FROMP-v1.0 User's Guide

Dhwani K Desai, Harald Schunck, Johannes Loeser, Julie LaRoche

February 12, 2015

Contents

1 About Fromp **2**

1.1 What FROMP is NOT! . . . 3

2 Background on hidden Markov model (HMM) profiles **4**

2.1 Building EC profiles . . . 4

2.2 PFAM protein family profile HMMs . . . 6

2.3 Using profile HMMs to annotate sequences . . . 6

3 Installation Instructions **7**

3.1 For Windows . . . 7

3.2 For Linux . . . 7

4 Using FROMP **8**

4.1 GUI FROMPing . . . 8

4.1.1 Input formats . . . 8

4.1.2 Output . . . 10

4.1.3 Start-up screen . . . 10

4.1.4 File Menu . . . 10

4.1.5 Project Menu . . . 11

4.1.5.1 Edit Samples . . . 11

4.1.5.2 Select Pathways . . . 13

4.1.5.3 Search . . . 13

4.1.6 Analyse Menu . . . 14

4.1.6.1 Pathway Completeness Score . . . 14

4.1.6.2 Pathway Activity . . . 17

4.1.6.3 EC Activity . . . 19

4.1.7 Create your own pathways using Pathway Designer . . . 20

4.2 Command line FROMPing . . . 23

Chapter 1

About Fromp

FROMP is an acronym for Fragment Recruitment On Metabolic Pathways. It is Java program which was designed for the following objectives:

- To map Enzyme Commission (EC) number and Protein Family (Pfam) assignments of meta-omic (metagenomic and metatranscriptomic) sequences to KEGG reference metabolic pathways or custom-designed using weights for ECs and a Pathway Completeness score, Pathway Activity Score or an odds-ratio for gene enrichment
- To display the KEGG reference pathways showing proportionate contributions from each meta-omic sample to each EC in the pathways
- To export metabolic profiles for the samples in a project in the form of matrices of either Pathway Completeness or Pathway Activity scores or Odds-ratios (for individual ECs) as text files.

Motivation: The sheer scale of the meta-omic datasets that are now available warrants the development of automated protocols for organising, annotating and comparing the samples in terms of their metabolic profiles. We describe a user-friendly java program FROMP (Fragment Recruitment on Metabolic Pathways) for mapping and visualising enzyme annotations onto the Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways or custom-designed pathways and comparing the samples in terms of either their Pathway Completeness Scores or their relative Activity Scores. This program along with our fully-configurable PERL based annotation organisation pipeline Meta2Pro (METAbolic PROfiling of META-omic data) offers a quick and accurate standalone solution for metabolic profiling of environmental samples. Apart from pictorial comparisons, FROMP can also generate score matrices for multiple meta-omics samples which can be used directly by other statistical programs.

Availability: The source code and documentation for FROMP can be downloaded from `https://sites.google.com/site/dhwanidesai/home/software` along with the Meta2Pro collection of PERL scripts.

Contact: Dhwani.Desai@Dal.Ca, jlo@informatik.uni-kiel.de, hschunck@geomar.de

1.1 What FROMP is NOT!

- FROMP is NOT a "black box" solution for a complete analysis of metagenomes or metatranscriptomes which can work on raw sequences: It is meant to be used as an analytical tool for comparing multiple meta-omic samples in terms of their PFAM and ModEnzA EC annotations.
- FROMP is NOT a statistical comparison tool: The matrices output by FROMP could be used as input for other statistical comparison program like STAMP, but there is no provision for statistical differences between groups of samples in FROMP.

Chapter 2

Background on hidden Markov model (HMM) profiles

The next three sections describe the construction and usage of profile hidden Markov models (HMMs) ([1], [2]) of the PFAM protein families ([3], also see 2.2 for details) and the ModEnzA EC numbers ([4]). Users familiar with these concepts can skip ahead to Installation of the program 3 and the Input formats that FROMP is able to read (See 4.1.1).

2.1 Building EC profiles

The EC number profiles can be generated using the program ModEnzA collection of scripts ([4]) which is available from –check `http://sites.google.com/site/dhwanidesai` check –. The following scripts have been tested with version 2.3.2 of the Hmmer package. The Hmmer 2.3.2 can be downloaded from `http://hmmer.janelia.org/software/archive`. Since 2010, Hmmer 2.3.2 (or just Hmmer 2.0) has been replaced by Hmmer 3.0. All the Pfam protein family HMMs have also, since been migrated to the newer format. ModEnzA is based on the HMM-ModE protocol [5]. The latest version of HMM-Mode using the HMMer 3.0 package can be dowloaded from `http://www.jnu.ac.in/Faculty/andrew/`.

Briefly, the ModEnzA protocol for constructing profiles can be described as follows:

1. Get a list of protein sequence IDs associated with each EC number from the Enzyme database. Retrieve protein sequences for each Ec number obtained above from the SwissProt database. The script `1_prepare-enzyme-files.pl` can be used to accomplish this step. The inputs for this script include the name of the Enzyme.dat file (which contains the Enzyme database in the flatfile format) and the SwissProt.fasta. The latest version of the Swissprot (now known as

Uniprot) database in the FASTA can be obtained from `http://www.uniprot.org/downloads` while the latest version of the enzyme.dat file can be downloaded from `ftp://ftp.expasy.org/databases/enzyme/`. The description of usage options for this script can be obtained by running the `1_prepare-enzyme-files.pl` script on the command prompt without any arguments as follows:

```
perl 1_prepare-enzyme-files.pl
press enter
```

2. Cluster the protein sequences retrieved for each EC number using either the Markov Clustering Algorithm (MCL, see `http://micans.org/mcl/`) or the CDHIT program (`http://weizhong-lab.ucsd.edu/cd-hit/`). You can use the script `2_cluster-enzyme-sequences.pl` for this purpose. The inputs for this script include a text file containing a list of the sequence files, e.g the files generated in the previous step and a choice of the clustering method (either Mcl or Cdhit). The list file can be generated as follows:

   ```
   ls -1 *.seq > listfile.txt
   ```

   The description of usage options for this script can be obtained by running the `2_cluster-enzyme-sequences.pl` script on the command prompt without any arguments as follows:

   ```
   perl 2_cluster-enzyme-sequences.pl
   press enter
   ```

3. Write out the clusters for each EC number as separate files. Use the script `3_separate-nonSinglet-clusters.pl`. Inputs to this script are again the text file containing a list of the sequence files, e.g the files generated in the previous step and a choice of the clustering method (either Mcl or Cdhit). Clustering with MCl has been extensively tested for accuracy ([4]), and hence is the recommended method for clustering. For each EC this script writes out the non-Singlet clusters (clusters having 3 or more sequences) to separate files. The clustering of the training sequences for an EC might result in multiple clusters corresponding to different sub-units of the enzyme or due to convergently evolved sequences which are not similar. In all these cases, the script will construct separate profiles for each of the clusters for a given EC. For example, MCL clusters the swissprot sequences for the DNA-directed RNA polymerase (EC 2.7.7.6) into 14 subgroups with 3 or more sequences. So EC 2.7.7.6 has 14 different profiles, each stored with a different name such as 2.7.7.6_1, 2.7.7.6_2,...upto 2.7.7.6_14. The files will be used in the next step to generate separate HMM-ModE profiles. For more information on options for this script, on the command prompt, type:

```
perl 3_separate-nonSinglet-clusters.pl
press enter
```

