

Supplementary Material for

“Mining co-regulated gene profiles for the detection of functional associations in gene expression data”

Attila Gyenesei, Ulrich Wagner, Simon Barkow, Etzard Stolte and Ralph Schlapbach

Content:

I.) Synthetic Gene-Expression Data Analysis

II.) Yeast80 and Compendium Data Sets Analysis

III.) Cluster Data

IV.) Gene Ontology Results

V.) Pathway Analysis

VI.) Software usage

I.) Synthetic Gene-Expression Data Analysis

Supplementary data for Fig.1:

The synthetic gene-expression data

(<http://fgcz-intranet.uzh.ch/Publish/publications/GyeneseiEtAl2007/SynthSampleData1.txt>):

Sample gene expression data for Fig.1.								
	c1	c2	c3	c4	c5	c6	c7	c8
gene A	3.55	2.57	2.28	2.02	3.53	3.55	3.30	2.91
gene B	3.52	2.23	2.06	2.67	3.69	3.39	3.92	2.78
gene C	3.13	2.30	2.47	2.56	3.12	3.42	3.70	2.67
gene D	3.38	2.37	2.46	2.53	3.65	3.21	3.22	2.81
gene E	-2.43	-1.54	-1.50	-1.13	-2.84	-2.09	-2.13	-1.99
gene F	-2.45	-1.04	-1.32	-1.44	-2.17	-2.20	-2.35	-1.90
gene G	-2.91	-1.64	-1.81	-1.78	-2.32	-2.80	-2.79	-1.15
gene H	-2.47	-1.40	-1.65	-1.11	-2.69	-2.06	-2.80	-1.23

- The MAP algorithm produces the following result (*SynthSampleData1.map*):

{1,2,3,4,5,6,7,8 : 1,2,3,4 (-) 5,6,7,8} (8,8)

- Traditional APD methods provide the following results:

{1,2,3,4,5,6,7,8 : 1,2,3,4} (8,4)

{1,2,3,4,5,6,7,8 : 5,6,7,8} (8,4)

- The PNCGC algorithm (Ji and Tan, 2004) gives the following result(given in <http://fgcz-intranet.uzh.ch/publish/Publications/GyeneseiEtAl2007/PNCGC.zip> (*data1_posneg.txt*)):

There are 3 Positive and Negative Co-regulated Gene Clusters in all.
 (Gene1 Gene5 Gene6)@(A1A3, A1A4, A2A5, A2A6, A3A5, A3A6, A3A7, A4A5, A4A6, A4A7, A4A8,)
 (Gene2 Gene5 Gene6)@(A1A2, A1A3, A2A5, A2A6, A2A7, A3A5, A3A6, A3A7, A3A8, A4A5, A4A7,)
 (Gene5 Gene6)@(A1A2, A1A3, A1A4, A2A5, A2A6, A2A7, A3A5, A3A6, A3A7, A3A8, A4A5, A4A6, A4A7, A4A8,)

- This synthetic data set has also been mined using the bi-clustering methods Bimax, CC, ISA and OPSM, as implemented in the BiCAT software (see section 2.7). The result files are collected in a folder “SynthSampleData1” and stored in a zipped folder, which can be downloaded from:

<http://fgcz-intranet.uzh.ch/publish/Publications/GyeneseiEtAl2007/MAPProfilesAndBiclusters.zip>. The respective result files (e.g. “SynthSampleDataSet1.MAP.txt” etc) for each algorithm as given by BiCAT consist of three rows per cluster: The first row corresponds to the number of genes and number of conditions, resp. The second row comprises the genes in the cluster and the third row the conditions. For each cluster, a graphical output was produced as a .png file. The naming of the file is composed of the number of genes, the number of conditions and the row number in the result file.

Supplementary data for Fig.2:

The synthetic gene-expression data

(<http://fgcz-intranet.uzh.ch/publish/publications/GyeneseiEtAl2007/SynthSampleData2.txt>):

Sample gene expression data for Fig.2.								
	c1	c2	c3	c4	c5	c6	c7	c8
gene A	3.18	3.12	3.67	-2.75	-2.01	-2.81	3.42	2.76
gene B	3.19	3.05	3.61	-2.19	-2.16	-2.80	3.76	2.59
gene C	3.63	3.85	3.91	-2.60	-2.56	-2.07	3.49	2.63
gene D	3.64	3.23	3.03	-2.09	-2.10	-2.06	3.33	2.96
gene E	-2.54	-2.98	-2.45	3.13	3.63	3.18	-2.25	-1.11
gene F	-2.48	-2.18	-2.28	3.48	3.90	3.60	-2.02	-1.32
gene G	-2.15	-2.08	-2.45	3.93	3.84	3.51	-2.66	-1.90
gene H	-2.15	-2.96	-2.90	3.54	3.22	3.72	-2.20	-1.05

- The MAP algorithm produces the following result (*SampleData2.map*):

{1,2,3,4,5,6,7,8 : 1,2,3,4 (-) 5,6,7,8} (8,8)

- Traditional APD methods do not provide any results for **all of the conditions**.

- The PNCGC algorithm (Ji and Tan, 2004) gives the following result(given in <http://fgcz-intranet.uzh.ch/publish/Publications/GyeneseiEtAl2007/PNCGC.zip> (*data2_posneg.txt*)):

There are 0 Positive and Negative Co-regulated Gene Clusters in all.

- This synthetic data set has also been mined using the bi-clustering methods Bimax, CC, ISA and OPSM, as implemented in the BiCAT software (see section 2.7). The result files are collected in a folder “SynthSampleData2” and stored in a zipped folder, which can be downloaded from:

<http://fgcz-intranet.uzh.ch/publish/Publications/GyeneseiEtAl2007/MAPProfilesAndBiclusters.zip>

.The respective result files (e.g. “SynthSampleDataSet2.MAP.txt” etc) for each algorithm

as given by BiCAT consist of three rows per cluster: The first row corresponds to the number of genes and number of conditions, resp. The second row comprises the genes in the cluster and the third row the conditions. For each cluster, a graphical output was produced as a .png file. The naming of the file is composed of the number of genes, the number of conditions and the row number in the result file.

