

eSOMet Manual v0.9 β

Isam Haddad, i.haddad@tu-bs.de

March 20, 2009

Contents

1	Introduction	2
1.1	What is this all about?	2
1.2	Installation	4
1.2.1	Starting eSOMet using Webstart	4
1.2.2	Using the executable Jar	5
2	Starting a project	6
2.1	Create and load a new project	6
2.2	Data management	7
2.3	Making a PCA	10
3	Clustering the data	11
3.1	Hierarchical cluster analysis	11
3.2	Emergent self organizing maps	13
3.3	Dendrograms and silhouettes	14
4	Docking to the KEGG database	15
5	Identification of significantly changed metabolites	17
5.1	Starting from a dendrogram	17
5.2	Observation of metabolite signal ratios	17
5.3	Visualization on the KEGG pathway maps	17

1 Introduction

1.1 What is this all about?

eSOMet is a novel software for the analysis of GC-MS metabolome data. The focus is on the highly accurate detection of relationships between different metabolic patterns. For this purpose, the here described software represents a novel analysis pipeline (Figure 1) which processes the data by the following steps:

1. Acquisition of the multivariate GC-MS data. The data set is organized such that the metabolite concentrations are described as statistical variables, where each experimental trial denotes one statistical observation.
2. Normalization and reduction of the data set by application of a principle component analysis. The PCA rearranges the data in the multidimensional space such that the first dimensions represent the largest variances within the data set, whereas the dimensions of higher order cover the noise only.
3. Clustering of the experimental trials according to their similarity, using either hierarchical cluster analysis (HCA) or emergent self-organizing maps (ESOMs).
4. In both cases, a dendrogram representing the relationships of the metabolic patterns of each experimental trial is calculated.
5. Different branches from the dendrograms representing distinct environmental conditions or genetic backgrounds can manually be selected and the metabolic changes statistically evaluated.
6. Finally, the fold change ratios of the metabolite concentrations and their statistical significances are displayed on the KEGG metabolic pathway maps using color codes, hence, presenting the results in a more biological fashion.

The general usage of the software and its application to recorded data will be lined out in this manual.

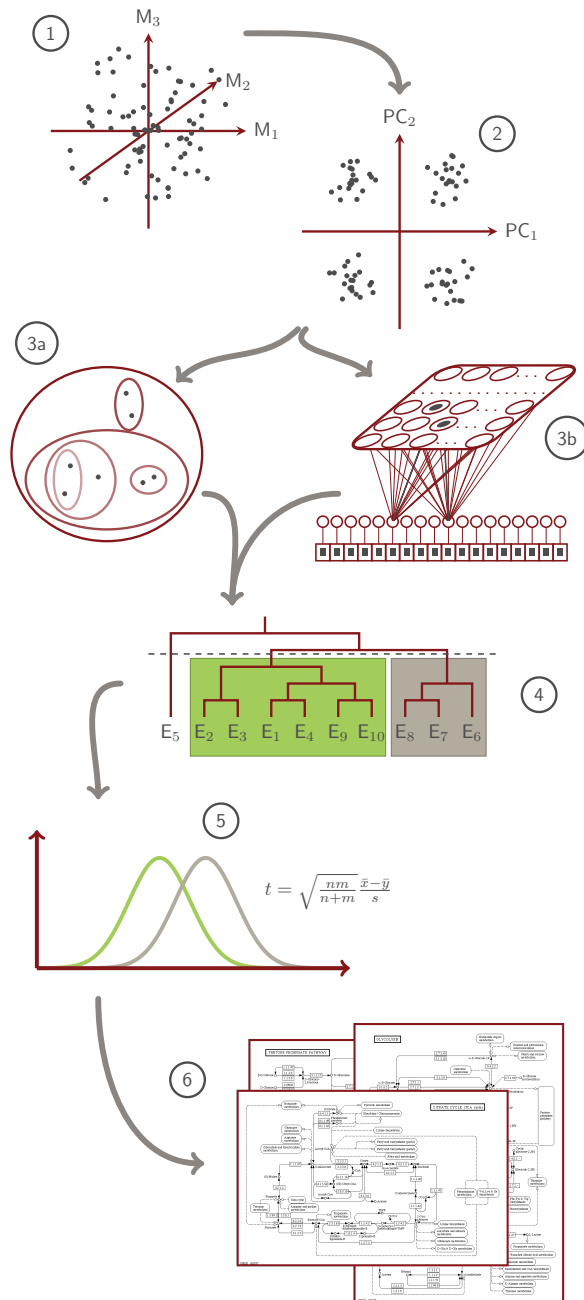


Figure 1: The general workflow of the data processing pipeline, which is compiled in the eSOMet software tool.

1.2 Installation

eSOMet was exclusively implemented using the Java technology. So basically, there is no need to *install* eSOMet. This section just helps you to get eSOMet started when you are unexperienced with the application of Java based software.