4. Generate HMM-ModE profiles for the EC numbers This step is performed by using the `4_hmmmode.pl` script. This also has the same inputs as the earlier script, namely, text file containing a list of the sequence files, e.g the files generated in the previous step and a choice of the clustering method (either Mcl or Cdhit). But additionally, you also have to provide a file containing the negative training sequences, a file name to store the optimized threshold scores for each EC and the paths to 2 directories which will store the training sequences an the test sequence samples for each EC for the cross-validation exercise. For more information:

```
perl 4_hmmmode.pl
press enter
```

2.2 PFAM protein family profile HMMs

The Pfam database ([3]) organizes protein families in the form of multiple protein sequence alignments and profile HMMs. There are two kinds of Pfam collections- Pfam-A and Pfam-B. Pfam-A HMMs represent high quality, manually curated families. Although these Pfam-A entries cover a large proportion of the sequences in the underlying sequence database, in order to give a more comprehensive coverage of known proteins, Pfam-A is supplemented by automatically generated entries are called Pfam-B. Although of lower quality, Pfam-B families can be useful for identifying functionally conserved regions when no Pfam-A entries are found. The latest Pfam-A HMMs can be downloaded from

2.3 Using profile HMMs to annotate sequences

Chapter 3

Installation Instructions

FROMP is Java program and hence requires an installed and working Java version. You can get the latest java version from `http://java.com/en/`.

3.1 For Windows

Unzip the .zip file to extract the Fromp-v1.0 folder and double click on the "FROMP.jar" file. YUP! It is that simple!

3.2 For Linux

The default java environment in Linux is the OpenJDK. However, if you are not sure which java environment came with your version of linux, you should get the latest Sun Java JDK and the Java Runtime Environment (JRE) and then congfigure your Linux so that it uses the Sun java instead of the default. You can find tips on how to do that for Fedora, RHEL and CentOS at `http://www.freetechie.com/blog/installing-sun-java-on-fedora-12/` and for Ubuntu at `https://help.ubuntu.com/community/Java`, although in our own experience the latest Open java also works without a problem.

Unzip the .zip file to extract the Fromp-v1.0 folder. Change directory to this folder and type:

```
java -Xms128m -Xmx256m -jar FROMP.jar start
```

This fixes the heapspace (or memory required to run FROMP). In case you get an "java.lang.OutOfMemoryError: Java heap space" error while running huge samples, you can try changing the -Xmx option to -Xmx512m depending upon the RAM capacity of your machine.

Chapter 4

Using FROMP

FROMP can be used in two modes - the Graphical Users Interface (GUI) as well as from the command line (Cl).

4.1 GUI FROMPing

4.1.1 Input formats

FROMP can read EC numbers and PFAM accession numbers from a variety of formats. PFAM names cannot be read in at present. The data can be prepared in any of the following formats:

1. The output of the *hmmscan* program from the HMMER package can be read as is. For details on how to construct the ModEnzA HMM profiles and to run the *hmmscan* refer to 2.1, 2.2 and 2.3.

   Example: (edited)

   ```
   Bac_DnaA          PF00308.11 GDDSEYSO2DHKTX_1_2   -         4.1e-19   68.7.....
   Relaxase          PF03432.7  GDDSEYSO2DOBSZ_5_1   -           5e-06   25.7.....
   HisKA_3           PF07730.6  GDDSEYSO2DBODZ_5_0   -         1.6e-07   31.2.....
   4.1.1.61_1        -          GDDSEYSO2DTH9N_1       -         1.3e-15   53.5.....
   3.1.26.4_1      -          GDDSEYSO2DXC98_2       -         4.2e-10   34.9.....
   ```

   Note: The EC numbers or Pfams can be repeated multiple times in the file as there can multiple copies of the same enzyme in the sample. All repetitions of an enzyme are counted as copies of

the enzyme by FROMP. There could, however, also be multiple enzymes associated with the same sequence ID, i.e a single sequence could have hits to more than one profile. In this case, the profile with the maximum score on that sequence is generally chosen. So for accurate interpretations of results, the input files need to be processed to remove multiple profile hits to the same sequence. You can use the script ..... available in the package for doing this task.

2. FROMP can also read in a simple list of EC numbers and Pfam accession numbers (one-column)

3. A tab or comma separated list of ECs and Pfams with counts (two-column)

   Example:

   ```
   1.1.1.1,2
   1.2.11.5,0
   PF00308,9
   PF03432,7
   PF07730,0
   ```

4. EC numbers and Pfams accession ID with counts and sequence Ids (three-column)

   Example:

   ```
   PF00308,1,GDDSEYS02DHKTX_1_2
   PF03432,1,GDDSEYS02DOBSZ_5_1
   PF07730,1,GDDSEYS02DBODZ_5_0
   4.2.1.24,1,BisonMetagenome_READ_00093361_1
   4.1.3.27,1,BisonMetagenome_READ_00086932_5
   4.1.1.61,1,BisonMetagenome_READ_00087042_4
   ```

5. It can also read in a comma or tab separated matrix with the EC and Pfam counts for multiple samples. The EC numbers or Pfam accessions should be in rows and the sample counts in columns

   Example:

   ```
   1.2.1.1,2,15,0
   1.1.1.1,0,1,0
   ```

```
i.e EC 1.2.1.1 has a count of 2 in sample 1, 15 in sample 2
    and 0 in sample 3 and so on.
```

Some example input files are provided in the `Real-Samples` folder. These include 10 metagenomes from chimneys and hydrothermal vents (located in `/Real-Samples/chimney-data`) and the transcriptomes of *T. oceanica* obtained under iron limitation and iron sufficiency ([6]) (located in `/Real-Samples/T-oceanica`). These files have been prepared by concatenating the output of the program *hmmscan* run with the PFAM and the ModEnzA EC profiles on the amino acid sequences translated from the metagenome or metatranscriptome sequences.

4.1.2 Output

The comparative recruitment of various samples on the reference pathways can be exported as PNG files. The various score matrices (including the EC count matrix) for the samples and the sequence IDs of the fragments mapping onto each EC or pathway can also be exported as text files.

All data in FROMP is organised in the form of projects. A project can have any number of samples added to it. All projects are saved as .frp files which can be exchanged easily between different computers and users.

4.1.3 Start-up screen

The FROMP start-up screen is shown in Figure 4.1. The user has the option to create a new project or open an existing one (from the `Project Options` list). The same options can also be accessed throught the File Menu (4.1.4). Additionally, the start-up screen also provides a list of the most recently created projects for quick access through `Recent Projects`.

4.1.4 File Menu

The File Menu can be used to create a new project, open an existing project or save an open project.

