

## Functional Analysis of Metabolomics Data

Mónica Chagoyen, Javier López-Ibáñez, and Florencio Pazos

### Abstract

Metabolomics aims at characterizing the repertory of small chemical compounds in a biological sample. As it becomes more massive and larger sets of compounds are detected, a functional analysis is required to convert these raw lists of compounds into biological knowledge. The most common way of performing such analysis is “annotation enrichment analysis,” also used in transcriptomics and proteomics. This approach extracts the annotations overrepresented in the set of chemical compounds arisen in a given experiment. Here, we describe the protocols for performing such analysis as well as for visualizing a set of compounds in different representations of the metabolic networks, in both cases using free accessible web tools.

**Key words** Metabolomics, Metabolic pathway, Metabolite, Functional enrichment, Metabolism, Bioinformatics

---

### 1 Introduction

The so-called omics technologies aim at characterizing, in a high-throughput way, the whole repertoires of different types of molecules in biological samples. Within the main omics technologies, we can cite genomics (the characterization of the gene content of an organism/sample), transcriptomics (the characterization of expression levels, generally of mRNAs), proteomics (characterization of the repertory of translated proteins), and metabolomics (characterization of the repertory of small molecules) [1]. These approaches complement each other since genes, mRNAs, proteins, and metabolites represent different, albeit somehow related, levels of the cellular complexity.

A common characteristic of these approaches is that, in general, the results they produce (i.e., long lists of expressed genes or identified proteins in a given sample) need some sort of “post-processing” in order to extract useful information from them. This is called “biological/functional analysis,” or “secondary analysis” to distinguish it from the “primary analysis” aimed at processing the original “raw” data of the experiment (e.g., intensity values, sequence reads, spectral peaks) so as to obtain the list of genes/proteins. This secondary

analysis can translate, for example, a long list of hundreds of genes, without an evident meaning by itself, into a reduced list of 2–5 biological pathways (those enriched in the genes/proteins of the original list) that do have a biological meaning for the researcher. Indeed, the most common form of secondary analysis of transcriptomics and proteomics data is called “annotation enrichment analysis” [2]. While there are tens of tools and web servers for performing enrichment analysis of transcriptomics and metabolomics data, the number of tools for performing such analysis over metabolomics data is much lower [3]. In part, this is due to the fact that metabolomics was one of the latest comers to the omics club. But another reason is that it is not as massive as its other omics counterparts, and in many cases, metabolomics experiments are targeted to the identification of a relatively low number of metabolites, and hence secondary analysis is not mandatory. But as metabolomics workflows are able to identify more and more metabolites, these analyses become more important. The goal of metabolomics functional analysis is the same as in transcriptomics: convert a long list of metabolites showing up in a given experiment into a reduced set of meaningful biological terms, such as the pathways/biological processes enriched in them. Consequently, the methodologies for performing this analysis are also the same: these generally look for keywords (i.e., pathway names, functional groups, associated genes, diseases) significantly overrepresented (according to some statistical test) in the set of metabolites with respect to a background set. In the case of metabolomics, this background set is also problematic since while in other omics it is naturally given by the gene content of the organism of interest or the set of genes assayed (e.g., those on the chip), the whole set of metabolites “used” by a given organism is not known.

Another way of interactively and qualitatively inferring the pathways or metabolic context of a set of metabolites is simply to visualize them in a representation of the metabolic network. In this way, one can easily grasp whether these compounds are clustered together and if so, in which pathways; infer other related metabolites not detected in the experiment, etc.

In the following, we describe in detail the protocols for using a freely available web server for performing functional (enrichment) analysis of metabolomics data. We also describe two other servers which allow visualizing a set of metabolites entered by the user in different representations of metabolic networks. Together, these tools allow obtaining functional knowledge from a raw list of metabolites coming from a metabolomics experiment.

---

## 2 Methods

This chapter explains how to analyze the biological context of a set of compounds, typically obtained in a metabolomics experiment, through the use of three web-accessible computational tools:

Interactive Pathways Explorer (iPath) [4]: <http://pathways.embl.de>.  
KEGG PATHWAY Database [5]: <http://www.genome.jp/kegg/pathway.html>.

Metabolites Biological Role (MBRole) [6]: <http://csbg.cnb.csic.es/mbrole>.

## 2.1 Data Preparation

The main input for our analysis is a set of compounds, given by their identifiers (IDs) in some database. In this chapter, we will use KEGG compound IDs ([www.genome.jp/kegg/compound/](http://www.genome.jp/kegg/compound/)) to perform the analysis. If you do not know the KEGG IDs of your compounds, you can use a compound ID conversion tool, like the Batch Conversion of the Chemical Translation Service ([cts.fiehnlab.ucdavis.edu](http://cts.fiehnlab.ucdavis.edu)), or the ID conversion utility of the MBRole server ([csbg.cnb.csic.es/mbrole](http://csbg.cnb.csic.es/mbrole)). See **Note 1** on how to share metabolomics results.

