FROMP-v1.0 User's Guide

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

July 29, 2015

Contents

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, Protein Family (Pfam), and InterPro 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

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:

```
ulimit -c unlimited
java -jar FROMP.jar start
```

This fixes any problems with memory allocation and ensures that FROMP will not run out of memory when running huge samples.

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, PFAM accession numbers, and InterPro 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 *InterProScan* program can be read as input to FROMP within a .ipr file. As well FROMP can accept *InterProScan* outputs from several samples seperated by the '>' character along with the sample name. FROMP will seperate each of these into seperate samples once an operation is done on the data.

   Note:

   (a) The data `will` need to be preprocessed to remove duplicate reads as FROMP works on the assumption that all data has been preprocessed.

   (b) When selecting samples, only one .ipr file can be loaded at once. There can also be no other samples present but the .ipr file so that the samples can be extracted properly.

   `Example: (edited)`

```
 >ERR770958_MERGED_FASTQ

HWI-M02024:110:000000000-A8H0K:1:1101:14372:19903-1:N:0:GTCTGCAA_1_348_+

e12d60578941a68b0b941c2d8ad416f6 115 Gene3D G3DSA:3.20.20.70

1 114 6.3E-14 T 03-03-2015 IPR013785 Aldolase-type TIM barrel

GO:0003824

HWI-M02024:110:000000000-A8H0K:1:1101:14372:19903-1:N:0:GTCTGCAA_1_348_+

e12d60578941a68b0b941c2d8ad416f6 115 Pfam PF04898 Glutamate synthase central domain

1 114 2.8E-22 T 03-03-2015 IPR006982 Glutamate synthase, central-N

GO:0006807|GO:0015930|GO:0055114 UniPathway: UPA00045

HWI-M02024:110:000000000-A8H0K:1:1101:15409:22699-1:N:0:GTCTGCAA_1_134_+

04979772e84a9bb7d80b4a03bcc33336 44 Gene3D G3DSA:3.40.50.1820

1 44 3.8E-9 T 03-03-2015 IPR029058 Alpha/Beta hydrolase fold

.....
```

2. 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 and 2.2

```
Example: (edited)


Bac_DnaA          PF00308.11 GDDSEYS02DHKTX_1_2   -          4.1e-19   68.7.....
Relaxase          PF03432.7  GDDSEYS02DOBSZ_5_1   -            5e-06   25.7.....
HisKA_3           PF07730.6  GDDSEYS02DBODZ_5_0   -          1.6e-07   31.2.....
4.1.1.61_1        -          GDDSEYS02DTH9N_1     -          1.3e-15   53.5.....
3.1.26.4_1        -          GDDSEYS02DXC98_2     -          4.2e-10   34.9.....
```

Note: The EC numbers and PFAM accession numbers 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.

3. FROMP can also read in a simple list of EC numbers, PFAM accession numbers, and InterPro
   accession numbers (one-column)

4. A tab or comma separated list of ECs, PFAMs and InterPro Ids with counts (two-column)

    ```
    Example:


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

5. EC numbers, PFAM accession numbers, and InterPro accession numbers 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
    ```

6. It can also read in a comma or tab separated matrix with the EC, Pfam, InterPro counts for multiple samples. The EC numbers, PFAM accession numbers, and InterPro accession numbers 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`

Save projects
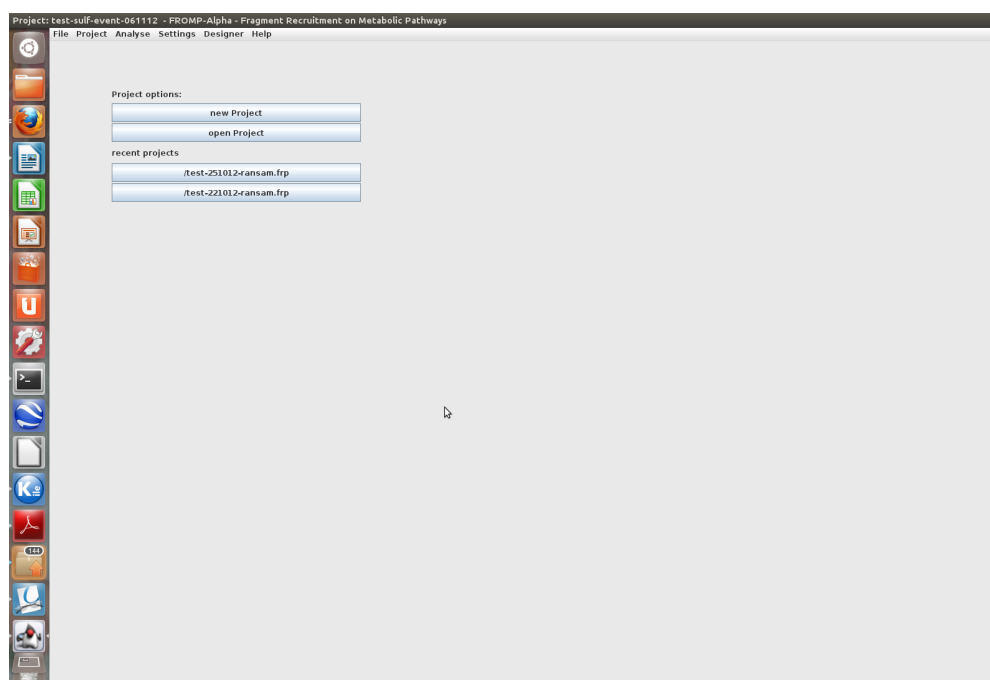
`File->Save Project/Save Project As`

Figure 4.1: The FROMP start-up screen

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 project 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:
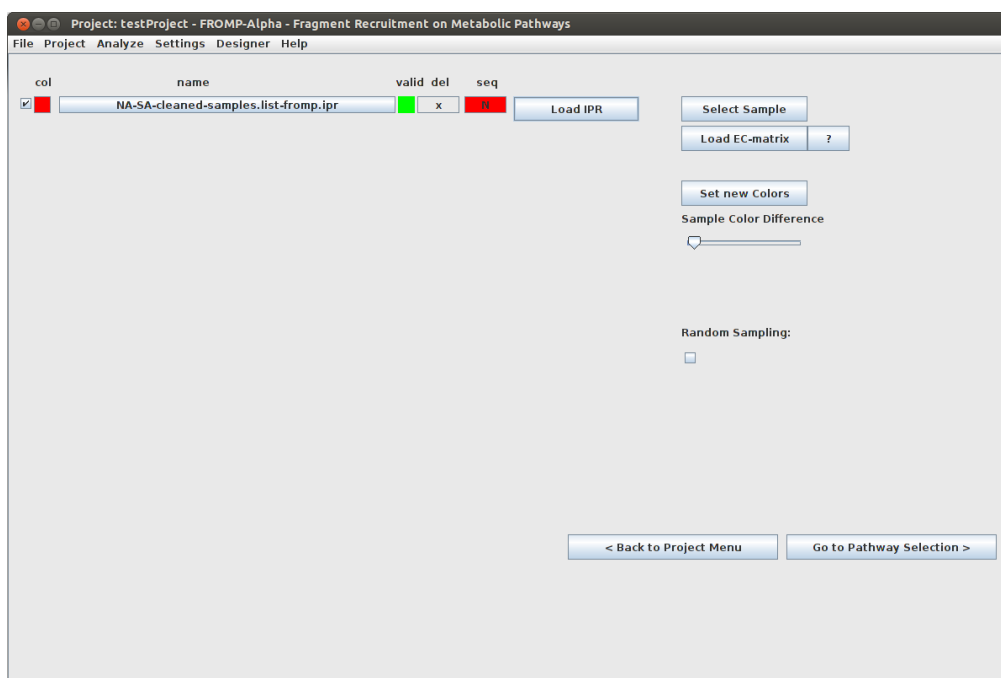
Figure 4.2: The "Edit Samples" window in FROMP

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

Add a matrix file for multiple samples   A matrix file containing EC counts, PFAM counts, and/or InterPro counts with the samples arranged in columns and the EC/PFAM/Interpro IDs in rows, can be uploaded and added to a project by clicking the `Load EC-matrix` button on the `Edit Samples` page. PFAMs and InterPros will be converted to EC numbers during a conversion step. For details on this input format check section 4.1.1.

