ClutrFree
Introduction and User Guide
Version 1.1

Ghislain Bidaut
G_Bidaut@fccc.edu
Fox Chase Cancer Center - Bioinformatics

1 Summary

ClutrFree has for primary function to display clusters (the output of any clustering algorithm) computed from experimental data in a convenient manner that will aid in their interpretation. We assume here that experiments are presented under the form of a large matrix where each lines represents a variable over a series of conditions (the columns). The data is usually clustered, and the obtained clusters are used for interpretation. Beside the choice of the clustering algorithm, program, parameters need to be adjusted (for instance, filtering parameters or numbers of nodes for Self-Organizing Maps). ClutrFree aids in determining those parameters.

ClutrFree features an integrated algorithm to group the clusters into a tree, each tree level being an experiment. In parallel, a secondary tree is constructed with the gene membership. As a primary goal, we were interested in analyzing the behavior of a clustering algorithm applied on experimental microarray data as we increased the number of basis vectors. To help drawing conclusions, genes can be displayed along with their IDs, descriptions, and ontology.

The cluster tree and the membership tree are built with the following procedure:

1. Each analysis is represented as a tree level, beginning with the experiment having the least number of clusters. For comparison of experiments having the same number of clusters, the order is not important.

2. The $n+1$ level is compared to the n level by calculating the *Pearson* correlation of all nodes at n with all nodes at $n+1$. The connections are made from highest to lowest. Any orphan nodes from level $n+1$ are connected to the level n by the same method.

3. Step 2 is repeated for each level until the full tree has been constructed.

This lead us to a cluster organization similar to he one in Figure 1.

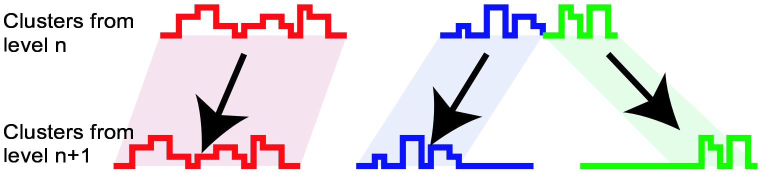


Figure 1: The tree construction algorithm showing how basis vectors are related for a decomposition in N basis vector to the decomposition in $N+1$ basis vectors.

2 Getting ClutrFree and the most recent version of this documentation

If you have access to the Internet, the latest version of ClutrFree and its documentation as well as test datasets can be downloaded from `http://bio-informatics.fccc.edu`. ClutrFree (as well as this documentation) has been licensed under the General Public License (GPL, see the license section in the Annex) and is available under the for of a JAR executable and source code. Tree test datasets are available on the same web site to provide examples of the data format used by ClutrFree. The two first datasets have been created from Bayesian Decomposition (Ochs *et al.* (1999)) analysis of the Rosetta Compendium data (Hughes *et al.* (2000)). One is under the form of plain Bayesian Decomposition output files (see on the same web site `data_demo_bd.zip`), and the other under the form of plain matrix files (see `data_demo_generic.zip`). The third dataset is a Bayesian Decomposition Analysis of the Cell cycle data (Cho *et al.*, 1998).

3 Data Preparation

3.1 Fundamental Matrices used in the analysis

ClutrFree has been primarily designed to view and analyze data from matrix factorization algorithms such as Bayesian Decomposition(BD) (Ochs *et al.* (1999)) or Singular Value Decomposition (SVD) (Holter *et al.* (2000)). However, visualizing clusters from other type of widely employed program (Hierarchical clustering or Self-Organizing Maps) is possible. The two first cited methods performs a matrix factorization of the data matrix **D** leading us to a pattern matrix **P** and an amplitude matrix **A**, verifying

$$\mathbf{D} = \mathbf{A}.\mathbf{P}. \tag{1}$$

A, **P** and **D** are structured as described in the Figure 2. Essentially, the **P** matrix groups a set of K vectors from R^N describing the process that underlie the data **D** and the **A** matrix quantify the amplitude necessary to reconstruct the data from the basis vectors.

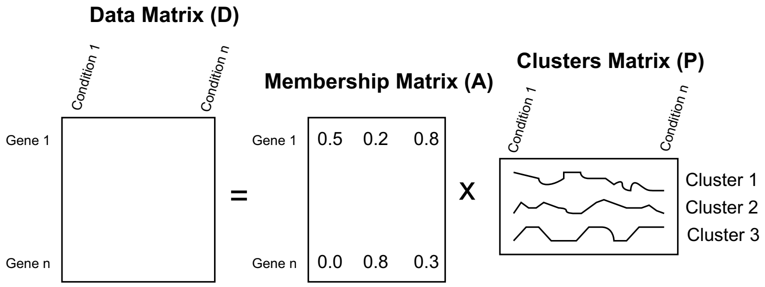


Figure 2: Matrices structures employed by ClutrFree in the case of Bayesian Decomposition and Principle Component Analysis

In the general case of a classic clustering algorithm such as Hierarchical clustering or SOM, the A matrix can be viewed as a binary matrix the membership of each studied gene as seen in Figure 3.

3.2 Uncertainty

ClutrFree take the uncertainty of the original data under the form of a standard deviation into account. When available, this uncertainty is used in the display and to binarize data: If the data point (either gene membership,

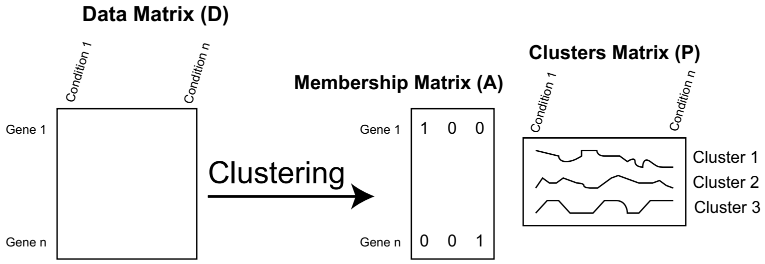


Figure 3: Matrices structures employed by ClutrFree in the case of analysis from Hierarchical clustering and analog algorithms

or cluster point) is greater than $n\sigma$, the value 1 is assigned to this point (member), otherwise, 0 (nonmember).

