

GEDI

Gene Expression Dynamics Inspector v2.1

Input Data Format

(refers to 3.1. in User Manual – Overview)

GEDI's input file is one *large tab-delimited ASCII file* that contains all the samples (gene expression experiments) and is only slightly different to most standard microarray output spreadsheets. The preparation analysis in GEDI, all you have to do is some modifications of the header rows, which will be discussed here.

As is conventionally used, the bulk of the file is a matrix of N_G rows (genes) by N_S columns (samples). Each **column** represents one gene microarray measurement of a profile (sample) across the set G , of genes g_i ($i = 1, \dots, N_G$); each gene is represented by **row**, giving rise to the N_G rows. This corresponds to current convention. But since GEDI can analyze multiple parallel time courses, for dynamic analysis, the set S , containing the samples can be partitioned into any number of J distinct classes S_j representing the J time courses: $S = \{ S_1, S_2 \dots S_J \}$. Each set S_j can contain any T_j number of time point samples: $S_j = \{ s_{1j}, s_{2j}, \dots, s_{T_j j} \}$ and is represented by a column, giving rise to the total N_S columns. Note: There is not limit on how many classes J , or how many time points in the class.

GEDI can be used to analyze two different types of gene expression profile data.

- *Dynamic analysis*: Each class has more than one time-series data
- *Static analysis*: Each class has only one data point. It is essentially a special case of dynamic analysis with only 1 time point.

The entry in the first column, first row indicates whether the data are to be analysed as *dynamic* or *static* data set.

1. General format

1.1. Dynamic analysis

In essence, the manual formatting of the header rows is required to tell GEDI which samples belong to the same time course. In *Dynamic* analysis, each sample will have a numeric >sample ID= that represent the time at which the sample was measured (for example, 0, 0.5, 1.2, 3, etc). Sample ID is unique in the same sample class. But of course, two time courses will have samples with the same ID '1h'. Hence, an additional row will contain the label indicating the time course (sample class J) a sample belongs to.

The file format

The input file that contains the data matrix must be annotated to indicate which of the samples belong to which sample class S_j , (i.e., time course). This is achieved by the additional header rows "Sample Header Rows". Thus, the rows in the data files are broken up into 3 sections:

- **Dataset Header Row**: defines the dataset as "Dynamic", additional information on the data set.
- **Sample Header Rows**: There are J such header rows, one for each sample class (time course). The content of the cells indicate for each column whether the respective sample belongs to the sample class indicated in the first column or not:
- **Gene Expression Rows**: contains the gene ID, gene description and the expression values.

For columns with data belonging to the sample class indicated in the first column, enter the *time point label*, otherwise enter *-1*

General format for GEDI dynamic analysis

The general format of the spreadsheet with the expression data, illustrated for a two-time course experiment is as follows:

}Dynamic	DESC	(Time Unit)	(Total Gene Number)	(Data set name)					
} (time course1)	DESC	T1	T2	T3	T4	-1	-1	-1	-1
} (time course2)	DESC	-1	-1	-1	-1	T1	T2	T3	T4
(Gene ID)	(Gene Descr)	D	D	D	D	D	D	D	D
(Gene ID)	(Gene Descr)	D	D	D	D	D	D	D	D

Entries in **bold** must be entered as such. The terms in parenthesis (...) are data-specific labels.

TX stands for time point values, or the sample ID of the sample

D stands for the numeric data for gene expression level

DESC heads the column with the gene description.

1.2. Static analysis

Static analysis is straightforward, since samples are not grouped into classes, hence, there is only one sample header row. The sample labels are directly above the respective columns.

The file format

- *Dataset Header Row*: Defines the dataset as “Static”.
- *Sample Header Row*: Contains label of each sample – for each column
- *Gene Expression Rows*.

General format for GEDI static analysis

}Static	DESC	N/A	(Total Gene Number)	(Data set name)				
}\$	DESC	S1	S2	S3	S4	S5	S6	S7
(Gene ID)	(Gene Descr.)	D	D	D	D	D	D	D
(Gene ID)	(Gene Descr.)	D	D	D	D	D	D	D
(Gene ID)	(Gene Descr.)	D	D	D	D	D	D	D

Entries in **bold** must be entered as such. The terms in parenthesis (...) are data-specific labels.

S stands for the sample name, or the sample ID of the sample

D stands for the numeric data for gene expression level

DESC heads the column with the gene description.

See the example below for more details.

2. Specific example and more details

As an illustration of GEDI=s input, we present a “dynamic analysis” of an experiment with $J=3$ time courses, each of them representing a sample class S_j .

EXAMPLE. Assume one is interested in comparing the time course of gene expression profile changes after treatment of cell cultures with three different cytokines, FGF, VEGF and EGF. These three cellular responses can be visualized in three animated windows displaying the change of the patterns of the

transcriptome represented as GEDI mosaics.

The three parallel time courses would represent the $J = 3$ classes of samples. Each can contain a different number of samples time points (0 to 10 hours). For example, for class EGF, the sample IDs are 0, 0.5, 5, 10:

	Samples for each sample class
S_1 : FGF	{0h, 0.5h, 5h, 10h}
S_2 : VEGF	{0h, 2h, 4h, 6h, 8h}
S_3 : EGF	{0h, 3h, 4h, 8h, 10h}

(1) Dataset Header Rows : This first header row contains information for whole dataset

The first line is the information about the whole data file. It has 4 columns.