New Project

```
File->New Project->Enter Project Name->Save
```

Open Project To open an exsiting project (.frp file)

```
File->Open Project->Browse files on computer
```

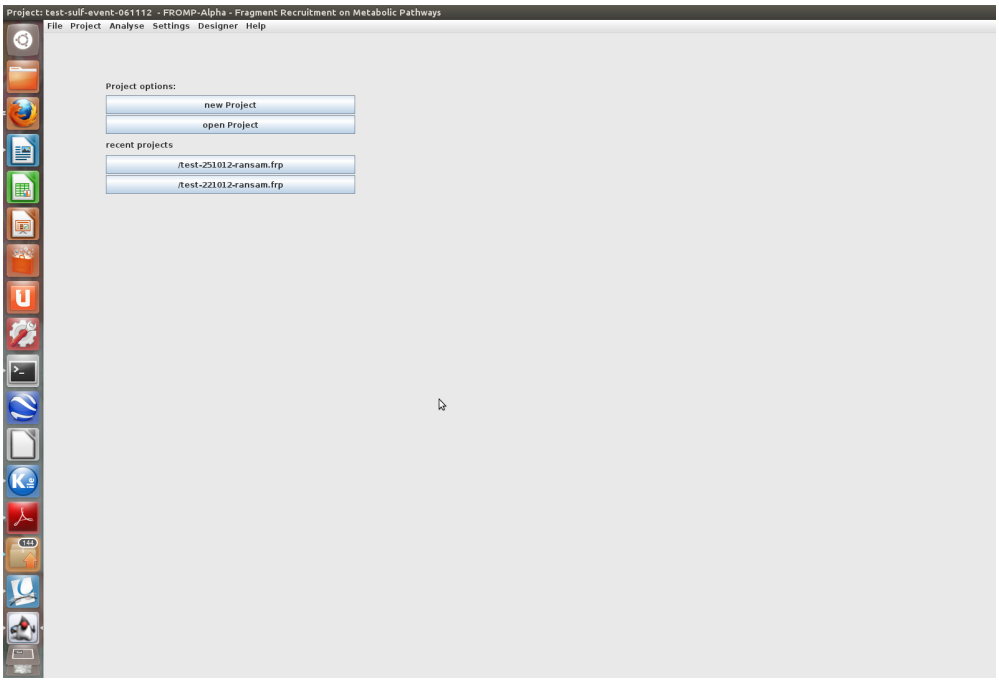

Figure 4.1: The FROMP start-up screen

Save projects

`File->Save Project/Save Project As`

4.1.5 Project Menu

Here, you can manage samples in a given project (add/remove them to a project, change colors for each sample etc).

4.1.5.1 Edit Samples

This page helps you to add/remove samples in project and adjust their colors. In additon to all the action buttons on the page, there is a button (`Back to Project menu`) that takes you back to the porject lists and another one (`Go to Pathway Selection`) that will take you to the next step in analysis. The following actions are possible throught this page:

Add samples Click on the "Select sample" button (Figure 4.2), browse for the sample files (Check the "Input formats" 4.1.1).

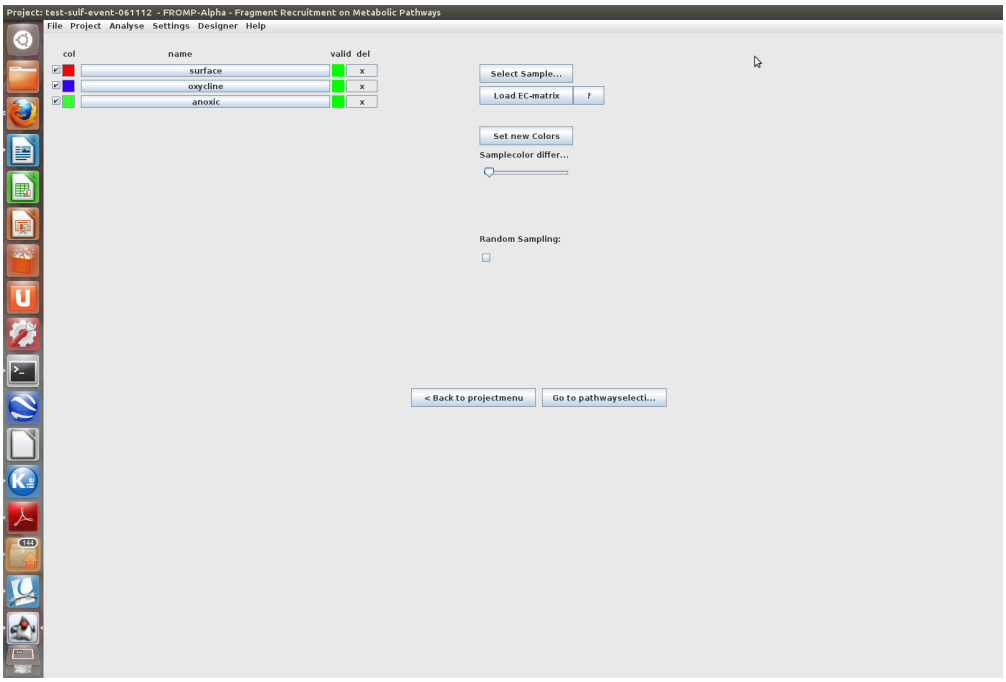

Figure 4.2: The "Edit Samples" window in FROMP

Add a matrix file of EC counts for multiple samples A matrix file containing EC counts, with the samples arranged in columns and the EC numbers in rows, can be uploaded and added to a project by clicking the `Load EC-matrix` button on the `Edit Samples` page. For details on this input format check section 4.1.1.

Edit sample names Click on the text boxes containing the sample names to edit the sample names (Figure 4.2). These names will be displayed in all results.

Set colors for samples By default FROMP automatically associates each sample with a varying shade of red

```
255,0,0 => 200,0,0 => 150,0,0
```

etc for each successive sample added. You can increase the difference by changing the "Sample Color difference" meter and then clicking "Set New Colors". Alternatively, you can click on the color square next to the sample name text box and choose color for each sample individually (Figure 4.2).

Equalize sample sizes Clicking the checkbox for "Random Sampling" equalizes the sample sizes. FROMP fixes the size of the smallest sample as the sample size and randomly samples equal numbers

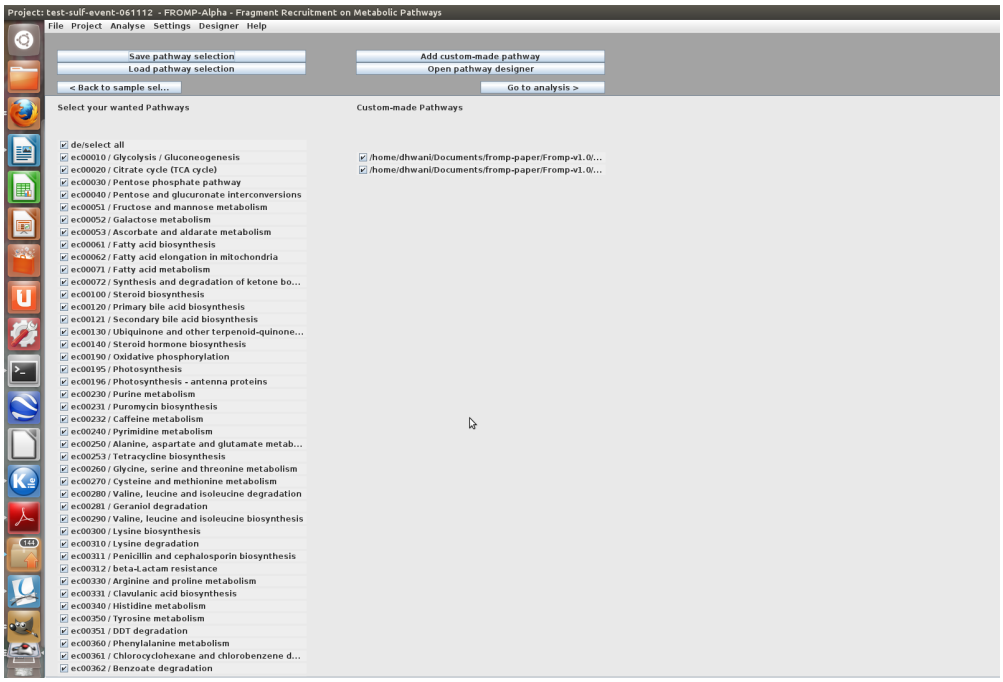

Figure 4.3: The `Select Pathways` page

of hits from the other samples.

