Metabolomic Data Analysis with MetaboAnalyst 6.0

Name: guest18175079373188657455

January 17, 2025

1 Background

Even with high mass accuracy afforded by current high-resolution MS platforms, it is often impossible to uniquely identify a given peak based on its mass alone. To get around this issue, a key concept is to shift the unit of analysis from individual compounds to individual pathways or a group of functionally related compounds (i.e. metabolite sets (PMID: 20457745)). The general assumption is that the collective behavior of a group is more robust against a certain degree of random errors of individuals. The mummichog algorithm is the first implementation of this concept to infer pathway activities from a ranked list of MS peaks identified by untargeted metabolomics. The original algorithm implements an over-representation analysis (ORA) method to evaluate pathway-level enrichment based on significant features. Users need to specify a pre-defined cutoff based on p-values. For further details about the original implementation, please refer to Li et al. 2013 (PMC3701697). A complementary approach is the Gene Set Enrichment Analysis (GSEA) method, a widely used method to extract biological meaning from a ranked gene list (PMID: 16199517). Unlike ORA, this method considers the overall ranks of MS peaks without using a significance cutoff. It is able to detect subtle and consistent changes which can be missed from ORA methods. Both the mummichog algorithm (Version 1.0.10), which has been carefully translated from the Python programming to R, and the adapted GSEA method have been implemented in the MS Peaks to Paths module. The module also includes an expanded knowledgebase of 21 organisms for pathway analysis.

2 Overview

The MS Peaks to Pathways module consists of three steps - uploading the user's data, selection of a pathway library, and pathway analysis.

3 Data Input

The MS Peaks to Pathways module accepts either a three column table containing the m/z features, p-values, and statistical scores, a two-column table containing m/z features and either p-values or t-scores, or a one-column table ranked by either p-values or t-scores. All inputted files must be in .txt format. If the input is a three column table, both the mummichog and GSEA algorithms can be applied. If only p-values (or ranked by p-values) are provided, only the mummichog algorithm will be applied. If only t-scores (or ranked by t-scores) are provided, only the GSEA algorithm will be applied.

A total of 1841 m/z features were found in your uploaded data. The instrument's mass accuracy is 5 ppm. The instrument's analytical mode is mixed . The uploaded data contains 4 columns. The column headers of uploaded data are m.z, p.value, t.score and mode . The range of m/z peaks is trimmed to 50-2000. 0 features have been trimmed. A total of 1841 input mz features were retained for further analysis.

3.0.1 Parameters

Users also need to specify the mass accuracy (ppm), the ion mode (positive or negative), and the p-value cutoff to delineate between significantly and non-significantly enriched pathways (pathway-level only), and pathway library used. Currently, MetaboAnalyst 6.0 is primarily designed to support peaks obtained from high-resolution MS instruments such as Orbitrap, or TOF instruments.

The selected p-value cutoff is: 1e-05.

3.0.2 Library

The knowledge-base for this module consists of five genome-scale metabolic models obtained from the original Python implementation which have either been manually curated or downloaded from BioCyc, an expanded library of 21 organisms derived from KEGG metabolic pathways, and 10 distinct metabolite set libraries. Users must select one of 21 KEGG pathway libraries, or one of five metabolic models.

The user's selected library is: hsa.kegg .

4 Mummichog Output

The aim of of this module is to leverage the power of known metabolic models/pathways to gain functional insight directly from m/z features. There are three steps for the mummichog algorithm, 1) Permutations: A list of metabolites (the same length as the number of significant m/z features) are inferred from the user's uploaded set of m/z features, considering all potential matches (isotopes/adducts). These tentative compounds are then mapped onto known metabolic pathways for the selected organism. For each pathway, a hypergeometric p-value is calculated. 2) Step 1 is repeated multiple times to calculate the null distribution of p-values for all pathways, and is modeled as a Gamma distribution. 3) Following this, the significant m/z features are used to calculate the p-values for each pathway (Step 1). These p-values are then adjusted for the permutations.

4.1 Mummichog Pathway Analysis Plot

The pathway summary plot below displays all matched pathways as circles. The color and size of each circle corresponds to its p-value and enrichment factor, respectively. The enrichment factor of a pathway is calculated as the ratio between the number of significant pathway hits and the expected number of compound hits within the pathway.