Randomized data of Fig.2:

The synthetic gene-expression data

(<http://fgcz-intranet.uzh.ch/publish/Publications/GyeneseiEtAl2007/SynthSampleData3.txt>):

:

Sample gene expression data for Fig.1.								
	c1	c2	c3	c4	c5	c6	c7	c8
gene A	-2.45	-2.25	-2.02	3.12	3.85	3.84	3.03	2.96
gene B	-2.45	-2.19	-2.75	3.6	2.59	-2.56	-2.16	2.63
gene C	3.64	-2.45	-2.81	3.76	3.61	3.48	-2.2	3.9
gene D	3.63	-2.09	3.18	-2.06	3.13	-2.98	3.49	-1.11
gene E	3.18	-1.05	3.67	-2.01	3.51	3.42	-2.08	2.76
gene F	-2.15	3.22	-2.8	-2.9	3.23	3.72	-2.07	3.33
gene G	3.19	-1.9	-2.96	3.54	3.63	-2.28	3.05	-2.45
gene H	-2.15	-1.32	3.91	-2.6	-2.18	-2.1	3.93	-2.66

- The MAP algorithm produces the following result (*SampleData1.map*):

```

{7,8 : 2,3,5,6 (-) 4,7,8} (2,7)
{6,8 : 1,3,5,6 (-) 4,7,8} (2,7)
{6,7 : 1 (-) 2} (2,2)
{6,7,8 : 3,5,6 (-) 4,7,8} (3,6)
{5,8 : 1,2,3,5,6 (-) 8} (2,6)
{5,7,8 : 2,3,5,6 (-) 8} (3,5)
{5,6,8 : 1,3,5,6 (-) 8} (3,5)
{5,6,7,8 : 3,5,6 (-) 8} (4,4)
{4,7 : 1,7 (-) 5,6} (2,4)
{4,6 : 2,7 (-) 5,6} (2,4)
{4,6,7,8 : 5,6 (-) 7} (4,3)
{4,5 : 4,5,6} (2,3)
{4,5,6,7,8 : 5,6} (5,2)
{3,8 : 1,2,3,6 (-) 4,8} (2,6)
{3,7,8 : 2,3,6 (-) 4,8} (3,5)
{3,6,8 : 1,3,6 (-) 4,8} (3,5)
{3,6,7,8 : 3,6 (-) 4,8} (4,4)
{3,5 : 1,2,3,6,7 (-) 8} (2,6)
{3,5,8 : 1,2,3,6 (-) 8} (3,5)
{3,5,7,8 : 2,3,6 (-) 8} (4,4)
{3,5,6,8 : 1,3,6 (-) 8} (4,4)
{3,5,6,7,8 : 3,6 (-) 8} (5,3)
{3,4 : 1,2,3,7 (-) 4,5,8} (2,7)
{3,4,8 : 1,2,3 (-) 4,8} (3,5)
{3,4,7 : 1,7 (-) 5} (3,3)
{3,4,7,8 : 2,3 (-) 4,8} (4,4)
{3,4,6 : 2,7 (-) 5} (3,3)
{3,4,6,8 : 1,3 (-) 4,8} (4,4)
{3,4,6,7,8 : 3 (-) 4,8} (5,3)
{3,4,6,7,8 : 5 (-) 7} (5,2)
{3,4,5 : 1,2,3,7 (-) 8} (3,5)
{3,4,5,8 : 1,2,3 (-) 8} (4,4)
{3,4,5,7,8 : 2,3 (-) 8} (5,3)
{3,4,5,6,8 : 1,3 (-) 8} (5,3)
{1,8 : 1,2,6 (-) 4,7} (2,5)
{1,7 : 1,8 (-) 3,5} (2,4)
{1,7,8 : 2,6 (-) 4,7} (3,4)
{1,6 : 2,8 (-) 3,5} (2,4)
{1,6,8 : 1,6 (-) 4,7} (3,4)
{1,5 : 3,4,5,7 (-) 8} (2,5)
{1,5,6,7,8 : 3,5 (-) 8} (5,3)
{1,4 : 3,7 (-) 6,8} (2,4)
{1,3 : 1,2,6 (-) 4,5} (2,5)
{1,3,8 : 1,2,6 (-) 4} (3,4)
{1,3,7,8 : 2,6 (-) 4} (4,3)
{1,3,6,8 : 1,6 (-) 4} (4,3)
{1,3,5,8 : 1,2,6} (4,3)
{1,3,5,7,8 : 2,6} (5,2)
{1,3,5,6,8 : 1,6} (5,2)
{1,3,4 : 1,2 (-) 4,5} (3,4)
{1,3,4,8 : 1,2 (-) 4} (4,3)
{1,3,4,7 : 1 (-) 5} (4,2)
{1,3,4,7,8 : 2 (-) 4} (5,2)