4 Data Input Format

To use ClutrFree, the series of experiment from the clustering runs must be properly organized on the file system using the directory tree organization from Figure 4 where each experiment is stored in a sub-directory containing tab-delimited files. Those files can have the following formats:

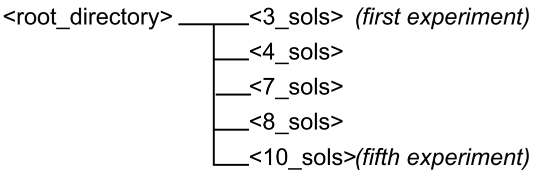


Figure 4: Directory Structure that must be respected for ClutrFree: Each Experiment is stored under a `*_sol` directory under the form of a tab-delimited text files.

- Plain tab delimited files (as in `data_demo_generic.zip` downloadable from `http:`
  `bioinformatics.fccc.edu` as an example) : The **P** and **A** matrices are respectively stored as a plain tab-delimited file containing only the numerical values using the extension `.pat` and the extension `.amp`. Their associated absolute uncertainties (if they have been computed)

must be stored in the `.sgp` and `.sga` files. If The uncertainties are not available, ClutrFree will use the value 0 instead, and the binarization will not be available. This format is ideal to import data from any program.

- Bayesian Decomposition Inference file (see `data_demo_bd.zip` as an example) : All the matrices are stored in the inference file generated by Bayesian Decomposition, including their uncertainty.

The format used for the experiment/descriptions is described in the GUI section.

5 System Requirements and Installation

ClutrFree has been written in java and runs under the Java Runtime Environment(JRE) version 1.4 from Sun Microsystems. Therefore, all the platforms such as Linux, Solaris, etc... currently supported by the Java development team can be used. The JRE can be downloaded free of charge from `http://java.sun.com`. The JAI (Java Advanced Imaging) version 1.1.2 must be also downloaded and installed to enable TIFF graphics exportation (see the JAI website at `http://java.sun.com/products/java-media/jai`). You might also need to get a copy of the Bayesian Decomposition binary from `http://bioinformatics.fccc.edu`, but ClutrFree can use data from any clustering program if properly formatted. ClutrFree (downloadable also from `http://bioinformatics.fccc.edu`) itself comes as a JAR file for convenience.

6 Using The Program

6.1 Launch ClutrFree

Under Linux and Unix, ClutrFree is launched by issuing the command line:

```
$ java -jar clutrfree.jar
```

The program accept also the syntax

```
$ java -jar clutrfree.jar <directory>
```

with `<directory>` being the directory root of the experiments to analyze. On other platforms a double-click on the jar file will launch the program. If you run the program from a text terminal, various informations on the annotations files and data files are displayed. From other platforms, a double-click on the jar file will launch the program.

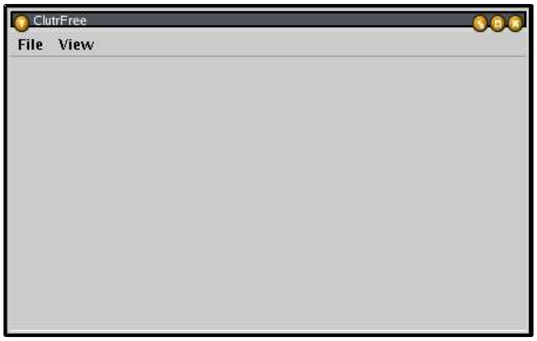
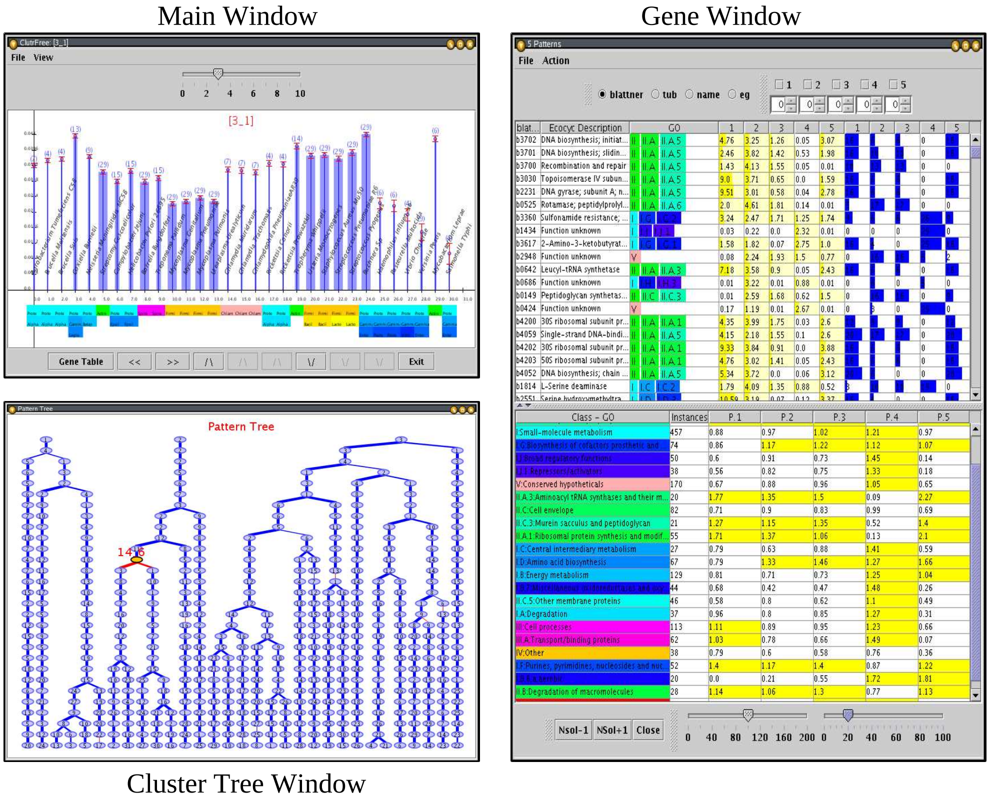


Figure 5: Main Window Of ClutrFree

6.2 The GUI Explained

At The first start of ClutrFree, an empty window appears (see Figure 5), unless the data directory has been given as an argument. From this window, the data can be imported using the menu [file][Import Data...]. The root directory of the experiment has to be chosen with the dialog box (for instance `"data_demo_bd"`). Once the data have been loaded, ClutrFree display a confirmation message. Once acknowledged, ClutrFree display tree windows: The first one is the main window that display the current cluster with its associated uncertainties, and persistence. The second and third one display respectively the Cluster Tree Window that aids the navigation in the Cluster Tree, and the Membership Tree. A click on the [gene table] button on the Main Window will display the gene table containing annotations, memberships, and ontology counts. The windows are displayed in Figure 6).