The only requirement to run the tool is a proper installation of Java 1.6 or higher. If you are uncertain whether you have Java installed or not, open a Shell (Unix / Linux), a Terminal (OS-X) or a Command Prompt (Windows) and type

```
java -version
```

If the response looks similar to this

```
java version "1.6.0_06"  
Java(TM) SE Runtime Environment (build 1.6.0_06-b02)  
Java HotSpot(TM) Server VM (build 10.0-b22, mixed mode)
```

everything is fine and you can proceed. Otherwise, please visit <http://www.java.com> and install the latest Java Runtime Environment (JRE) on your system.

We provide two options to use eSOMet: A Webstart application and an executable Jar package.

1.2.1 Starting eSOMet using Webstart

The easiest way to start eSOMet is to launch it via the Java Webstart technology. Simply visit the eSOMet website¹ and click the Webstart button. If your browser is properly configured the Webstart launcher should right away start to download all required resources and bring up the software within less than a minute (depending on your internet bandwidth).

Otherwise, if your browser offers you to download a file called `esomet.jnlp`, just save it to your harddisk and start the software by typing into a Shell, Terminal or Command Prompt:

```
javaws </path/to/esomet.jnlp>
```

¹<http://esomet.tu-bs.de>

where `</path/to/esomet.jnlp>` defines the relative or absolute location of the downloaded file. Equivalently, it is also possible to start the software directly using the Shell, etc. by typing

```
javaws http://esomet.tu-bs.de/ressources/esomet.jnlp
```

1.2.2 Using the executable Jar

If you do not want to use the first option – which actually should be favored, since Webstarts manages version updates automatically for you – it is also possible to download eSOMet as executable Java package (Jar file). For this purpose, simply download the package from the website and double click the downloaded file `esomet.jar`. The program should start immediately. If your system is not configured to support this action, you can still type in a Shell, Terminal, whatever and type

```
java -jar </path/to/esomet.jar>
```

where `</path/to/esomet.jar>` defines the relative or absolute location of the downloaded file.

Finally, if you see a program window on your screen that looks similar to the one depicted in figure 2 you got eSOMet started and may proceed with this manual.

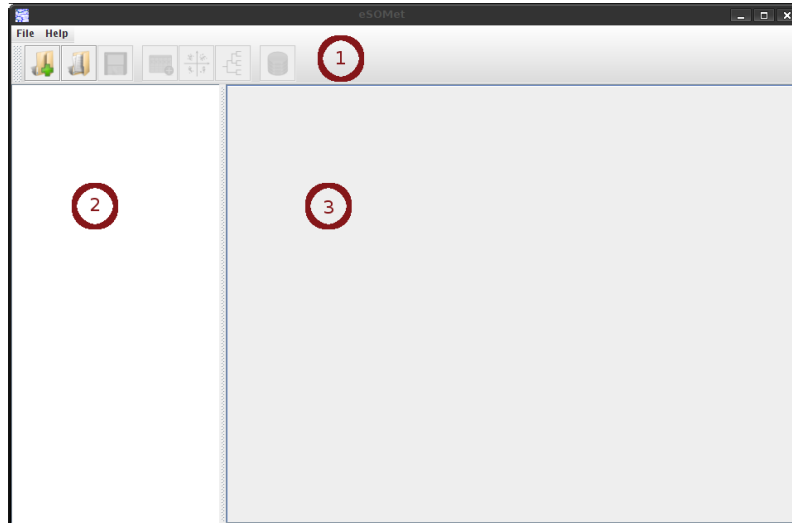





Figure 2: Representation of the plain window of eSOMet after the startup of the program. The top part (1) of the window is called the *toolbar*, the left panel (2) is called the *project panel*, the right panel (3) is called the *content panel*.

2 Starting a project

2.1 Create and load a new project

The first step after launching eSOMet is either to open a new project or, otherwise, load an existing one.

For opening a project, click the -button. A file chooser dialog will pop up and ask you for a location, where the project file will be saved. The eSOMet project files have the data extension `*.eso`. The project contains the intermediate results that are acquired during procession via the analysis pipeline. Each analysis will be displayed in the right hand panel of the main program window, the so called *project panel* (figure 2). After each analysis step, you should click on the -button to store the results permanently to your harddisk.


If you want to load an existing project, click the -button. A file chooser dialog will pop up and ask you for the location of a previously saved `*.eso`-file. On the website,

there is a test case analysis stored in the *.eso format. You can download², unzip and load it into eSOMet.

2.2 Data management

After you have created a new project, the first thing you should do is importing the multivariate metabolome data. Each project can contain one dataset only. The data is organized in a multivariate data matrix. Each column describes one particular experiment or measurement – the so called metabolic pattern (in statistical terms, each column denotes one observation). Each row describes one particular metabolite that has been recorded (in statistical terms each row denotes one variable). The matrix entries contain the concentrations or signal strengths of each metabolite measured in a certain observation.

Hence, the imported data have to be in a specific tabular format: The first row must contain distinct names for each experiment, the first column distinct names for each metabolite. The table cells contain the measured values with floating point precision. Optionally, the second column may contain the KEGG identifiers. If one metabolite is assigned by more than one KEGG identifier, they can be given in the same table cell, separated by a slash (/) each.