4.1.5.2 Select Pathways

You can select just a subset of interesting pathways that you want to analyse using this option. The selected pathways can be saved as a text file with extension .pcg (`Save pathway selection` button) and a saved selecion can be loaded into the program using the `Load pathway selection` button (Figure 4.3). The user can also add custom designed pathways for recruiting meta-omic sequences. The custom made pathways can be designed using the **Pathway Designer** feature by clicking on the `Open Pathway Designer` button. For details on how to create a custom pathway using the Pathway Designer, jump to section 4.1.7. A set of custom designed pathways can be imported into an existing project using the `Add custom-made pathway` button and then selecting the appropriate .pwy files from your computer. Additionally, there are two buttons to help you navigate to the previous step (`Back to Sample selection`) or to the next step in analysis (`Go to analysis`).

4.1.5.3 Search

The Pathways can searched by using either of Pathway ID (e.g ec00010) or Pathway name (e.g Glycolysis) or EC number (e.g 1.1.1.100). Inputting just a part of the name or EC number also works.

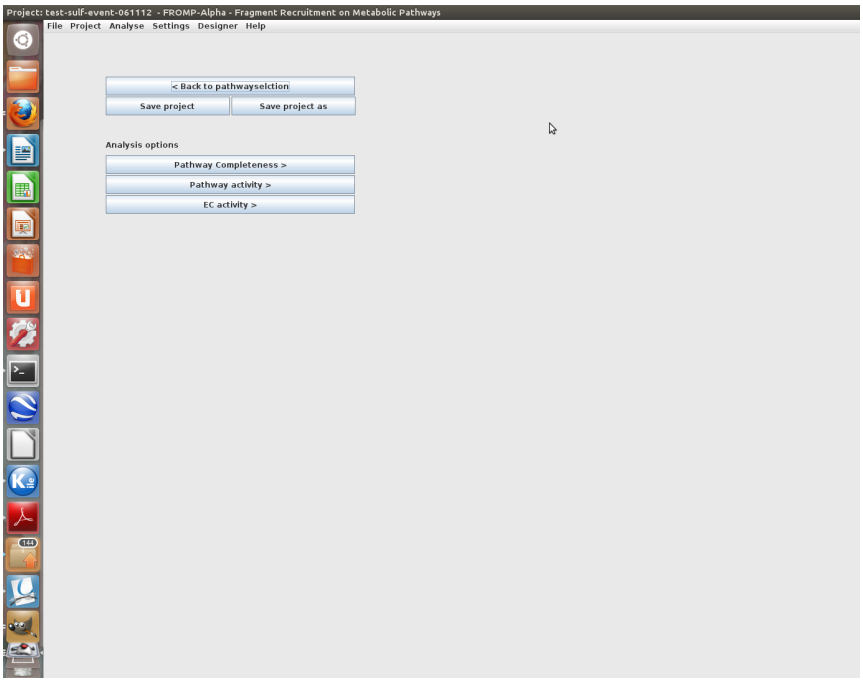

Figure 4.4: The `Analyse` options

4.1.6 Analyse Menu

Once you have successfully created a project, added samples and selected/added pathways, the next step is the actual calculation of Pathway Completeness, pathway activity or EC activity scores. Clciking the `Go to analysis` button on the `Select pathways` page (see 4.1.5.2) opens up the Analysis page (Figure 4.4). The analysis options can also be accessesed by using the `Analyse` menu from the Menu bar. Using this page you can analyse the samples for their Pathway Completeness Scores (what percentage of a given pathway is recoverable in the sample), Pathway Activity Score (the total hits to all ECs in a given pathway) and the EC Activity (total hits to all ECs in the samples). At this point, you can also save the project using the `Save project` or `Save project as` buttons. This will save the entire project along with any custom made pathways that you added in the previous step. The various Analysis options are described in detail in the next sections.

4.1.6.1 Pathway Completeness Score

This refers to the extent to which a pathway can be recovered in a given sample. The purpose of FROMP is to map the EC numbers to Pathways. Some EC numbers participate in multiple pathways and hence are not good indicators of the pathway being actually present in the sample. So we assign weights to ECs based on the number of pathways that they participate in as described earlier (). The

weights for the ECs for a given pathway are then summed up For each EC i the weight

$W_i = \left((N_{(T,i)}/N_{(U,i)})/N_{(P,i)}\right) * \sqrt{L_{(UBC,r)}}$

where N(T,i) is the total number of ECs in all pathways that have EC i, N(U,i) is the number of unique ECs in all the pathways that have EC i and N(P,i) is the total number of pathways where EC i is present and L(UBC,r) is the total edge-length of the unbranched chain containing EC i in the reference pathway. The pathway completeness score for a pathway p is then

$$C_P = \left(\frac{\sum_{i \in EC_p} W_i * I_i * \sqrt{L_{(UBC,s)}}}{\sum_{i \in EC_p} W_i} \right)$$

where Wi is the specificity weight of each EC i in pathway p, and Ii is 1 if the EC number is detected in the sample and L(UBC,s) is the edge-length of the unbranched chain containing EC i in the sample. There are three options for Pathway Completeness score calculation.

1. The score without using the unbranched chain information (default)
2. The chain information is used while calculating the weights, i.e the unbranched chains of the reference pathways are identified and if an EC in such a chain is present in the sample the weight is multiplied by $\sqrt{L_{(UBC,r)}}$. This can be acheived by clicking the `Use chaining mode 1` checkbox in the Pathway Completeness Analysis window (4.5)
3. The chain information used both in the weights as well as the score calculations. (click the `Use chaining mode 2` check box in 4.5

Show Pathway Scores The `Show pathway scores` tab shows the Pathway Completeness Scores for individual samples. The `next Sample` or `prev. Sample` buttons can be used to navigate between the samples.Clicking on a pathway button for any sample opens the mapping display of the ECs in the sample mapped to this pathway along with the number of hits for each EC. This tab also shows and overall Pathway competeness score where all the pathways in all samples are pooled together. Clicking any pathway button in the Overall window displays the mapping of all the samples on the pathway. The EC boxes in the KEGG pathway maps are colored by the sample colors proportionate to the number of hits for the EC in each sample . The number of hits in all samples for each EC are also displayed as bar-charts and number matrices (4.7).

The pathways can be sorted according to the Pathway Completeness Scores by checking the `Sort by score` box. A minimum score limit can be set by entering the cutoff score in the Min.shown score box. For example, entering 30 in this box will display only those pathway buttons where the score is ¿= 30 (Figure 4.5).