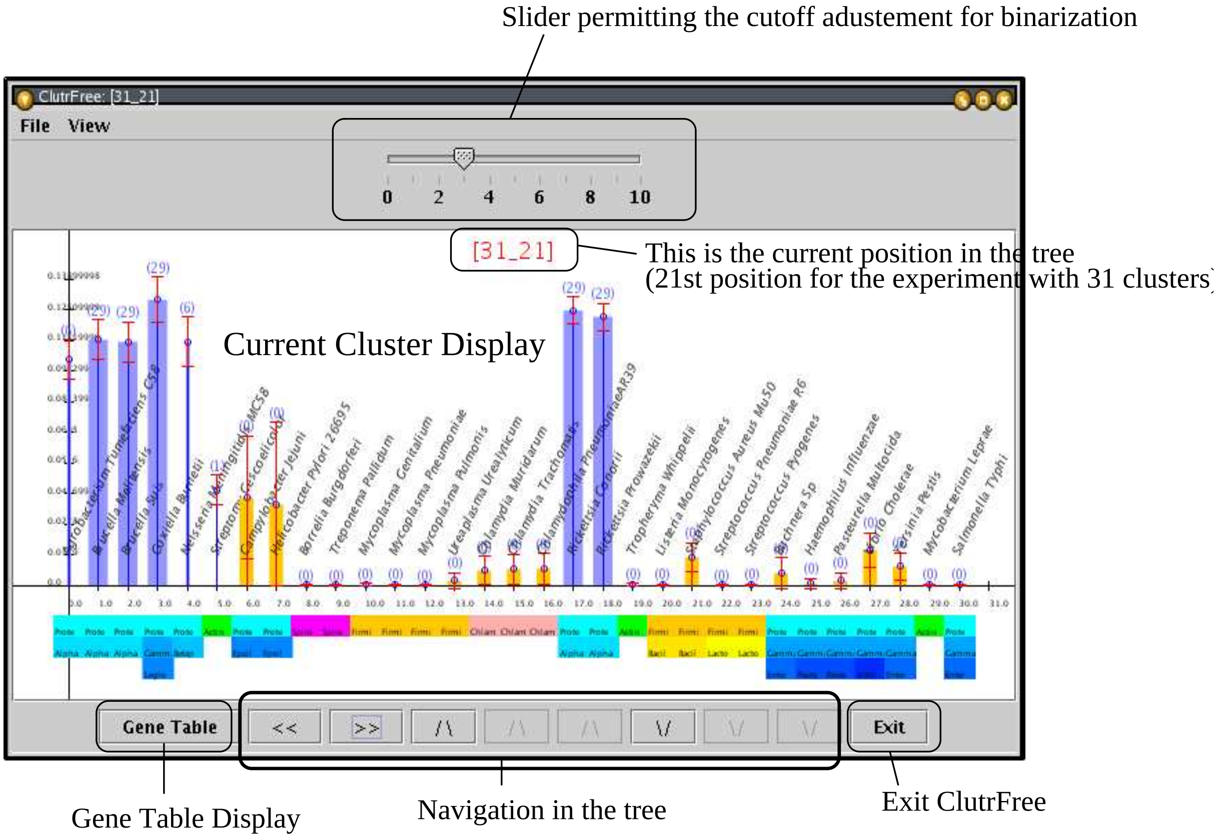
blat...	Ecocyc Description	GO			1	2	3	4	5	1	2	3	4	5
b3702	DNA biosynthesis; initiat...	II	II.A	II.A.5	4.76	3.25	1.26	0.05	3.07	16	[illegible]	[illegible]	0	18
b3701	DNA biosynthesis; slidin...	II	II.A	II.A.5	2.46	3.82	1.42	0.53	1.98	16	10	10	0	18
b3700	Recombination and repair	II	II.A	II.A.5	1.43	4.13	1.55	0.05	0.01	10	13	13	0	0
b3030	Topoisomerase IV subun...	II	II.A	II.A.5	9.0	3.71	0.65	0.0	1.59	16	[illegible]	[illegible]	0	18
b2231	DNA gyrase; subunit A; n...	II	II.A	II.A.5	9.51	3.01	0.58	0.04	2.78	16	6	6	0	18
b0525	Rotamase; peptidylprolyl...	II	II.A	II.A.6	2.0	4.61	1.81	0.14	0.01	[illegible]	17	17	0	0
b3360	Sulfonamide resistance; ...	I	I.G	I.G.2	3.24	2.47	1.71	1.25	1.74	9	6	6	26	9
b1434	Function unknown	I	I.J	I.J.1	0.03	0.22	0.0	2.32	0.01	0	0	0	29	0
b3617	2-Amino-3-ketobutyrat...	I	I.G	I.G.1	1.58	1.82	0.07	2.75	1.0	16	4	0	25	18
b2948	Function unknown	V			0.08	2.24	1.93	1.5	0.77	0	16	16	6	2
b0642	Leucyl-tRNA synthetase	II	II.A	II.A.3	7.18	3.58	0.9	0.05	2.43	16	6	6	0	18
b0686	Function unknown	I	I.H	I.H.3	0.01	3.22	0.01	0.88	0.01	0	6	0	5	0
b0149	Peptidoglycan synthetas...	II	II.C	II.C.3	0.01	2.59	1.68	0.62	1.5	0	16	16	0	9
b0424	Function unknown	V			0.17	1.19	0.01	2.67	0.01	0	3	0	29	0
b4200	30S ribosomal subunit pr...	II	II.A	II.A.1	4.35	3.99	1.75	0.03	2.6	12	9	9	0	12
b4059	Single-strand DNA-bindi...	II	II.A	II.A.5	4.15	2.18	1.55	0.1	2.6	20	17	17	0	20
b4202	30S ribosomal subunit pr...	II	II.A	II.A.1	9.33	3.84	0.91	0.0	3.88	16	[illegible]	[illegible]	0	18
b4203	50S ribosomal subunit pr...	II	II.A	II.A.1	4.76	3.02	1.41	0.05	2.43	16	[illegible]	6	0	18
b4052	DNA biosynthesis; chain ...	II	II.A	II.A.5	5.34	3.72	0.0	0.06	3.12	20	[illegible]	0	0	20
b1814	L-Serine deaminase	I	I.C	I.C.2	1.79	4.09	1.35	0.88	0.52	3	10	10	18	0
b2551	Serine hydroxymethyltra...	I	I.D	I.D.2	10.59	3.19	0.07	0.12	3.37	16	4	0	0	18