For importing the data, click the -button. A wizard (figure 3) will pop up and guide you through the import process. You can either

- import the data from a *.csv-file
- or paste the data, after they have been copied to the clipboard from tools like Microsoft Excel or OpenOffice Calc. In this case, do not select and copy the entire spreadsheet rather select only the part of the spreadsheet that contains the data.

An exemplary *.csv-file is comprised by the test case data, available on the website. After the successful data import, the tree in the project panel contains a new entry called *Data*, which consists of three items:

²<http://esomet.tu-bs.de/resources/TestCase.zip>

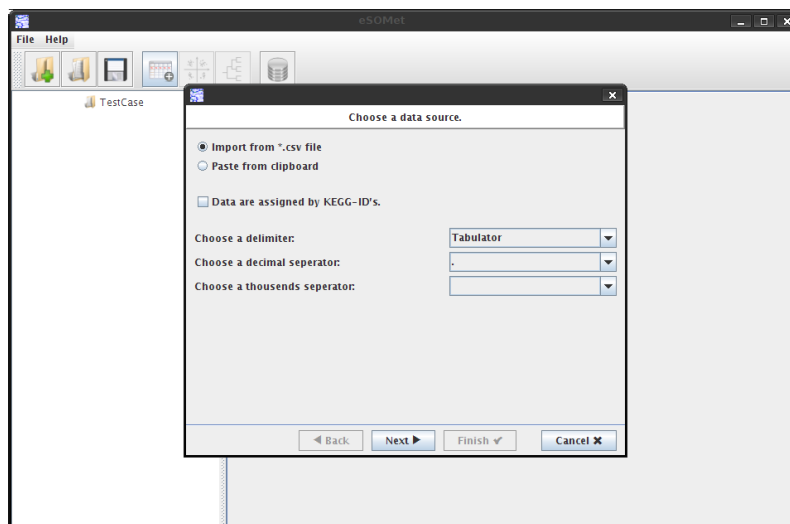


Figure 3: The first dialog of the data import wizard.

Original data By double clicking on this item, a table is opened in the content panel that displays the imported data.

Z-score scaled data A table containing the data after a z-score scaling is opened after clicking this item. It also contains two columns with the values of the standard means and standard deviations of the original data (figure 4).

Range scaled data A table containing the data after a range scaling is opened after clicking this item. It also contains one column with the values of the standard means and two columns with the minimum and maximum value of the original data.

Pareto scaled data A table containing the data after a pareto scaling is opened after clicking this item. It also contains two columns with the values of the standard means and the square roots of the standard deviation of the original data.

Vast scaled data A table containing the data after a vast scaling is opened after clicking this item. It also contains two columns with the values of the standard means and the standard deviations of the original data.

Level scaled data A table containing the data after a level scaling is opened after clicking this item. It also contains one column with the values of the standard means of the original data.

Total signal scaled data A table containing the data after a total signal scaling is opened after clicking this item. It contains also one row of the λ -values for each column, which are derived by the normalization.


The description of the mathematical basis of all methods here described, as well as the explanation of differences, advantages and disadvantages of the various methods employed in this utilization are beyond the scope of this manual. Instead we refer to the publication which is currently under submission and the references listed therein. You will keep informed through the eSOMet website about the publishing process.

However, these tables comprised by the data object are all read-only. Currently it is not intended to manipulated or directly enter raw data into eSOMet.