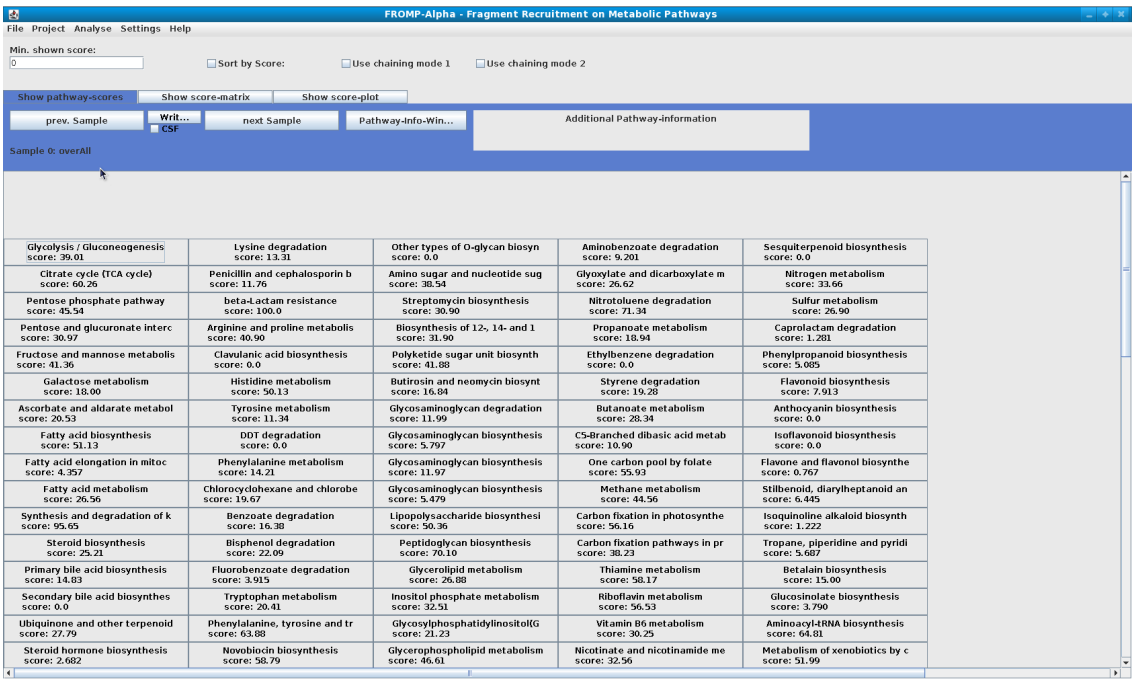

Figure 4.5: The `Pathway Completeness Analysis` window in FROMP

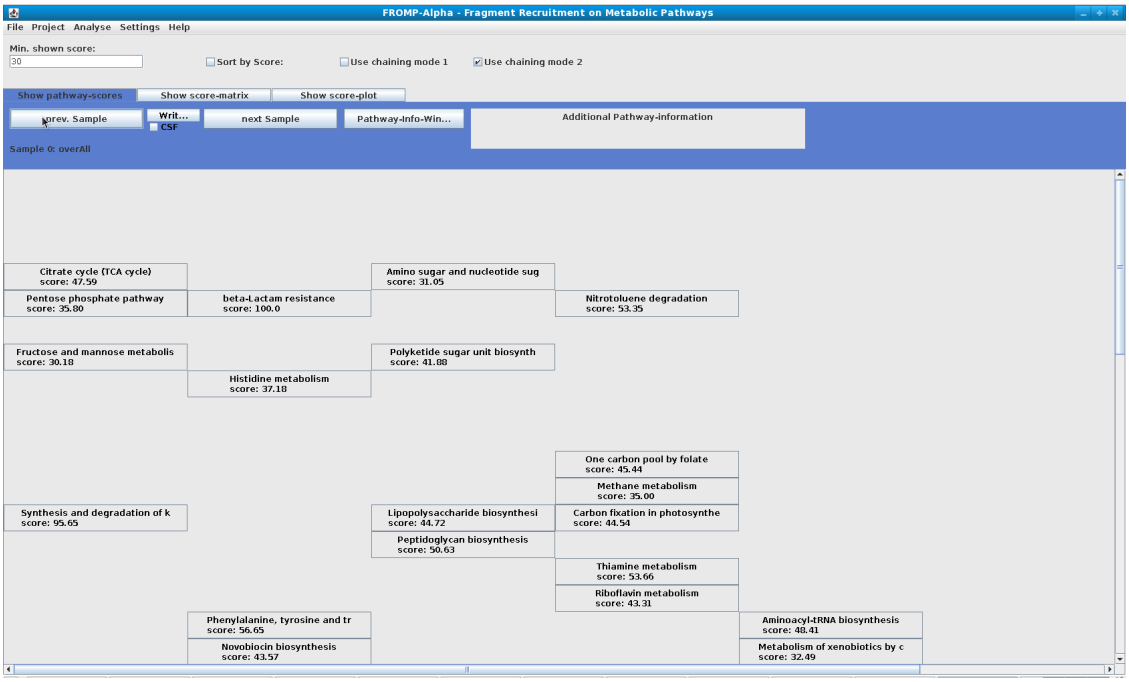

Figure 4.6: The `Minimum score cutoff` feature in "Pathway Completeness Analysis"

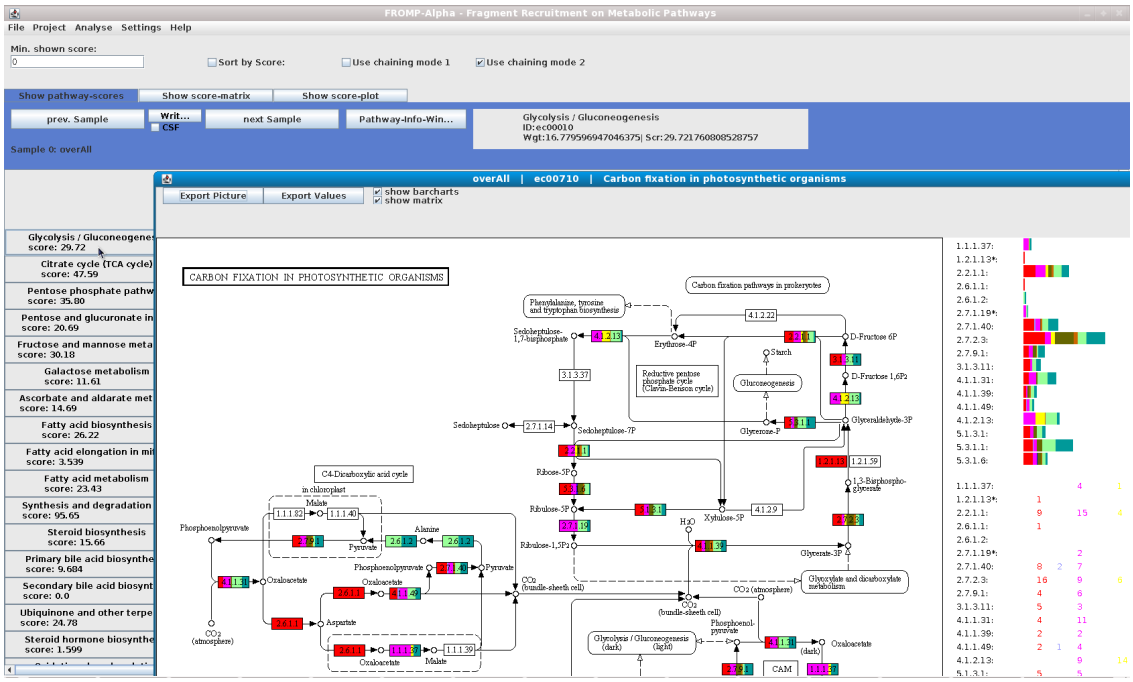

Figure 4.7: Pathway mapping visualization in "Pathway Completeness Analysis"

Show Score Matrix The `Show score matrix` tab shows the pathway completeness scores of all pathways for all samples as a matrix (Figure 4.8). Again, clicking any pathway button in any sample here will open up the pathway map for that sample. Clicking a pathway in the overall column (the first column, black color by default) will open up the mapping of all the samples on that pathway (similar to Figure 4.7).

On both the `Show Pathway Scores` or `Show score matrix` tabs, there is an option (clickning the `Write to file` button) to write out the corresponding matrices as TAB separated (default) or Comma separated (by clicking the CSF check-box) text files.

Show Score Plot The `Show Score plot` tab shows a graphical plot of the Pathway scores. This plot can be enlarged or minimized (`Scale up` or `Scale down` buttons) and exported as a PNG file.