Class - GO	Instances	P.1	P.2	P.3	P.4	P.5
I:Small-molecule metabolism	457	0.88	0.97	1.02	1.21	0.97
I.G:Biosynthesis of cofactors prosthetic and ...	74	0.86	1.17	1.22	1.12	1.07
I.J:Broad regulatory functions	50	0.6	0.91	0.73	1.45	0.14
I.J.1:Repressors/activators	38	0.56	0.82	0.75	1.33	0.18
V:Conserved hypotheticals	170	0.67	0.88	0.96	1.05	0.65
II.A.3:Aminoacyl tRNA synthases and their m...	20	1.77	1.35	1.5	0.09	2.27
II.C:Cell envelope	82	0.71	0.9	0.83	0.99	0.69
II.C.3:Murein sacculus and peptidoglycan	21	1.27	1.15	1.35	0.52	1.4
II.A.1:Ribosomal protein synthesis and modif...	55	1.71	1.37	1.06	0.13	2.1
I.C:Central intermediary metabolism	27	0.79	0.63	0.88	1.41	0.59
I.D:Amino acid biosynthesis	67	0.79	1.33	1.46	1.27	1.66
I.B:Energy metabolism	129	0.81	0.71	0.73	1.25	1.04
I.B.7:Miscellaneous oxidoreductases and oxy...	44	0.68	0.42	0.47	1.48	0.26
II.C.5:Other membrane proteins	46	0.58	0.8	0.62	1.1	0.49
I.A:Degradation	37	0.96	0.8	0.85	1.27	0.31
III:Cell processes	113	1.11	0.89	0.95	1.23	0.66
III.A:Transport/binding proteins	62	1.03	0.78	0.66	1.49	0.07
IV:Other	38	0.79	0.6	0.58	0.76	0.36
I.F:Purines, pyrimidines, nucleosides and nuc...	52	1.4	1.17	1.4	0.87	1.22
I.B.8.a aerobic	20	0.0	0.21	0.55	1.72	1.81
II.B:Degradation of macromolecules	28	1.14	1.06	1.3	0.77	1.13

Figure 6: ClutrFree GUI with the Main Window, the Gene Table, and the Cluster Tree.

6.3 The Cluster window

The cluster window (Figure 7) permits the navigation among the tree using the button panel, the adjustment of the binarization cutoff and the display of the current cluster. It has a complex graphics display grouping most of the clusters related informations in one single resource that is detailed below.



Slider permitting the cutoff adustement for binarization

This is the current position in the tree (21st position for the experiment with 31 clusters)

Current Cluster Display

Gene Table Display

Navigation in the tree

Exit ClutrFree

Figure 7: Main Window Of ClutrFree with data loaded

Each cluster point (Figure 8) is plotted along with the following information;

- Values, Uncertainty.
- The point is painted in blue if considered a member, or yellow for a nonmember.
- The upper value is the *persistence* measure and the thickness of the blue rectangle is proportional to this persistence (see Figure 11.
- Each point is annotated with a string, and a color annotations if hierarchical annotations are provided (see section on the file formats).

