Installation instructions

1. Extract the file "**alignPaths-pkg_Bundle.tar.gz**" to any location at your system. The folder "**alignPaths-pkg_Bundle**" which will be created will be referred to as "**BASE_DIR**" for the rest of this manual.

2. In a console window, with "**../BASE_DIR/mcl-10-201**" as the current directory, execute the following commands:

```
> ./configure
> make
> make install
```

3. Copy the file "**mcl**", which has been created in the folder "**../BASE_DIR/mcl-10-201/src/shmcl**", to the folder "**../BASE_DIR/bin**".

4. Install the MCR runtime engine in your system. You can obtain the MCR installation package from the following address:

http://www.mathworks.com/products/compiler/mcr/index.html

5. Copy the file "**javabuilder.jar**" from the "**../MATLAB_Compiler_Runtime/v711/toolbox/javabuilder/jar**" folder to the "**BASE_DIR/lib**" folder (the location of the "**MATLAB_Compiler_Runtime**" folder is defined during the MCR's installation at step 4).

6. Dowload the BLAST+ installer that fits to your system from the following address:
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST
Untar the latest installer file and copy the files "**balstp**" and "**makeblastdb**" from the "**bin**" folder to the "**../BASE_DIR/bin**" folder.

6. Set/update the following Environment variables:

>AXIS_PATH:
```
export AXIS_PATH=../BASE_DIR/lib/axis-1_4
```

>CLASSPATH:
```
export   CLASSPATH=$CLASSPATH:$AXIS_PATH/axis.jar:$AXIS_PATH/commons-
discovery-0.2.jar:$AXIS_PATH/commons-logging-
1.0.4.jar:$AXIS_PATH/javax.jms.jar:$AXIS_PATH/jaxrpc.jar:$AXIS_PATH/
mailapi_1_3_1.jar:$AXIS_PATH/servlet.jar:$AXIS_PATH/wsdl4j-
1.5.1.jar:/BASE_DIR/lib/keggapi.jar:/BASE_DIR/lib/javaml-
0.1.5.jar:/BASE_DIR/lib/weka.jar:/BASE_DIR/lib/javabuilder.jar:/BASE
_DIR/lib/phylogeny_pkg.jar:.
```

>PATH:
```
export PATH=$PATH:/BASE_DIR/bin:.
```

> JAVA_HOME:
e.g. in Mac OS X:
```
export JAVA_HOME=/Library/Java/Home
```

How-to-Use instructions

1. Input

The input parameters for the program are read through the "**input.txt**" file, which is (and must be) located at the **BASE_DIR** folder. The "**input.txt**" file must comply with the following format:

```
> Pathway Map Id:
String
> MCL:
boolean
> Inflation parameter:
int
> EM:
boolean
> e-Value:
int,int
> Number of genomes:
int
> Genomes:
String
String
String
...
```

Fig. 1. "input.txt" file format.

Each line with **bold** formatting (called as a field) declares the parameter type whose value/values follow in the next line/lines until the next field with *italic* formatting.

- The field **"Pathway Map id"** accepts as a value a KEGG pathway map identifier, e.g. *00010* for the Glycolysis/Gluconeogenesis pathway.
- The field **"MCL"** accepts the value *true* if the MCL algorithm is going to be used for the gene clustering. Otherwise, it must be set to *false*.
- The field **"Inflation parameter"** accepts as a value an *integer* that defines the *inflation parameter* for the execution of the MCL algorithm (optimal value: 12).
- The field **"EM"** accepts the value *true* if the EM algorithm is going to be used for the gene clustering. Otherwise, it must be set to *false*.
- The field **"e-Value"** accepts as a value two comma-separated *integers.* Let *a,-b* be the set of values for the **"e-Value"** field. Then, the e-Value parameter's value will be: a^{-b}.
- The field **"Number of genomes"** accepts as a value an *integer* that defines the number of genomes participating in the current run.
- The field **"Genomes"** accepts as values a newline-separated list of *Strings*, which are the KEGG identifiers of the genomes participating in the current run.

A sample **"input.txt"** file is given below:

```
> Pathway Map Id:
00010
> MCL:
true
> Inflation parameter:
12
> EM:
false
> e-Value:
10,-5
> Number of genomes:
3
> Genomes:
ath
eco
hsa
```

Fig. 2. Sample **"input.txt"** file.