```

{1,3,4,6 : 2 (-) 5} (4,2)
 {1,3,4,6,8 : 1 (-) 4} (5,2)
 {1,3,4,5 : 3,7 (-) 8} (4,3)
 {1,3,4,5,6,7,8 : 3 (-) 8} (7,2)
 {2,7 : 1,4,7,8 (-) 6} (2,5)
 {2,6 : 2,4,7,8 (-) 6} (2,5)
 {2,6,7,8 : 4,7,8 (-) 6} (4,4)
 {2,5 : 1,2,3,4,5,7} (2,6)
 {2,5,8 : 1,2,3,5} (3,4)
 {2,5,7 : 1,4,7} (3,3)
 {2,5,7,8 : 2,3,5} (4,3)
 {2,5,6 : 2,4,7} (3,3)
 {2,5,6,8 : 1,3,5} (4,3)
 {2,4 : 1,2,3,7 (-) 6} (2,5)
 {2,4,7 : 1,7 (-) 6} (3,3)
 {2,4,6 : 2,7 (-) 6} (3,3)
 {2,3 : 4,5,8 (-) 6} (2,4)
 {2,3,6,7,8 : 4,8 (-) 6} (5,3)
 {2,3,5,6,7,8 : 6 (-) 8} (6,2)
 {2,3,4 : 4,5,8} (3,3)
 {2,3,4,6,7,8 : 4,8} (6,2)
 {2,3,4,5 : 1,2,3,7} (4,4)
 {2,3,4,5,8 : 1,2,3} (5,3)
 {2,3,4,5,7 : 1,7} (5,2)
 {2,3,4,5,7,8 : 2,3} (6,2)
 {2,3,4,5,6 : 2,7} (5,2)
 {2,3,4,5,6,8 : 1,3} (6,2)
 {1,2 : 1,2,8} (2,3)
 {1,2 : 3,4,5,7 (-) 6} (2,5)
 {1,2,7 : 1,8} (3,2)
 {1,2,6 : 2,8} (3,2)
 {1,2,6,7,8 : 4,7 (-) 6} (5,3)
 {1,2,5 : 3,4,5,7} (3,4)
 {1,2,5,6,7,8 : 3,5} (6,2)
 {1,2,5,6,7,8 : 4,7} (6,2)
 {1,2,4 : 3,7 (-) 6} (3,3)
 {1,2,4,6,7,8 : 6 (-) 7} (6,2)
 {1,2,3 : 4,5 (-) 6} (3,3)
 {1,2,3,6,7,8 : 4 (-) 6} (6,2)
 {1,2,3,4,5 : 3,7} (5,2)
 {1,2,3,4,5 : 4,5} (5,2)
 {1,2,3,4,5,8 : 1,2} (6,2)

- The PNCGC algorithm (Ji and Tan, 2004) gives the following result(given in <http://fgcz-intranet.uzh.ch/publish/Publications/GyeneseiEtAl2007/PNCGC.zip> (*data1_posneg.txt*)):
- This synthetic data set has also been mined using the bi-clustering methods Bimax, CC, ISA and OPSM, as implemented in the BiCAT software (see section 2.7). The result files are collected in a folder “SynthSampleData3” and stored in a zipped folder, which can be downloaded from: <http://fgcz-intranet.uzh.ch/publish/Publications/GyeneseiEtAl2007/MAPPProfilesAndBiclusters.zip>. The respective result files (e.g. “SynthSampleDataSet3.MAP.txt” etc) for each algorithm as given by BiCAT consist of three rows per cluster: The first row corresponds to the number of genes and number of conditions, resp. The second row comprises the genes in the cluster and the third row the conditions. For each cluster, a graphical output was produced as a .png file. The naming of the file is composed of the number of genes, the number of conditions and the row number in the result file.
-