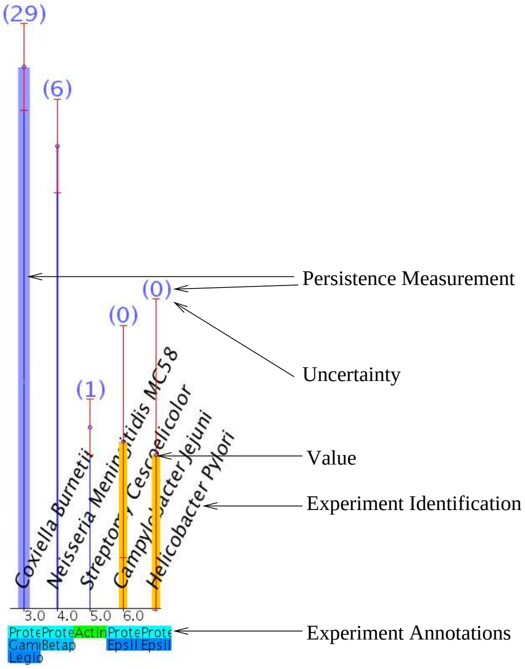
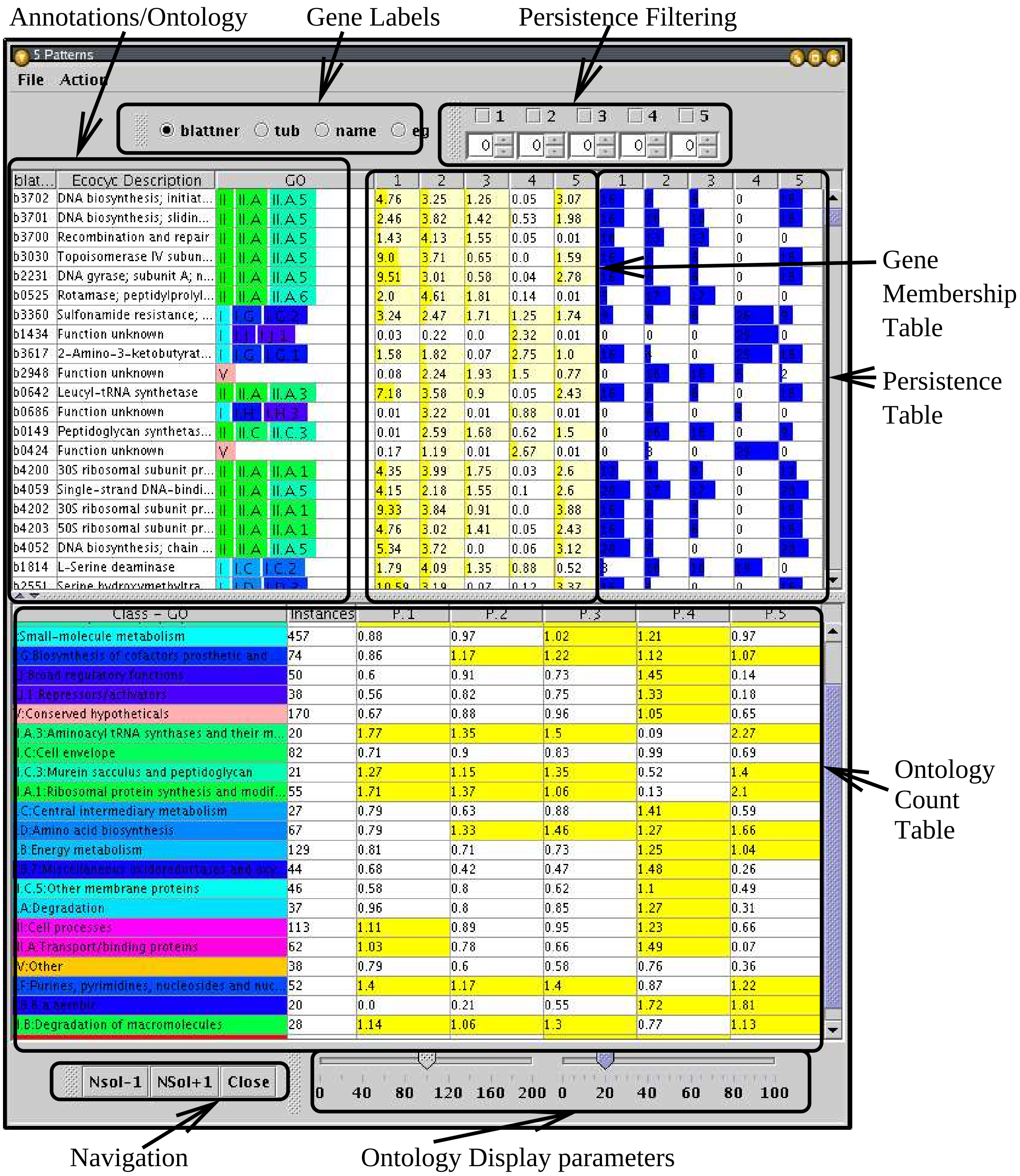


Figure 8: Cluster Stem Plot

6.4 The Gene Table

The Gene table can be accessed by the [Gene Table] button from the cluster window. As seen in Figure 9, the gene table shows essentially the membership of each gene for each cluster as well as advanced annotations to help the researcher to find coherent group of genes. The information displayed depends essentially on the annotations provided in the annotations genes files described hereafter.



blat...	Ecocyc Description	GO			1	2	3	4	5	1	2	3	4	5
b3702	DNA biosynthesis; initiat...	II	II.A	II.A.5	4.76	3.25	1.26	0.05	3.07	16	4	4	0	18
b3701	DNA biosynthesis; slidin...	II	II.A	II.A.5	2.46	3.82	1.42	0.53	1.98	16	10	10	0	18
b3700	Recombination and repair	II	II.A	II.A.5	1.43	4.13	1.55	0.05	0.01	10	13	13	0	0
b3030	Topoisomerase IV subun...	II	II.A	II.A.5	9.0	3.71	0.65	0.0	1.59	16	4	4	0	18
b2231	DNA gyrase; subunit A; n...	II	II.A	II.A.5	9.51	3.01	0.58	0.04	2.78	16	4	4	0	18
b0525	Rotamase; peptidylprolyl...	II	II.A	II.A.6	2.0	4.61	1.81	0.14	0.01	3	17	17	0	0
b3360	Sulfonamide resistance; ...	I	I.G	I.G.2	3.24	2.47	1.71	1.25	1.74	6	4	4	25	9
b1434	Function unknown	I	I.J	I.J.1	0.03	0.22	0.0	2.32	0.01	0	0	0	29	0
b3617	2-Amino-3-ketobutyrat...	I	I.G	I.G.1	1.58	1.82	0.07	2.75	1.0	16	4	0	25	18
b2948	Function unknown	V			0.08	2.24	1.93	1.5	0.77	0	16	16	6	2
b0642	Leucyl-tRNA synthetase	II	II.A	II.A.3	7.18	3.58	0.9	0.05	2.43	16	4	4	0	18
b0686	Function unknown	I	I.H	I.H.3	0.01	3.22	0.01	0.88	0.01	0	4	0	3	0
b0149	Peptidoglycan synthetas...	II	II.C	II.C.3	0.01	2.59	1.68	0.62	1.5	0	16	16	0	9
b0424	Function unknown	V			0.17	1.19	0.01	2.67	0.01	0	3	0	29	0
b4200	30S ribosomal subunit pr...	II	II.A	II.A.1	4.35	3.99	1.75	0.03	2.6	12	8	8	0	12
b4059	Single-strand DNA-bindi...	II	II.A	II.A.5	4.15	2.18	1.55	0.1	2.6	20	17	17	0	20
b4202	30S ribosomal subunit pr...	II	II.A	II.A.1	9.33	3.84	0.91	0.0	3.88	16	4	4	0	18
b4203	50S ribosomal subunit pr...	II	II.A	II.A.1	4.76	3.02	1.41	0.05	2.43	16	4	4	0	18
b4052	DNA biosynthesis; chain ...	II	II.A	II.A.5	5.34	3.72	0.0	0.06	3.12	20	4	0	0	20
b1814	L-Serine deaminase	I	I.C	I.C.2	1.79	4.09	1.35	0.88	0.52	3	10	10	19	0
b2551	Serine hydroxymethyltra...	I	I.D	I.D.2	10.59	3.19	0.07	0.12	3.37	16	3	0	0	16

