Metabolomic Data Analysis with MetaboAnalyst 4.0

Name: guest6387830825194072215

October 7, 2019

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 Compound Name Mapping

	Query	Match	HMDB	PubChem	KEGG	SMILES
1	C00263	L-Homoserine	HMDB0000719	12647	C00263	C(CO)[C@@H](C(=O)O)N
2	C19636	Turanose	HMDB0011740	5460935	C19636	C([C@@H]1[C@H]([C@@H]([C@H]([C@H](O1)O[C@@H]([C@@H]([C@@H
3	C00159	D-Mannose	HMDB0000169	18950	C00159	C([C@@H]1[C@H]([C@@H]([C@@H](C(O1)O)O)O)O)O
4	C05402	Melibiose	HMDB0000048	440658	C05402	C([C@@H]1[C@@H]([C@@H]([C@H]([C@H](O1)OC[C@@H]2[C@H]([C@@
5	C01384	Maleic acid	HMDB0000176	444266	C01384	C(=C\C(=O)O)\C(=O)O

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 **ath** (KEGG organisms abbreviation).

4.2 Pathway Enrichment Analysis

Pathway enrichment analysis usually refers to quantitative enrichment analysis directly using the compound concentration values, as compared to compound lists used by over-representation analysis. As a result, it is more sensitive and has the potential to identify **subtle but consistent** changes amongst compounds involved in the same biological pathway.

Many procedures have been developed in the last decade for quantitative enrichment analysis, the most famous being the Gene Set Enrichment Analysis. Many new and improved methods have been implemented since. The enrichment analysis is based on GlobalTest and GlobalAncova. Both methods support enrichment analysis with binary, multi-group, as well as continuous phenotypes. The p-values can be approximated based on the asymptotic distribution without using permutations which is computationally very intensive and is not suitable for web applications. Please note, when sample sizes are small, the approximated p values may be slightly less accurate compared to p values obtained by using a permutation-based method (for details, please refer to the paper by Goeman, J.J. et al. [1] and by

[1] Jelle J. Goeman and Peter Buhlmann. *Analyzing gene expression data in terms of gene sets: methodological issues*, Bioinformatics 2007 23(8):980-987

Hummel, M. et al. [2]) However, since our focus is to identify the most relevant pathways within the pathways in the library, we are more interested in the rank of the pathway, not its absolute p-value. Therefore, this disadvantage may be tolerated.

The selected pathway enrichment analysis method is **Globaltest**.

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. [3]

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.

[2]Manuela Hummel, Reinhard Meister and Ulrich Mansmann. *GlobalANCOVA: exploration and assessment of gene group effects*, Bioinformatics 2008 24(1):78-85