Metabolite	KeggID	Standard	Variance	Glu-0h-a	Glu-0h-b	Glu-0h-c	Glu-1h-a	Glu-1h-b	Glu-1h-c	Glu-2h-a	Glu-2h-b	Glu-2h-c
Z-Aminoac.	C00334	228E002	376E003	-551E-003	-211E-003	-211E-003	-1495E-003	-177E-003	-514E-003	-538E-003	-602E-003	-538E-003
Z-Diacylg.	C00673	126E001	223E001	-567E-003	-459E-003	-503E-003	-471E-003	-503E-003	-493E-003	-531E-003	-515E-003	-492E-003
Z-Hydroxy.	C01087	953E002	984E002	-54E-002	-232E-002	-203E-002	-186E-002	-220E-002	-116E-002	-253E-002	-179E-002	-144E-002
Z-Phosph.	C02502	155E002	195E002	-821E-003	-415E-003	-238E-003	-495E-003	-185E-003	-207E-003	-272E-003	-532E-003	-510E-003
Z-Oxoglut.	C00026	137E003	224E003	-231E-002	-217E-002	-293E-002	-844E-004	-840E-004	-954E-004	-572E-003	-305E-003	-551E-003
Z-Phosph.	C00931	118E003	172E003	-442E-003	-442E-003	-442E-003	-442E-003	-442E-003	-442E-003	-442E-003	-442E-003	-442E-003
Z-Phosph.	C00197	705E003	558E003	-118E-002	-109E-002	-104E-002	-112E-002	-107E-002	-101E-002	-111E-002	-118E-002	-111E-002
Z-Amino.	C00430	337E000	476E000	-707E-003	-854E-004	-191E-003	-263E-003	-405E-003	-292E-003	-638E-003	-465E-003	-286E-003
Z-Oxopro.	C01873	195E003	148E003	-83E-002	-941E-003	-896E-003	-867E-003	-839E-003	-216E-002	-298E-003	-291E-003	-106E-002
Z-Phosph.	C00345	370E001	618E001	-571E-003	-580E-003	-567E-003	-557E-003	-529E-003	-494E-003	-526E-003	-495E-003	-517E-003
Alanine	C00147	101E003	120E003	-340E-003	-416E-003	-446E-003	-541E-003	-534E-003	-120E-002	-178E-003	-174E-003	-174E-003
Alanine	C01401	117E004	133E004	-820E-003	-834E-003	-821E-003	-864E-003	-855E-003	-866E-003	-877E-003	-870E-003	-881E-003
alpha.am.	C01083	113E004	638E003	-185E-002	-143E-002	-134E-002	-245E-002	-351E-002	-178E-002	-871E-003	-128E-002	-100E-002
AMP	C00020	149E003	232E003	-603E-003	-603E-003	-603E-003	-603E-003	-603E-003	-603E-003	-603E-003	-603E-003	-603E-003
Arginine	C02385	704E000	139E001	-547E-003	-547E-003	-547E-003	-547E-003	-547E-003	-547E-003	-547E-003	-547E-003	-547E-003
Beta-Alan.	C00099	539E002	509E002	-893E-003	-743E-003	-700E-003	-102E-002	-103E-002	-103E-002	-954E-003	-955E-003	-922E-003
C15acid	-	408E002	623E002	-355E-002	-296E-002	-366E-002	-933E-003	-298E-004	-354E-002	-138E-002	-463E-002	-527E-002
C16acid	-	430E003	139E003	-807E-003	-136E-002	-166E-002	-112E-002	-408E-003	-310E-003	-359E-003	-108E-002	-750E-003
C18acid	-	384E001	304E001	-218E-002	-218E-002	-226E-002	-126E-002	-126E-002	-126E-002	-126E-002	-126E-002	-126E-002
C19acid	-	552E001	310E001	-478E-003	-296E-004	-530E-003	-124E-003	-378E-003	-354E-003	-134E-003	-704E-003	-550E-003
crotonact.	-	402E000	385E000	-687E-003	-687E-003	-687E-003	-687E-003	-687E-003	-687E-003	-687E-003	-687E-003	-687E-003
Citrate	C00158	114E003	781E002	-142E-002	-146E-002	-149E-002	-135E-002	-135E-002	-130E-002	-131E-002	-98E-002	-117E-002
Cyclohexa.	C00854	835E001	804E001	-455E-003	-288E-004	-964E-004	-114E-003	-503E-003	-435E-003	-193E-003	-103E-003	-86E-004
Cyclitol.	-	472E000	308E001	-111E-001	-101E-002	-141E-002	-447E-004	-641E-004	-731E-004	-148E-004	-370E-004	-418E-004
Cystathion.	C00542	988E001	969E001	-613E-003	-613E-003	-613E-003	-613E-003	-613E-003	-613E-003	-613E-003	-613E-003	-613E-003
Cystine	C00380	115E002	130E002	-135E-003	-141E-003	-137E-003	-186E-003	-181E-003	-220E-003	-410E-003	-359E-003	-453E-003
D-Fructos.	C00354	866E002	149E002	-575E-003	-968E-003	-572E-003	-374E-003	-356E-003	-547E-003	-573E-003	-573E-003	-574E-003
D-Fructos.	C00085	116E003	131E003	-648E-003	-539E-003	-499E-003	-471E-003	-543E-003	-357E-003	-599E-003	-516E-003	-525E-003
D-Galacta.	C02262	553E001	120E002	-239E-003	-295E-003	-295E-003	-295E-003	-295E-003	-295E-003	-295E-003	-295E-003	-295E-003
D-Glucos.	C00198	591E002	366E003	-358E-003	-164E-003	-164E-003	-164E-003	-164E-003	-164E-003	-156E-003	-157E-003	-157E-003
D-Glucos.	C00352	204E001	210E001	-897E-003	-822E-003	-827E-003	-371E-003	-197E-003	-267E-003	-354E-003	-154E-003	-453E-003
D-Glycerate	C00528	418E001	218E001	-373E-003	-346E-003	-351E-003	-873E-003	-646E-003	-555E-003	-723E-003	-170E-003	-346E-003
D-Lyxose.	C00475	138E001	443E001	-101E-003	-175E-003	-135E-003	-437E-003	-651E-004	-354E-003	-826E-004	-115E-003	-263E-003
D-Mannitol	C00392	101E002	138E002	-522E-003	-426E-003	-426E-003	-548E-003	-343E-003	-510E-003	-477E-003	-577E-003	-612E-003
D-Ribulose	C00129	121E002	138E002	-778E-003	-873E-003	-608E-003	-493E-003	-559E-003	-401E-003	-450E-003	-536E-003	-611E-003
D-Xylof.	C00310	896E001	834E001	-681E-003	-295E-003	-237E-003	-109E-003	-337E-003	-361E-003	-678E-003	-732E-003	-673E-003
D-Xylof.	C00231	228E002	246E002	-557E-003	-561E-003	-589E-004	-278E-003	-367E-003	-262E-003	-524E-003	-402E-003	-488E-003
Dicyclohexyl	-	369E001	164E001	-124E-002	-705E-003	-904E-003	-224E-002	-760E-003	-861E-003	-531E-003	-234E-002	-490E-003
Erythritol	C00503	312E002	219E002	-122E-002	-124E-002	-117E-002	-976E-003	-102E-002	-345E-003	-714E-003	-808E-003	-75E-002
Fructose	C01498	687E003	183E004	-337E-003	-134E-003	-149E-003	-145E-003	-148E-003	-154E-003	-158E-003	-152E-003	-157E-003