Loading multiple samples within one .ipr file   An interpro file containing the results of multiple samples can also be loaded by pressing the `Load IPR` button on the `Edit Samples` page once an .ipr file has been selected.

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.
Note: It is currently not possible to load a .ipr file, separate the samples and then change their names. In order to change the samples names for .ipr files samples the samples must be loaded first and the project saved. The project must then be loaded and the samples names can now be editied.

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).

Add sequence files It is possible to add the fasta sequence files to your samples for the purpose of outputting sequences mapping to particular ec numbers.

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 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 ṗwy 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`).
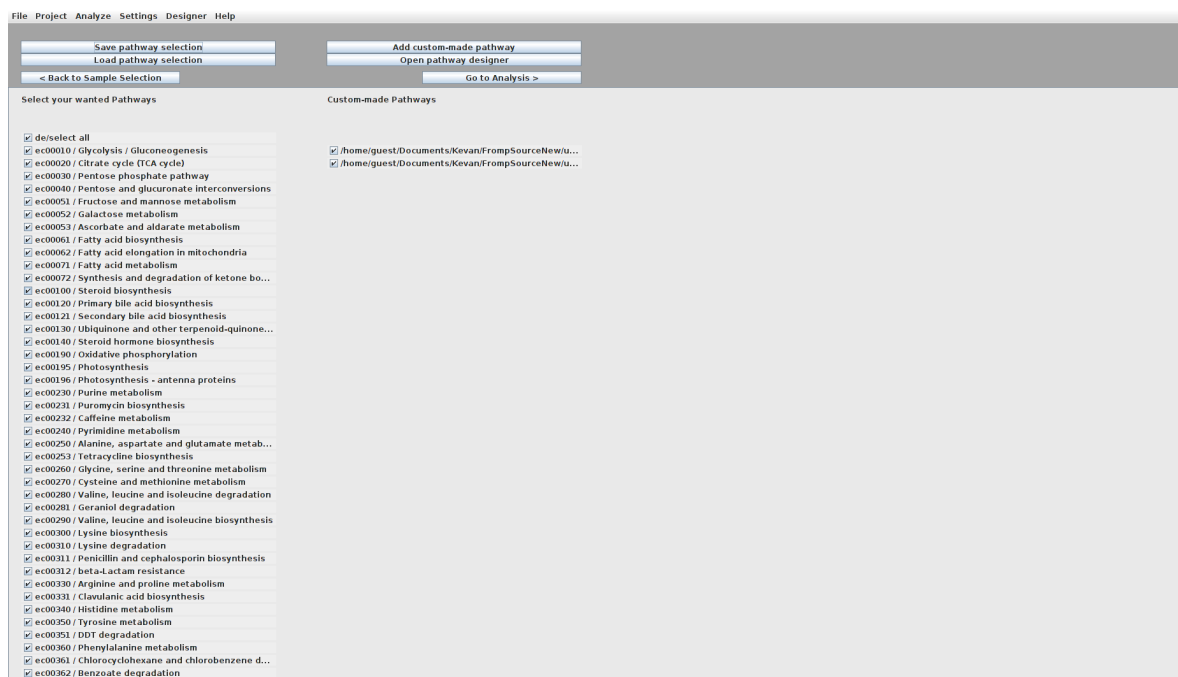
Figure 4.3: The `Select Pathways` page

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). If the desired search method is Pathway name or number the search option `Search Pathways` should be chosen. If the desired search method is EC number then the search option `Search EC number` should be chosen (Figure: 4.4). The search will complete even if just part of pathway name, number or EC number is entered.
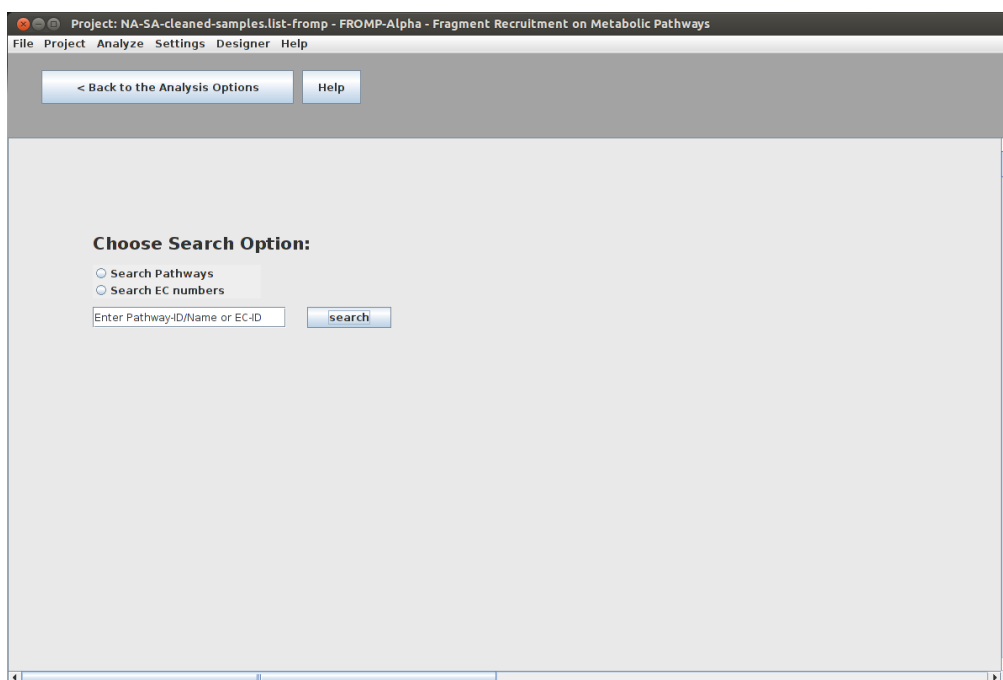
Figure 4.4: The `Search` page which can be used to find specific Pathways

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. Clicking the `Go to analysis` button on the `Select pathways` page (see 4.1.5.2) opens up the Analysis page (Figure 4.5). 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.
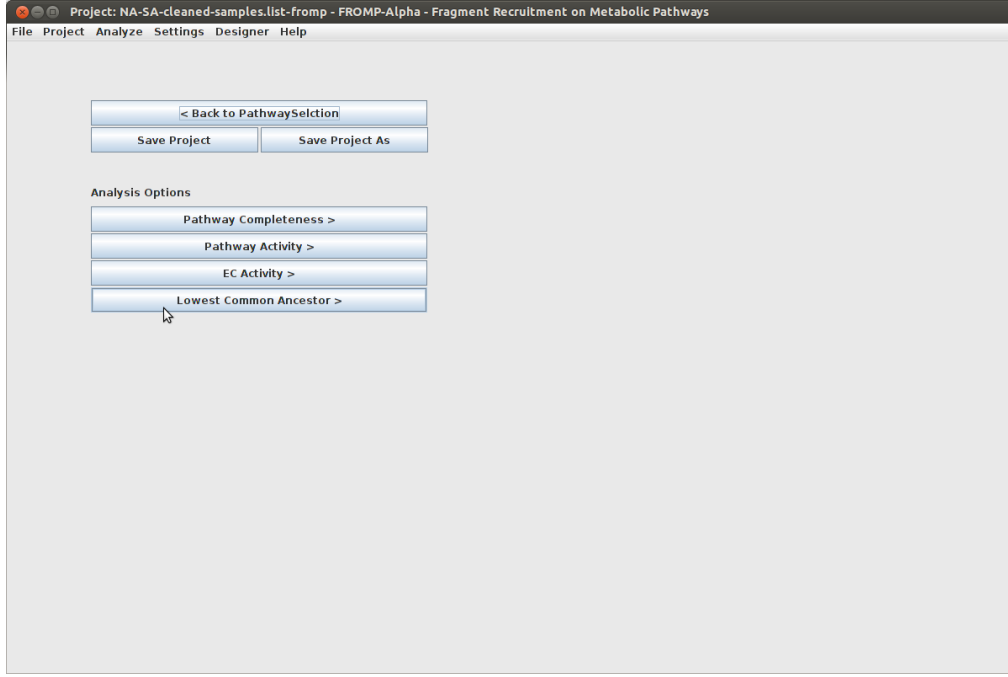
Figure 4.5: The `Analyse` options

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 each EC in a given pathway are then summed.

