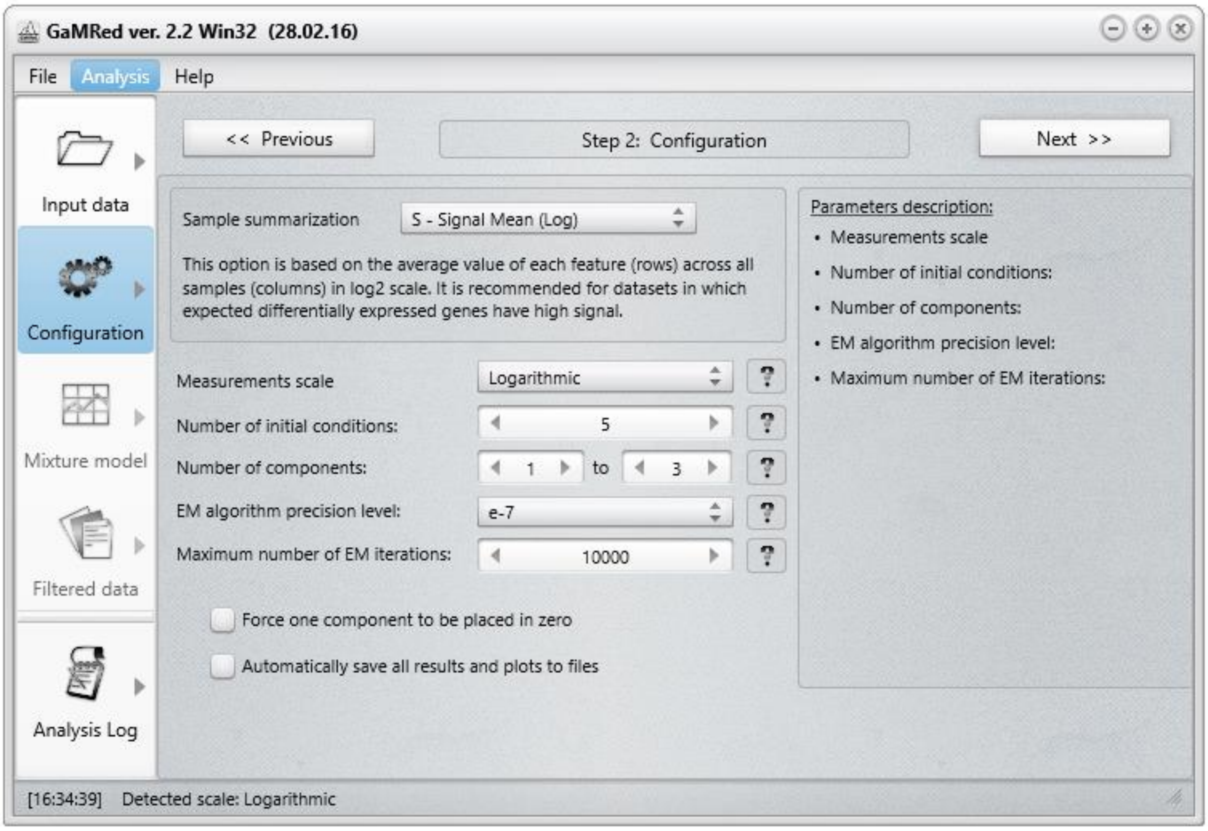
Probeset	GSM1764...	GSM1764...	GSM1764...	GSM1764...	GSM17
AFFX-BioB-5...	7,3538837	7,3541727	7,3814006	6,9516201	8,427
AFFX-BioB-...	8,0895281	8,1613636	8,1801682	7,5313935	9,155
AFFX-BioB-3...	7,2140269	7,2018466	7,038435	6,6937723	8,212
AFFX-BioC-5...	8,838603	8,9759426	9,0028896	8,3767586	10,03
AFFX-BioC-3...	9,4021435	9,4397497	9,4700203	8,8682642	10,45
AFFX-BioDn...	10,059458	10,056962	10,192268	9,5563946	10,90
AFFX-BioDn...	11,228698	11,330462	11,311278	11,049366	12,12
AFFX-CreX-5...	12,773239	12,793247	12,83289	12,494873	13,36
AFFX-CreX-3...	13,131567	13,160667	13,22813	12,973229	13,65
AFFX-DapX-...	4,7956686	4,8807716	4,8747287	4,8047299	4,825
AFFX-DapX-...	4,8416061	4,8112097	4,8524671	4,7903709	4,980
AFFX-DapX-...	4,8667054	4,9905324	4,9229789	4,7721138	5,056
AFFX-LysX-5...	4,6984305	4,611805	4,5887084	4,6640697	4,800
AFFX-LysX-...	5,7764254	5,9198055	5,802969	5,7268271	6,000
AFFX-LysX-3...	4,7916665	4,8204222	4,8862543	4,7576141	4,865
AFFX-PheX-...	4,7935405	4,90064	4,9625535	4,9132118	4,985

```
[16:32:46]   Table loaded in <1s  with 11 column(s) and 45101 row(s). Detected 10 data column(s)
```

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'.