Figure 4: Test case data were imported into eSOMet. The content panel shows the tabular arrangement of the normalized multivariate data. KEGG-identifiers are also imported (column 2) and standard means and standard deviations were calculated (columns 3 and 4).

2.3 Making a PCA

As next step of the analysis pipeline you can now perform a principle component analysis of the data. For this purpose click the -button. A new entry in the project panel will appear containing the results of the PCA. This entry contains three items:

PCA_2D-PCA A plot representing the projection of the z-score scaled variables on the first two principle components 5. However, the rank of the components can be switched, giving a projection of the data on two arbitrarily chosen components.

PCAVariances A plot representing the recovered variances per principle component.

PCACumulativeVariances A plot representing the percentage of recovered variance by the first n principle components.

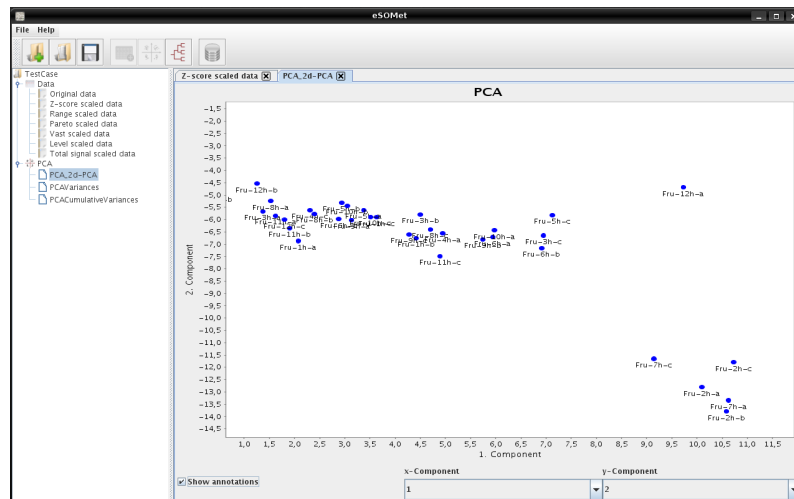



Figure 5: The PCA 2d-projection of the test-case data are loaded in the content panel. Only a zoomed selection of the actual plot is displayed showing one relevant area.



It is possible to zoom into the plots by left-clicking with the mouse in the upper left corner of the area to be selected, dragging to the bottom corner and releasing the mouse button. In order to zoom out, left click and move the mouse to the left and release the button.

3 Clustering the data

The central part of the analysis pipeline is the hierarchical clustering of the metabolic patterns according to their similarity. For this purpose eSOMet provides two quite different approaches: The hierarchical cluster analysis (HCA) and emergent self-organizing maps (ESOMs).

3.1 Hierarchical cluster analysis

No matter if you either want to perform a HCA or an ESOM analysis, you start the clustering analysis by clicking the -button. A wizard will pop up and guide you through the parameter adjustments. The first two of the following steps are equal between the two analysis methods:

1. First you have to give a unique label to this analysis. This is important, because you can perform more than one cluster analysis with varying parameter settings. Furthermore, you have to choose a normalization or reduction you want to perform on the data. Data reduction is carried out through a principle component analysis where you can either choose the degree of variance covered or, alternatively, a number of principle components on which you want to project the data (figure 6). If no PCA was performed yet, the results will be added to the project panel and indicated by the -symbol.
2. As next step you have to choose a clustering methods. Since the next section deals with ESOMs, we assume that you choose the HCA method this time.
3. Finally, the linkage method and the metric are chosen, according to which the HCA is performed (figure 7).
4. After pressing the Next-button the calculation starts and usually within an instant – depending on your data and hardware, of course – the results are displayed in the project panel, indicated by the -symbol.