4.1.6.2 Pathway Activity

This is simply the sum of counts for the ECs in a given pathway. There is also an option to multiply this score by the EC weight (click on the `Include weights` check-box) to get a weighted Pathway activity score. The Pathway Activty matrix can be sorted by the total Pathway Activity for each pathway (i.e the sum of the rows) by clicking the `Sort by Linesum` check-box (Figure 4.10).

Figure 4.8: The matrix display in "Pathway Completeness Analysis"

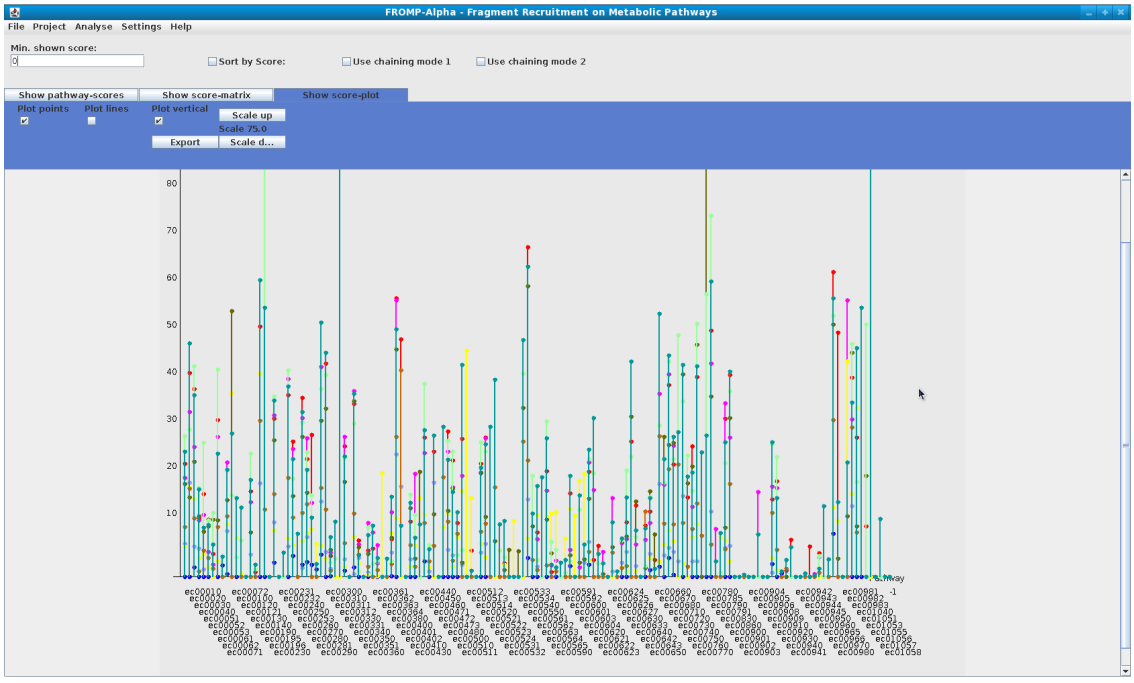

Figure 4.9: The Pathway Completeness Score plot

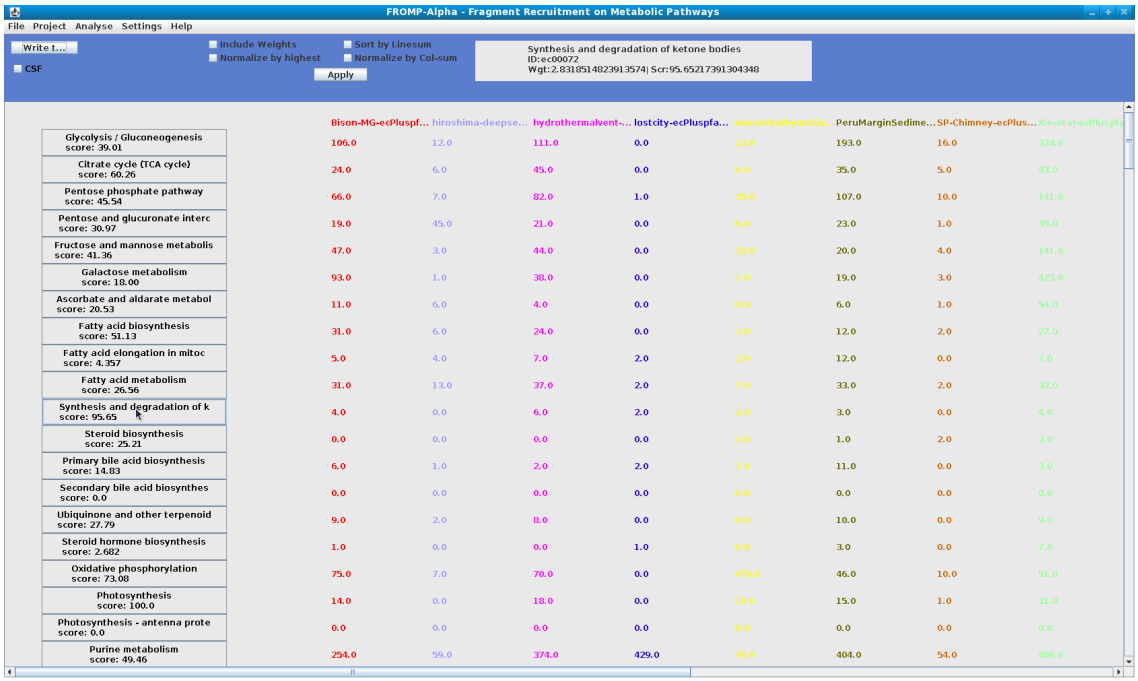

Figure 4.10: The Pathway Activity Analysis

Using the `Normalize by col-sum` or `Normalize by highest` options, the score can be normalized by the Sample total (total for each column) or the Pathway total (row total) to get a percentage contribution of each pathway to the Sample or a percentage contribution of each sample to the overall pathway activity respectively (Figure 4.10.

Like the Pathway Completeness analysis this activity matrix can also be written out as a TAB seperated (default) or Comma separated (by clicking the `Write to file` button.

4.1.6.3 EC Activity

This analysis window displays the hit counts for each EC number in each of the samples.

EC orientated view This is the default view in the EC activity analysis where the hit counts for each EC are displayed for each sample. By default the matrix is sorted by EC number (Figure 4.11). The following features are available here

- The `sort by sum` option sort the entire matrix by the row sum.
- The `Unmapped at end of list` option is always checked by default. This keeps the ECs which cannot be mapped to any pathway at the end of the EC activty matrix. Un-checking this box results in the mapped as well as the unmapped ECs being sorted by EC number.

- `Odds Ratio` calculates the Odds ratio for enrichment of EC numbers in samples as described in (Gill, et al., 2006). If A and C are the occurrence counts of a given EC in sample i and all other comparison samples j, respectively, and B and D are occurrence counts of all other ECs in sample i and comparison samples j respectively, then the odds ratio for the given EC in sample i is $\left(\frac{A/B}{C/D}\right)$.
- The hit counts can be linked to the corresponding sequence IDs of the hits by checking the `Include Repseq IDs` option. Left-clicking on any of the ECs while the `Include Repseq IDs` option is selected will open a `RepSeq` window which displays the sequence IDs for the given EC.
- While the `Include Repseq IDs` option is selected, these sequence IDs can be exported to the `RepSeqIDs` folder in your `FROMP` directory by right clicking the EC in question and selecting the `export` option, or by selecting the `export` option from the drop-down `File` menu in the `RepSeq` window.
- To display the hits to incomplete EC numbers (those ECs where all 4 digits in the EC number are not present) in the EC activity matrix, click on the `Display incomplete ECs` check-box.
- Clicking on a Sample Name (the column heading) will sort the matrix by the hit counts in that sample.