Running Example: Sample data set: *SynthSampleData4.txt*

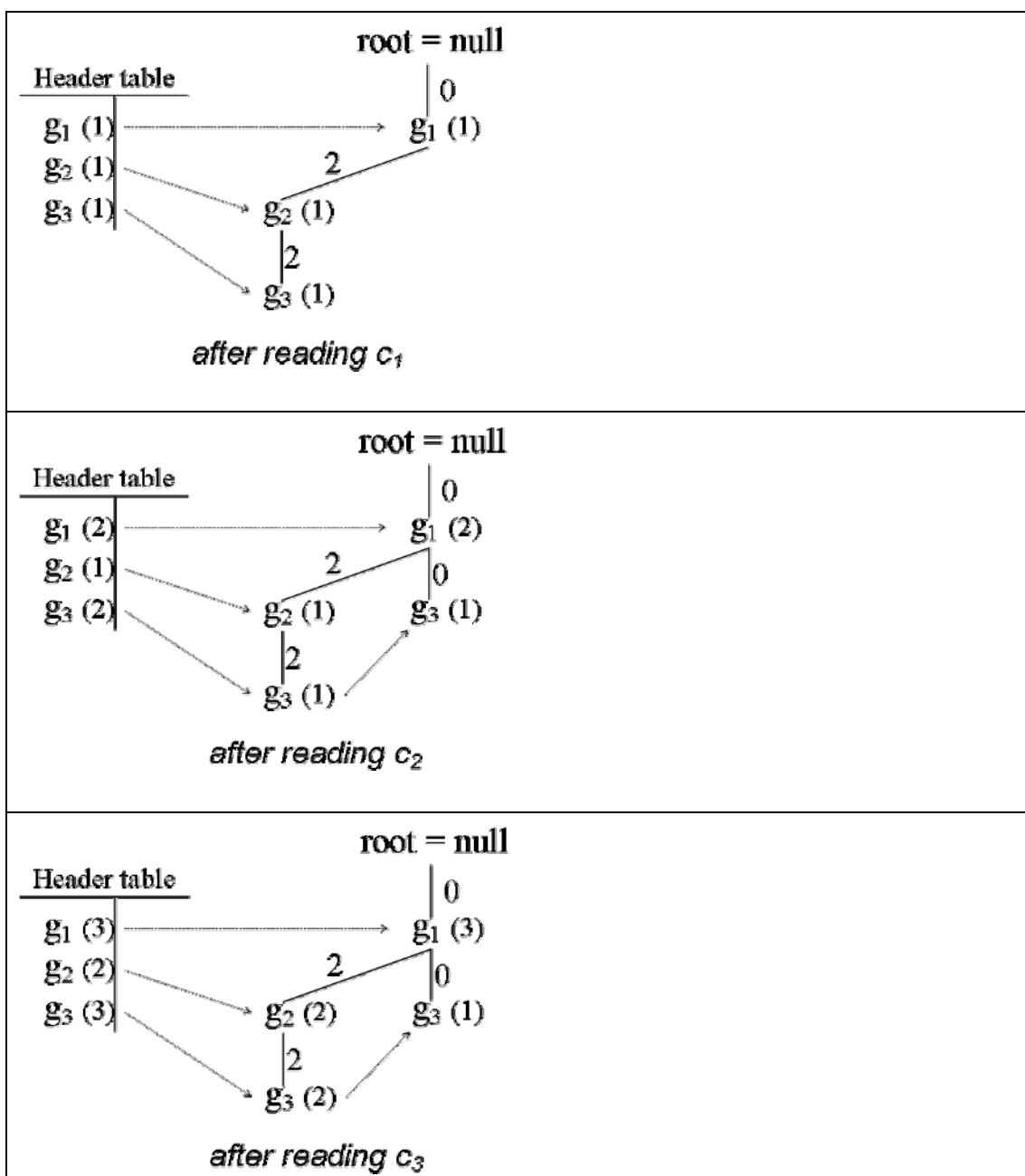
	g1	g2	g3
c1	-1	1	-1
c2	1	0	1
c3	1	-1	1

The mining is carried out in two steps in which the first step constructs a compact data structure called Gene Profile tree (or GP-tree), and the second step extracts the frequent co-regulated gene profiles directly from the GP-tree structure.

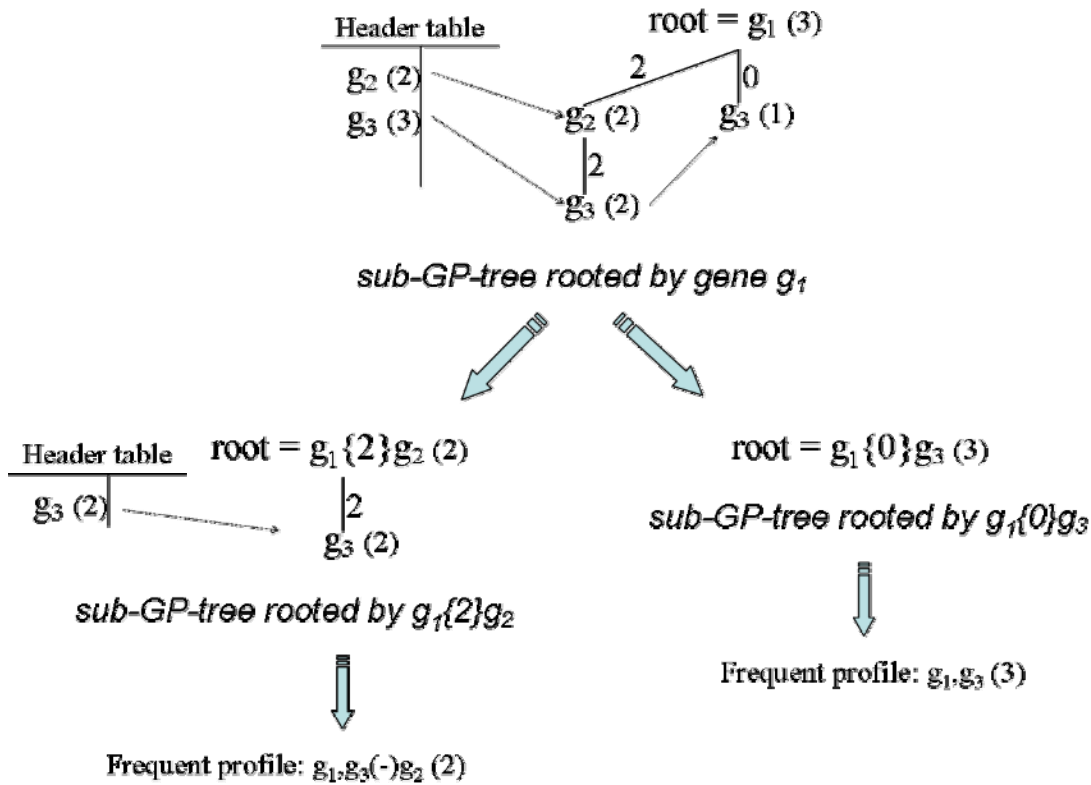
Step 1: Constructing a GP-tree. A GP-tree is constructed by reading the expression data condition by condition and mapping each condition onto a path in the GP-tree. A path compression occurs when two or more conditions have the same gene profile starting from the first gene in the tree. More overlapped paths result in a more compressed dataset and a smaller tree. As a consequence, the mining algorithm needs less time to extract the frequent co-regulated profiles from the GP-tree structure.

The next figure illustrates how the GP-tree is constructed and updated for the first, second and last conditions from the sample data given above.

Note that genes are followed by their supports in both the header table and the tree. The parent distances (or gene distances) are shown on the edges between the child and its parent. For more detail, see Gyenesei et al., 2006.



Step 2: Mining Co-Regulated Gene Profiles Using the GP-tree. The developed mining algorithm generates co-regulated gene profiles from the constructed GP-tree by exploring the tree in a top-down and recursive manner. It splits the problem into sub-problems by decomposing the GP-tree into disjoint sub-GP-trees, and then calls the recursion again with the sub-trees. If the constructed sub-GP-tree has only a single branch then all co-regulated gene profiles are enumerated directly from the single branch.



II.) Yeast80 and Compendium Data Sets Analysis

Supplementary data for Table 2:

The capability of our MAP implementation for various support thresholds (yeast80):

YEAST80				COMPENDIUM			
Condition support threshold	Gene support threshold	Running time (in sec.)	Number of patterns	Condition support threshold	Gene support threshold	Running time (in sec.)	Number of patterns
10	10	3	340	10	10	6	73831
	9	3	608		9	7	82433
	8	3	1097		8	8	91246
	7	4	1998		7	8	99958
	6	4	3434		6	9	108362
	5	5	5776		5	9	116223
	4	5	9419		4	13	123686
	3	6	14615		3	15	129753
	2	6	19751		2	17	133228

Supplementary Table for Chapter 3.1, Parameter settings for MAP and the different biclustering methods:

For MAP and Bimax both, up and down regulated genes were considered. For ISA, OPSM and CC the parameter values recommended by the authors were used. For an explanation of the different parameters, see the original literature.