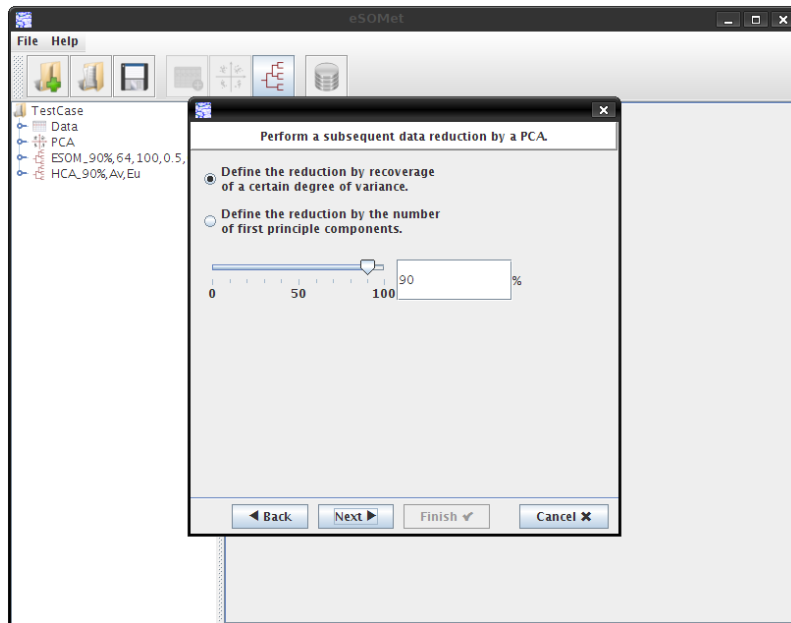


Figure 6: The data reduction dialog as one step of the cluster wizard.

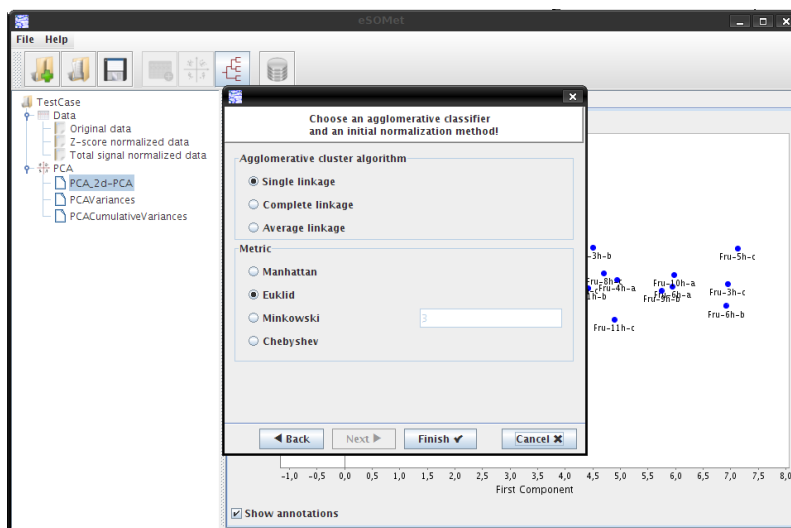


Figure 7: The dialog asking for the linkage and metric of the HCA.

3.2 Emergent self organizing maps

If you decide to carry out a ESOM clustering instead a HCA the last described dialogue is replaced by three ESOM specific dialogs:

1. First you have the option to choose one of three predefined settings, for a very fast but sketchy analysis, a very slow but very accurate analysis and a compromise between both. However, we advise to adjust the settings manually, in order to know which parameters you apply.
2. The first parameters to be set concern the size of the neural network to be trained and the number of training cycles. Keep in mind, that both parameter influence the runtime of the algorithm proportionally, each.
3. Finally, you have to specify the training parameters, as the training rate, the initial training radius and the parameter for the random initialization (figure 8).

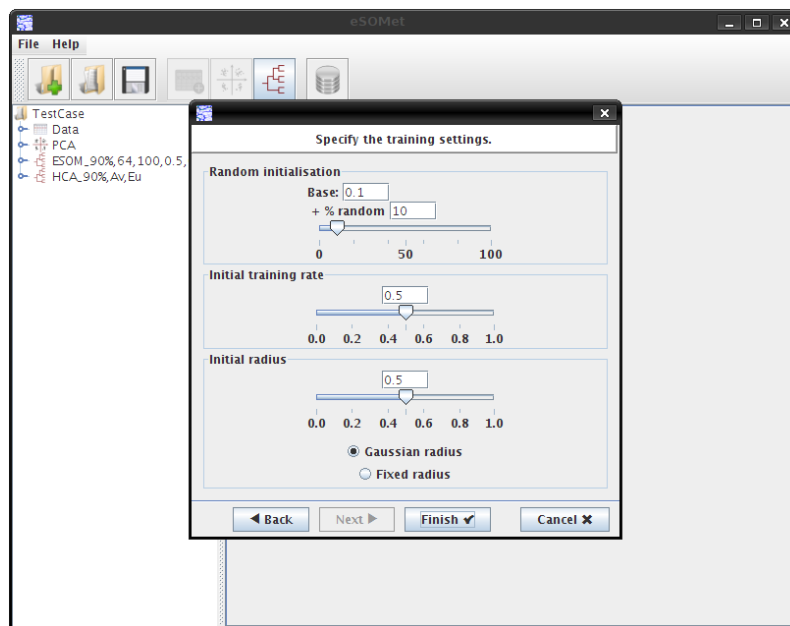




Figure 8: ESOM clustering dialog for the specification of training parameters.