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

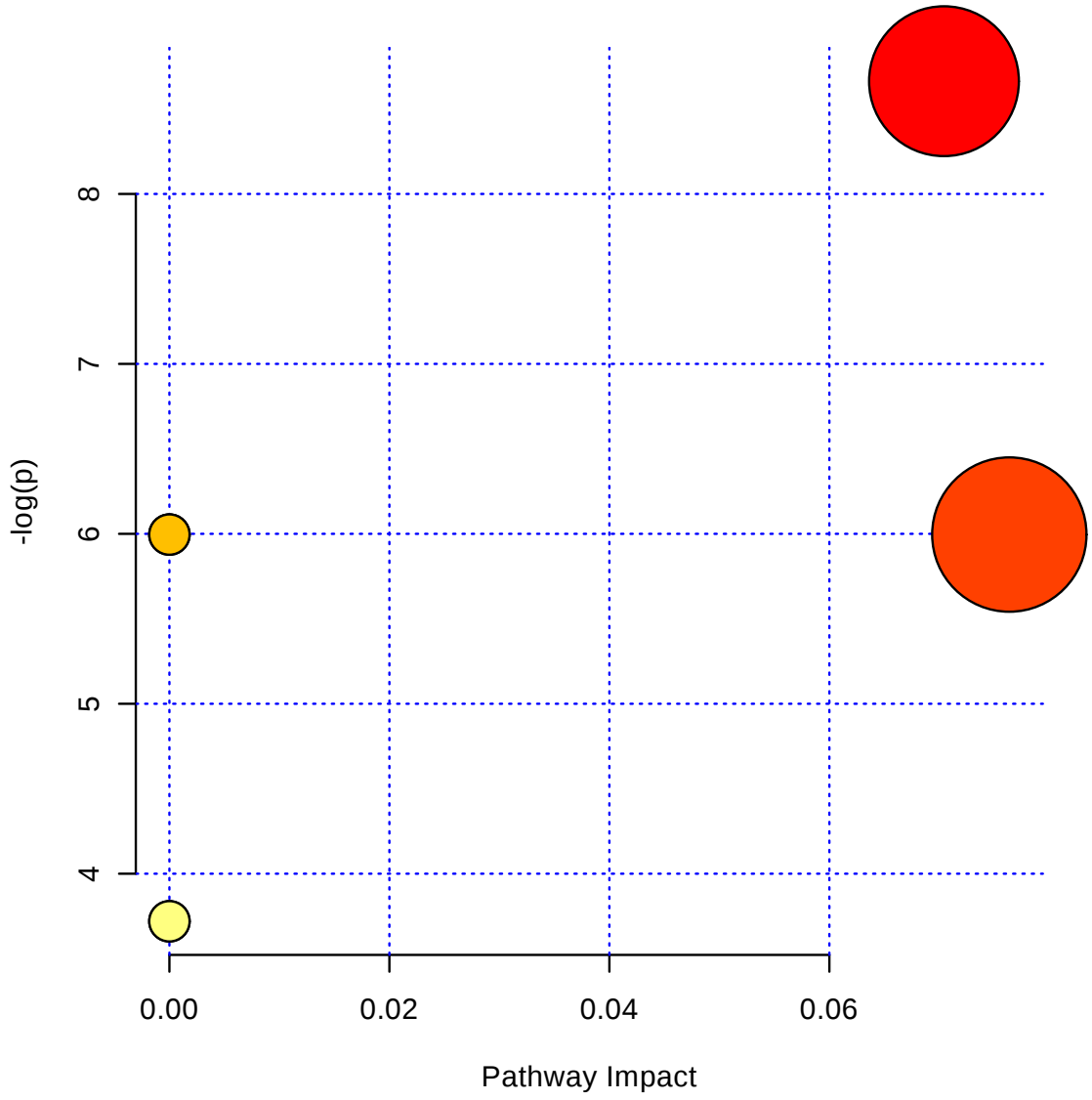

Figure 1: Summary of Pathway Analysis

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 Cmpd	Hits	Raw p	-log(p)	Holm adjust	FDR	Impact
Galactose metabolism	26	2	1.73E-04	8.66E+00	1.04E-03	1.04E-03	0.07
Glycine, serine and threonine metabolism	30	1	2.49E-03	6.00E+00	1.24E-02	3.73E-03	0.08
Cysteine and methionine metabolism	34	1	2.49E-03	6.00E+00	1.24E-02	3.73E-03	0.00
Lysine biosynthesis	10	1	2.49E-03	6.00E+00	1.24E-02	3.73E-03	0.00
Fructose and mannose metabolism	16	1	2.42E-02	3.72E+00	4.85E-02	2.42E-02	0.00
Amino sugar and nucleotide sugar metabolism	41	1	2.42E-02	3.72E+00	4.85E-02	2.42E-02	0.00

6 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"pathqea\", FALSE)"
 [2] "mSet<-Read.TextData(mSet, \"Replacing_with_your_file_path\", \"rowu\", \"disc\");"
 [3] "mSet<-CrossReferencing(mSet, \"kegg\");"
 [4] "mSet<-CreateMappingResultTable(mSet)"
 [5] "mSet<-SanityCheckData(mSet)"
 [6] "mSet<-ImputeVar(mSet, method=\"min\")"
 [7] "mSet<-PreparePrenormData(mSet)"
 [8] "mSet<-Normalization(mSet, \"NULL\", \"NULL\", \"ParetoNorm\", ratio=FALSE, ratioNum=20)"
 [9] "mSet<-PlotNormSummary(mSet, \"norm_0_\", \"png\", 72, width=NA)"
[10] "mSet<-PlotSampleNormSummary(mSet, \"snorm_0_\", \"png\", 72, width=NA)"
[11] "mSet<-SetKEGG.PathLib(mSet, \"ath\")"
[12] "mSet<-SetMetabolomeFilter(mSet, F);"
[13] "mSet<-CalculateQeaScore(mSet, \"rbc\", \"gt\")"
[14] "mSet<-PlotPathSummary(mSet, \"path_view_0_\", \"png\", 72, width=NA)"
[15] "mSet<-PlotKEGGPath(mSet, \"Galactose metabolism\",528, 480, \"png\", NULL)"
[16] "mSet<-RerenderMetPAGraph(mSet, \"zoom1570467370948.png\",528.0, 480.0, 100.0)"
[17] "mSet<-PlotKEGGPath(mSet, \"Glycine, serine and threonine metabolism\",528, 480, \"png\", NULL)
[18] "mSet<-SaveTransformedData(mSet)"
[19] "mSet<-PreparePDFReport(mSet, \"guest6387830825194072215\")\n"
```

The report was generated on Mon Oct 7 16:56:58 2019 with R version 3.5.1 (2018-07-02).