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

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

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

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 (4.6)

3. The chain information used both in the weights as well as the score calculations. (click the `Use chaining mode 2` check box in 4.6

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 (4.9). Within the pathway mapping visualization, all the highlighted ec numbers can be clicked on in order to open a webpage containing more information about that ec number (Figure: 4.10).
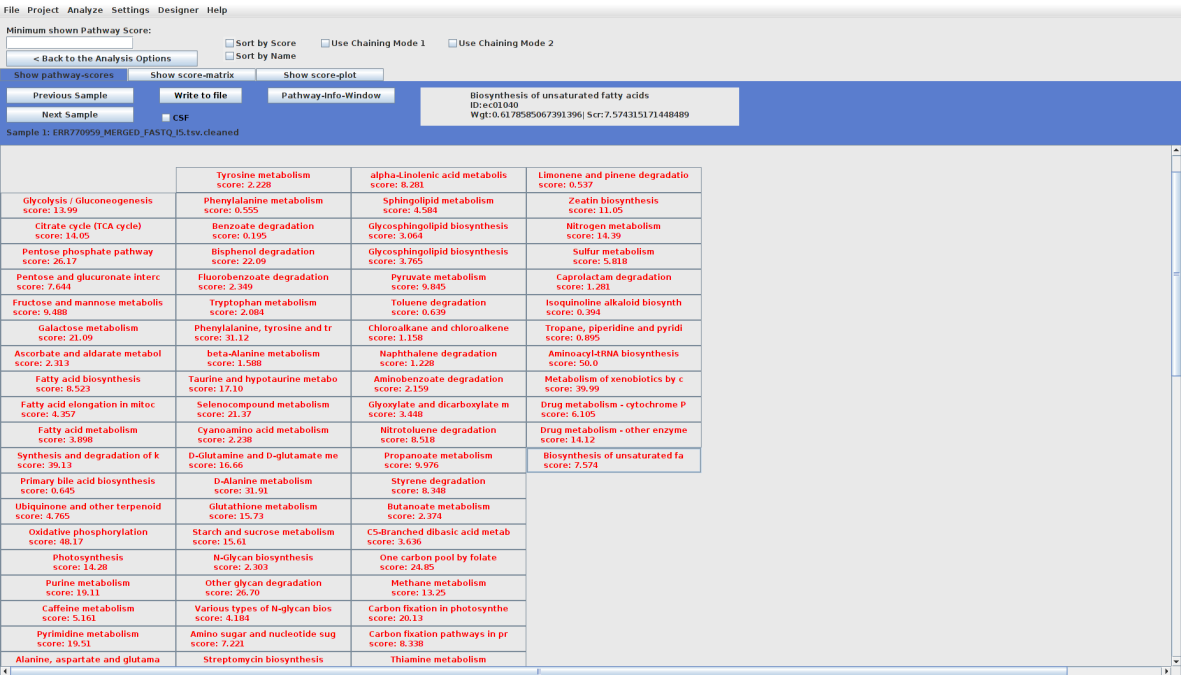
Figure 4.6: The `Pathway Completeness Analysis` window in FROMP

The pathways can be sorted according to the Pathway Completeness Scores by checking the `Sort by score` box or by alpabetical order by checking the `Sort by name` box (Figure 4.8). A minimum score limit can be set by entering the cutoff score in the Min.shown score box. For example, entering 10 in this box will display only those pathway buttons where the score is $>= 10$ (Figure 4.6).
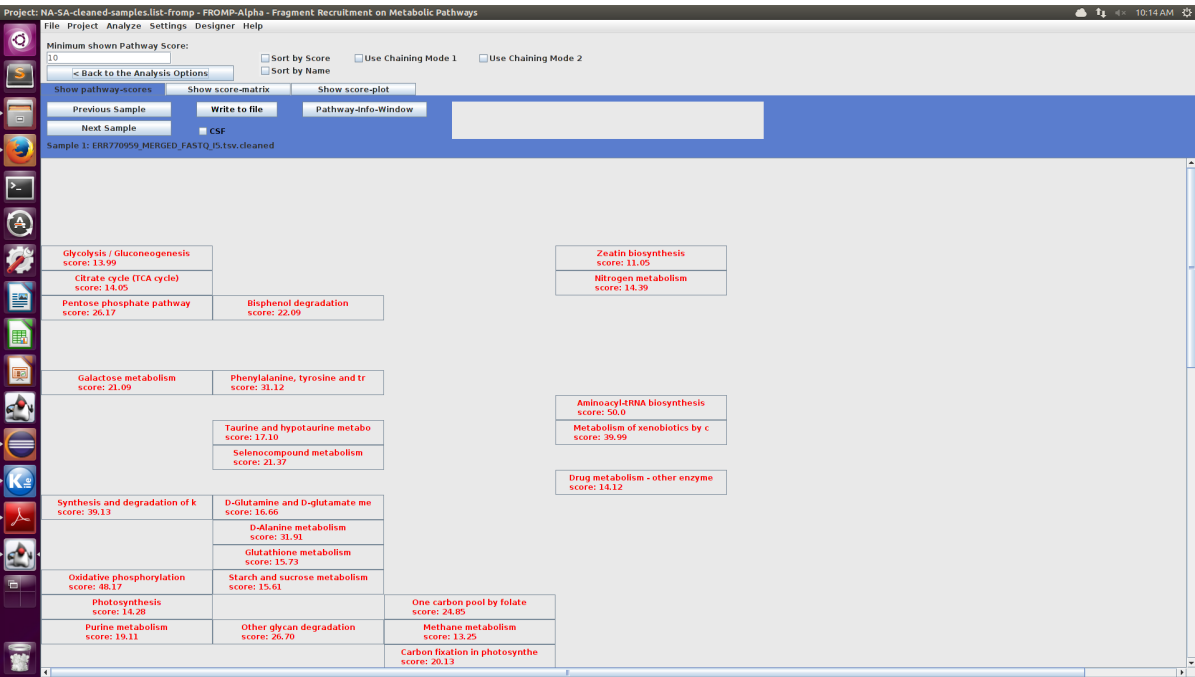
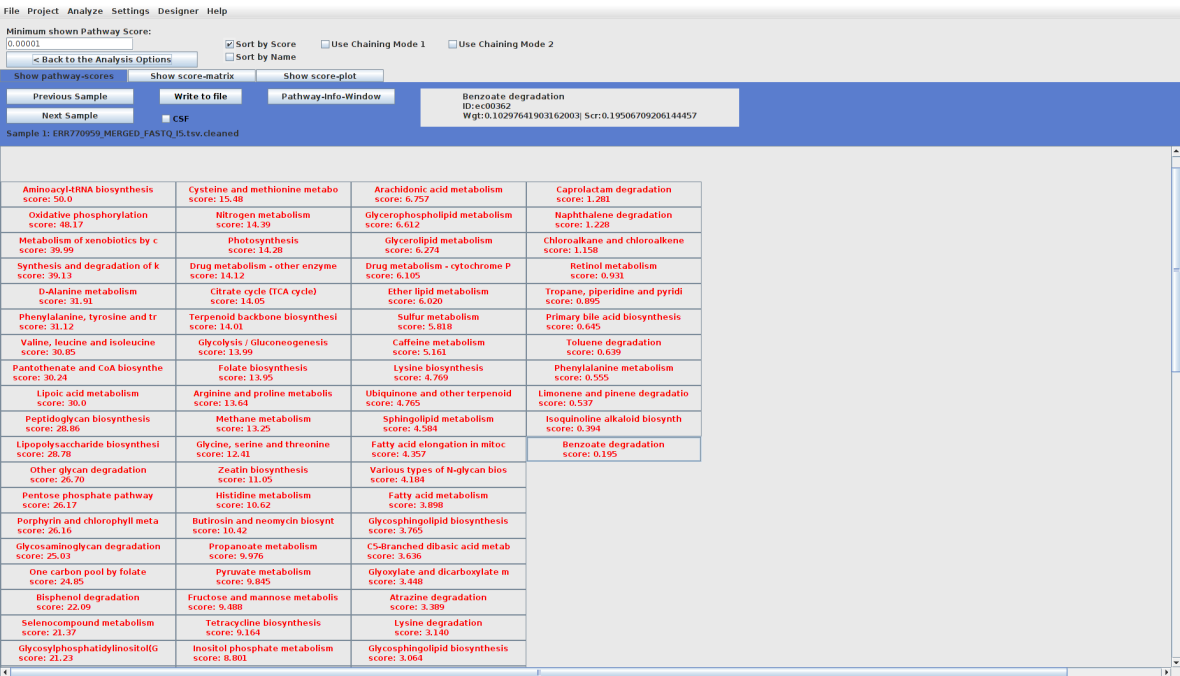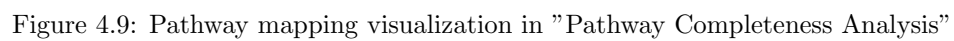Figure 4.7: The `Minimum score cutoff` feature in "Pathway Completeness Analysis"