Algorithm	parameter settings
MAP	Discretization threshold = 1 (-1), min cluster size = 10x10
Bimax	Discretization threshold = 1 (-1), min cluster size = 10x10
ISA	Threshold for genes = 2.0, threshold for conditions = 2.0
OPSM	Number of passed models for each iteration = 10
CC	$\delta = 0.5$, $\alpha = 1.2$, number of bicluster: 50

Supplementary Results for Chapter 3.1

- This Yeast80 data set has also been mined using the bi-clustering methods Bimax, CC, ISA and OPSM, as implemented in the BiCAT software (see section 2.7). The result files are collected in a folder “Yeast80” and stored in a zipped folder, which can be downloaded from:
<http://fgcz-intranet.uzh.ch/publish/Publications/GyeneseiEtAl2007/MAPProfilesAndBiclusters.zip>. The respective result files (e.g. “MAPoutput.YEAST80.txt” etc.) for each algorithm as given by BiCAT consist of three rows per cluster: The first row corresponds to the number of genes and number of conditions, resp. The second row comprises the genes in the cluster and the third row the conditions. For each cluster, a graphical output was produced as a .png file. The naming of the file is composed of the number of genes, the number of conditions and the row number in the result file, e.g. 55_10_31.txt

III.) Cluster Data

The results produced by MAP can be presented as a data matrix of patterns versus genes. The hierarchical clustering of patterns and of genes results based on the metric described in section 2.6 can be visualized as in Fig.6. The MAP software produced files that can be used by the software MapleTree (<http://rana.stanford.edu>) for display. The respective output files from MAP for the Yeast80 data set can be downloaded from here:

<http://fgcz-intranet.uzh.ch/publish/Publications/GyeneseiEtAl2007/ClusterData.zip>

IV.) Gene Ontology Results

For each MAP profile and for each bi-cluster of all bi-clustering methods, a Gene Ontology (GO) analysis was carried out using the ErmineJ software. For each single analysis, a file was created. The files are store in a zipped folder and can be downloaded from here:

<http://fgcz-intranet.uzh.ch/publish/Publications/GyeneseiEtAl2007/GOAnalysisYeast80.zip>

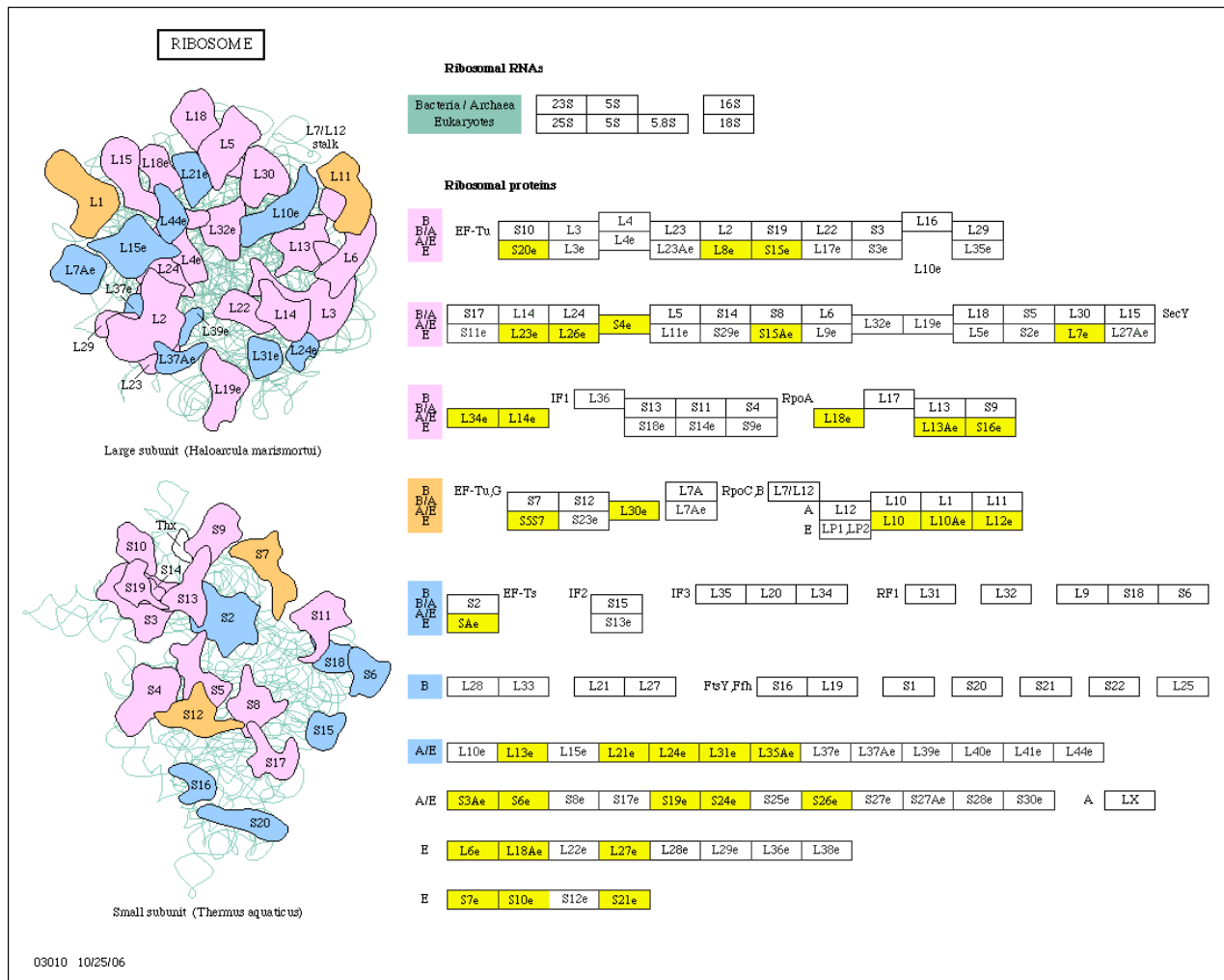
The naming of the GO result files is based on the naming of the respective bicluster or profile as indicated above in the section “Yeast80 and Compendium Data Set Analysis”, in the subsection: “Supplementary Results for Chapter 3.1”

