GED PRO TOOLS: Gene Expression Data prePROcessing TOOLS.

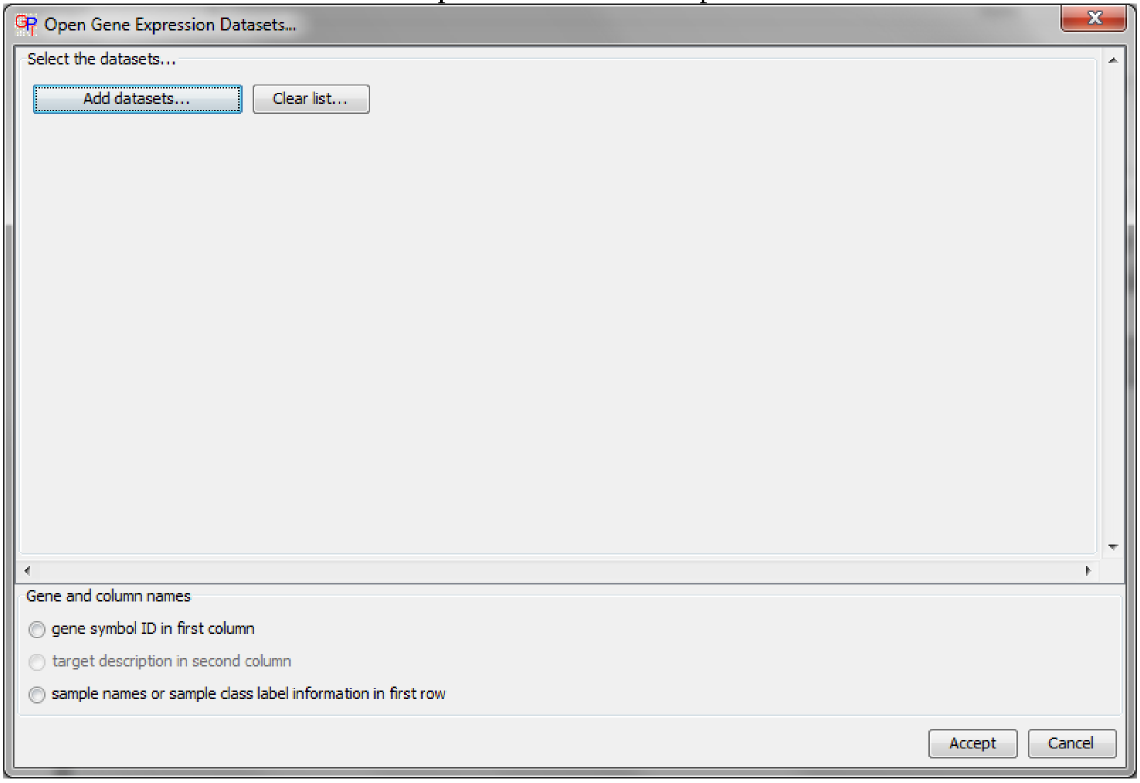
Requirements

GEDPROTOOLS is a java application and requires a Java Runtime Environment (JRE) 1.6 or greater in order to be executed. Just double click in the gedprotools.jar file to run the application. It will run in both **Windows** and **Linux** based systems (make sure the .jar file has executable permissions and that the Java Virtual Machine is properly configured). It can also be executed from a command line with: java –jar gedprotools.jar