Figure 4.8: The `Sort by score` feature in "Pathway Completeness Analysis"

Figure 4.9: Pathway mapping visualization in "Pathway Completeness Analysis"

Figure 4.10: Kegg webpage result from clicking on the highlighted ec number "4.3.1.19"

Show Score Matrix The `Show score matrix` tab shows the pathway completeness scores of all pathways for all samples as a matrix (Figure 4.11). 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 4.9).
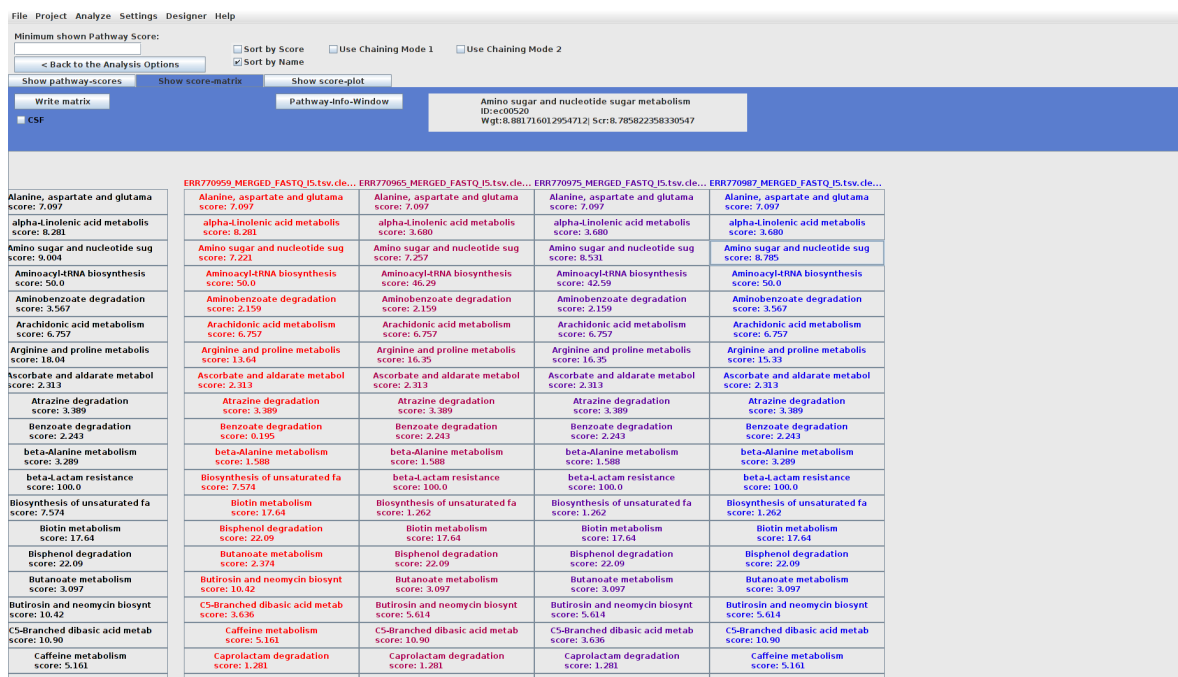
Figure 4.11: The matrix display in "Pathway Completeness Analysis"

