Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest8993539346205070843

October 5, 2021

1 Background

The Pathway Analysis module combines results from powerful pathway enrichment analysis with pathway topology analysis to help researchers identify the most relevant pathways involved in the conditions under study.

There are many commercial pathway analysis software tools such as Pathway Studio, MetaCore, or Ingenuity Pathway Analysis (IPA), etc. Compared to these commercial tools, the pathway analysis module was specifically developed for metabolomics studies. It uses high-quality KEGG metabolic pathways as the backend knowledgebase. This module integrates many well-established (i.e. univariate analysis, over-representation analysis) methods, as well as novel algorithms and concepts (i.e. Global Test, GlobalAncova, network topology analysis) into pathway analysis. Another feature is a Google-Map style interactive visualization system to deliver the analysis results in an intuitive manner.

2 Data Input

The Pathway Analysis module accepts either a list of compound labels (common names, HMDB IDs or KEGG IDs) with one compound per row, or a compound concentration table with samples in rows and compounds in columns. The second column must be phenotype labels (binary, multi-group, or continuous). The table is uploaded as comma separated values (.csv).

3 Compound Name Matching

The first step is to standardize the compound labels used in user uploaded data. This is a necessary step since these compounds will be subsequently compared with compounds contained in the pathway library. There are three outcomes from the step - exact match, approximate match (for common names only), and no match. Users should click the textbfView button from the approximate matched results to manually select the correct one. Compounds without match will be excluded from the subsequently pathway analysis.

Table 1 shows the conversion results. Note: *1* indicates exact match, *2* indicates approximate match, and *0* indicates no match. A text file contain the result can be found the downloaded file *name_map.csv*

Table 1: Result from

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	Glutamine	L-Glutamine	HMDB0000641	5961	C00064	C(CC(=O)N)[C@@H](C(=O)O)N
2	Asparagine	L-Asparagine	HMDB0000168	6267	C00152	C([C@@H](C(=O)O)N)C(=O)N
3	Alanine	L-Alanine	HMDB0000161	5950	C00041	C[C@@H](C(=O)O)N
4	Dihydroorotate	4,5-Dihydroorotic acid	HMDB0000528	648	C00337	C1C(NC(=O)NC1=O)C(=O)O
5	Ornithine	Ornithine	HMDB0000214	6262	C00077	C(C[C@@H](C(=O)O)N)CN
6	NADPH	NADPH	HMDB0000221	22833512	C00005	C1C=CN(C=C1C(=O)N)[C@@H]2[C
7	Arginine	L-Arginine	HMDB0000517	6322	C00062	C(C[C@@H](C(=O)O)N)CN=C(N)N
8	Homoarginine	Homo-L-arginine	HMDB0000670	9085	C01924	C(CCN=C(N)N)C[C@@H](C(=O)O)
9	2-Aminoethylphosphonate	Ciliatine	HMDB0011747	339	C03557	C(CP(=O)(O)O)N

10	Glycine	Glycine	HMDB0000123	750	C00037	C(C(=O)O)N
11	4-Guanidinobutanoic acid	4-Guanidinobutanoic acid	HMDB0003464	500	C01035	C(CC(=O)O)CN=C(N)N
12	Pyrroline-5-carboxylic acid	NA	NA	NA	NA	NA
13	NG-dimethyl-L-arginine	Asymmetric dimethylarginine	HMDB0001539	123831	C03626	CN(C)C(=NCCC[C@@H](C(=O)O)N
14	UMP	Uridine 5'-monophosphate	HMDB0000288	6030	C00105	C1=CN(C(=O)NC1=O)[C@H]2[C@@
15	Acetyl Proline	NA	NA	NA	NA	NA
16	Isoleucine	L-Isoleucine	HMDB0000172	6306	C00407	CC[C@H](C)[C@@H](C(=O)O)N
17	O-Phosphorylethanolamine	O-Phosphoethanolamine	HMDB0000224	1015	C00346	C(COP(=O)(O)O)N
18	Cellobiose	Cellobiose	HMDB0000055	10712	C06422	C([C@@H]1[C@H]([C@@H]([C@H]([C
19	Histidine	L-Histidine	HMDB0000177	6274	C00135	C1=C(NC=N1)C[C@@H](C(=O)O)N
20	Lysine	L-Lysine	HMDB0000182	5962	C00047	C(CCN)C[C@@H](C(=O)O)N
21	Sucrose	Sucrose	HMDB0000258	5988	C00089	C([C@@H]1[C@H]([C@@H]([C@H]([C
22	Aminoadipic acid	Aminoadipic acid	HMDB0000510	469	C00956	C(CC(C(=O)O)N)CC(=O)O
23	Uric acid	Uric acid	HMDB0000289	1175	C00366	C12=C(NC(=O)N1)NC(=O)NC2=O
24	Threonine	L-Threonine	HMDB0000167	6288	C00188	C[C@H]([C@@H](C(=O)O)N)O
25	Malate	L-Malic acid	HMDB0000156	222656	C00149	C([C@@H](C(=O)O)O)C(=O)O
26	Phosphocholine	Phosphorylcholine	HMDB0001565	8691	C00588	C[N+](C)(C)CCOP(=O)(O)O
27	Citrate	Citric acid	HMDB0000094	311	C00158	C(C(=O)O)C(CC(=O)O)(C(=O)O)O
28	Hydroxyproline	4-Hydroxyproline	HMDB0000725	5810	C01157	C1[C@H](CN[C@@H]1C(=O)O)O
29	Raffinose	Raffinose	HMDB0003213	10542	C00492	C([C@@H]1[C@@H]([C@@H]([C@H](