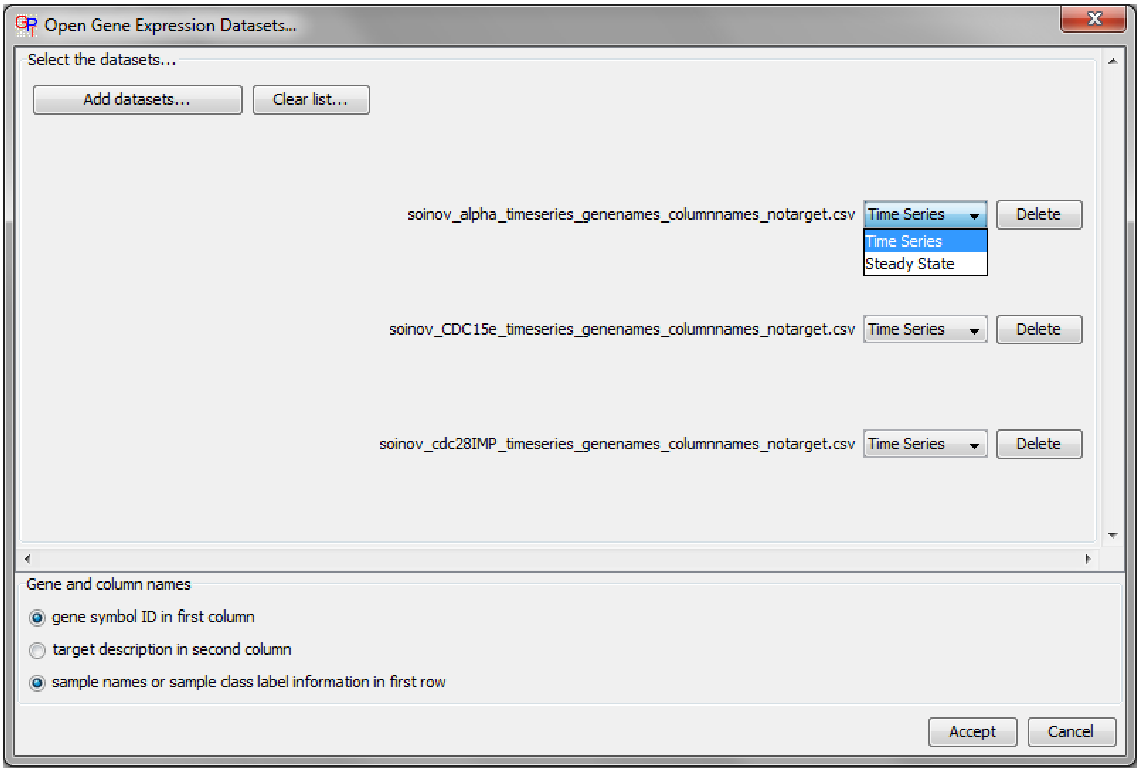
Open Gene Expression Datasets

To open the loading GED dialog box, select File -> Open GED in the menu or press the button on the options bar.

Once the dialog is shown, you can add datasets by pressing the "Add datasets..." button in the left upper corner of the dialog. The application allows multiple file selection so you can select all the data at once or you can press the button and choose the files one by one. The datasets must be in CSV format in order to be compatible with GEDPROTOOLS. Check the sample data for an example.


Open Gene Expression Datasets...
Select the datasets...
Add datasets...
Clear list...
Gene and column names
○ gene symbol ID in first column
○ target description in second column
○ sample names or sample class label information in first row
Accept
Cancel

When you have selected the datasets, they will be displayed in the dialog box and, you can choose for each one if they correspond to Time Series data or Steady State data. Also, you can delete it from the list individually or you can delete all the files selected by pressing the "Clear list" button.



Select the datasets...

soinov_alpha_timeseries_genenames_columnnames_notarget.csv: Time Series (Time Series / Steady State)

soinov_CDC15e_timeseries_genenames_columnnames_notarget.csv: Time Series

soinov_cdc28IMP_timeseries_genenames_columnnames_notarget.csv: Time Series

Gene and column names

● gene symbol ID in first column

○ target description in second column

● sample names or sample class label information in first row

Finally, you need to specify if the data has the gene names in the first column and/or if they have the sample names (or class label information) in the first row. Also, you can select if the genes have annotations in the second column on the datasets. If any of these options were not elected an anonymous name will be assigned to the non-selected one. In order to discretize the datasets with supervised approaches, class label information is required for each column. Thereby, each column must be annotated with an integer specifying the class corresponding to the column, and the "sample names or sample class label information in first row" option must be selected.