On both the `Show Pathway Scores` or `Show score matrix` tabs, there is an option (clicking 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 line plot of the Pathway scores (Figure 4.12). This plot can be enlarged or minimized (`Scale up` or `Scale down` buttons). The pathway scores can also be represented as either a scatter plot (Figure 4.13) or a stacked bar chart (Figure: 4.14) by choosing the `Show Scatter Plot` or `Show Bar Chart` buttons. These charts can be displayed with or without legends by selecteding the `Show legend` checkbox before pressing the `Show Scatter Plot` or `Show Bar Chart` buttons. All of these graphical representations of the Pathway scores can be exported as a .PNG file by the `Export All` button.
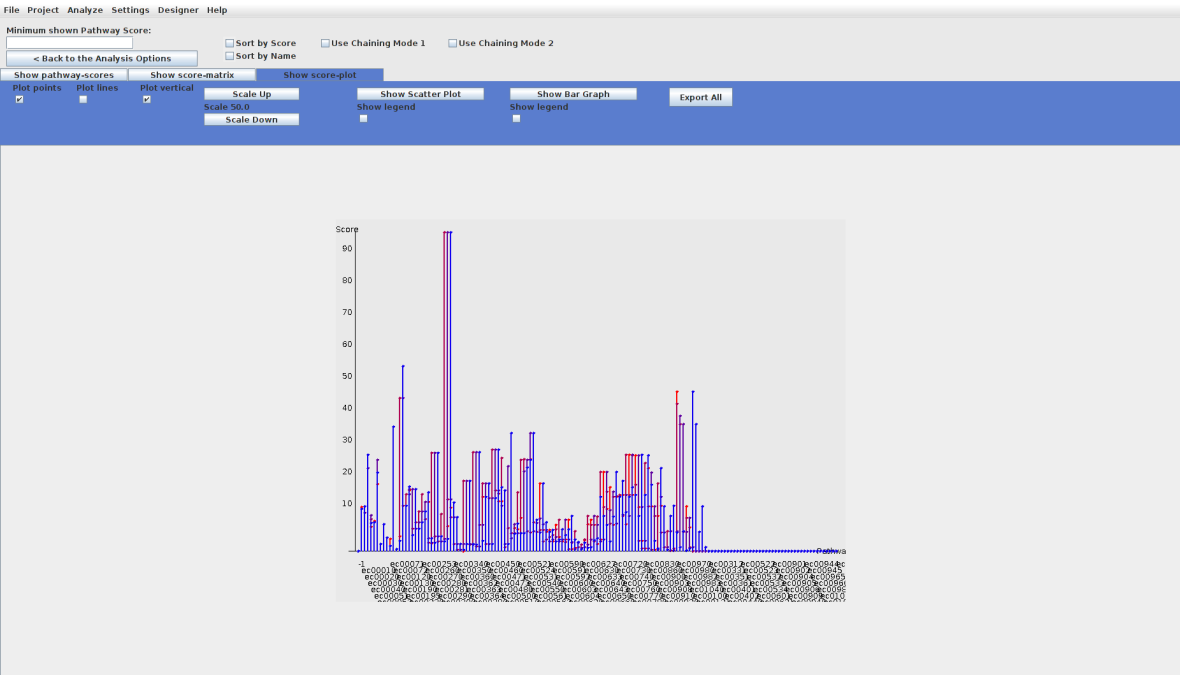
Figure 4.12: The Pathway Completeness Score line plot
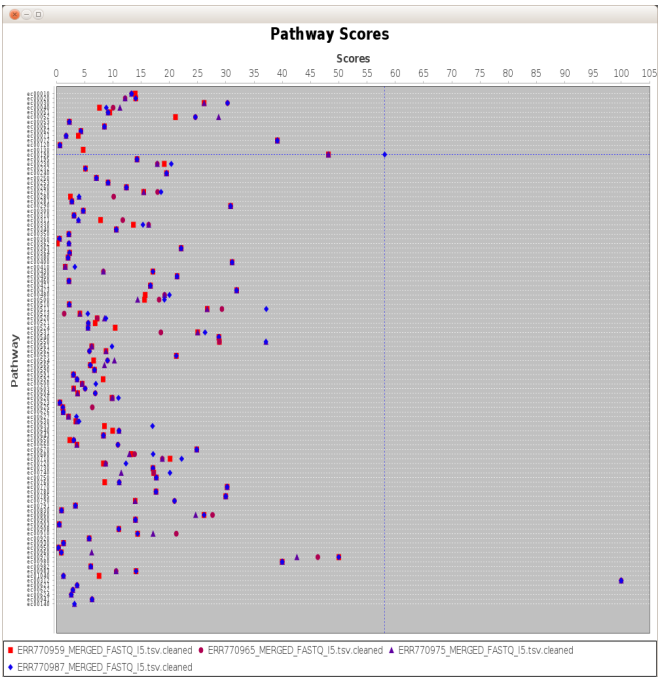
Figure 4.13: The Pathway Completeness Score scatter plot
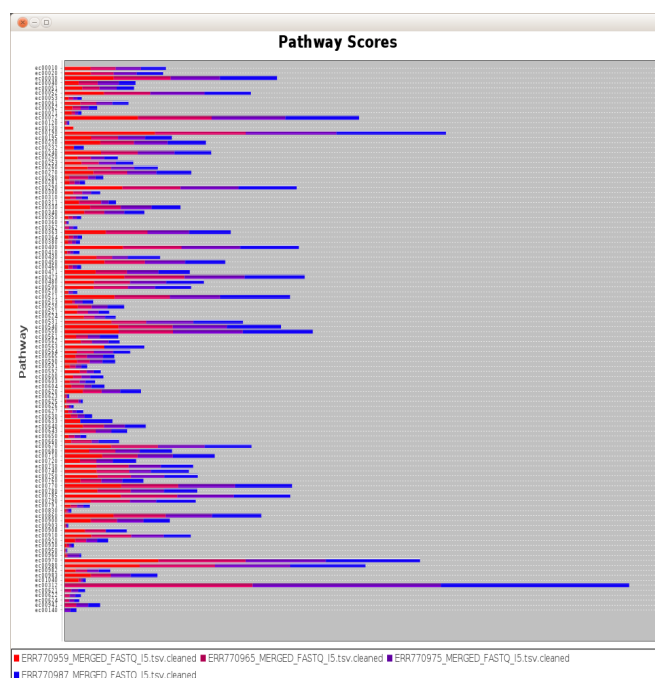
Figure 4.14: The Pathway Completeness Score bar chart

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 4.15).

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 4.15.

If the screen gets too cluttered and the user would only like to view a subset of the samples they can do so. By pressing the sample name of each of the unwanted samples before pressing the rebuild button, the user will remove those samples from the screen. Pressing rebuild again will bring those samples back to the screen.

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).

Figure 4.15: The Pathway Activity Analysis

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 4.16). 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 (Figure 4.16) 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.

- From the `RepSeq` window it is also possible to export the sequences mapping to a particular EC number from the drop-down `File` menu if you have added the sequence file for that sample. This is made possible thanks to an implementation of BioJava [7].They will be exported to the `Sequences` folder in your `FROMP` directory.
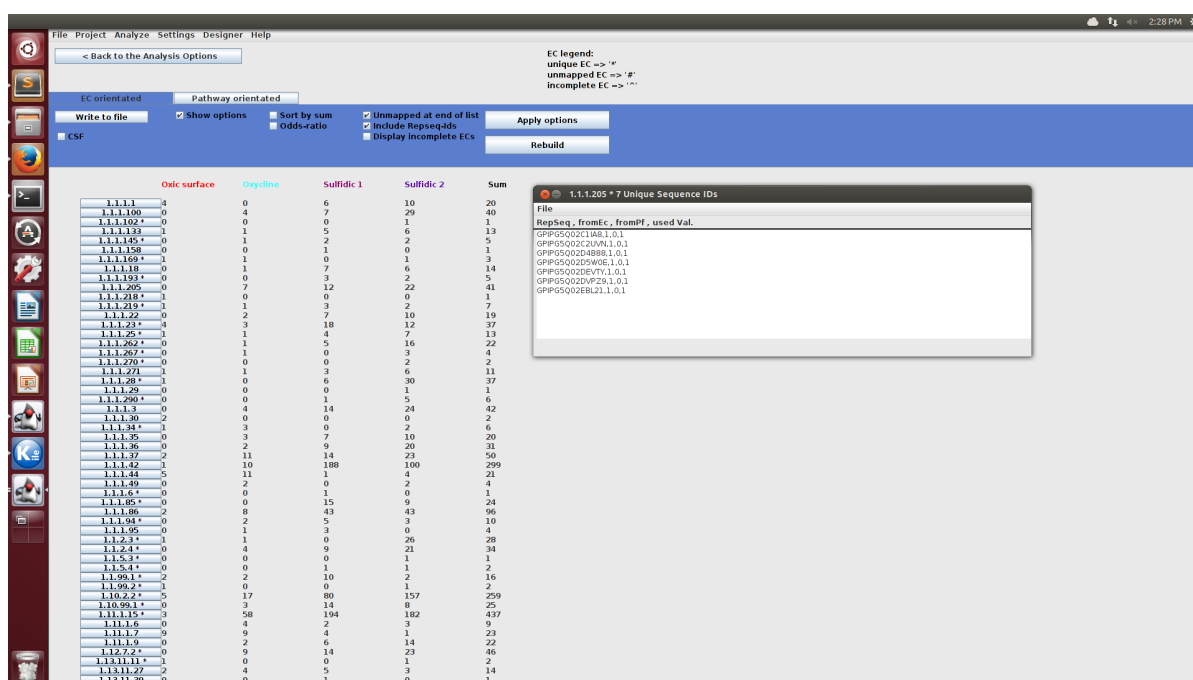
Figure 4.16: The EC orientated tab in EC Activity Analysis with one of the RepSeq windows open

- 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), like in the Pathway Activity analysis, will

de-select that sample from the users view. Clicking the rebuild button will hide all samples whose names have been clicked. Pressing rebuild again will re-display the samples.