## 2.2 Pathway Mapping and Visualization

### 2.2.1 Global View (with iPath)

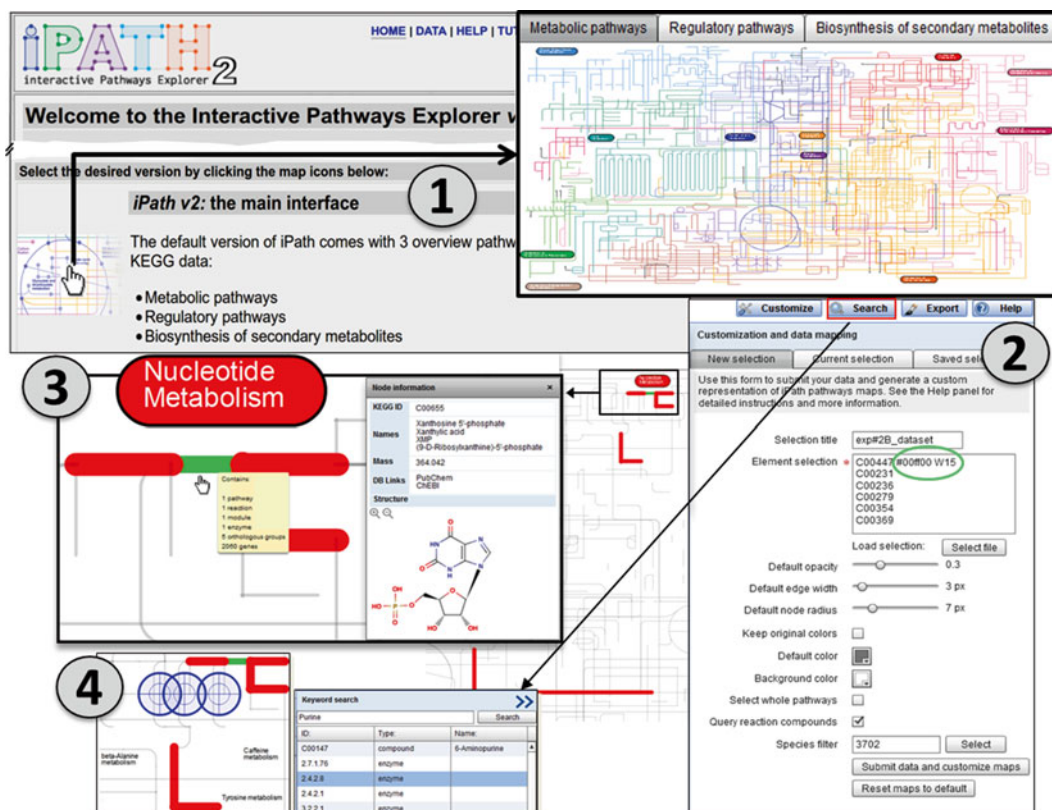
iPath allows visualizing a set of chemical compounds in the context of three global pathway maps: “metabolic pathways,” “regulatory pathways,” and “biosynthesis of secondary metabolites.” These global pathway maps are based on the information provided by KEGG.

### 2.2.2 Simple Visualization

To open the iPath interface, point your web browser to the iPath website ([pathways.embl.de](http://pathways.embl.de)) and click on the image next to the “iPath v2: the main interface” label (Fig. 1). A global pathway map appears that corresponds to the “Metabolic pathways” section. You can navigate to the other two sections (“Regulatory pathways” and “Biosynthesis of secondary metabolites”) by selecting the corresponding tabs at the top of the pathway map (Fig. 1).

To highlight your compounds of interest in the pathway maps, click on the “Customize” button at the top right corner. In the form which shows up (Fig. 1), enter the following data:

- Write a name in the “Selection title” to identify your set of compounds (e.g., name of the experiment). First paste the list of compound IDs in the “Element selection” or, alternatively, load a file containing this list (by clicking the “Select file” button).
- Activate the “Query reaction compounds” checkbox.
- Optionally, you can restrict your analysis to the pathways of a particular organism. Do it by entering the NCBI taxonomy ID or the KEGG three-letter code of your organism in the “Species filter” field. If you don’t know this information, you can search by organism name by clicking the “Select” button. A “Species search” window will appear. Write the name of the organism, and select from the list of matches. The NCBI taxonomy ID of the selected organism will appear on the Species filter. Close the “Species search” window.
- Click on “Submit data and customize maps.” Now, the compounds entered are highlighted in the pathway map, by default as thick red lines marking the reactions they are involved in (Fig. 1). You can zoom in, zoom out, and drag the pathway



**Fig. 1** Screenshots of the iPath system. (1) Main entry page taking to zoomable/navigable global maps. (2) The “Customize” tab contains the main form for entering the list of compounds that are going to be highlighted in the maps (3). This list can include codes associated to the individual compounds to differentially change their color, line width, etc. (e.g., green circle). The “Search” tab allows to look for items in the maps (4)

map using the gray navigation controls (shown on the upper left corner of the map) or the mouse wheel.