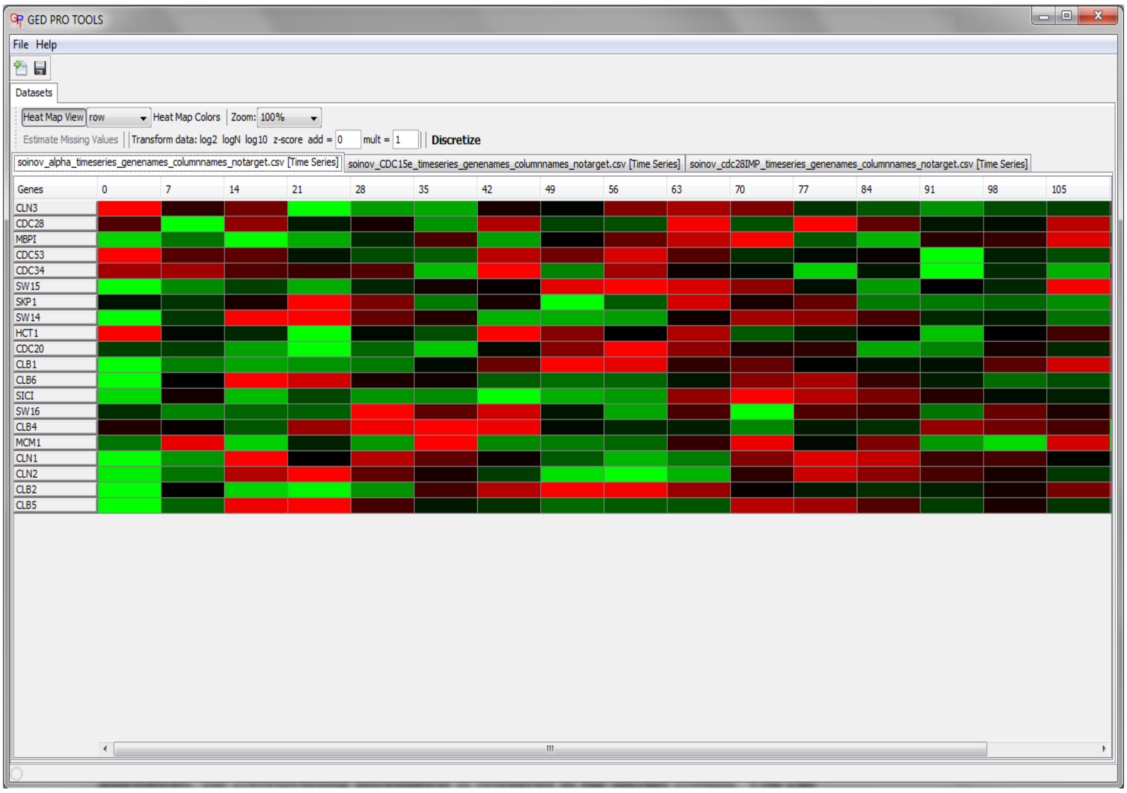
Note: all the data must be consistent with the selected "Gene and column names" options, i.e., if you have selected that the gene names are in the first column, then all the datasets must meet this convention. This allows the automatic resize and reordering of the data rows in the case that they were provided with different number of genes and in different genes order. In the case that the data does not have the gene names in the first column, they must be provided with the same number of genes and in the same order.

Save a GED

To save a GED, select File -> Save As in the main menu or press the 💾 on the options bar. The GED will be saved in CSV format.