Class - GO	Instances	P.1	P.2	P.3	P.4	P.5
:Small-molecule metabolism	457	0.88	0.97	1.02	1.21	0.97
.G:Biosynthesis of cofactors prosthetic and	74	0.86	1.17	1.22	1.12	1.07
J:Broad regulatory functions	50	0.6	0.91	0.73	1.45	0.14
J.1:Repressors/activators	38	0.56	0.82	0.75	1.33	0.18
V:Conserved hypotheticals	170	0.67	0.88	0.96	1.05	0.65
I.A.3:Aminoacyl tRNA synthases and their m...	20	1.77	1.35	1.5	0.09	2.27
I.C:Cell envelope	82	0.71	0.9	0.83	0.99	0.69
I.C.3:Murein sacculus and peptidoglycan	21	1.27	1.15	1.35	0.52	1.4
I.A.1:Ribosomal protein synthesis and modif...	55	1.71	1.37	1.06	0.13	2.1
.C:Central intermediary metabolism	27	0.79	0.63	0.88	1.41	0.59
.D:Amino acid biosynthesis	67	0.79	1.33	1.46	1.27	1.66
.B:Energy metabolism	129	0.81	0.71	0.73	1.25	1.04
.B.7:Miscellaneous oxidoreductases and oxy	44	0.68	0.42	0.47	1.48	0.26
I.C.5:Other membrane proteins	46	0.58	0.8	0.62	1.1	0.49
.A:Degradation	37	0.96	0.8	0.85	1.27	0.31
II:Cell processes	113	1.11	0.89	0.95	1.23	0.66
II.A:Transport/binding proteins	62	1.03	0.78	0.66	1.49	0.07
V:Other	38	0.79	0.6	0.58	0.76	0.36
.F:Purines, pyrimidines, nucleosides and nuc...	52	1.4	1.17	1.4	0.87	1.22
.B.6.a aerobic	20	0.0	0.21	0.55	1.72	1.81
I.B:Degradation of macromolecules	28	1.14	1.06	1.3	0.77	1.13

Figure 9: This is the gene window with all its majors components. The gene IDs and their annotationsontologies, the tool bars, the ontology counting table, and the navigation arrows.

The gene membership are displayed as a series of values in a table, each cell being plotted as a yellow bar. If the gene is considered as a member, the background will be highlighted in white, otherwise, the background will be set to white, as seen in the Figure 10.

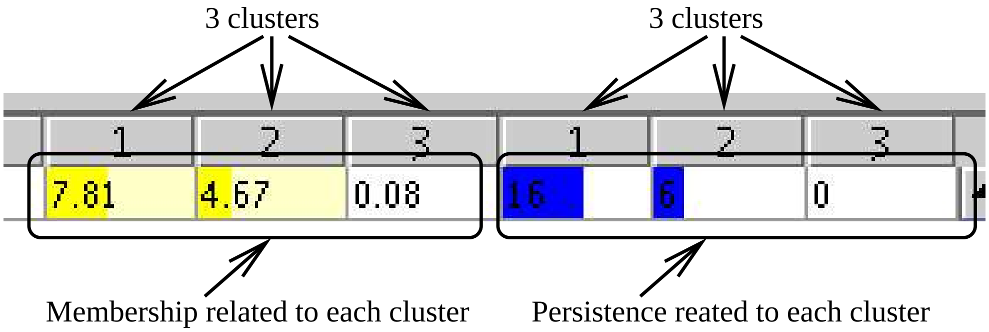

3 clusters
3 clusters
1
2
3
1
2
3
7.81
4.67
0.08
16
6
0
Membership related to each cluster
Persistence reated to each cluster

Figure 10: Magnification on the gene membership section of the Gene Table. In this example, the current gene is a member of the cluster 1 and 2. Its persistence is greater in cluster 1.

Along with the gene membership are plotted the *persistence* values. We defined as *persistence* the measurement of a gene to belong consecutively to the same tree branch. This is illustrated in Figure 11. The upper tool bar of the Gene Table (Figure 9) permits the filtering of the genes that do not meet a certain level of persistence ant therefore that are not stable along a tree branch.

On the bottom part of the genes window, ClutrFree displays a table of Ontology counts (if ontologies are provided) constructed by the following method (see Ochs (2003)):

For all patterns, ClutrFree determines the number of genes within the pattern assigned to each gene ontology term and compare this to the number of genes assigned to the term within the full data set. In order to avoid false enhancement due to only a few genes being assigned to the GO term, it is possible to eliminated all terms represented by less than a certain number genes in the data set. We computed the enhancement of the term within the pattern using the formula

$$Enhancement = \frac{N_{patt}^{GO}/N_{patt}^{TOTAL}}{N_{data}^{GO}/N_{data}^{TOTAL}} \tag{2}$$

where the numerator gives the normalized number of genes with the GO term

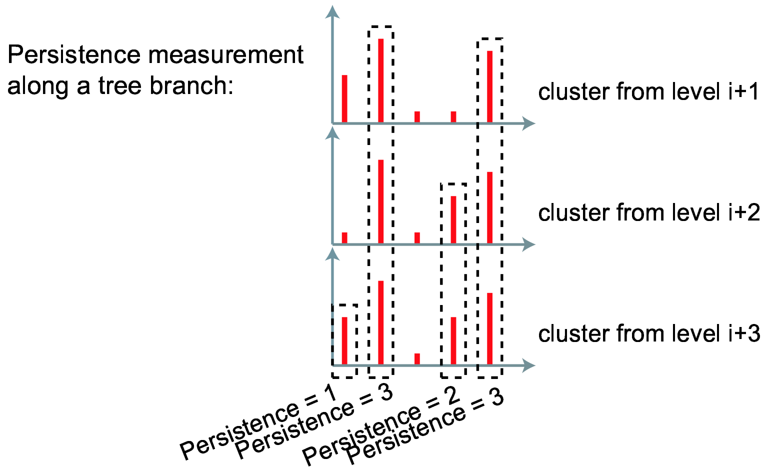


Figure 11: Principle of the persistence measure.

in the pattern, and the denominator is the normalized number of genes with the GO term in the full data set.