Column 1: Start with symbol “}”, “Dynamic” for time series analysis, “Static” for static analysis.

Column 2: “DESC”, for gene description, this column is optional.

Column 3: Time Unit – it can be a unit of the time point (hour or minute or day) for dynamic analysis. For static analysis, it is not applicable, so “N/A” is used.

Column 4: Total number of genes, same as the rows of gene expression data

Column 5: Name of the data set

(2) Sample Header Row: Header rows for sample classes

The next Header Rows contain the column labels describing

- the sample class S (one of the J time courses) to which a column belongs.
- the individual sample s (*time points* in our example) it represents

For dynamic analysis, the data matrix contains J header rows on the top, one for each class of sample S_j . Thus, in our example there will be three header rows for the 3 time series.

Each Header Row contains the sample labels of a class S_i . The *first column* in each header rows contains the name of the sample class S_i (*time courses in the example: VEGF, FGF, EGF*).

To distinguish the header lines from the beginning of the gene expression data lines, the user must insert a *special character* “}” before the time series’ description.

The columns in the header row on the right of the first column describe the samples (*time point*) q_{ij} represented by a particular column. The label of a column must be entered in that header row which describes the sample class (*time course*) to which that column belongs. In the positions header row / column) where a particular column does *not* correspond to a particular sample class, a value of ‘-1’ is entered as a space filler. For this reason, all of the sample labels (time values) must be equal to or greater than 0, i.e., no sample time point label may be defined with a negative number.

Note: It is possible and sometime necessary to assign one common column to several sample classes. For instance, one might want to do this for a *reference sample*, e.g., the 0 hour time point (pre-treatment control) which should be part of all the time courses. To do this, simply enter the same time point value (0) into all of the header rows of the sample classes (time series) which use that sample time point, 0. (See the following example).

There is no maximum limit of number of time courses (J) in GEDI

For *static* analysis, there will be only one single sample header row, with samples names in each column. Again, the first column need to have string “} \$”.

(3) Data Rows

Below the header rows are the data rows. These rows must have a gene description or name in the first column, followed by the time point values for the time course. The *gene IDs* need to be **unique** for each gene. The *gene description/name* column is optional and does not need to be unique. The data point values in each column correspond to expression levels for the respective gene of the samples (time points).

There is no maximum limit of input genes in GEDI.

The table of data should be formatted as a matrix (e.g. in Excel), so that each row represents a gene g_i ($i = 1, \dots, N_G$) and each column represents a sample s_{ij} ($i = 1, \dots, N_G, j = 1, \dots, J$) (time point). The input file may contain blank cells. In this case, NaNs will be inserted for all calculations. Aside from the first column of data (gene ID) and optional second column (gene description), the entire data part of spreadsheet should contain only numeric values.

In our EXAMPLE, the Header Rows of three time series ‘FGF’, ‘VEGF’, and ‘EGF’ is shown. Because the Header Rows identifies each sample column, the actual order of the columns does not matter. The first header line indicates it is a dynamic dataset. Then it gives the name of the dataset (“Drug dynamics”), total gene count (“5”), the time unit (“Hour”) of the time points. 2nd column corresponds to the 0 hour time point of all of the time courses. 4th column represents the 3rd hour time point (at hour 5) of FGF. A ‘-1’ has been placed in the 2nd and 3rd rows of 3rd column so that GEDI recognizes this column of data to be exclusively for ‘FGF’. 6th column represents the 2th hour time point (at hour 2) of VEGF. Again, note that the first and third lines of the Header Rows have a ‘-1’ in 6th column so that GEDI knows that the data in 6th column belongs solely to the ‘VEGF’ time course. This pattern continues throughout the rest of the columns.

}Dynamic	Hour	5	Drug dynamics									
}FGF	0	0.5	5	10	-1	-1	-1	-1	-1	-1	-1	-1
}VEGF	0	-1	-1	-1	2	4	6	8	-1	-1	-1	-1
}EGF	0	-1	-1	-1	-1	-1	-1	-1	3	4	8	10
Gene ABC1	-0.9	-1	-1	-2.4	-5.5	0.1	0	-0.4	-0.9	-1.3	0	-0.5
Gene ABC2	-0.8	-0.4	-0.9	-1	-0.4	-0.4	0.2	-0.6	-0.7	-0.5	0.6	2.2
Gene ABC3	0.6	0.9	0.6	0.3	0.7	1.1	0.9	1	0.4	0.7	0.7	0.7
Gene ABC4	-0.3	-0.1	0.3	0.4	0.6	0	0.1	0.3	0.3	0.8	0.4	-0.1
Gene ABC5	-1.4	-0.5	-0.2	-0.4	-0.6	-0.2	-0.7	-0.1	-0.7	-1	-0.6	-0.1

Note

The file format should be in the form of a Text (Tab delimited) (*.txt)@ file. This format is exportable from Excel or other spreadsheet applications. Occasionally excel will insert stray characters in the columns to the right of your data matrix, or in the rows below your matrix. If this happens, GEDI may fail to load the data file. To fix the problem, it is recommended that you simply highlight your entire data matrix, and copy it over into a new excel spreadsheet.