Example: GO_55_10_31.txt means a bi-cluster or profile with 55 genes and 10 conditions. In the result file, the entry for this cluster starts at line 31

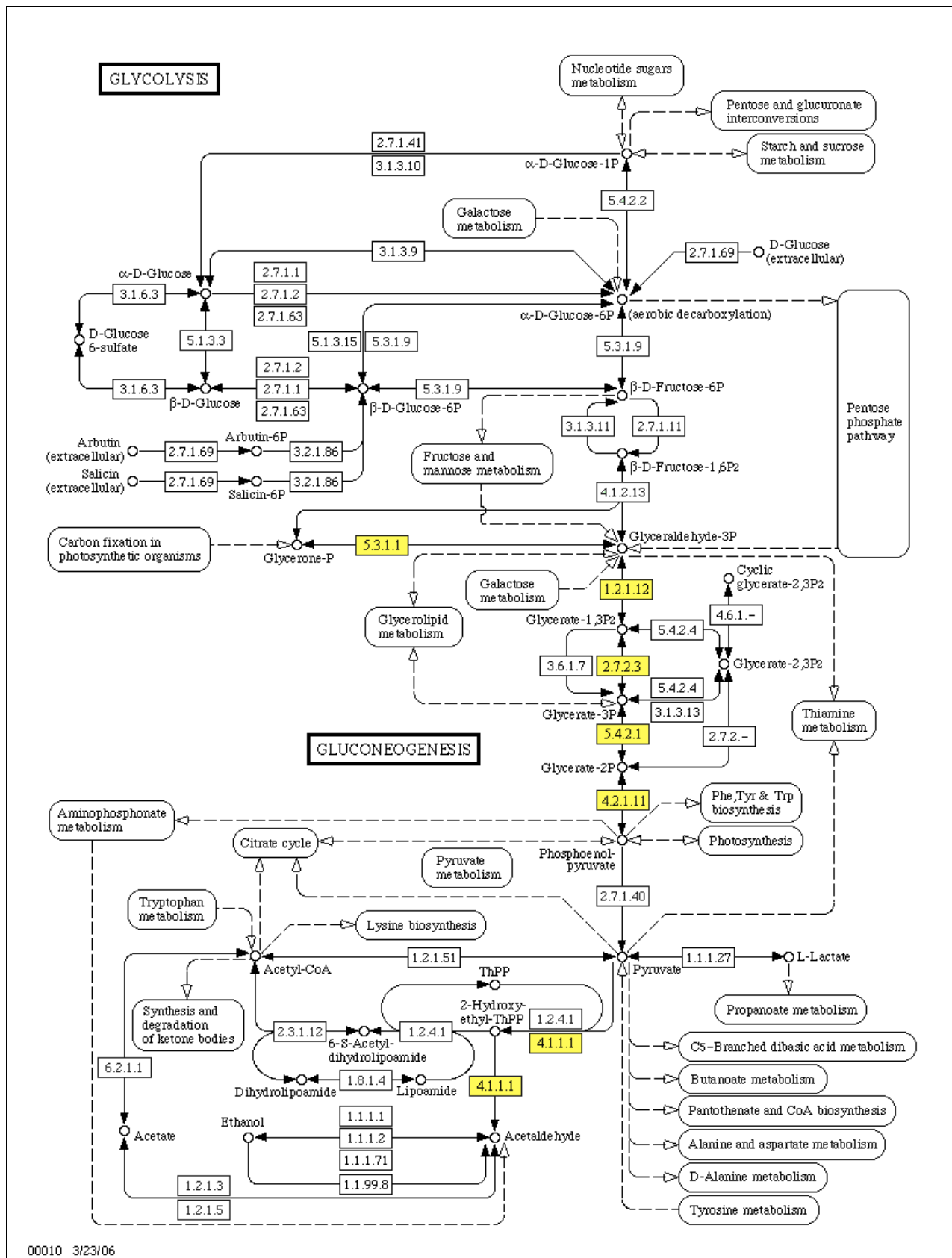
V.) Pathway Analysis

Two profiles found by MAP in the Yeast80 data were biologically especially interesting and were mapped onto two different KEGG pathways. The genes found in the respective profiles and pathways are coloured yellow.