Pathway orientated view In this view the EC hits are arranged according to the pathways (Figure 4.12. Here one can sort the Pathways according to their pathway activity scores (`Sort pathes by sum`) and within each pathway table the EC sub-matrix can be sorted according to the row sum (`Sort ECs by sum`).

The EC matrix in both these tabs can be exported out using the `Write to file` button as described earlier.

4.1.7 Create your own pathways using Pathway Designer

The **Pathway Designer** interface can be opened either from the `Pathway Selection` page (4.1.5.2) or from the `Designer` menu in the Menu bar. Opening the Pathway Designer opens up a grid layout. The spacing of the grid can be controlled by the `+` and `-` buttons on the left. The EC numbers and metabolite names can be adde as nodes on the grid and connected by links. You can also add any kind of information to a node.

The following actions are possible in the Pathway Designer interface:

- **Adding new nodes** Left-clicking anywhere on the grid introduces a new node.

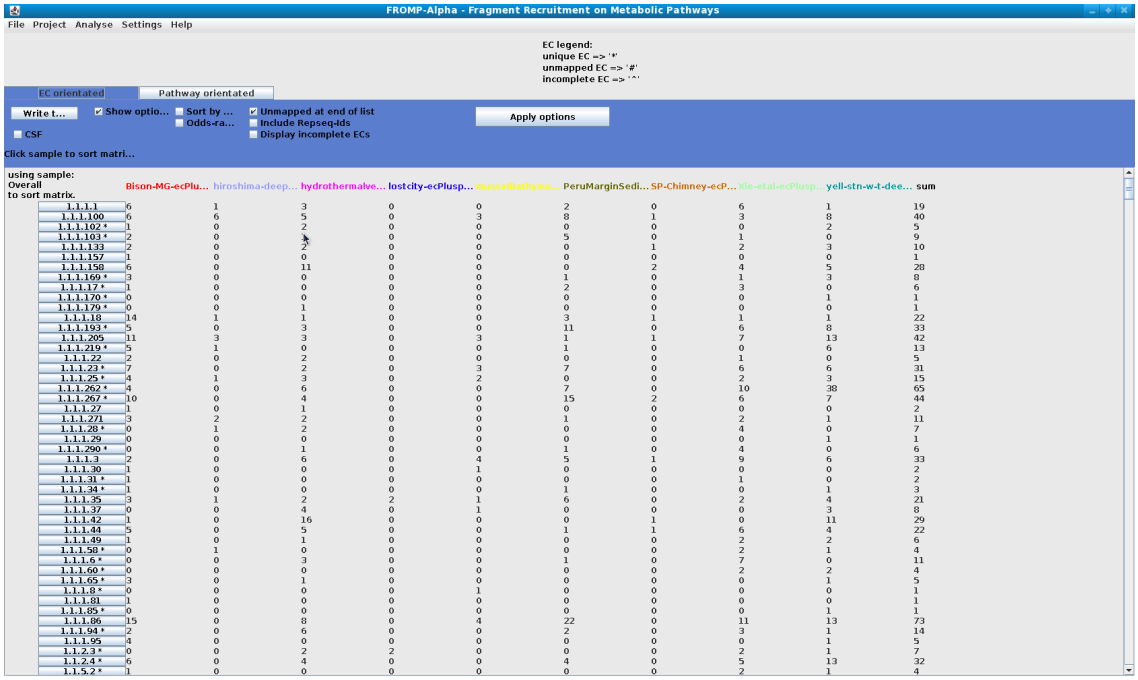

Figure 4.11: The EC orientated tab in EC Activity Analysis

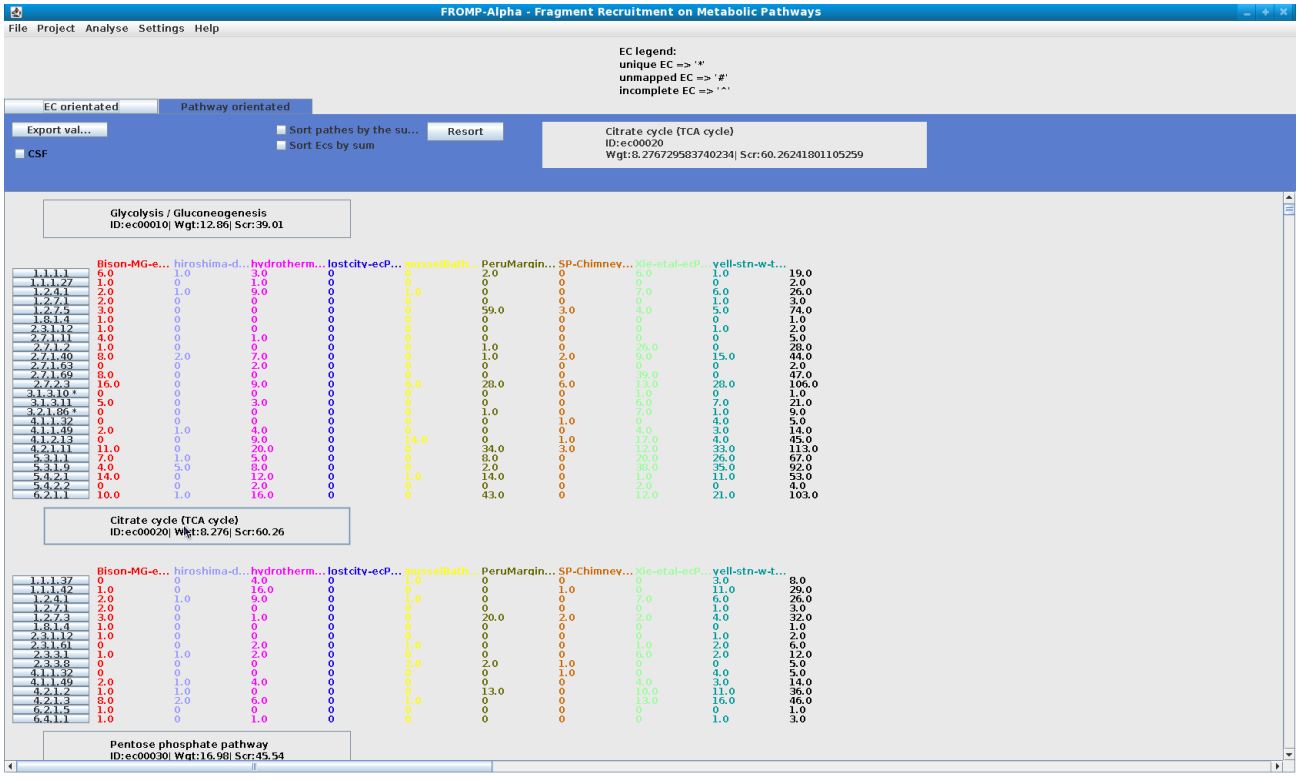

Figure 4.12: The Pathway orientated tab in EC Activity Analysis

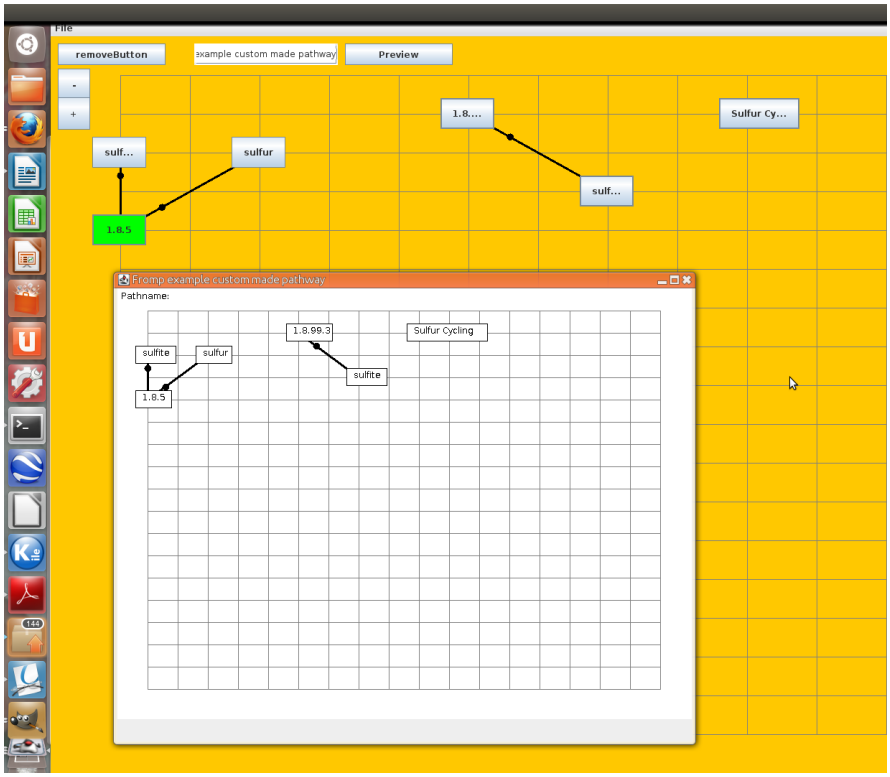

Figure 4.13: The Pathway Designer interface showing a custom made pathway and a preview window showing the final layout

- **Naming/editing a node** The name of a new node can be edited by typing in the node `Name` box.
- **Moving nodes around** Left-clicking on an existing node will highlight the node and change its color to green. Once a node is green, you can move it around on the grid by Left-clicking at the new location. A selected or highlighted node can be un-highlighted by Left-clicking on it again.
- **Adding links between nodes** Left-click to highlight a node and then Left-click on the node that you want to link with it. The link will have a direction from the first node to the second node. A dot which is placed at the 3rd quarter on the link line signifies the direction of the link. For example, in Figure 4.13, on the link between Sulfite and EC 1.8.99.3, a dot is placed on the 3rd quarter of the line towards the EC node. This means that the link is directed from Sulfite to EC 1.8.99.3 or, in other words, Sulfite is the substrate for EC 1.8.99.3.
- **Examine a preview of the layout** Once you have added a few nodes and links, you might want to see the preview of the pathway. Click on the `Preview` button to open up a preview window.
- **Remove nodes** Highlight a node by Left-clicking and then click the `Remove button` button
- **The File Menu** The `File Menu` provides options to save the current pathway or open an existing one for editing etc. The custom designed pathways are saved with a .pwy extension and can be added to any project file.

4.2 Command line FROMPing

The command line version of the program can be used as following:

```
java -jar FROMP.jar 'inputFilePath' 'outputFilePath' options EC
```

The inputFile could be either of the following:

- A fromp-project-file (.frp)
- A normal fromp-sample-textfile (*.out or *.txt etc)
- A list of .out-Files (.lst) example: A textfile 'test.lst' containing the paths to the sample files as follows:

```
-C:\Users\nukota\FrompAlpha\Real-Samples\Xie-etal-ecPluspfam.txt
-C:\Users\nukota\FrompAlpha\Real-Samples\Bison-MG-ecPluspfam.out
-C:\Users\nukota\FrompAlpha\Real-Samples\yell-stn-w-t-deep-ecPluspfam.out
```

