Metabolomic Data Analysis with MetaboAnalyst 6.0

Name: guest5546627935843046589

October 7, 2024

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	1-carboxyethylleucine	NA	NA	NA	NA	NA
2	1-methylguanine	1-Methylguanine	HMDB0003282	70315	C04152	CN1C(N)=NC2=C(
3	1-ribosyl-imidazoleacetate	Imidazoleacetic acid riboside	HMDB0002331	440569	C05131	OC[C@H]1O[C@H](
4	2-keto-3-deoxy-gluconate	2-Keto-3-deoxy-D-gluconic acid	HMDB0001353	194024	C01216	OC[C@@H](O)[C@
5	2-methylhexanoylcarnitine	NA	NA	NA	NA	NA
6	3-hydroxyadipate	3-Hydroxyadipic acid	HMDB0000345	151913		OC(CCC(O)=O)CC
7	3-hydroxyphenylacetoylglutamine	4-Hydroxyphenylacetylglutamic acid	HMDB0006061	61152160	C05595	OC(=O)CC[C@H](
8	3-hydroxyproline	NA	NA	NA	NA	NA
9	4-allylphenol sulfate	4-Vinylphenol sulfate	HMDB0062775	6426766		OS(=O)(=O)OC1=

10	4-hydroxyglutamate	4-Hydroxy-L-glutamic acid	HMDB0002273	440854	C05947	N[C@@H](C[C@@H
11	adenosine	Adenosine	HMDB0000050	60961	C00212	NC1=C2N=CN([C
12	alpha-ketoglutarate	Oxoglutaric acid	HMDB0000208	51	C00026	OC(=O)CCC(=O)C
13	asparagine	L-Asparagine	HMDB0000168	6267	C00152	N[C@@H](CC(N)=
14	choline	Choline	HMDB0000097	305	C00114	C[N+](C)(C)CCO
15	citrulline	Citrulline	HMDB0000904	9750	C00327	N[C@@H](CCCNC(
16	cystathionine	L-Cystathionine	HMDB0000099	439258	C02291	N[C@@H](CCSC[C
17	cysteinylglycine disulfide	L-Cysteinylglycine disulfide	HMDB0000709	22833544		N[C@@H](CSSCC(
18	dimethylglycine	Dimethylglycine	HMDB0000092	673	C01026	CN(C)CC(O)=O
19	enterolactone sulfate	NA	NA	NA	NA	NA
20	etiocholanolone glucuronide	Etiocholanolone glucuronide	HMDB0004484	443078	C11136	[H][C@@]12CCC(=
21	glucose 6-phosphate	Glucose 6-phosphate	HMDB0001401	5958	C00092	OC1O[C@H](COP(
22	glucuronide of C8H14O2 (3)	NA	NA	NA	NA	NA
23	glucuronide of C8H18O2 (2)	NA	NA	NA	NA	NA
24	glutamine conjugate of C9H16O2 (1)	NA	NA	NA	NA	NA
25	guaiacol sulfate	O-methoxycatechol-O-sulphate	HMDB0060013	22473		COC1=CC=CC=C
26	guanosine	Guanosine	HMDB0000133	6802	C00387	NC1=NC2=C(N=C
27	homoarginine	Homo-L-arginine	HMDB0000670	9085	C01924	N[C@@H](CCCCNC
28	kynurenine	Kynurenine	HMDB0000684	161166	C00328	N[C@@H](CC(=O)
29	lanthionine	Lanthionine	HMDB0240656	98504		N[C@@H](CSC[C@
30	levulinoylcarnitine	NA	NA	NA	NA	NA
31	methylurea	N-Methylurea	METPA1296		C16363	
32	N2-acetyl,N6,N6-dimethyllysine	NA	NA	NA	NA	NA
33	N4-acetylcytidine	N4-Acetylcytidine	HMDB0005923	107461		CC(=O)NC1=NC(
34	oxindolylalanine	NA	NA	NA	NA	NA
35	prolylglycine	Prolylglycine	HMDB0011178	6426709		OC(=O)CNC(=O)[
36	tartarate	Tartaric acid	HMDB0000956	444305	C00898	O[C@H]([C@@H](O
37	taurine	Taurine	HMDB0000251	1123	C00245	NCCS(O)(=O)=O
38	tryptamine	Tryptamine	HMDB0000303	1150	C00398	NCCC1=CNC2=C1
39	uric acid ribonucleoside	beta-D-3-Ribofuranosyluric acid	HMDB0029920	131750925	C05513	OCC1OC(C(O)C1C
40	ursocholate	Ursocholic acid	HMDB0000917	122340	C17644	[H][C@@]12CC[C@

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 **hsa** (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 'Hypergeometric 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