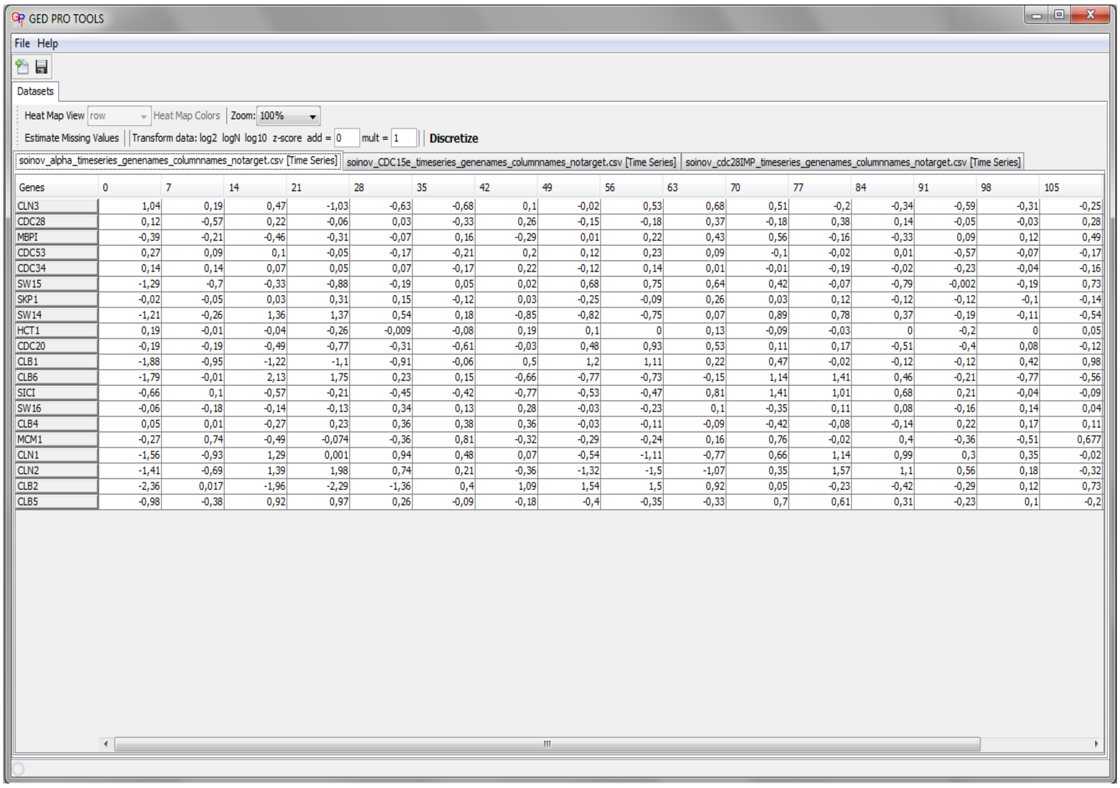
Datasets views and manipulation

The framework allows for different views of the datasets and provides useful tools to manipulate it.


GED PRO TOOLS
File
Help
Datasets
Heat Map View
row
Heat Map Colors
Zoom: 100%
Estimate Missing Values
Transform data: log2 logN log10 z-score add = 0 mult = 1
Discretize
soinov_alpha_timeseries_genenames_columnnames_notarget.csv [Time Series]
soinov_CDC15e_timeseries_genenames_columnnames_notarget.csv [Time Series]
soinov_cdc28IMP_timeseries_genenames_columnnames_notarget.csv [Time Series]
Genes
0
7
14
21
28
35
42
49
56
63
70
77
84
91
98
105
CLN3
CDC28
MBPI
CDC53
CDC34
SW15
SKP1
SW14
HCT1
CDC20
CLB1
CLB6
SICI
SW16
CLB4
MCM1
CLN1
CLN2
CLB2
CLB5

The datasets view shows each data file in a different tab and it also indicates if they correspond to a Time Series or a Steady State sampling. If the loaded data contain gene annotations, the corresponding information is displayed in the second column. You can choose for a Heat Map view or a Numeric view (default) of the data by clicking on the "Heat Map View" button. In the Heat Map view, the framework allows the selection of the colors representing the above average and the below average values of the heat map by pressing the "Heat Map Colors" button. Also, you can select how the Heat Map is calculated, i.e., by row, by column or by the whole matrix.

If the button "Heat Map View" is unselected, the view of the datasets turns into the Numeric view mode. In this mode, you can modify by hand the values of the datasets on each sample by double clicking with the left mouse button into the desired cell. Also, it is possible to estimate the missing values of the selected dataset by pressing the "Estimate Missing Values" button in the secondary option bar. This function will replace the cells marked as missed (represented with a 999 value) with an estimation of the value obtained by a Bayesian Principal Component Analysis Method (BPCA).