Sample summarization: S - Signal Mean (Log)

This option is based on the average value of each feature (rows) across all samples (columns) in log2 scale. It is recommended for datasets in which expected differentially expressed genes have high signal.

Measurements scale: Logarithmic

Number of initial conditions: 5

Number of components: 1 to 3

EM algorithm precision level: e-7

Maximum number of EM iterations: 10000

☐ Force one component to be placed in zero

☐ Automatically save all results and plots to files

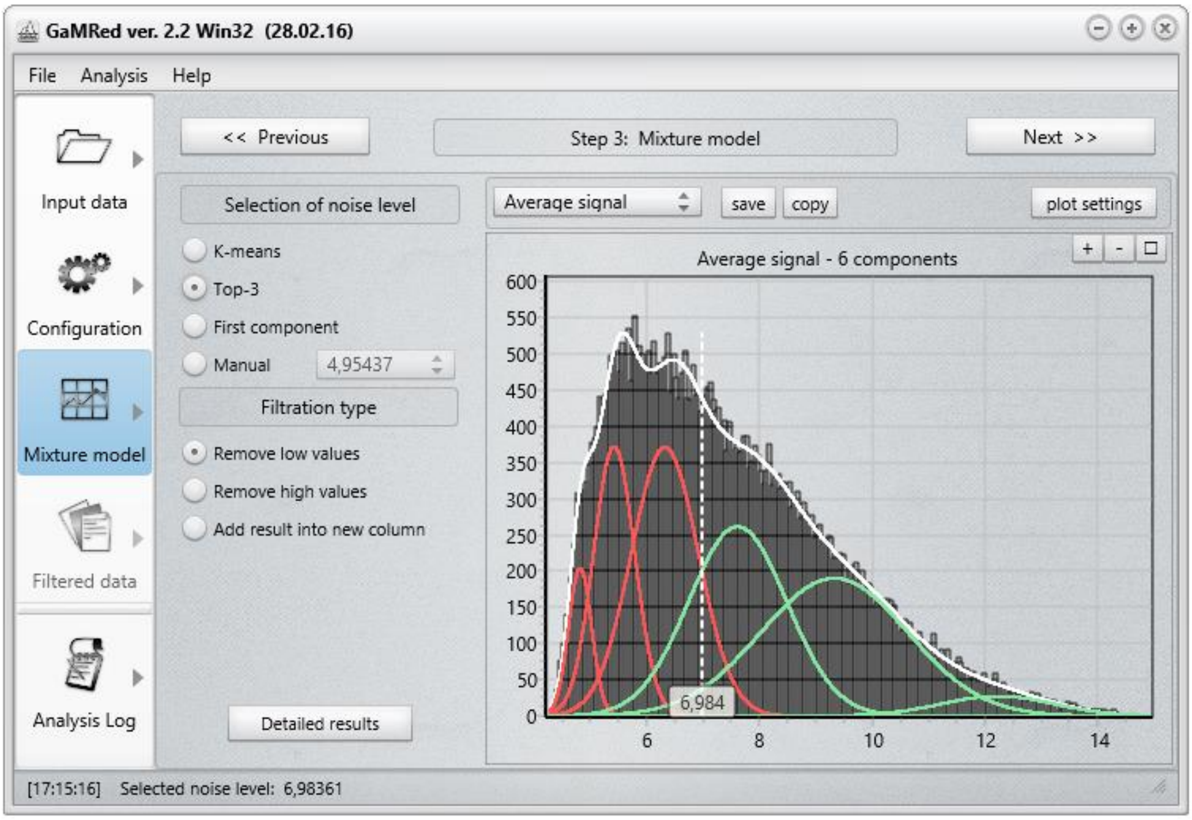
Parameters description:
- Measurements scale
- Number of initial conditions:
- Number of components:
- EM algorithm precision level:
- Maximum number of EM iterations:

```
[16:34:39]   Detected scale: Logarithmic
```

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'.