- The sequences of a given EC can be exported by left clicking on the EC number and selecting the desired format for the sequences to be saved in. The sequences can be saved in one file per EC number or a file per sample EC number. An option to calculate lowest common ancestor right away can also be chosen at this time (Figure: 4.17).

Figure 4.17: Left-click sequence export options. Can choose just to export the sequences for the specific EC or to also calculate the lowest common ancestor

Pathway orientated view In this view the EC hits are arranged according to the pathways (Figure 4.18). 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`).

File Project Analyze Settings Designer Help

< Back to the Analysis Options

EC legend:
unique EC => '*'
unmapped EC => '#'
incomplete EC => '^'

EC orientated | Pathway orientated

Export values ☑ Sort paths by the sum of ECs ☐ CSF ☐ Sort ECs by sum Resort Rebuild

Glyoxylate and dicarboxylate m
ID:ec00630| Wgt:13.25| Scr:19.73

	Oxic surface	Oxycline	Sulfidic 1	Sulfidic 2	
1.1.1.29	0	0	0	1.0	1.0
1.1.1.36	0	2.0	9.0	20.0	31.0
1.1.1.37	2.0	11.0	14.0	23.0	50.0
1.2.1.2	3.0	4.0	17.0	12.0	36.0
2.3.3.1	5.0	8.0	15.0	37.0	65.0
2.3.3.9	4.0	4.0	12.0	48.0	68.0
2.7.1.31	1.0	0	0	0	1.0
2.1.3.18 *	0	2.0	26.0	45.0	73.0
3.5.1.10	0	0	3.0	5.0	8.0
3.5.1.9	0	0	0	1.0	1.0
4.1.1.39	41.0	38.0	398.0	349.0	826.0
4.1.1.8 *	0	1.0	0	1.0	2.0
4.1.3.1 *	5.0	57.0	49.0	641.0	752.0
4.1.3.16	0	2.0	0	0	2.0
4.2.1.3	6.0	24.0	62.0	146.0	238.0
5.1.99.1	0	1.0	3.0	0	4.0
5.4.99.2	2.0	3.0	11.0	2.0	18.0

Oxidative phosphorylation
ID:ec00190| Wgt:11.32| Scr:73.08

	Oxic surface	Oxycline	Sulfidic 1	Sulfidic 2	
1.10.2.2 *	5.0	17.0	80.0	157.0	259.0
1.6.5.3 *	16.0	23.0	111.0	84.0	234.0
1.6.99.3 *	0	0	2.0	1.0	3.0
1.6.99.5 *	4.0	2.0	24.0	14.0	44.0
1.9.3.1	70.0	75.0	114.0	239.0	498.0
2.7.4.1 *	0	0	0	2.0	2.0
3.6.1.1 *	37.0	71.0	80.0	260.0	448.0
3.6.3.14	30.0	52.0	76.0	93.0	251.0

Carbon fixation pathways in pr
ID:ec00720| Wgt:10.78| Scr:33.44

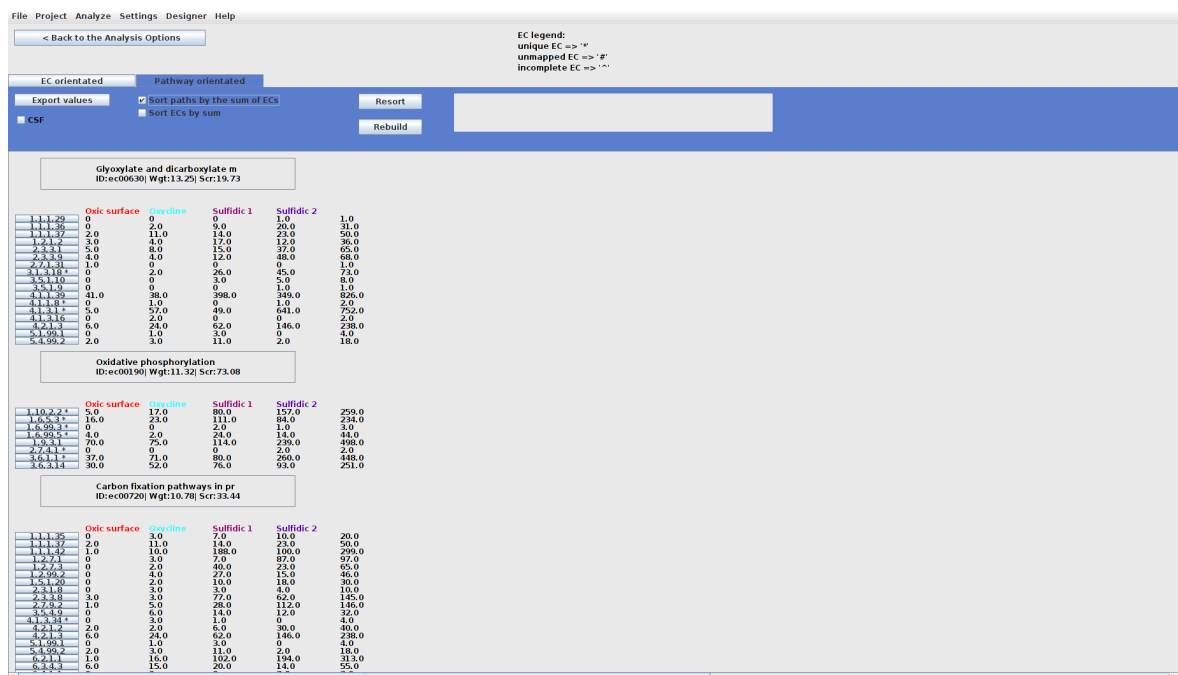|	Oxic surface	Oxycline	Sulfidic 1	Sulfidic 2	
1.1.1.35	0	3.0	7.0	10.0	20.0
1.1.1.37	2.0	11.0	14.0	23.0	50.0
1.1.1.42	1.0	10.0	188.0	100.0	299.0
1.2.7.1	0	3.0	7.0	87.0	97.0
1.2.7.5	0	2.0	40.0	23.0	65.0
1.2.99.2	0	4.0	27.0	15.0	46.0
1.5.1.20	0	2.0	10.0	18.0	30.0
2.3.1.8	0	3.0	3.0	4.0	10.0
2.3.3.8	3.0	3.0	77.0	62.0	145.0
2.7.9.2	1.0	5.0	28.0	112.0	146.0
3.5.4.9	0	6.0	14.0	12.0	32.0
4.1.3.34 *	0	3.0	1.0	0	4.0
4.2.1.2	2.0	2.0	6.0	30.0	40.0
4.2.1.3	6.0	24.0	62.0	146.0	238.0
5.1.99.1	0	1.0	3.0	0	4.0
5.4.99.2	2.0	3.0	11.0	2.0	18.0
6.2.1.1	1.0	16.0	102.0	194.0	313.0
6.3.4.3	6.0	15.0	20.0	14.0	55.0

Figure 4.18: The Pathway orientated tab in EC Activity Analysis

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

4.1.6.4 Lowest Common Ancestor

This window allows the user to find the lowest common ancestors for a given sequence file. The sequences within the file undergo a tryptic peptide digest. The resulting peptide segements are sent to `http://unipept.ugent.be/` using get requests. The response contains wheter or not the sent peptide sequence is recgonized to have a taxonomic lowest common ancestor. The results are then collected for all the peptide sequences sent for the sequence file into a summary response table and visual piechart.