2. Run

Having set the desired parameters for the current run in the **"input.txt"** file, in a console window, with "**../BASE_DIR/src**" as the current directory, execute the following commands:

```
> javac *.java
> java MainClass
```

3. Ouput

The program's output is generated at the **"../BASE_DIR/output"** folder and includes the following files:

3.1 Pathway images:

a. only with clusters marking:
Saved at the "BASE_DIR/output/***NUM_OF_GENOMES***-genomes/***TIMESTAMP***/MCL/images" folder.

b. with clusters & Groups marking:
Saved at the "BASE_DIR/output/***NUM_OF_GENOMES***-genomes/***TIMESTAMP***/MCL/images_with_groups" folder.

3.2 FASTA files for each cluster:

Saved at the "BASE_DIR/output/***NUM_OF_GENOMES***-genomes/***TIMESTAMP***/MCL/fasta" folder.

Each file's name complies with the following naming convention:

"cluster_***NUM_OF_CLUSTER***_FastaFile.txt"

e.g.

"cluster_1_FastaFile.txt"

3.3 Phylogenetic trees

The phylogenetic tree image files are saved at the "BASE_DIR/output/***NUM_OF_GENOMES***-genomes/***TIMESTAMP***/MCL/phylo_trees/ ***NUM_OF_GENOMES***_genomes" folder.

Each file's name complies with the following naming convention:

"Group_***NUM_OF_GROUP***.Tree.Cl_***Cluster_Id***.ec_***EC_Id***. koId_ ***KO_Id***.png"

e.g.

Group_2.Tree.Cl_1.ec_2.7.1.1.koId_K00844.png.

3.4 Output Log File (outputLog.txt)

The **"outputLog.txt"** file's format is given below:

```
Log for MCL Algorithm.

> Pathway: String. (e.g. map00052.)

> Genomes examined: String, String, ... (e.g. ath, dme, hsa.)

> Total number of genes: int. (e.g. 103.)

> Number of clusters: int. (e.g. 4)

> Number of genes in each cluster:

- Cluster int: int [String(int), String(int), ...].
...
```

(e.g.
- Cluster 1: 77 [ath(34), dme(20), hsa(23)].
- Cluster 2: 12 [ath(0), dme(10), hsa(2)].
)

> List of clusters with genes belonging only to a single genome:

- Cluster *int*:
 # Genome: *String*
 # Genes:
String **(***String***)**
...

(e.g.
- Cluster 3:
Genome: ath
Genes:
ath:AT1G12240 (ec:3.2.1.26)
ath:AT1G55740 (ec:2.4.1.82)
...
)

> Clustering validation:

- Average ~Similarity~ between clusters:

AVG_SIMILARITY_MATRIX

(e.g.

Clusters	1	2	3	4
1	1.0	0.417	0.333	0.333
2	0.417	2.156	0.375	0.625
3	0.333	0.375	3.0	0.5
4	0.333	0.625	0.5	3.0

)

- Maximum ~Similarity~ between clusters:

MAX_SIMILARITY_MATRIX

(e.g.

Clusters	1	2	3	4
1	1.0	0.667	0.333	0.333
2	0.667	3.0	0.5	1.0
3	0.333	0.5	3.0	0.5
4	0.333	1.0	0.5	3.0

)

- Minimum ~Similarity~ between clusters:

MIN_SIMILARITY_MATRIX

(e.g.

Clusters	1	2	3	4
1	1.0	0.333	0.333	0.333
2	0.333	1.0	0.0	0.5
3	0.333	0.0	3.0	0.5
4	0.333	0.5	0.5	3.0

)

- Standard deviation of ~Similarity~ between clusters:

STD_SIMILARITY_MATRIX

(e.g.

Clusters	1	2	3	4
1	0.0	0.144	0.0	0.0
2	0.144	0.967	0.217	0.219
3	0.0	0.217	0.0	0.0
4	0.0	0.219	0.0	0.0

)

****

- ~Homologies/Gene~ between clusters:

MIN_SIMILARITY_MATRIX

(e.g.

Clusters	1	2	3	4
1	0.184	0.04	0.039	0.039
2	0.077	0.569	0.0	0.0
3	0.105	0.0	0.45	0.0
4	0.097	0.0	0.0	0.5

)

{Total number of 'BLACK' elements: *2*}
- Clusters/Genes extracted from 'BLACK' elements:
> *String*:
Cluster *int* -> [*String String ...*]

(e.g.
{Total number of 'BLACK' elements: 2}
- Clusters/Genes extracted from 'BLACK' elements:
> ec:3.2.1.20:
Cluster 1 -> [ath:AT3G45940 ath:AT5G11720 dme:Dmel_CG11909 hsa:2548 hsa:2595 hsa:8972]
Cluster 2 -> [dme:Dmel_CG11669 dme:Dmel_CG14934 dme:Dmel_CG14935 dme:Dmel_CG7685 dme:Dmel_CG8690 dme:Dmel_CG8693 dme:Dmel_CG8694 dme:Dmel_CG8695 dme:Dmel_CG8696]
> ec:3.2.1.23:
Cluster 1 -> [ath:AT1G72990 ath:AT3G52840 dme:Dmel_CG3132 dme:Dmel_CG9092 hsa:2720]
Cluster 3 -> [ath:AT3G54440]
)

Cluster: *int*, **EC:** *String*

- Average phylogenetic distances:
AVG_PHYLO_DIST_MATRIX
...