- If you have a list of custom defined pathway file (.pwy files), these can also be added to the list file like so:

```
test.lst:
"
C:\Users\nukota\FrompAlpha\Real-Samples\Xie-etal-ecPluspfam.txt
C:\Users\nukota\FrompAlpha\Real-Samples\Bison-MG-ecPluspfam.out
C:\Users\nukota\FrompAlpha\Real-Samples\yell-stn-w-t-deep-ecPluspfam.out
<userP>C:\Users\nukota\userPaths\namepath.pwm
<userP>C:\Users\nukota\Sulfur oxidation.pwm
"
The first 3 file paths are samples and the last 2 are userPathways
```

The outputFilePath should be just a folder where the user wants to put the output-files
Options:

- 'h' print help or summary of available options
- 'p' for all pathway pictures of all samples (e.g if there are 5 samples and 150 pathways, it will generate 150x5 png images)
- 's' for the pathway-score-matrix
- 'm' for the pathway-activity-matrix
- 'e' for the EC-activity-matrix
- 'am' for the pathway-score-matrix, pathway-activity-matrix, and EC-activity-matrix
- 'op' for the png images of all KEGG pathways for the overall sample (i.e all samples together, concatenated into one. These are the multiple comparison pictures)
- 'up' to only print out user designed Pathways (only works with a project-file or a .lst-file)
- 'a' for all of the above simultaneously

Sequence IDs:

- in cmdFROMP you are able to export sequence IDs of a particular EC

- to do so, simply add the EC number (ie. 1.1.1.1) you are looking for to the end of your command and text files contain the sequence IDs for the EC in each of the samples will be generated
- if a particular sample doesn't contain the EC in question, no file will be generated
- **NOTE:** to export sequence IDs one of the above options (excluding the 'help' option) `MUST` be included

Examples:

- Calculate the Pathway completeness score matrix for all samples and store it in a file called out

```
java -jar FROMP.jar c:\Users\nukota\test.lst c:\Users\nukota\out s
```

- Print the help - description of all available options

```
java -jar FROMP.jar h
```

- Print out pngs of meta-omic sequences recuited onto just the custom designed Pathways (only works with a project-file or a .lst-file)

```
java -jar FROMP.jar c:\Users\nukota\test.lst c:\Users\nukota\out up
```

- Build the EC activity matrix and export all sequence IDs for the EC 1.1.1.1 in all samples to the folder `RepSeqIDs` in the `FROMP` directory

```
java -jar FROMP.jar c:\Users\nukota\test.lst c:\Users\nukota\out e 1.1.1.1
```

Bibliography

[1] Durbin R, Eddy S, Krogh A and Mitchison G (1998) in Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge.

[2] Eddy SR (1998) HMMER: Profile hidden Markov models for biological sequence analysis. [ftp://selab.janelia.org/pub/software/hmmer/CURRENT/Userguide.pdf]

[3] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam Protein Families Database. *Nucleic Acids Res.* **(Database 32)**, D138-D141.

[4] Desai DK, Nandi S, Srivastava PK, and Lynn AM (2011) ModEnzA: Accurate Identification of Metabolic Enzymes Using Function Specific Profile HMMs with Optimised Discrimination Threshold and Modified Emission Probabilities. *Adv. Bioinformatics* **vol. 2011, Article ID 743782, 12 pages**, doi:10.1155/2011/743782.

[5] Srivastava PK, Desai DK, Nandi S, and Lynn AM (2007) HMM-ModE improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. *BMC Bioinformatics* **8:104**, doi:10.1186/1471-2105-8-104.

[6] Lommer M, Specht M, Roy AS, Kraemer L, Andreson R, Gutowska MA, Wolf J, Bergner SV, Schilhabel MB, Klostermeier UC *et al.* (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol.* **13(7)**, R66..