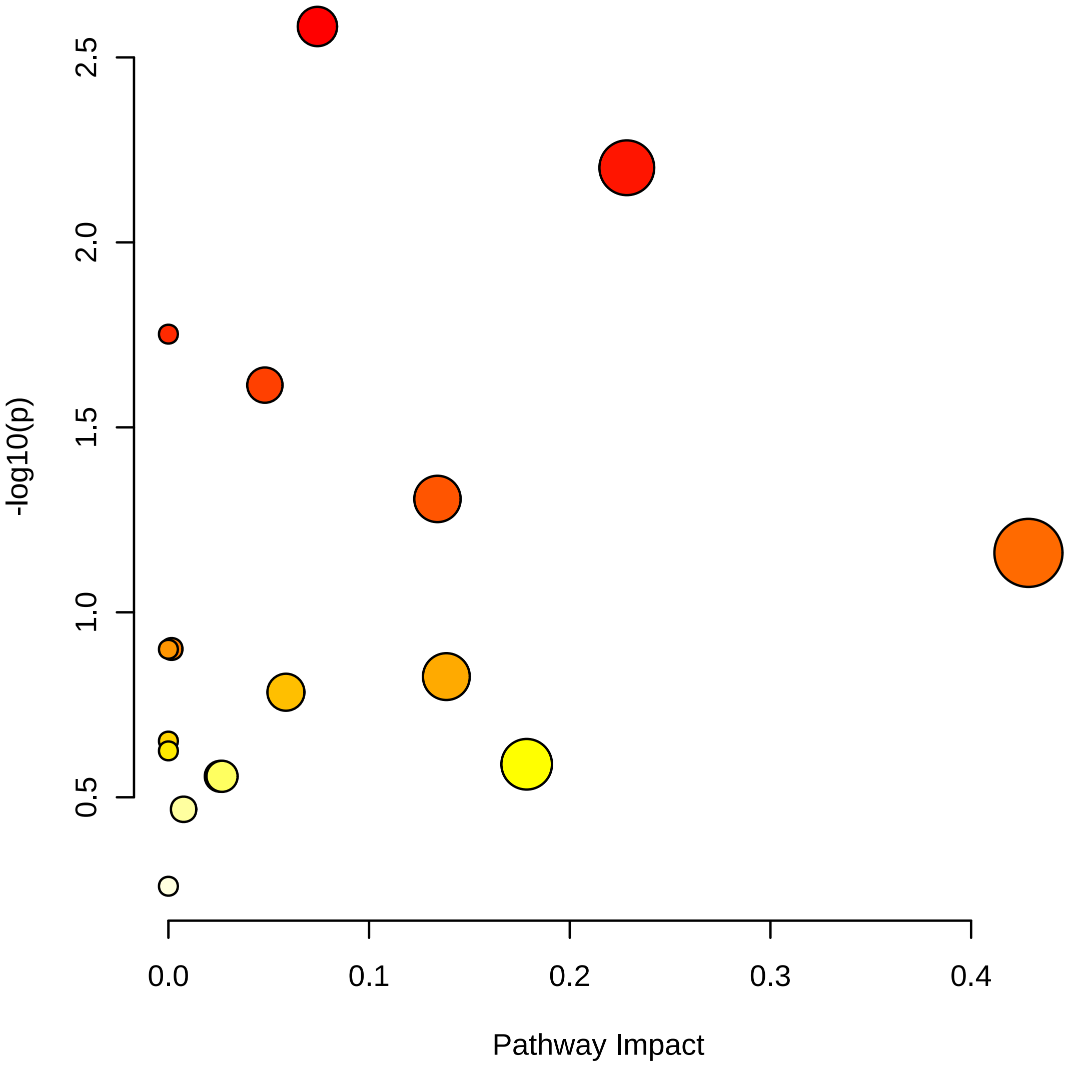

Figure 1: Summary of Pathway Analysis

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
Glycine, serine and threonine metabolism	33	0.29	3	2.61E-03	2.58E+00	2.09E-01	2.09E-01	0.07
Arginine biosynthesis	14	0.12	2	6.28E-03	2.20E+00	4.96E-01	2.51E-01	0.23
Neomycin, kanamycin and gentamicin biosynthesis	2	0.02	1	1.77E-02	1.75E+00	1.00E+00	4.72E-01	0.00
Alanine, aspartate and glutamate metabolism	28	0.25	2	2.43E-02	1.61E+00	1.00E+00	4.86E-01	0.05
Tryptophan metabolism	41	0.36	2	4.94E-02	1.31E+00	1.00E+00	7.90E-01	0.13
Taurine and hypotaurine metabolism	8	0.07	1	6.91E-02	1.16E+00	1.00E+00	9.21E-01	0.43
Purine metabolism	70	0.62	2	1.26E-01	9.01E-01	1.00E+00	1.00E+00	0.00
Butanoate metabolism	15	0.13	1	1.26E-01	9.00E-01	1.00E+00	1.00E+00	0.00
Starch and sucrose metabolism	18	0.16	1	1.49E-01	8.26E-01	1.00E+00	1.00E+00	0.14
Citrate cycle (TCA cycle)	20	0.18	1	1.64E-01	7.84E-01	1.00E+00	1.00E+00	0.06
Lipoic acid metabolism	28	0.25	1	2.23E-01	6.52E-01	1.00E+00	1.00E+00	0.00
Inositol phosphate metabolism	30	0.27	1	2.37E-01	6.25E-01	1.00E+00	1.00E+00	0.00
Cysteine and methionine metabolism	33	0.29	1	2.57E-01	5.89E-01	1.00E+00	1.00E+00	0.18
Glycerophospholipid metabolism	36	0.32	1	2.78E-01	5.57E-01	1.00E+00	1.00E+00	0.03
Arginine and proline metabolism	36	0.32	1	2.78E-01	5.57E-01	1.00E+00	1.00E+00	0.03
Primary bile acid biosynthesis	46	0.41	1	3.41E-01	4.68E-01	1.00E+00	1.00E+00	0.01
Steroid hormone biosynthesis	87	0.77	1	5.50E-01	2.59E-01	1.00E+00	1.00E+00	0.00