GED PRO TOOLS
File Help
Datasets

Heat Map View: row | Heat Map Colors | Zoom: 100%

Estimate Missing Values | Transform data: log2 logN log10 z-score add = 0 mult = 1 | Discretize

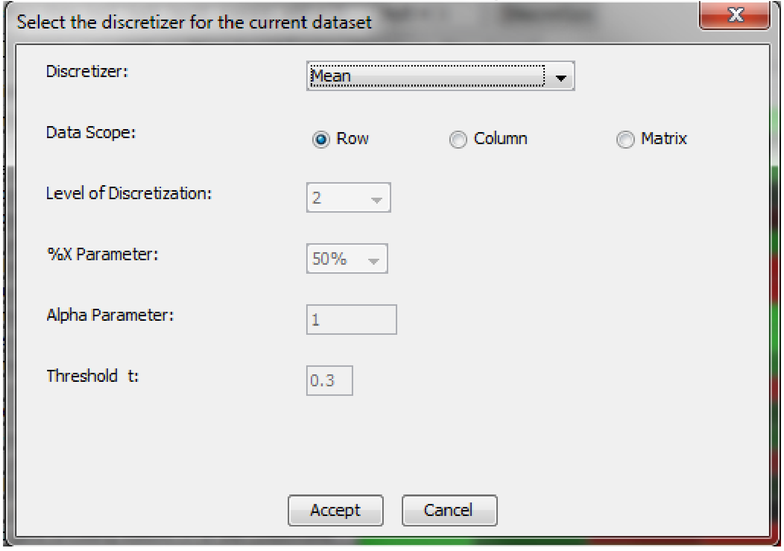
soinov_alpha_timeseries_genenames_columnnames_notarget.csv [Time Series]
soinov_CDC15e_timeseries_genenames_columnnames_notarget.csv [Time Series]
soinov_cdc28IMP_timeseries_genenames_columnnames_notarget.csv [Time Series]