4 Pathway Analysis

In this step, users are asked to select a pathway library, as well as specify the algorithms for pathway enrichment analysis and pathway topology analysis.

4.1 Pathway Library

There are 15 pathway libraries currently supported, with a total of 1173 pathways :

- Homo sapiens (human) [80]
- Mus musculus (mouse) [82]
- Rattus norvegicus (rat) [81]
- Bos taurus (cow) [81]
- Danio rerio (zebrafish) [81]
- Drosophila melanogaster (fruit fly) [79]
- Caenorhabditis elegans (nematode) [78]
- Saccharomyces cerevisiae (yeast) [65]
- Oryza sativa japonica (Japanese rice) [83]
- Arabidopsis thaliana (thale cress) [87]
- Escherichia coli K-12 MG1655 [87]
- Bacillus subtilis [80]
- Pseudomonas putida KT2440 [89]
- Staphylococcus aureus N315 (MRSA/VSSA)[73]
- Thermotoga maritima [57]

Your selected pathway library code is **cel** (KEGG organisms abbreviation).

4.2 Over Representation Analysis

Over-representation analysis tests if a particular group of compounds is represented more than expected by chance within the user uploaded compound list. In the context of pathway analysis, we are testing if compounds involved in a particular pathway are enriched compared to random hits. MetPA offers two of the most commonly used methods for over-representation analysis:

- Fishers'Exact test
- Hypergeometric Test

Please note, MetPA uses one-tailed Fisher's exact test which will give essentially the same result as the result calculated by the hypergeometric test.

The selected over-representation analysis method is **Fishers' exact test**.

4.3 Pathway Topology Analysis

The structure of biological pathways represent our knowledge about the complex relationships among molecules within a cell or a living organism. However, most pathway analysis algorithms fail to take structural information into consideration when estimating which pathways are significantly changed under conditions of study. It is well-known that changes in more important positions of a network will trigger a more severe impact on the pathway than changes occurred in marginal or relatively isolated positions.

The pathway topology analysis uses two well-established node centrality measures to estimate node importance - **degree centrality** and **betweenness centrality**. Degree centrality is defined as the number of links occurred upon a node. For a directed graph there are two types of degree: in-degree for links come from other nodes, and out-degree for links initiated from the current node. Metabolic networks are directed graph. Here we only consider the out-degree for node importance measure. It is assumed that nodes upstream will have regulatory roles for the downstream nodes, not vice versa. The betweenness centrality measures the number of shortest paths going through the node. Since the metabolic network is directed, we use the relative betweenness centrality for a metabolite as the importance measure. The degree centrality measure focuses more on local connectivities, while the betweenness centrality measure focuses more on global network topology. For more detailed discussions on various graph-based methods for analyzing biological networks, please refer to the article by Tero Aittokallio, T. et al. [1]

Please note, for comparison among different pathways, the node importance values calculated from centrality measures are further normalized by the sum of the importance of the pathway. Therefore, the total/maximum importance of each pathway is 1; the importance measure of each metabolite node is actually the percentage w.r.t the total pathway importance, and the pathway impact value is the cumulative percentage from the matched metabolite nodes.

Your selected node importance measure for topological analysis is **relative betweenness centrality**.

5 Pathway Analysis Result

The results from pathway analysis are presented graphically as well as in a detailed table.

A Google-map style interactive visualization system was implemented to facilitate data exploration. The graphical output contains three levels of view: **metabolome view**, **pathway view**, and **compound view**. Only the metabolome view is shown below. Pathway views and compound views are generated dynamically based on your interactions with the visualization system. They are available in your downloaded files.