4. After all information is collected a progress window will pop up and inform you

about the state of the calculation. Depending on the chosen parameter and your data, this calculation can be accomplished between seconds and hours. Once the calculation terminated, it is indicated by the appearance of a new -symbol in the project panel.

3.3 Dendrograms and silhouettes

No matter if you performed a HCA or ESOM clustering, as result of both procedures you'll find in the -sections a symbol labeled as dendrogram, which opens the clustered tree of your data in the content panel. Here you can investigate the relationship between the different observations. When you click on a branching-point you'll mark the complete underlying cluster as marked. When you have marked more than one cluster and click the *Display Silhouette widths*-button, the silhouette widths for the selected observations are calculated, based on the selection and displayed next to the observation name (figure 9). You can also deselect a selected cluster by clicking on the branching point again.

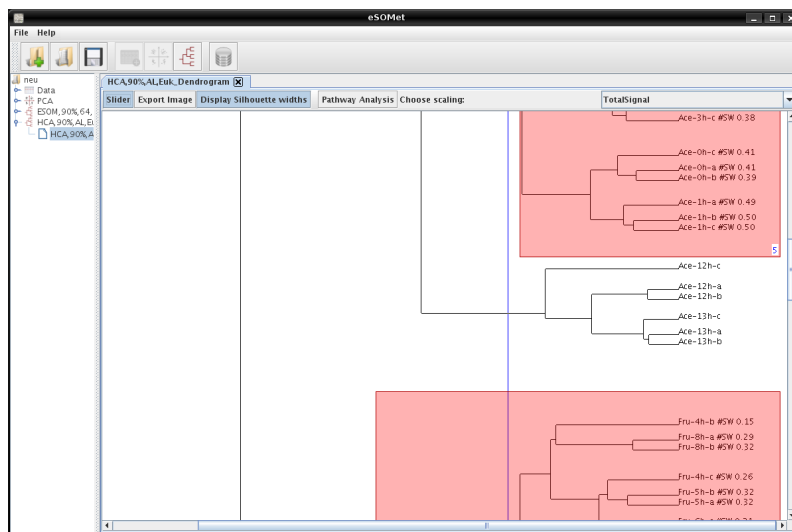




Figure 9: A calculated dendrogram is loaded into the content panel. Additionally, two clusters were manually selected and the silhouette widths are displayed.

4 Docking to the KEGG database

So far, this was the first part of the intended analysis pipeline. The results of the clustering give already valuable insights into the relationship and structure of the recorded data. However, the second part of the analysis pipeline even refines the concept of structure driven metabolome analysis. For this purpose it becomes necessary to give *things* names. Hence, eSOMet implements the ability to deal with the enormous amount of information, stored in the Kyoto Encyclopedia of Genes and Genomes (KEGG), and allows scientists to interpret the results of previous analyzes in a biological well known fashion.

eSOMet deals with the KEGG data by downloading all the information via the internet and builds client-sided a fully functional SQL-database containing all important KEGG features. To start integrating the database into an existing project, click the -button. The appearing wizard offers you three methods of integration:

- You can download all relevant data using the built-in FTP-client. The data are per default copied to an temporary directory (specified as *Local download folder*, picture 10), meaning that most probably they are gone after the next reboot. If you want to keep them, change their location. Note that the entire procedure can take up to an hour. During this procedure the wizard dialog will remain open. After the wizard finished it's work you'll see the -symbol in the project panel. That means, that a database was registered. This database can used only by one instance of eSOMet. Trying to load an already loaded database into another prjgramm instance may cause unforeseeable results.
- If you have already downloaded all the files but no database created and registered yet, the wizard offers you also to work on your already downloaded files. For this purpose, you only have to specify the folder containing the KEGG files and let the wizard do its work. This may take up to 10 minutes and again, the dialog remains open until the database is created and registered. A collection of already downloaded files is provided with the test case data. When you want to use your own data, the folder -naming and -hierarchy must be the same as with the exemplary data.
- If you have already an existing database – either from another, not running project, or the pre-created database from the test case data – the wizard finally