Now, you are ready to navigate through the global maps to visualize the highlighted pathways/compounds in detail.

Move the mouse over each red line to show a summary of its content in terms of Nodes (compounds) and Edges (pathways, reactions, modules, and enzymes) related to your compounds (Fig. 1).

Move the mouse to locate each compound. When clicking a compound, a “Node information” window will show its Names, Mass, DB Links, and Structure.

Move the mouse to locate Edge information. A list of matching pathways, reactions, enzymes, orthologous groups, and genes is shown. Click and an “Edge information” window will appear.

To save the current visualization, use the “Export” button on the top right corner. Enter a title in the “Export title” and check the global maps you want to include. Finally, select the “Output format” and click the “Export maps” button. You can save the map in scalable vector format (SVG), encapsulated postscript (EPS), postscript (PS),

portable document format (PDF), or portable network graphics (PNG).

You can look for different items in the pathway maps by clicking the “Search” button on the top right corner. A text search, to search for entities in the maps (pathways, enzymes, reactions, compounds, etc.) shows up. Introduce the text you want to search and select an entity from the matches provided. This entity will be highlighted in the map as a blue telescopic sight sign (Fig. 1).

### 2.2.3 Advanced Customization

You can customize the representation of your set of compounds in the pathway visualization. It is possible to change the color, width, and opacity for each compound independently. This is possible by providing some extra labels next to the compound IDs in the input file (Fig. 1).

To indicate colors, you should provide a color code in either hexadecimal, RGB, or CMYK notations. *See Note 2* for help on how to obtain color codes. For example, green should be indicated as **#00ff00** (hexadecimal), **RGB(0,255,0)**, or **CMYK(100,0,100,0)**.

To change line width, write *W* and a number (e.g., *W20*).

To indicate opacity, just write a number in the range 0–1 (from fully transparent to fully opaque), for example, 0.5 (for a 50 % opacity).

This panel also allows changing the representation for the items not included in your selection (default values).

### 2.2.4 Detailed View in KEGG

To analyze in detail the roles of a set of compounds in an organism, this section will guide you through the KEGG PATHWAY Database. KEGG PATHWAY contains graphical representations for metabolic, genetic information processing, environmental information processing, and some cellular as well as organismal systems pathways. It also contains information on various human diseases and drugs.

Go to the KEGG PATHWAY website ([www.genome.jp/kegg/pathway.html](http://www.genome.jp/kegg/pathway.html)), and follow the “Search&Color Pathway” link (in the Pathway Mapping section).

If you want to restrict your analysis to a particular organism, click the “org” button in the “Search against” section. This will open a window to enter a 3-letter KEGG organism code (if you already know it) or to search by organism name (in case you do not know the KEGG code).

Paste the list of compound IDs in the “Enter objects...” text area, or alternatively upload a file containing the data (by clicking the “Browse” button).

Enter the color you want to use for highlighting compounds in the “default bgcolor” section (by default, they will be painted in pink).

Click the “Exec” button. A summary page with a list of all the pathways that contain at least one of the compounds entered appears. Now you can follow each link to show a map of the corresponding pathway, in which your input compounds will be highlighted (in pink by default or in the color specified in the previous form).

Detailed information on the compounds, enzymes, and related pathways can be accessed by clicking on the corresponding elements in the pathway image.

You can change the size of the image (with the % menu at top of the map). To save the pathway image, right-click and select “Save Image As...” from the menu.

As with iPath (previous section), it is possible to customize the colors used to highlight your compounds. To do so, add the color you want to use right after the compound ID. You can use color names (in English), as well as hexadecimal codes (e.g., #00ff00 for green). *See Note 2* for help on obtaining color codes.

---

### 3 Enrichment Analysis

Functional enrichment analysis (or overrepresentation analysis) detects the functional annotations that are significantly associated with our set of compounds. This type of statistical analysis was originally developed for the interpretation of transcriptomics experiments, and it is now widely used in both genomics and proteomics experiments [2]. In the last years, it was first adapted for the analysis of human metabolites [7] and it is increasingly used in the field of metabolomics [3].

Annotations of chemical compounds are keywords of different vocabularies representing different aspects of them: they can refer to biological functions (metabolic pathways, enzyme interactions, etc.), intended uses (drug pharmacological actions, chemical applications, etc.), biomedical associations (disease biomarkers, sample localization, etc.), or physicochemical characteristics (chemical taxonomies, functional groups, etc.).