Genes	0	7	14	21	28	35	42	49	56	63	70	77	84	91	98	105
CLN3	1,04	0,19	0,47	-1,03	-0,63	-0,68	0,1	-0,02	0,53	0,68	0,51	-0,2	-0,34	-0,59	-0,31	-0,25
CDC28	0,12	-0,57	0,22	-0,06	0,03	-0,33	0,26	-0,15	-0,18	0,37	-0,18	0,38	0,14	-0,05	-0,03	0,28
MBPI	-0,39	-0,21	-0,46	-0,31	-0,07	0,16	-0,29	0,01	0,22	0,43	0,56	-0,16	-0,33	0,09	0,12	0,49
CDC53	0,27	0,09	0,1	-0,05	-0,17	-0,21	0,2	0,12	0,23	0,09	-0,1	-0,02	0,01	-0,57	-0,07	-0,17
CDC34	0,14	0,14	0,07	0,05	0,07	-0,17	0,22	-0,12	0,14	0,01	-0,01	-0,19	-0,02	-0,23	-0,04	-0,16
SW15	-1,29	-0,7	-0,33	-0,88	-0,19	0,05	0,02	0,68	0,75	0,64	0,42	-0,07	-0,79	-0,002	-0,19	0,73
SKP1	-0,02	-0,05	0,03	0,31	0,15	-0,12	0,03	-0,25	-0,09	0,26	0,03	0,12	-0,12	-0,12	-0,1	-0,14
SW14	-1,21	-0,26	1,36	1,37	0,54	0,18	-0,85	-0,82	-0,75	0,07	0,89	0,78	0,37	-0,19	-0,11	-0,54
HCT1	0,19	-0,01	-0,04	-0,26	-0,009	-0,08	0,19	0,1	0	0,13	-0,09	-0,03	0	-0,2	0	0,05
CDC20	-0,19	-0,19	-0,49	-0,77	-0,31	-0,61	-0,03	0,48	0,93	0,53	0,11	0,17	-0,51	-0,4	0,08	-0,12
CLB1	-1,88	-0,95	-1,22	-1,1	-0,91	-0,06	0,5	1,2	1,11	0,22	0,47	-0,02	-0,12	-0,12	0,42	0,98
CLB6	-1,79	-0,01	2,13	1,75	0,23	0,15	-0,66	-0,77	-0,73	-0,15	1,14	1,41	0,46	-0,21	-0,77	-0,56
SICI	-0,66	0,1	-0,57	-0,21	-0,45	-0,42	-0,77	-0,53	-0,47	0,81	1,41	1,01	0,68	0,21	-0,04	-0,09
SW16	-0,06	-0,18	-0,14	-0,13	0,34	0,13	0,28	-0,03	-0,23	0,1	-0,35	0,11	0,08	-0,16	0,14	0,04
CLB4	0,05	0,01	-0,27	0,23	0,36	0,38	0,36	-0,03	-0,11	-0,09	-0,42	-0,08	-0,14	0,22	0,17	0,11
MCM1	-0,27	0,74	-0,49	-0,074	-0,36	0,81	-0,32	-0,29	-0,24	0,16	0,76	-0,02	0,4	-0,36	-0,51	0,677
CLN1	-1,56	-0,93	1,29	0,001	0,94	0,48	0,07	-0,54	-1,11	-0,77	0,66	1,14	0,99	0,3	0,35	-0,02
CLN2	-1,41	-0,69	1,39	1,98	0,74	0,21	-0,36	-1,32	-1,5	-1,07	0,35	1,57	1,1	0,56	0,18	-0,32
CLB2	-2,36	0,017	-1,96	-2,29	-1,36	0,4	1,09	1,54	1,5	0,92	0,05	-0,23	-0,42	-0,29	0,12	0,73
CLB5	-0,98	-0,38	0,92	0,97	0,26	-0,09	-0,18	-0,4	-0,35	-0,33	0,7	0,61	0,31	-0,23	0,1	-0,2

Both views allow the zooming in and out of the datasets with several scales to improve the visualization. Also, it is possible to transform the current data by means of $\log_2$, $\log_e$, $\log_{10}$, z-score, translation (add) and escalation (mult).

Discretizing the Datasets

In order to run a discretization approach, it is necessary to have at least one dataset loaded in the framework. To select a discretization method press the "Discretize" button, which will open the next dialog box:


Select the discretizer for the current dataset

Discretizer: Mean

Data Scope: ● Row ○ Column ○ Matrix

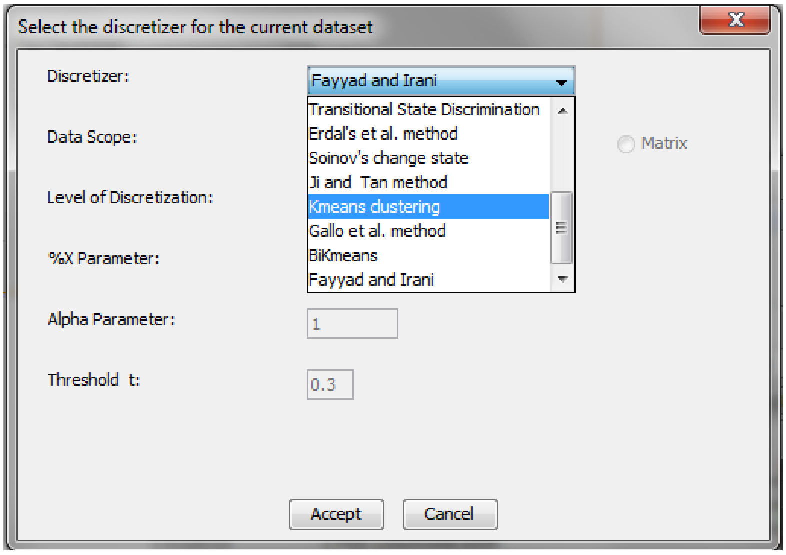
Level of Discretization: 2

%X Parameter: 50%

Alpha Parameter: 1

Threshold t: 0.3

Besides the selection of the discretization method, this dialog box allows to set the parameters corresponding to each method. In the Combo Box **Discretizer**, you can select between sixteen different discretization methods.