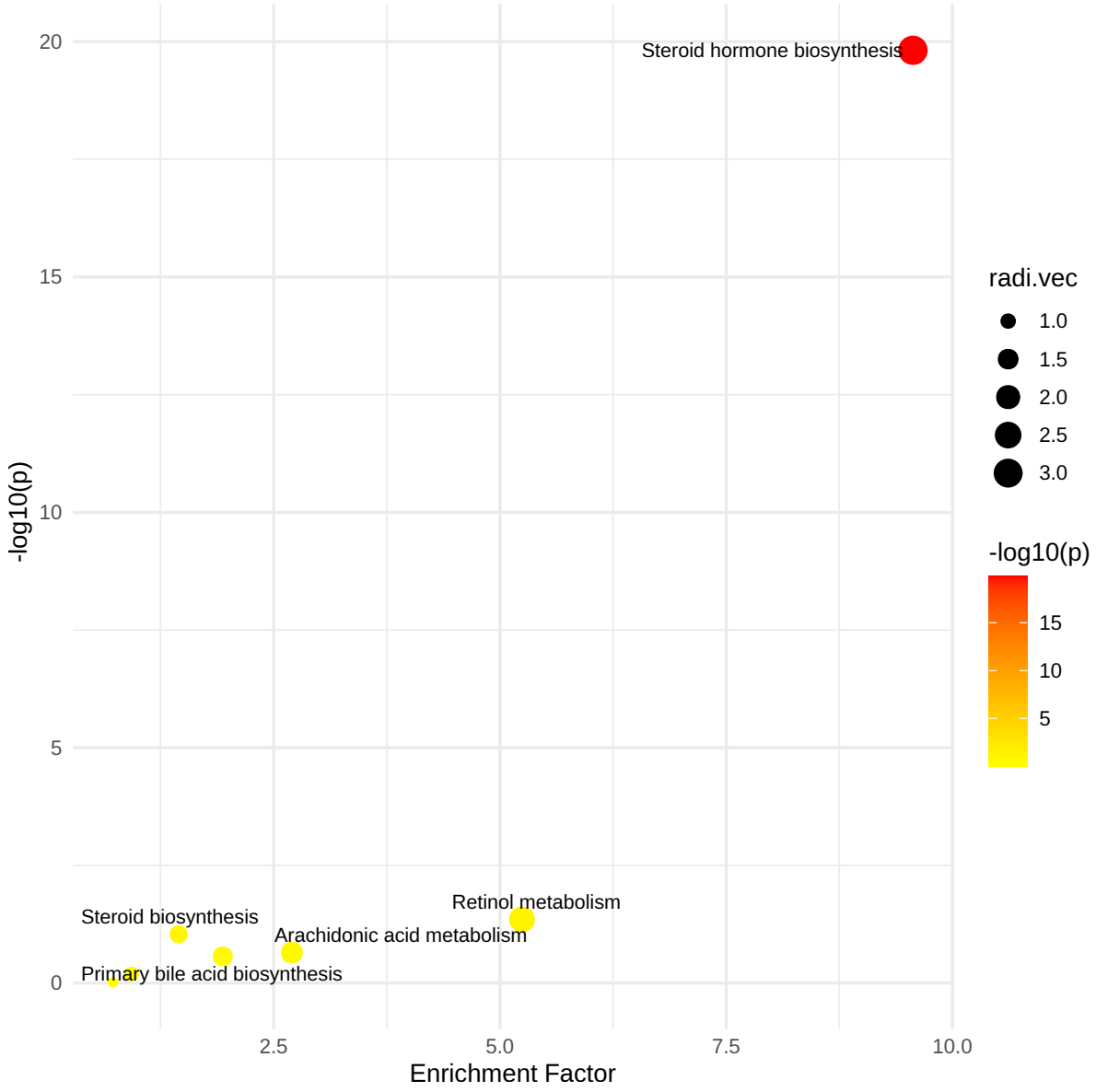

Figure 1: Summary of Pathway Analysis

4.2 Mummichog Pathway Analysis Results Table

The output of the mummichog analysis consists of a table of results containing ranked pathways that are enriched in the user-uploaded data. The table includes the total number of hits per pathway (all, significant, and expected), the raw p-values (Hypergeometric), and the p-value modeled on user data using a Gamma distribution.

Table 1: Results of the Mummichog Pathway Analysis

	Pathway total	Hits.total	Hits.sig	Expected	FET	EASE	Total	Sig	Gamma	Emp.Hits
Steroid hormone biosynthesis	87.00	43.00	28.00	2.93	0.00	0.00	43.00	28.00	0.00	0.00
Retinol metabolism	17.00	8.00	3.00	0.57	0.04	0.20	8.00	3.00	0.00	0.00
Steroid biosynthesis	41.00	5.00	2.00	1.38	0.09	0.43	5.00	2.00	0.01	1.00
Arachidonic acid metabolism	44.00	23.00	4.00	1.48	0.23	0.44	23.00	4.00	0.01	0.00
Primary bile acid biosynthesis	46.00	17.00	3.00	1.55	0.27	0.55	17.00	3.00	0.01	3.00
Tryptophan metabolism	41.00	27.00	1.00	1.38	0.96	1.00	27.00	1.00	1.00	74.00
Sphingolipid metabolism	32.00	9.00	1.00	1.08	0.64	1.00	9.00	1.00	1.00	12.00

5 Compound Matching Table

The output of the MS Peaks to Pathways module also consists of a comprehensive table containing the compound matching information for all user-uploaded m/z features. The table has four columns, containing the Query.Mass of each feature, the predicted Matched.Compound for each feature, the Matched.Form, and the Mass.Diff. As the file can be very long (>40 pages), please download it separately on the Downloads page of MetaboAnalyst.

6 Network Visualization

The MS Peaks to Pathways module also allows users to interactively view their data in a global KEGG metabolic network. Users will be able to their network as a SVG or PNG file on the Network Viewer page of MetaboAnalyst.

7 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"mass_all\", \"mummichog\", FALSE)"
 [2] "mSet<-SetPeakFormat(mSet, \"rmp\")"
 [3] "mSet<-UpdateInstrumentParameters(mSet, 5.0, \"mixed\", \"yes\", 0.02);"
 [4] "mSet<-Read.PeakListData(mSet, \"Replacing_with_your_file_path\");"
 [5] "mSet<-SanityCheckMummichogData(mSet)"
 [6] "mSet<-SetPeakEnrichMethod(mSet, \"mum\", \"v2\")"
 [7] "mSet<-SetMummichogPval(mSet, 1.0E-5)"
 [8] "mSet<-PerformPSEA(mSet, \"hsa_kegg\", \"current\", 2 , 100)"
 [9] "mSet<-PlotPeaks2Paths(mSet, \"peaks_to_paths_0_\", \"png\", 72, width=NA)"
[10] "mSet<-SaveTransformedData(mSet)"
[11] "mSet<-PreparePDFReport(mSet, \"guest18175079373188657455\")\n"
```

The report was generated on Fri Jan 17 10:04:40 2025 with R version 4.4.0 (2024-04-24), OS system: Linux.