Calculating Lowest Common Ancestor To find the lowest common ancestors for a sequence file, the file must first be chosen by clicking on the 'Find Lowest Common Ancestor' button (Figure: 4.19). The file can be in a format of either a single sequence file or a text file containing the filepaths to multiple sequence files. Once a sequence file is selected, the lowest common ancestors are calculated and the results displayed onto the window (Figure: 4.20). The screen can be cleared at any time by pressing the 'clear' button.
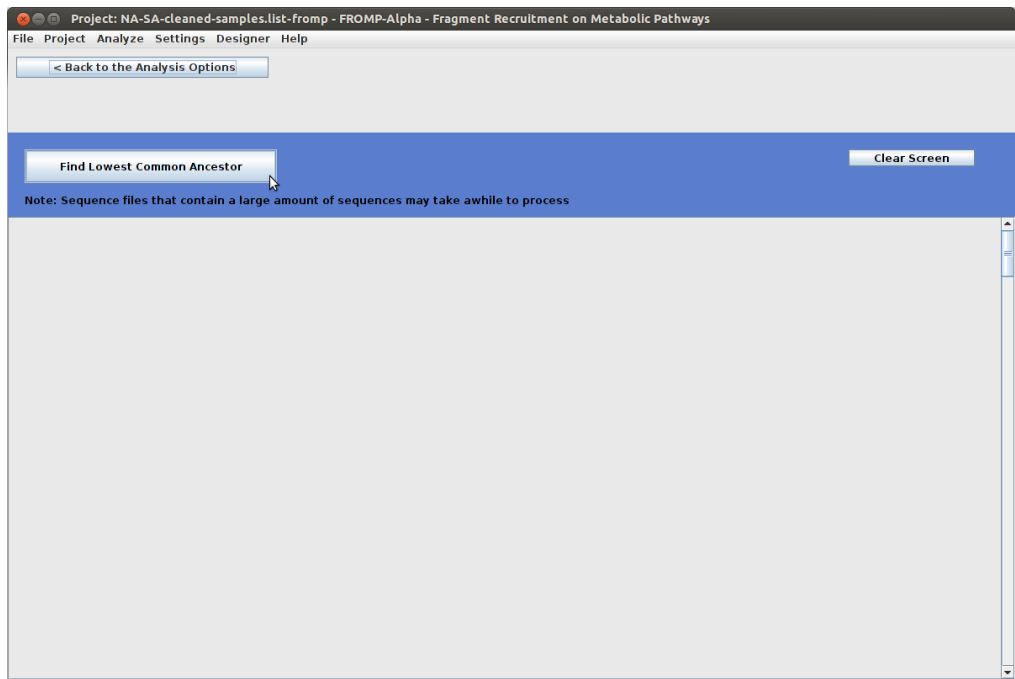
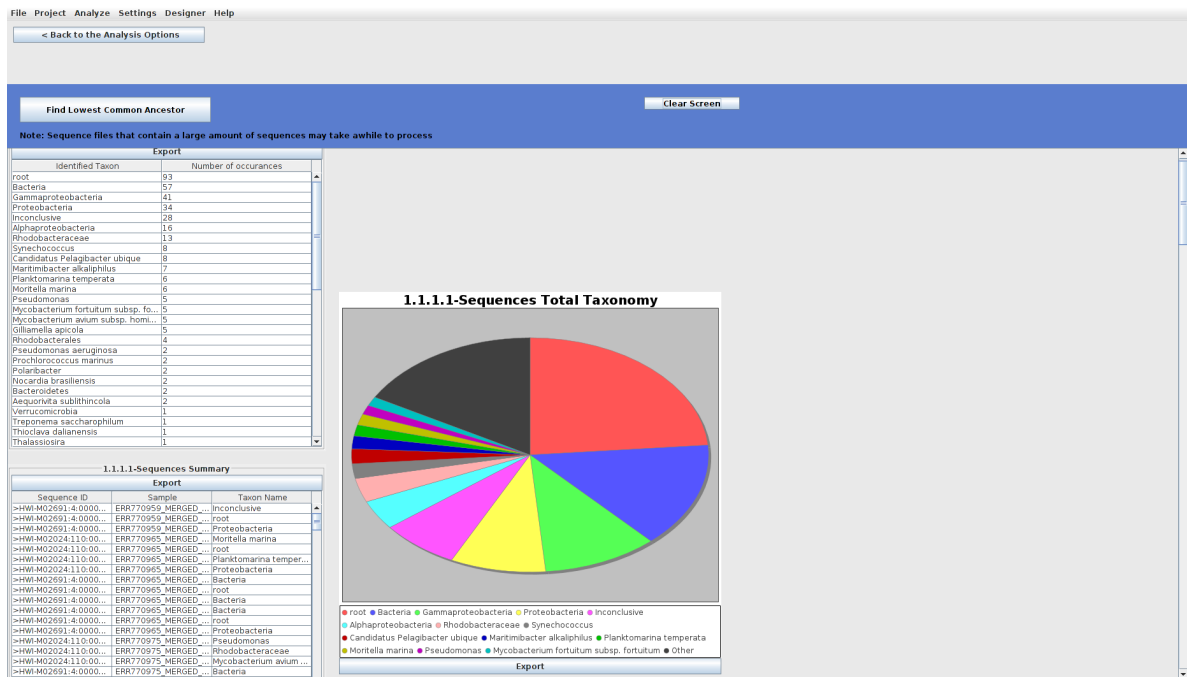Figure 4.19: Find Lowest Common Ancestor window

Figure 4.20: The displayed results of lowest common ancestor anaysis on a sequence file.

Exporting Lowest Common Ancestor It is possible to export the displayed results of the lowest common ancestor calculation by pressing the 'export' button near the desired figure to be exported (Figure: 4.21). Exported tables are saved as an excel file while exported piecharts are saved as a .png file.
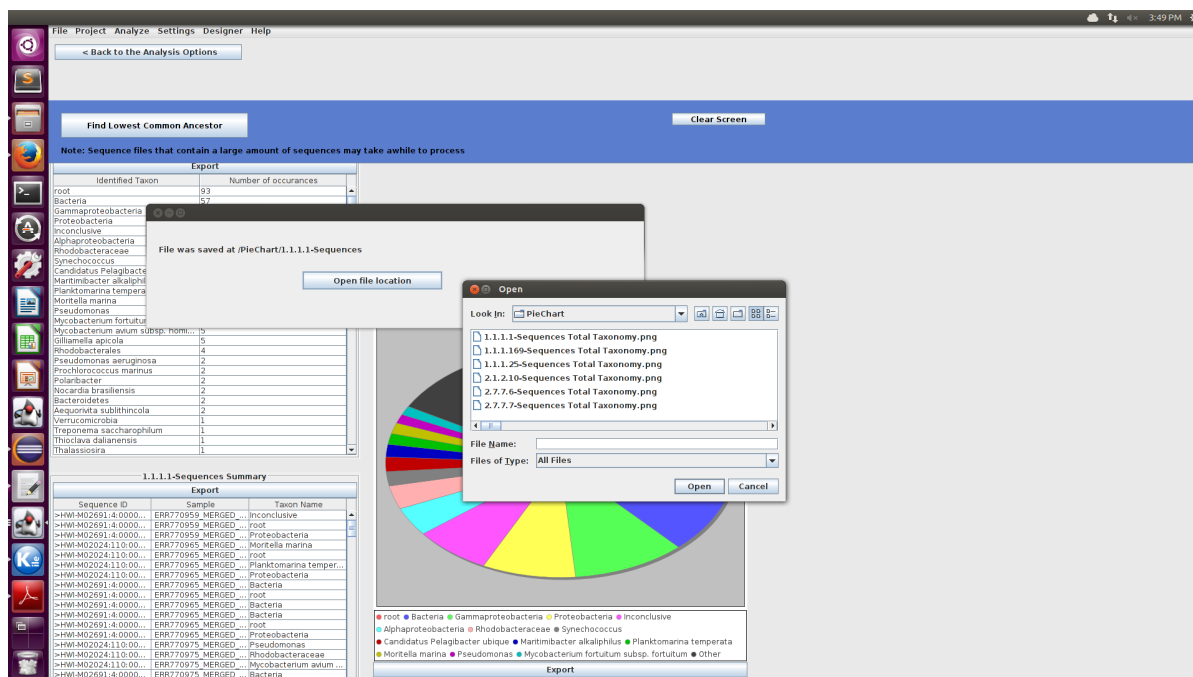

Figure 4.21: Pressing the 'Export' buttons results in the displayed figure to be saved to the users computer. The saved file location is displayed to the used once the save is completed.