Select the discretizer for the current dataset

Discretizer: Fayyad and Irani

Data Scope: ○ Matrix

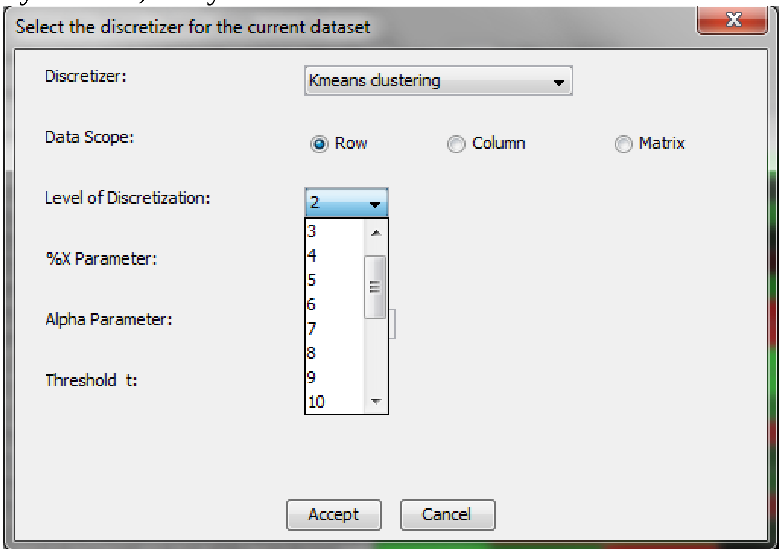
Level of Discretization:

%X Parameter:

Alpha Parameter: 1

Threshold t: 0.3

The **Data Scope** parameter, if allowed by the discretization approach, selects the data scope used in the discretization to compute the discrete states of the genes, i.e., by row, by column, or by matrix.



Select the discretizer for the current dataset

Discretizer: Kmeans clustering

Data Scope: ● Row ○ Column ○ Matrix

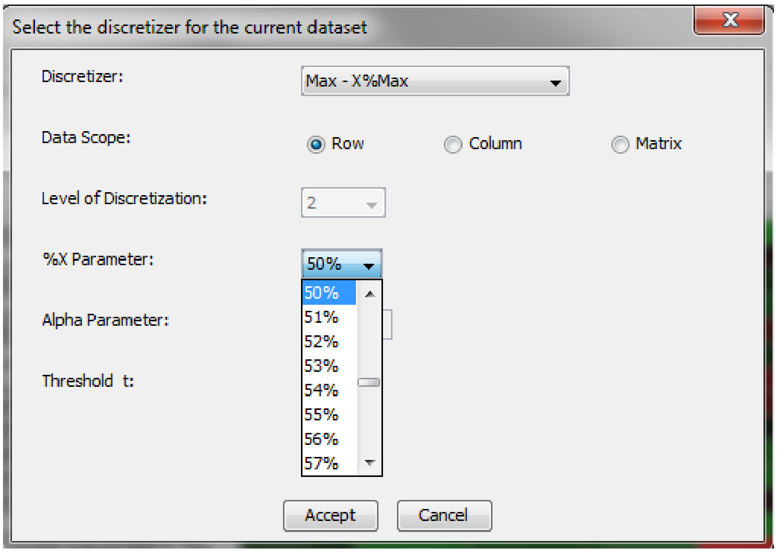
Level of Discretization: 2

%X Parameter:

Alpha Parameter:

Threshold t:

The **Level of Discretization**, if allowed by the discretization approach, selects the number of states used in the discretization. The max allowed value is calculated regarding the selected Data Scope. If the selected Data Scope is “row”, then the max value is the number of columns. If the selected Data Scope is “column”, then the max value is the number of rows. And if the selected Data Scope is “matrix”, then the max allowed value is the max number between the number of rows and the number of columns in the current dataset.