(e.g.
Cluster: 1, EC: ec:1.1.1.21
- Average phylogenetic distances:

Genomes	ath	dme	hsa
ath			
dme	0.995		
hsa	0.897	0.766	

)

***** GROUP:** *int* *******
PHYLO_DISTANCES_FROM_MATLAB_String
> Cluster: *int*, **EC:** *String*
...

(e.g.
*** GROUP: 1 ***
(Arabidopsis thaliana \(thale cress\) \(ath\),(Drosophila melanogaster \(fruit fly\) \(dme\),Homo sapiens \(human\) \(hsa\)));
> Cluster: 1, EC: ec:1.1.1.21
> Cluster: 1, EC: ec:2.7.1.11
> Cluster: 1, EC: ec:2.7.1.6
> Cluster: 1, EC: ec:2.7.7.12
> Cluster: 1, EC: ec:2.7.7.9
> Cluster: 1, EC: ec:3.2.1.23
> Cluster: 1, EC: ec:5.1.3.2

*** GROUP: 2 ***
(Homo sapiens \(human\) \(hsa\),(Arabidopsis thaliana \(thale cress\) \(ath\),Drosophila melanogaster \(fruit fly\) \(dme\)));
> Cluster: 1, EC: ec:2.7.1.1

- (Genomes - Clusters) Matrix:

GENOMES-CLUSTERS_MATRIX

(e.g.

Genomes	Cluster 1	Cluster 2	Cluster 3	Cluster 4
ath	1	0	1	0
dme	1	1	0	0
hsa	1	1	0	1

)

- (EC numbers - Clusters) Matrix:

ECNumbers-CLUSTERS_MATRIX

(e.g.

EC Numbers	Cluster 1	Cluster 2	Cluster 3	Cluster 4
ec:1.1.1.21	1.0	0.0	0.0	0.0
ec:2.4.1.22	0.0	0.0	0.0	1.0
ec:2.4.1.67	0.0	0.0	1.0	0.0
ec:2.4.1.82	0.0	0.0	1.0	0.0
ec:2.7.1.1	1.0	0.0	0.0	0.0
ec:2.7.1.11	1.0	0.0	0.0	0.0
ec:2.7.1.2	1.0	0.0	0.0	0.0

)

>> Appendix.
> Clusters generated.

- Cluster *int*:
String String String ...
...

(e.g.
- Cluster 3:
ath:AT1G12240 ath:AT1G55740 ath:AT1G62660 ath:AT2G36190
ath:AT3G13790 ath:AT3G54440 ath:AT3G57520 ath:AT4G01970
ath:AT5G20250 ath:AT5G40390

- Cluster 4:
hsa:2683 hsa:3906 hsa:3938 hsa:8704
)

> EC numbers -> Genes mapping:
String **->** **[***String, String, String, ...***]**
...

(e.g.
ec:1.1.1.21 -> [ath:AT2G37790, dme:Dmel_CG10863, dme:Dmel_CG12766,
dme:Dmel_CG6083, dme:Dmel_CG6084, dme:Dmel_CG9436, hsa:231, hsa:57016]
ec:2.4.1.22 -> [hsa:2683, hsa:3906, hsa:8704]
ec:2.4.1.67 -> [ath:AT4G01970]
ec:2.4.1.82 -> [ath:AT1G55740, ath:AT3G57520, ath:AT5G20250,
ath:AT5G40390]
ec:2.7.1.1 -> [ath:AT1G47840, ath:AT1G50460, ath:AT2G19860,
ath:AT3G20040, ath:AT4G29130, ath:AT4G37840, dme:Dmel_CG3001,
dme:Dmel_CG32849, dme:Dmel_CG33102, dme:Dmel_CG8094, hsa:3098, hsa:3099,
hsa:3101, hsa:80201]
)

Total time elapsed: *int* **sec**

(e.g.
Total time elapsed: 266 sec

{Revised by Dimitrios Vitsios, June 3, 2012}