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.
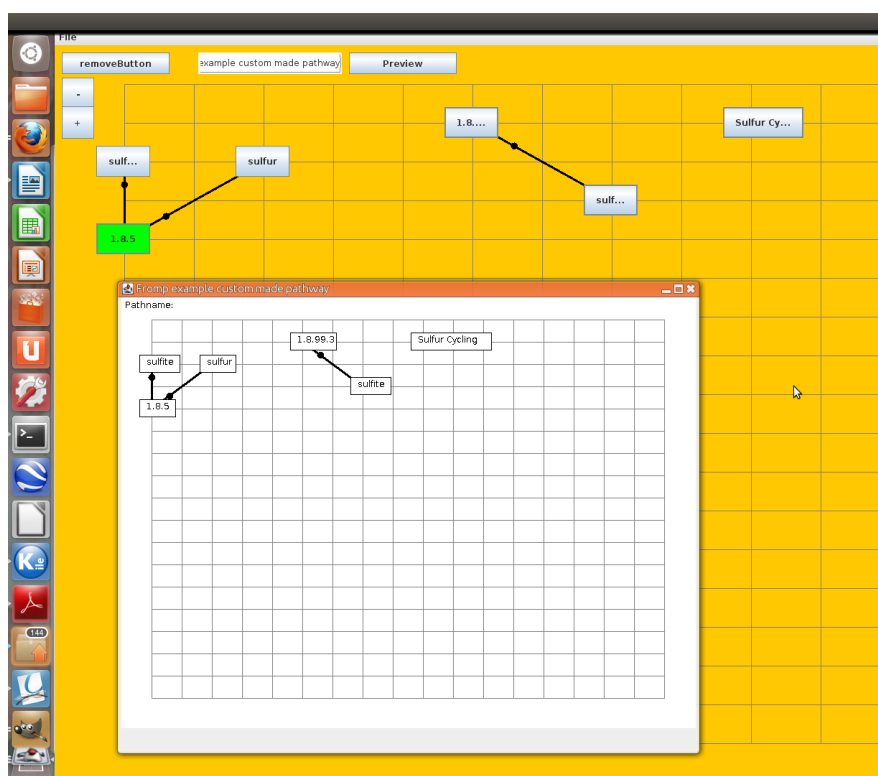
Figure 4.22: The Pathway Designer interface showing a custom made pathway and a preview window showing the final layout

The following actions are possible in the Pathway Designer interface:

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

- **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 4.22, 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' option 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

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' h

- 'p' for all pathway pictures of all samples (e.g if there are 5 samples and 150 pathways, it will generate 150x5 png images).

  Note: If the project contains a large amount of samples, exporting all the pathway pictures at once may cause a "java.lang.OutOfMemoryError: Java heap space" error. The error can be attempted to be solved by adding -Xmx option to -Xmx512m (or greater depending upon the RAM capacity of your machine) infront of the -jar section of the run code. If this does not solve the problem it is best to export the project and load it using the GUI Fromp in order to export the wanted pictures.

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' p

- 's' for the pathway-score-matrix

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' s

- 'm' for the pathway-activity-matrix

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' m

- 'e' for the EC-activity-matrix

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' e

- 'f' to export the project as a .frp file

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' f

- 'a' for all of the above simultaneously

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' a

- 'am' for the pathway-score-matrix, pathway-activity-matrix, and EC-activity-matrix

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' am

- '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)

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' op

- 'up' to only print out user designed Pathways (only works with a project-file or a .lst-file)

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' up

- 'ec' to export sequence IDs for a set of ECs. For this option, rather than including an output file path, you include the path to a file which contains a list of EC numbers.

  Usage: java -jar FROMP.jar 'input filepath' 'ec list filepath' ec

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' ec 'ec number'

- 'seq' to export sequences for a set of ECs into a file per sample. For this option you can either include the path to a file which contains a list of EC numbers in place of the output path, or simply include the input path 'seq' and the ECs you are interested in.

  Note: The input file must be a project file that already has a the sequence files add or else this command will not work.

  Usage: java -jar FROMP.jar 'input filepath' 'ec list filepath' seq

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' seq 'ec number'

- 'seqall' to export sequences for a sdet of ECs into a file per EC number. For this option you can either include the path to a file which contains a list of EC numbers in place of the output path, or simply include the input path 'seqall' and the ECs you are interested in.

  Note: The input file must be a project file that already has a the sequence files add or else this command will not work.

  Usage: java -jar FROMP.jar 'input filepath' 'ec list filepath' seqall

  Usage: java -jar FROMP.jar 'input filepath' 'output filepath' seqall 'ec number'

- 'lca' to find the lowest common ancestor for a given sequence file (or list of sequence file paths). Once the lowest common ancestor is determined the results are exported as both an excel file and a pie chart.

  Note: The input file must be a project file that already has a the sequence files add or else this command will not work.

  Usage: java -jar FROMP.jar 'input filepath' 'sequence filepath' lca

  Usage: java -jar FROMP.jar 'input filepath' 'sequence list filepath' lca

Sequence IDs: In cmdFROMP you are able to export sequence IDs of a particular set of ECs without including a file containing said list. To do so, simply add the EC numbers (ie. 1.1.1.1) you are looking for, either at the end of your command, or accompanied by an input file path if you have no need to preform a command along with it and text files contain the sequence IDs for the ECs in each of the samples will be generated. If a particular sample doesn't contain the EC in question, no files will

be generated

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 ec 1.1.1.1
  ```

- Export all sequence IDs for the ECs listed in the file eclist.txt

  ```
  java -jar FROMP.jar c:\Users\nukota\test.lst eclist.txt ec
  ```

- Export all sequence IDs for the ECs 1.1.1.1 and 4.1.1.39

  ```
  java -jar FROMP.jar c:\Users\nukota\test.lst 1.1.1.1 4.1.1.39
  ```

- Export all sequences for the EC 2.7.7.6 into their own sample files

  ```
  java -jar FROMP.jar c:\Users\nukota\test-sulfidic.frp seq 1.1.1.1 4.1.1.39
  ```

- Export all sequences for the EC's within the EC list file into one EC sequence file

  ```
  java -jar FROMP.jar /home/guest/Jennifer/JenTest.frp /home/guest/Jennifer/FrompSourceNew/Sequ
  ```

- Find the lowest Common Ancestor of a given list of Sequences

  ```
  java - jar FROMP.jar /home/guest/Jennifer/JenTest.frp /home/guest/Jennifer/FrompSourceNew/Seq
  ```

- Find the lowest Common Ancestor of a given Sequence filem

  ```
  java -jar FROMP.jar /home/guest/Jennifer/JenTest.frp /home/guest/Jennifer/FrompSourceNew/Sequ
  ```

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..

[7] Prlić A, Yates A, Bliven SE, Rose PW, Jacobsen J, Troshin PV, Chapman M, Gao J, Koh CH, Foisy S *et al.* (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics* **vol. 28 no. 20 2012, pages 2693-2695**, doi:10.1093/bioinformatics/bts494..