Next sections describes the files necessary to store this annotation data.

7 Annotation

7.1 Experiment Annotation File

The experiment annotation can takes two forms:

- For simple names description one can use a single row tab-delimited file describing each single experiment stored in `<root>/expnames.txt` formated as follow:

  ```
  expname_1   expname_2   ...
  ```

- For more complexes experiment annotations with color scheme, one can use a tab-delimited file stored in `<root>/expannot.txt` formatted the following way (`hipath` meaning "hierarchichal pathway", and `String` being arbitrary.):

```
ID:string  hipath:string   hipath:string
expname_1  I:Main_Class_1   I.1:Sub_Class_1
expname_2  I:Main_Class_1   I.1:Sub_Class_1
expname_3  I:Main_Class_1   I.1:Sub_Class_1
expname_4  II:Main_Class_2  I.1:Sub_Class_1
```

will lead to the graph shown in Figure 12. ("String" can be arbitrary fixed by the user. For The color scheme to be created by ClutrFree, the first class must be in roman number.

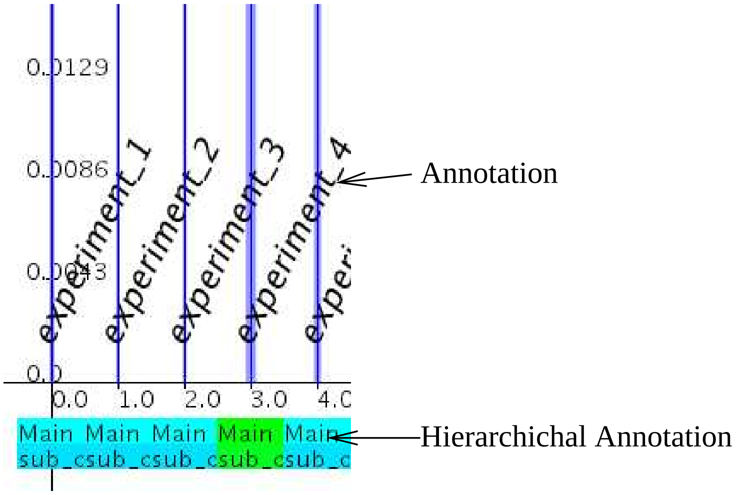


Figure 12: Example of cluster graph annotation

7.2 Gene Annotation File

The possible annotations "objects" that have been implemented so far, are the following: (The string in brackets is the header to be used in the description file).

- [ID], or identifiers, such as UniGene, or any short strings to identify the genes. If multiple IDs are used, only one at a time is displayed by the program.

- [Desc], or Descriptions, are plain string without particular structure or vocabulary.

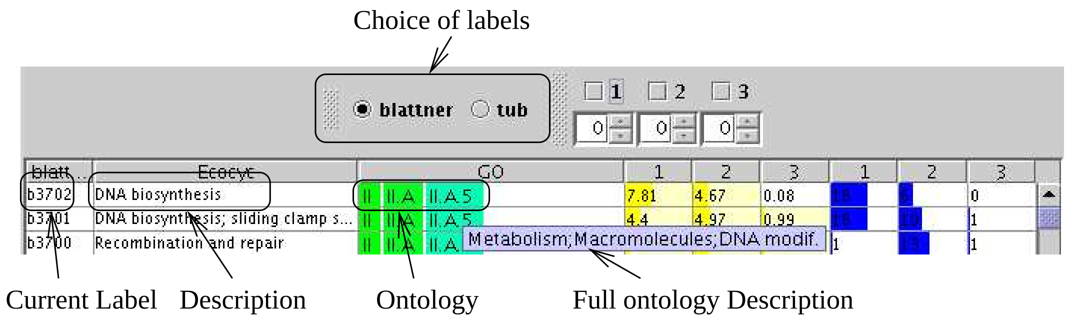
- [Path], or Pathways, are a series of string describing an ontology. The ontology hierarchy is not present in the IDs. The ontologies can takes many forms:
  - The template defined by the Gene Ontology Consortium (see Ashburner *et al.* (2000)):
    `"GO:number - Description"`.
  - A similar template to the one defined by the Consortium:
    `"number:description"`.
- [Hipath], or "hierarchichal pathways" are also a series of string describing an ontology, but with the hierarchy information embedded in the IDs:
  - A hierarchical form such as
    `"II.2.4b:Description"`.
    In this case, the hierarchy (`II.2.4b` being a sub-class of `II.2` being a sub-class of `II`) is taken into account and a color scheme is used to display the ontologies, improving considerably the readability. An example of such an ontology is given in figure 13.
  - A hierarchichal form such as `"II.2.4b:Description"`, identical to the previous case, but having multiple functions at the same level. The color scheme is also supported.
  - The form used in the Ecocyc (Karp *et al.*, 2002) database such as `"metabolism BC-1"` is also supported. In this case, a gene can also have multiple functions, and the color scheme is implemented as well. An example of such an ontology is given in figure 14.

The genes annotations are stored into two files:

- The IDs, descriptions, and ontologies are stored in `<root>/annot.txt` as follow. The file header identify the data type used in the annotation. The genes must be in the same order than in the membership matrix. For instance, a file formatted like this (`"Ecocyc"`, `"class tub"` being arbitrary strings)

  ```
  ID: blattner   ID: tub   Desc:  Ecocyc   hipath:class tub   hipath:class tub   hipath:class tub
  b3702          Rv0001    DNA biosynt.    II:metabolism      II.A:macromol.     II.A.5:  DNA repair
  ```

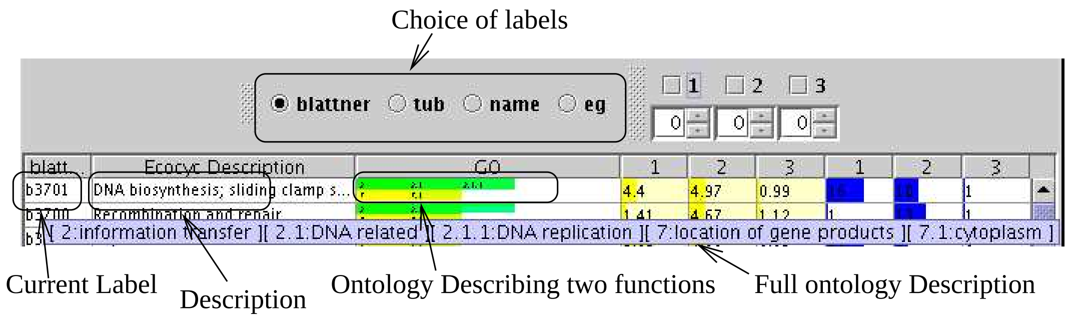
  will lead to the display seen in Figure 13.