Selection of noise level

○ K-means
● Top-3
○ First component
○ Manual 4,95437

Filtration type

● Remove low values
○ Remove high values
○ Add result into new column

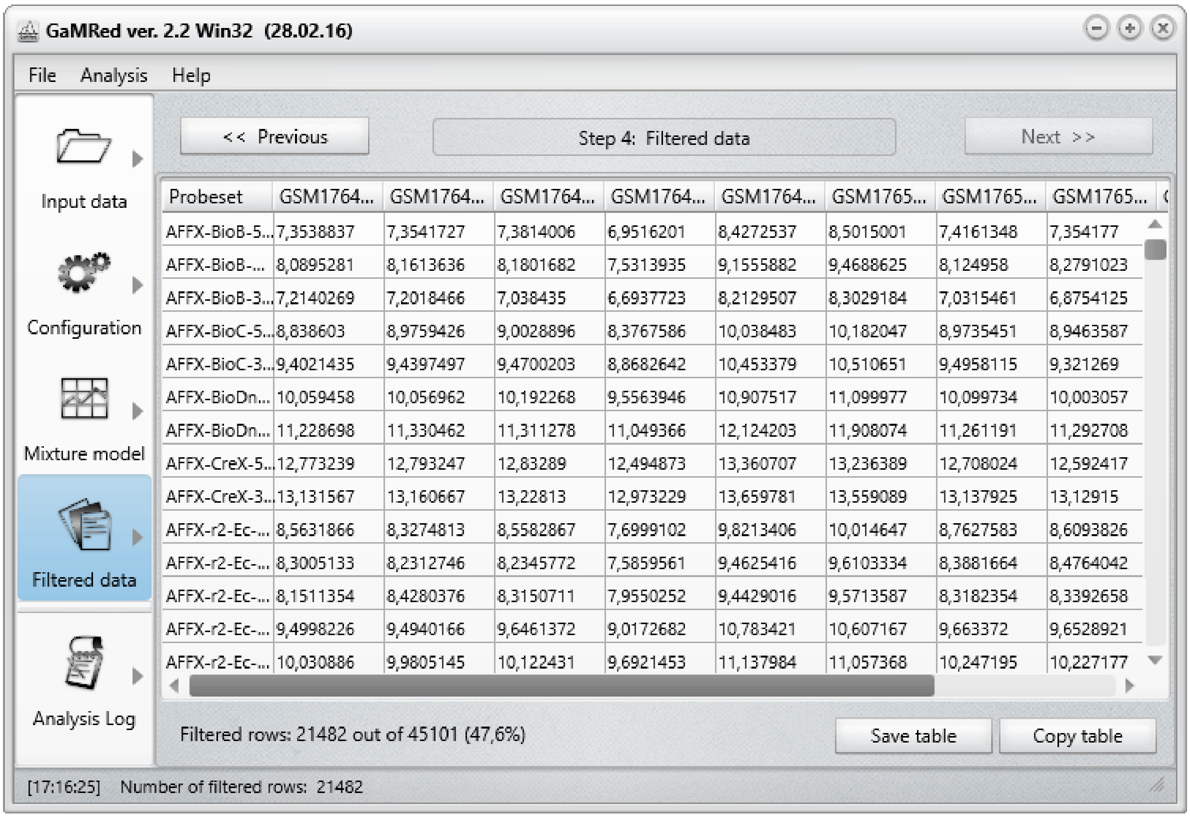
Average signal

```
[17:15:16]   Selected noise level: 6,98361
```

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.



Probeset	GSM1764...	GSM1764...	GSM1764...	GSM1764...	GSM1764...	GSM1765...	GSM1765...	GSM1765...
AFFX-BioB-5...	7,3538837	7,3541727	7,3814006	6,9516201	8,4272537	8,5015001	7,4161348	7,354177
AFFX-BioB-...	8,0895281	8,1613636	8,1801682	7,5313935	9,1555882	9,4688625	8,124958	8,2791023
AFFX-BioB-3...	7,2140269	7,2018466	7,038435	6,6937723	8,2129507	8,3029184	7,0315461	6,8754125
AFFX-BioC-5...	8,838603	8,9759426	9,0028896	8,3767586	10,038483	10,182047	8,9735451	8,9463587
AFFX-BioC-3...	9,4021435	9,4397497	9,4700203	8,8682642	10,453379	10,510651	9,4958115	9,321269
AFFX-BioDn...	10,059458	10,056962	10,192268	9,5563946	10,907517	11,099977	10,099734	10,003057
AFFX-BioDn...	11,228698	11,330462	11,311278	11,049366	12,124203	11,908074	11,261191	11,292708
AFFX-CreX-5...	12,773239	12,793247	12,83289	12,494873	13,360707	13,236389	12,708024	12,592417
AFFX-CreX-3...	13,131567	13,160667	13,22813	12,973229	13,659781	13,559089	13,137925	13,12915
AFFX-r2-Ec-...	8,5631866	8,3274813	8,5582867	7,6999102	9,8213406	10,014647	8,7627583	8,6093826
AFFX-r2-Ec-...	8,3005133	8,2312746	8,2345772	7,5859561	9,4625416	9,6103334	8,3881664	8,4764042
AFFX-r2-Ec-...	8,1511354	8,4280376	8,3150711	7,9550252	9,4429016	9,5713587	8,3182354	8,3392658
AFFX-r2-Ec-...	9,4998226	9,4940166	9,6461372	9,0172682	10,783421	10,607167	9,663372	9,6528921
AFFX-r2-Ec-...	10,030886	9,9805145	10,122431	9,6921453	11,137984	11,057368	10,247195	10,227177

Filtered rows: 21482 out of 45101 (47,6%)

[17:16:25] Number of filtered rows: 21482

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.