6 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"pathora\", FALSE)"
 [2] "cmpd.vec<-c(\"1-carboxyethylleucine\",\"1-methylguanine\",\"1-ribosyl-imidazoleacetate\",\"2-k
 [3] "mSet<-Setup.MapData(mSet, cmpd.vec);"
 [4] "mSet<-CrossReferencing(mSet, \"name\");"
 [5] "mSet<-CreateMappingResultTable(mSet)"
 [6] "mSet<-PerformDetailMatch(mSet, \"1-ribosyl-imidazoleacetate\");"
 [7] "mSet<-GetCandidateList(mSet);"
 [8] "mSet<-SetCandidate(mSet, \"1-ribosyl-imidazoleacetate\", \"Imidazoleacetic acid riboside\");"
 [9] "mSet<-PerformDetailMatch(mSet, \"2-keto-3-deoxy-gluconate\");"
[10] "mSet<-GetCandidateList(mSet);"
[11] "mSet<-SetCandidate(mSet, \"2-keto-3-deoxy-gluconate\", \"2-Keto-3-deoxy-D-gluconic acid\");"
[12] "mSet<-PerformDetailMatch(mSet, \"2-methylhexanoylcarnitine\");"
[13] "mSet<-GetCandidateList(mSet);"
[14] "mSet<-PerformDetailMatch(mSet, \"4-allylphenol sulfate\");"
[15] "mSet<-GetCandidateList(mSet);"
[16] "mSet<-SetCandidate(mSet, \"4-allylphenol sulfate\", \"4-Vinylphenol sulfate\");"
[17] "mSet<-PerformDetailMatch(mSet, \"3-hydroxyproline\");"
[18] "mSet<-GetCandidateList(mSet);"
[19] "mSet<-PerformDetailMatch(mSet, \"3-hydroxyphenylacetoylglutamine\");"
[20] "mSet<-GetCandidateList(mSet);"
[21] "mSet<-SetCandidate(mSet, \"3-hydroxyphenylacetoylglutamine\", \"4-Hydroxyphenylacetylglutamic
[22] "mSet<-PerformDetailMatch(mSet, \"cysteinylglycine disulfide\");"
[23] "mSet<-GetCandidateList(mSet);"
[24] "mSet<-SetCandidate(mSet, \"cysteinylglycine disulfide\", \"L-Cysteinylglycine disulfide\");"
[25] "mSet<-PerformDetailMatch(mSet, \"enterolactone sulfate\");"
[26] "mSet<-GetCandidateList(mSet);"
[27] "mSet<-PerformDetailMatch(mSet, \"guaiacol sulfate\");"
[28] "mSet<-GetCandidateList(mSet);"
[29] "mSet<-SetCandidate(mSet, \"guaiacol sulfate\", \"O-methoxycatechol-O-sulphate\");"
[30] "mSet<-PerformDetailMatch(mSet, \"levulinoylcarnitine\");"
[31] "mSet<-GetCandidateList(mSet);"
[32] "mSet<-PerformDetailMatch(mSet, \"N2-acetyl,N6,N6-dimethyllysine\");"
[33] "mSet<-GetCandidateList(mSet);"
[34] "mSet<-PerformDetailMatch(mSet, \"oxindolylalanine\");"
[35] "mSet<-GetCandidateList(mSet);"
[36] "mSet<-SetKEGG.PathLib(mSet, \"hsa\", \"current\")"
[37] "mSet<-SetMetabolomeFilter(mSet, F);"
[38] "mSet<-CalculateOraScore(mSet, \"rbc\", \"hyperg\")"
[39] "mSet<-PlotPathSummary(mSet, F, \"path_view_0_\", \"png\", 72, width=NA, NA, NA )"
[40] "mSet<-SaveTransformedData(mSet)"
[41] "mSet<-PreparePDFReport(mSet, \"guest5546627935843046589\")\n"
```

The report was generated on Mon Oct 7 16:20:44 2024 with R version 4.2.2 (2022-10-31), OS system: Linux.