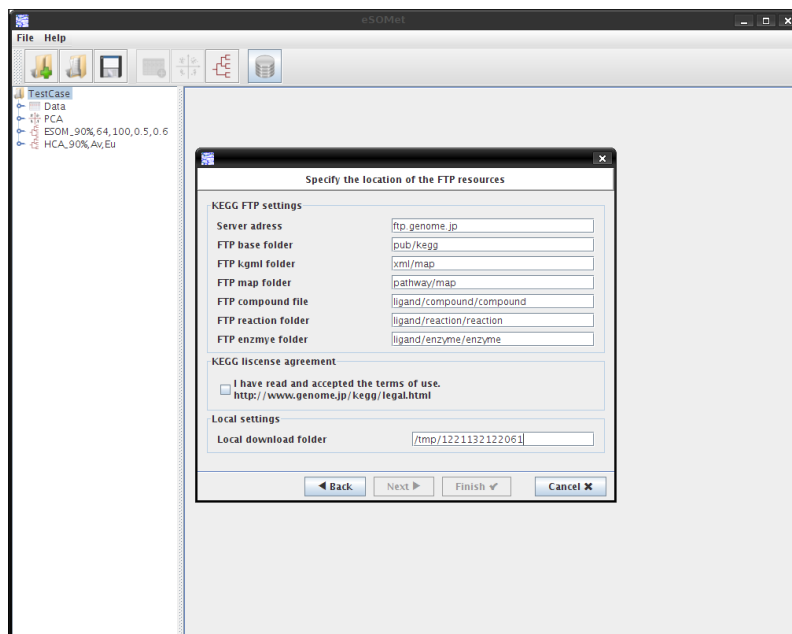


Figure 10: The FTP settings dialog of the database creator wizard.

offers you to simply specify the location of the database and registers it. This can still take up to 5 minutes, so don't worry when the wizard dialog does not vanish directly.

There are two more things to note:

- Once a database has been created, it is read-only. No data are ever written to it, meaning that you can expect from each repeat of an analysis the same result.
- The KEGG database has its own license model, which you'll find under <http://www.genome.jp/kegg/legal.html>. We do not grant any warranty for the used data, nor are we responsible if you use the data without fulfilling the requirements.

5 Identification of significantly changed metabolites

5.1 Starting from a dendrogram

Two subsections earlier you were taught, how to open a dendrogram and make cluster selections. Basically, that's all you need to now to perform the last analysis step, as well. You open a dendrogram of your choice, select exactly two clusters (even though more are possible, this feature is highly experimental), choose a normalization and then just press the *Pathway analysis*-button. This feature works only if the database has been registered to the current project before you load the dendrogram into the project panel. As soon as the calculation is finished, two new tabs will appear in the content panel: The analysis results for the metabolites and for the affected pathways.

5.2 Observation of metabolite signal ratios

After pressing the *Pathway analysis*-button, eSOMet will calculate for each metabolite per cluster the signal distribution and determine the ratio (cluster 1 vs cluster 2) of the mean signals together with the level of significance. These information are displayed in a color-coded fashion as shown in figure 11 in the metabolite tab.

5.3 Visualization on the KEGG pathway maps

Depending on how many metabolites exhibit significantly changed signal ratios, those pathways associating these metabolites are different affected. This is indicated by the significance level, calculated through an Fishers-Exact-Test. These pathway specific information are also color-coded shown in the pathway tab. Furthermore, it is possible to open the pathway map and see in a more biochemical context which metabolite concentrations exhibit a change between the two, previously from the dendrogram selected conditions.

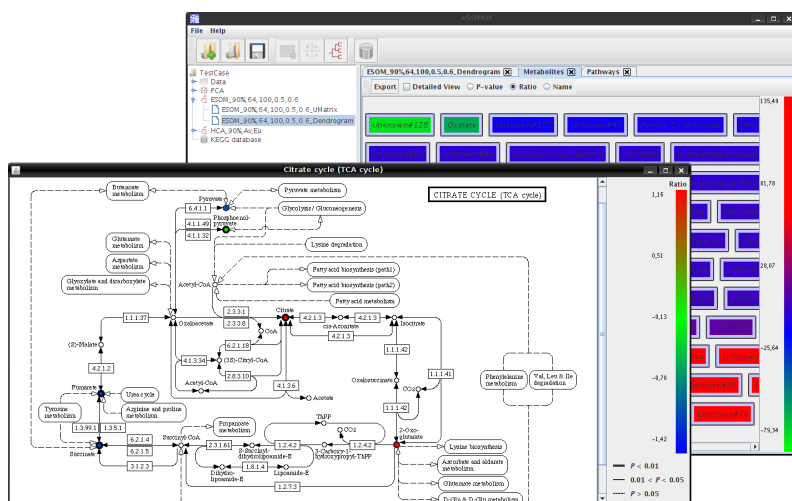


Figure 11: Exemplary output of the comparison between two clusters. In the background the signal ratios of the metabolites are shown in a color coded fashion. In the foreground one affected pathway map is shown.

6 Third party libraries

eSOMet was developed using a couple of third-party libraries:

Classifier4J <http://classifier4j.sourceforge.net/>

The Colt Project <http://acs.lbl.gov/hoscchek/colt/>

Apache Commons <http://commons.apache.org/>

Flanagan's Java Scientific Library <http://www.ee.ucl.ac.uk/mflanaga/java/>

HSQldb <http://hsqldb.org/>

iText <http://www.lowagie.com/iText/>

JDOM <http://jdom.org>

JFreeChart <http://www.jfree.org/jfreechart/>