Profile 55_10_31: Ribosomal Proteins



Profile: 11_10_232: Glycolysis-Pathway



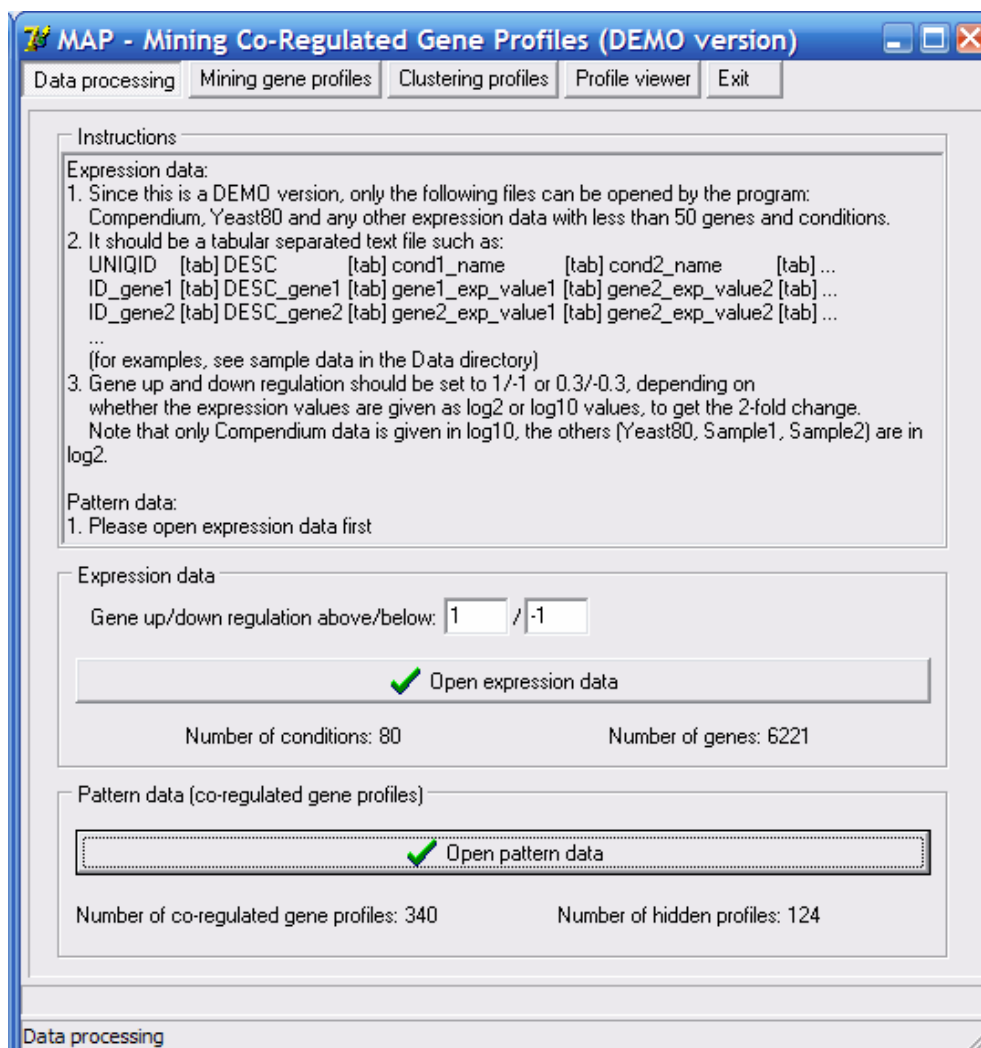
VI.) Software usage

A demo version of the software can be downloaded from:

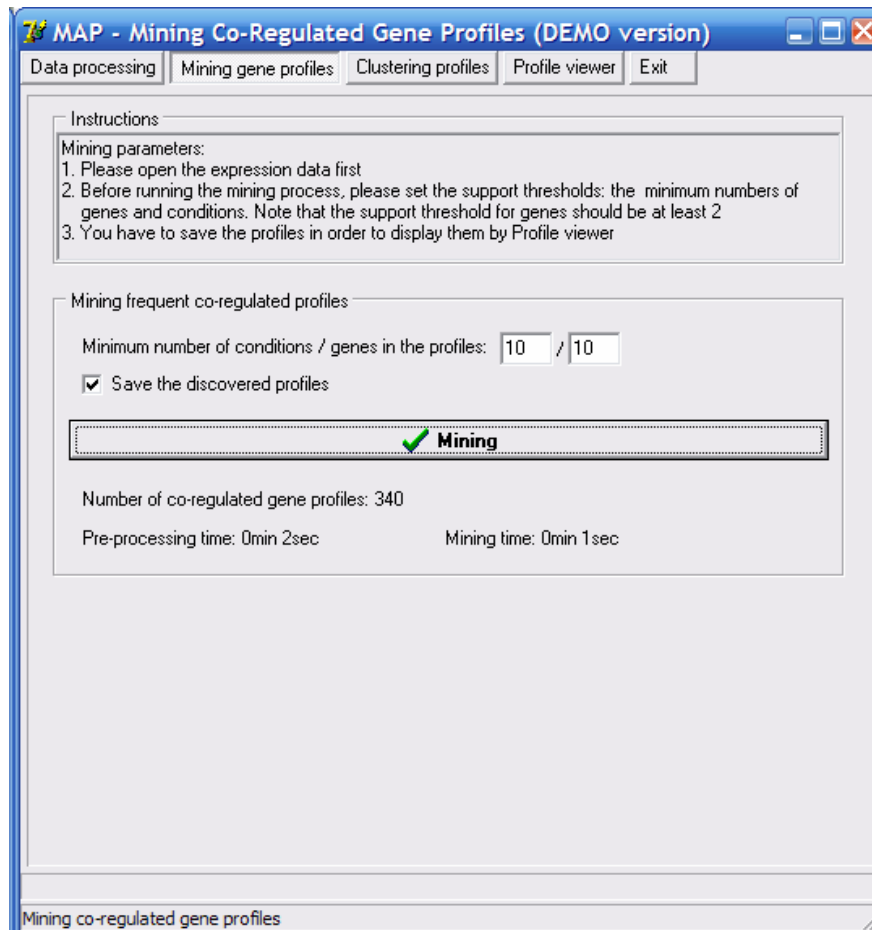
<http://fgcz-intranet.uzh.ch/publish/Publications/GyenesiEtAl2007/ MAPsoftware.zip>

In the following a short manual is added on how to use the software. The respective files can be downloaded from the supplementary data web-site (<http://www.fgcz.ch/publications/map>):