Select the discretizer for the current dataset

Discretizer: Max - X%Max

Data Scope: ● Row ○ Column ○ Matrix

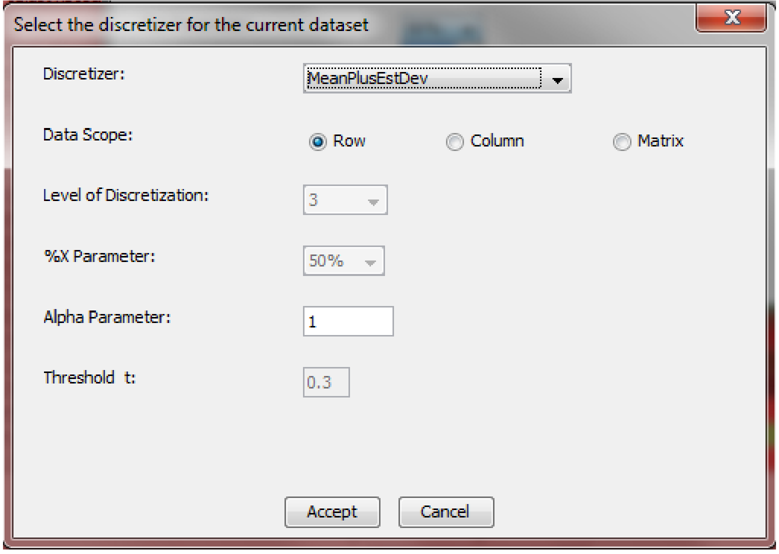
Level of Discretization: 2

%X Parameter: 50%

Alpha Parameter:

Threshold t:

The **%X Parameter** is used in the “Top %X” and the “Max – X% Max” Discretization approaches. This parameter allows to tune the desired percentage value in these discretization techniques.



Select the discretizer for the current dataset

Discretizer: MeanPlusEstDev

Data Scope: ● Row ○ Column ○ Matrix

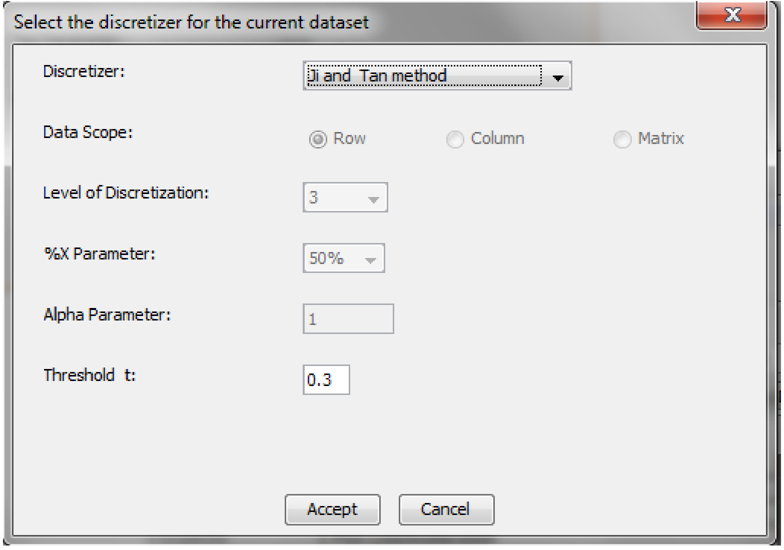
Level of Discretization: 3

%X Parameter: 50%

Alpha Parameter: 1

Threshold t: 0.3

The **Alpha Parameter** is used in the “MeanPlusEstDev” and in the “Erdal *et al.*” approaches.



Select the discretizer for the current dataset

Discretizer: Ji and Tan method

Data Scope: ● Row ○ Column ○ Matrix

Level of Discretization: 3

%X Parameter: 50%

Alpha Parameter: 1

Threshold t: 0.3

Finally, the **Threshold t** parameter is specific for the “Ji and Tan method.”

To perform the selected discretization just press the “Accept” Button. The discretization will always be performed over the selected dataset in its current form.