[1] Tero Aittokallio and Benno Schwikowski. *Graph-based methods for analyzing networks in cell biology*, Briefings in Bioinformatics 2006 7(3):243-255

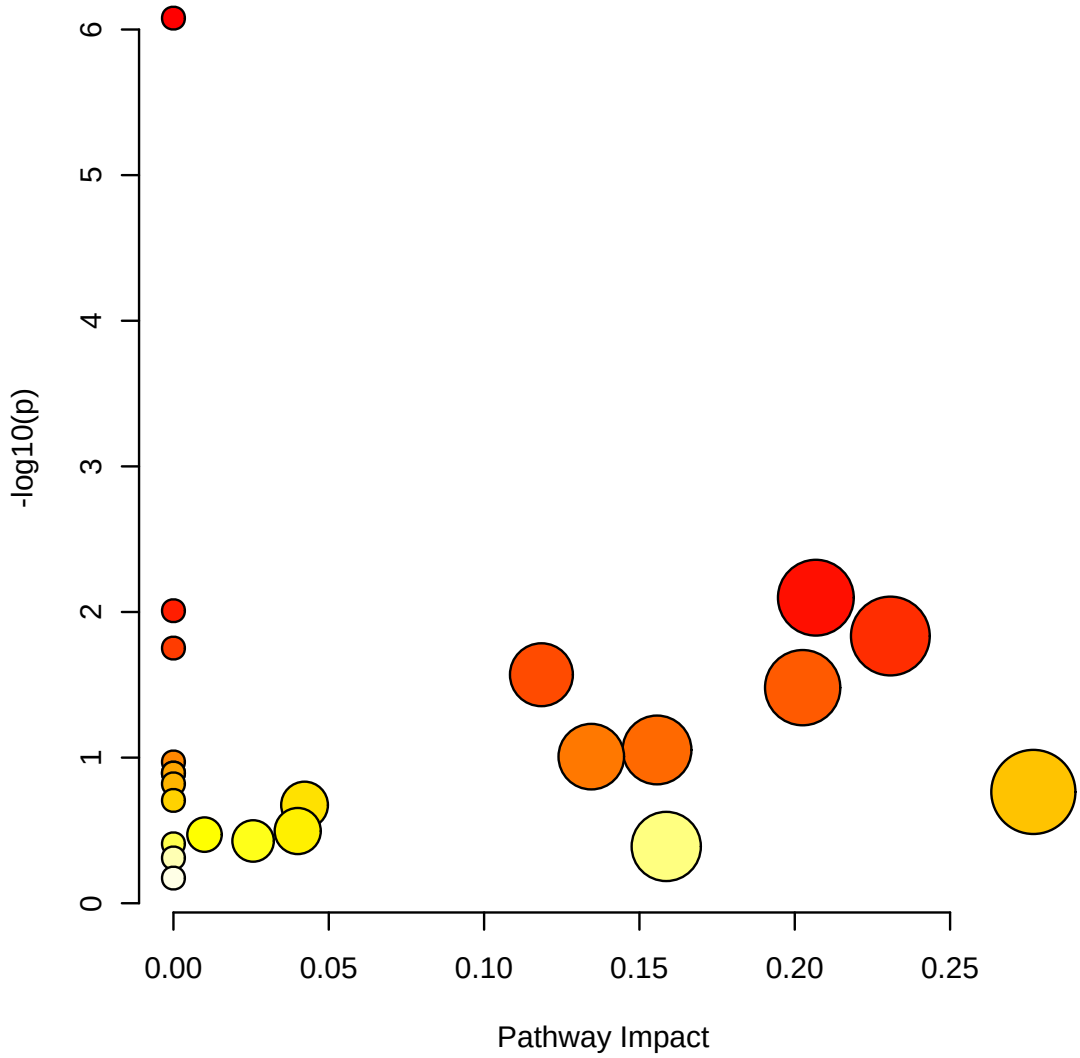

Figure 1: Summary of Pathway Analysis

The table below shows the detailed results from the pathway analysis. Since we are testing many pathways at the same time, the statistical p values from enrichment analysis are further adjusted for multiple testings. In particular, the **Total** is the total number of compounds in the pathway; the **Hits** is the actually matched number from the user uploaded data; the **Raw p** is the original p value calculated from the enrichment analysis; the **Holm p** is the p value adjusted by Holm-Bonferroni method; the **FDR p** is the p value adjusted using False Discovery Rate; the **Impact** is the pathway impact value calculated from pathway topology analysis.