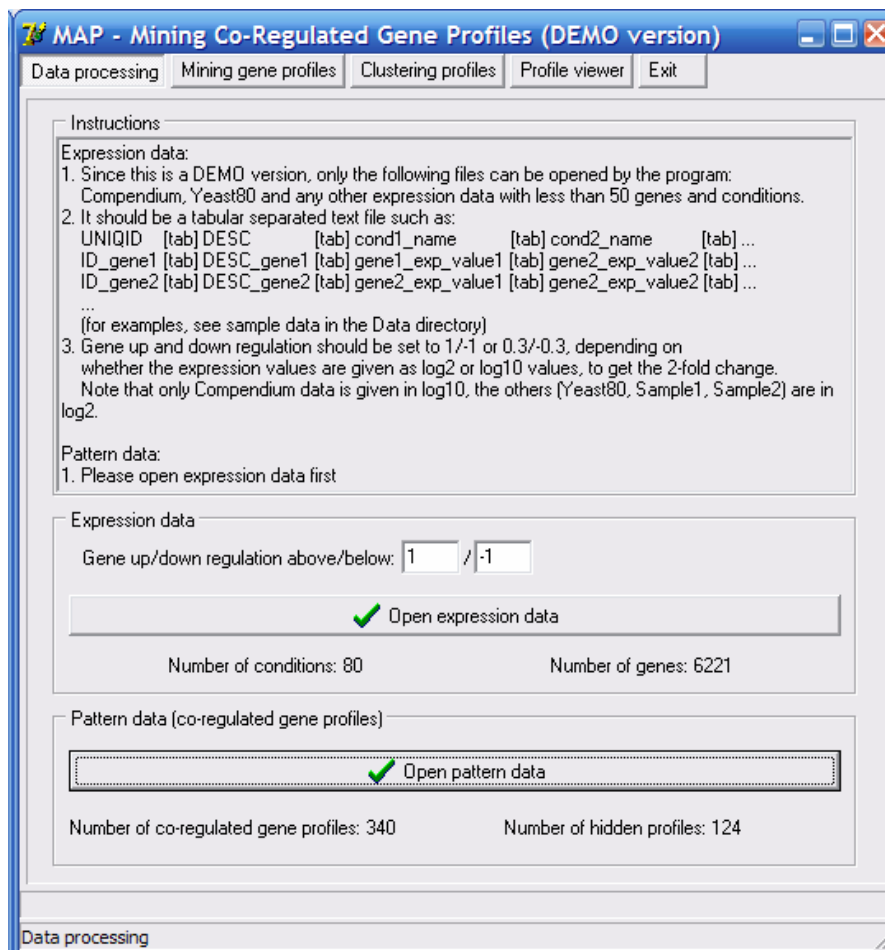
1.) Go to the tab “Data processing”, click on “Open expression data” and upload a gene expression data matrix like “Yeast80.txt”. Give an upper and a lower threshold for regulation.



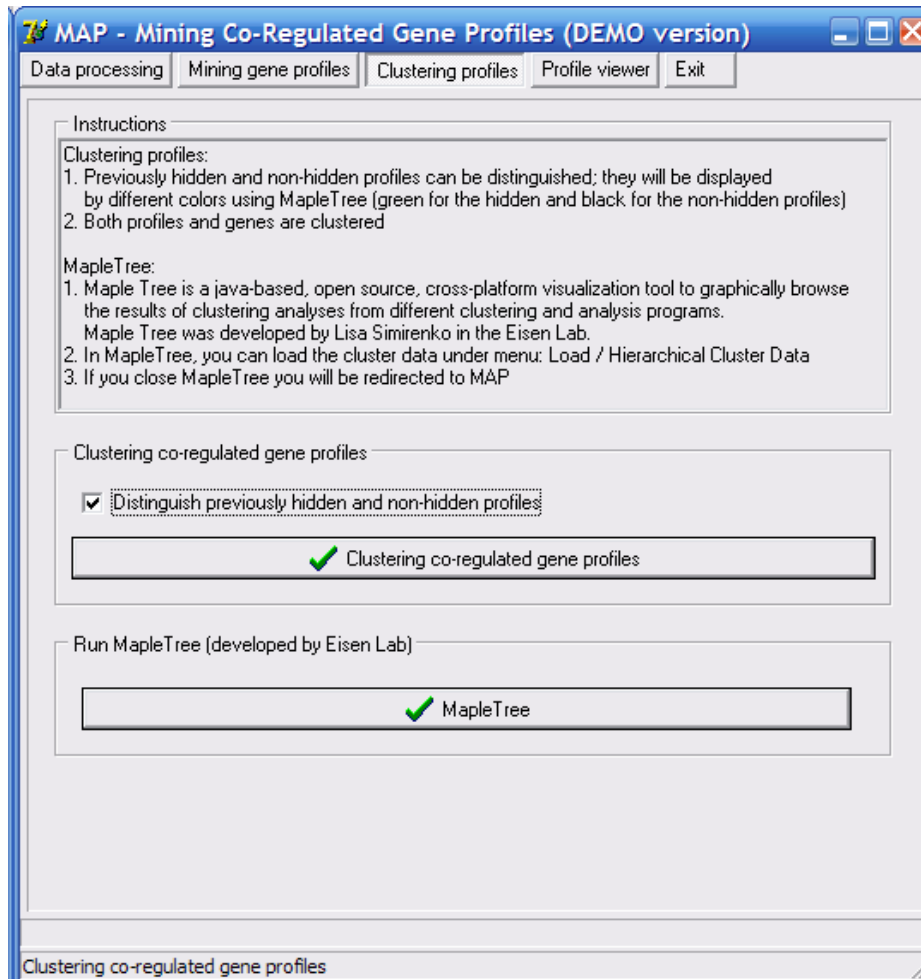
2.) Go to the tab “mining gene profiles” and generate a .dat file containing co-regulated profiles. Give the minimum size of the profiles in terms of number of conditions and number of genes.



3.) Go back to the tab “Data processing and open the pattern data produced in 2.) or download the respective file from our supplementary data web site (“ProfileData.zip”)



4.) Cluster the profiles and the respective gene in the tab “Clustering profiles”. Start MapleTree to view the resulting clusters



5.) View the profiles resulting from 3.) in the tab “Profile view”. You can choose the profile based on the number