blatt…	Ecocyc	GO	1	2	3	1	2	3
b3702	DNA biosynthesis	II II.A II.A.5	7.81	4.67	0.08	18	6	0
b3701	DNA biosynthesis; sliding clamp s...	II II.A II.A.5	4.4	4.97	0.99	18	10	1
b3700	Recombination and repair	II II.A II.A				1	19	1

Metabolism;Macromolecules;DNA modif.

Figure 13: Example of gene annotation: The full ontology description appears as a tooltip.


Choice of labels
● blattner ○ tub ○ name ○ eg
□ 1 □ 2 □ 3
0 0 0
Current Label
Description
Ontology Describing two functions
Full ontology Description

blatt…	Ecocyc Description	GO	1	2	3	1	2	3
b3701	DNA biosynthesis; sliding clamp s...		4.4	4.97	0.99	16	10	1
b3700	Recombination and repair		1.41	4.67	1.12	1	13	1

[2:information transfer][2.1:DNA related][2.1.1:DNA replication][7:location of gene products][7.1:cytoplasm]

Figure 14: Second example of gene annotation where genes are described by multiple functions. The full ontology description appears as a tooltip.

- Alternatively, Ontologies can be stored in `<root>/ontology.txt` in the format defined by the ASAP (Automated Sequence Annotation Pipeline, see Kossenkov *et al.* (2003)). The format is the following (The file is tab-delimited).

  ```
  gene_1 Ontology_1 Ontology_2
  gene_1 Ontology_1 Ontology_3
  gene_2 Ontology_1 Ontology_2 Ontology_3.
  ```

  With each ontology being defined as previously. The main advantage here is that the genes can be in any order and the same gene with the same ontology can appear several times.

7.3 Menus

This section summarized the menus of the GUI: The menus from the main window are the following:

- [File][Import Data] gives access to a file dialog box that allow the selection of the root directory containing the data.
- [File][Export Pattern Tree] brings up a dialog to export the pattern tree in a picture (JPEG, PNG and TIFF are the supported formats), or in a vector (SVG format, importable in Adobe Illustrator ©), or in the Graph Description format Dot used by the Graphviz package (made by AT&T). This package allows conversion to various formats, including Adobe Postscript ©.
- [File][Export Current Graphics] brings up a dialog to export the current displayed pattern in a picture (JPEG, PNG and TIFF are the supported formats), or in a vector (SVG format, importable in Adobe Illustrator ©).
- [File][About] brings up an about box with the author names and version number.
- [File][Exit] exits ClutrFree.
- [View][Stem] is the default view mode, usable when cluster points are independents.
- [View][Plot] set the view as a classic plot (similar to a time -series view).

The menus from the gene table are the following:

- [File][Export Membership Tree] export the Membership Tree the same way that the Pattern Tree can be exported (see [File][Export Pattern Tree]).
- [File][Export Membership Matrix] exports the membership matrix along with the gene IDs into a tab-delimited file.
- [File][Export Gene List] exports the gene IDs list into a tab-delimited file.

- [File][Export Enhancement Table] exports the gene counting table.
- [File][Close] closes the gene table.
- [Action][Cluster Binary Table] sorts the gene table according the the binary values of their membership.
- [Action][Search Value Table] search the annotations of the main table for a matching string in the data.
- [Action][Search Ontology Table] search the ontology table for a matching string.
- [Window][View The Membership Tree] re-open the membership window.
- [Persistence][Calculated on Membership Tree] Computes the persistence on the membership tree.
- [Persistence][Calculated on Pattern Tree] Computes the persistence on the pattern tree.

8 Licensing

ClutrFree and its documentation has been copyrighted under the General Public License (GPL, see details at `www.gpl.org` and in the `LICENCE.txt` file distributed with the source code). That essentially means that everybody is free to use ClutrFree, modify its source code and distribute it on a free basis, as long as the original authors names and institution are granted. It is strictly forbidden to prevent other from distributing it on the same basis.

9 History

9.1 This Documentation

- 25 Sep 2003 First Release (Version 1.0)
- 06 Jan 2004 This Release (Version 1.1)

9.2 ClutrFree Code

(See `HISTORY.txt` in the source tree for details)

- 23 Sep 2003: Version 1.0 released.
- 16 Oct 2003: Version 1.01
- 06 Jan 2004: Version 1.1

References

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, A., Matese, J., Richardson, J., Ringwald, M., Rubin, G. & Sherlock, G. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics,* **25**, 25–29.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. & Davis, R. W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell,* **2** (1), 65–73.

Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R. & Fedoroff, N. V. (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proceeding of the National Academy of Science,* **97** (15), 8409–8414.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. & Friend, S. H. (2000) Functional discovery via a compendium of expression profiles. *Cell,* **102**, 109–126.

Karp, P. D., Riley1, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., Pellegrini-Toole1, A., Bonavides, C. & Gama-Castro, S. (2002) The EcoCyc Database. *Nucleid Acid Research,* **30** (1), 56–58.

Kossenkov, A., Manion, F., Korotkov, E., Moloshok, T. D. & Ochs, M. F. (2003) Asap: automated sequence annotation pipeline for web-based updating of sequence information with a local dynamic database. *Bioinformatics,* **19**, 675–676.

Ochs, M. F. (2003) Bayesian decomposition. In *The Analysis of Gene Expression data, Methods and Software* New York Springer Verlag.

Ochs, M. F., Stoyanova, R. S., Arias-Mendoza, F. & Brown, T. R. (1999) A new method for spectral decomposition using a bilinear bayesian approach. *Journal of Magnetic Resonance,* **137**, 161–176.