Table 2: Result from Pathway Analysis

	Total	Expected	Hits	Raw p	-log10(p)	Holm adjust	FDR	Impact
Aminoacyl-tRNA biosynthesis	45	1.21	9	8.35E-07	6.08E+00	6.76E-05	6.76E-05	0.00
Glyoxylate and dicarboxylate metabolism	31	0.83	4	7.97E-03	2.10E+00	6.38E-01	2.65E-01	0.21
Arginine biosynthesis	6	0.16	2	9.80E-03	2.01E+00	7.74E-01	2.65E-01	0.00
Alanine, aspartate and glutamate metabolism	20	0.54	3	1.46E-02	1.83E+00	1.00E+00	2.87E-01	0.23
Valine, leucine and isoleucine biosynthesis	8	0.22	2	1.77E-02	1.75E+00	1.00E+00	2.87E-01	0.00
Glutathione metabolism	25	0.67	3	2.70E-02	1.57E+00	1.00E+00	3.64E-01	0.12
Arginine and proline metabolism	27	0.73	3	3.31E-02	1.48E+00	1.00E+00	3.83E-01	0.20
Pyrimidine metabolism	40	1.08	3	8.86E-02	1.05E+00	1.00E+00	8.63E-01	0.16
Citrate cycle (TCA cycle)	20	0.54	2	9.85E-02	1.01E+00	1.00E+00	8.63E-01	0.13
Lysine degradation	21	0.57	2	1.07E-01	9.70E-01	1.00E+00	8.63E-01	0.00
Phosphonate and phosphinate metabolism	5	0.13	1	1.28E-01	8.94E-01	1.00E+00	8.63E-01	0.00
D-Glutamine and D-glutamate metabolism	5	0.13	1	1.28E-01	8.94E-01	1.00E+00	8.63E-01	0.00
Nitrogen metabolism	6	0.16	1	1.51E-01	8.20E-01	1.00E+00	9.43E-01	0.00
Glycine, serine and threonine metabolism	28	0.75	2	1.72E-01	7.64E-01	1.00E+00	9.95E-01	0.28
Histidine metabolism	8	0.22	1	1.97E-01	7.06E-01	1.00E+00	1.00E+00	0.00
Glycerophospholipid metabolism	32	0.86	2	2.12E-01	6.74E-01	1.00E+00	1.00E+00	0.04
Starch and sucrose metabolism	14	0.38	1	3.19E-01	4.96E-01	1.00E+00	1.00E+00	0.04
Galactose metabolism	15	0.40	1	3.38E-01	4.71E-01	1.00E+00	1.00E+00	0.01
Sphingolipid metabolism	17	0.46	1	3.74E-01	4.28E-01	1.00E+00	1.00E+00	0.03
Selenocompound metabolism	18	0.48	1	3.91E-01	4.08E-01	1.00E+00	1.00E+00	0.00
Pyruvate metabolism	19	0.51	1	4.07E-01	3.90E-01	1.00E+00	1.00E+00	0.16
Purine metabolism	60	1.62	2	4.89E-01	3.11E-01	1.00E+00	1.00E+00	0.00
Valine, leucine and isoleucine degradation	40	1.08	1	6.72E-01	1.73E-01	1.00E+00	1.00E+00	0.00

6 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"pathora\", FALSE)"
 [2] "cmpd.vec<-c(\"Glutamine\",\"Asparagine\",\"Alanine\",\"Dihydroorotate\",\"Ornithine\",\"NADPH\
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"name\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-PerformDetailMatch(mSet, \"Acetyl Proline\");"
 [7] "mSet<-GetCandidateList(mSet);"
 [8] "mSet<-PerformDetailMatch(mSet, \"NG-dimethyl-L-arginine\");"
 [9] "mSet<-GetCandidateList(mSet);"
[10] "mSet<-SetCandidate(mSet, \"NG-dimethyl-L-arginine\", \"Asymmetric dimethylarginine\");"
[11] "mSet<-PerformDetailMatch(mSet, \"Pyrroline-5-carboxylic acid\");"
[12] "mSet<-GetCandidateList(mSet);"
[13] "mSet<-SetCandidate(mSet, \"Pyrroline-5-carboxylic acid\", \"1-Pyrroline-5-carboxylic acid\");"
[14] "mSet<-SetKEGG.PathLib(mSet, \"cel\", \"current\")"
[15] "mSet<-SetMetabolomeFilter(mSet, F);"
[16] "mSet<-CalculateOraScore(mSet, \"rbc\", \"fisher\")"
[17] "mSet<-PlotPathSummary(mSet, F, \"path_view_4_\", \"png\", 72, width=NA)"
[18] "mSet<-PlotKEGGPath(mSet, \"Glycine, serine and threonine metabolism\",576, 480, \"png\", NULL)
[19] "mSet<-RerenderMetPAGraph(mSet, \"zoom1633471015107.png\",576.0, 480.0, 100.0)"
[20] "mSet<-PlotKEGGPath(mSet, \"Glycine, serine and threonine metabolism\",576, 480, \"png\", NULL)
[21] "mSet<-SaveTransformedData(mSet)"
[22] "mSet<-PreparePDFReport(mSet, \"guest8993539346205070843\")\n"
```

The report was generated on Tue Oct 5 17:57:23 2021 with R version 4.0.2 (2020-06-22).