This section will show you how to do enrichment analysis with MBRole.

Go to the MBRole website (<http://csbg.cnb.csic.es/mbrole>) and follow the “Analysis” link (Fig. 2).

First, paste the list of compound IDs in the “Compound set” section, or alternatively upload a file with them (by clicking on the “Browse...” button in the same section).

Select the annotations you want to analyze (in the “Annotations” section). *See Note 3* on the input IDs accepted by MBRole and corresponding annotations.

Select a “Background set” from those provided or upload your own (Fig. 2):

- In case you want to analyze KEGG annotations, you need to select an organism from the menu (check “Pre-compiled” option), or provide a list of compounds for background (check “Provided by user,” and enter or upload the background set).

**MBRole** Home Analysis ID conversion

**Metabolites Biological Role**

**1. Compound set**  
Provide a list of compounds IDs. Currently we support KEGG compounds, HMDB metabolites, PubChem compounds and ChEBI 3star entities. You can also use our ID conversion utility.  
Upload file: Browse... No file selected

**2. Annotations**  
Select one compound type and the annotations to analyze. You can also use our ID conversion utility.  
 KEGG compounds  
 KEGG pathways  
 enzyme interactions  
 other interactions  
 biological role  
 chemical groups  
 HMDB metabolites  
 PubChem compound  
 ChEBI 3star compound  
 SMILES

**3. Background set**  
Statistics can be computed using a pre-compiled reference, or alternatively you can provide a background set.  
 Pre-compiled  
 Arabidopsis thaliana (thale c...  
 Provided by user  
 Upload file: Browse... No file selected

**MBRole results for:**  
**Pathways**  
 Pathways (14)  
 Chemical groups (9)  
 Pathways (set: 6 background: 3358)  
 Show compounds in results table  
 Show annotations with p-value ≤ 0.05 Set filter

id	label	p-val	adjusted p-val	in bckgnd	in set	%	Compounds
ath00710	Carbon fixation in photosynthetic organisms	5.1E-14	7.13E-13	23	6		C00447 C00231
ath01100	Metabolic pathways	6.58E-3	3.07E-2	1455	6		C00231 C00354
ath00030	Pentose phosphate pathway	1.29E-3	9.02E-3	32	2	33.3	C00279 C00231
ath01063	Biosynthesis of alkaloids derived from shikimate pathway	2.26E-2	6.32E-2	138	2	33.3	C00236 C00279
ath01061	Biosynthesis of phenylpropanoids	1.15E-2	4.02E-2	97	2	33.3	C00279 C00236

Export to .csv

Send request Reset Load example

**Fig. 2** Screenshots of the MBRole web interface. In the “Analysis” form (*left*), the user has to provide the list of compounds, the annotations (vocabulary) he/she wants to analyze and select a background set. The results page (*right*) contains tables with the list of enriched keywords and outgoing links to other databases

- In case you want to analyze any other annotation, you can choose to use the “Pre-compiled” background set (i.e., the full database) or provide your own set.

Click “Send request.” MBRole will then search for the annotations in your compound list and compute statistics.

The information generated by MBRole is a list of annotations (from the types of annotations selected – vocabularies) and their corresponding statistical estimates (namely,  $p$ -value and adjusted  $p$ -value). MBRole generates a table for each type of annotation (vocabulary) requested. These are available in the left column of the results page (Fig. 2). The number of top-scoring annotations shown can be changed by modifying the  $p$ -value threshold of the statistical test (“Set filter”). You can add to the table the list of compounds associated to each annotation with the corresponding checkbox at the top of the table (Fig. 2).

You can download the table by clicking on “Export to .csv.” This will generate a comma-separated file that can be saved to your computer and opened with a text editor or a spreadsheet (like MS Excel).

When the annotations of this table are KEGG pathways, these are active links to the corresponding pathway diagrams, where the compounds entered by the user are highlighted as red circles.

## 4 Notes

1. Although in reports/publications we often refer to chemical compounds by their names (once we know their chemical identity), it is always convenient to provide a list of public database IDs, to avoid ambiguities and to facilitate re-usage of your data by future studies. Providing a table with that information as supplementary material or submitting results to public databases like MetaboLights [8] is always a good practice.
2. If you are not familiar with color codes (like hexadecimal, RGB, or CMYK), you can use a visual color picker (e.g., <https://www.colorcodehex.com/html-color-picker.html>). Select the color from the visual palette, and obtain the corresponding color code.
3. The current version of MBRole needs as input a list of compound IDs from a given database (KEGG, HMDB, PubChem, and ChEBI). If you want to analyze a mixture of IDs from several databases, you need to convert the IDs and run the analysis for each of them. (This will be much simpler in the next version of MBRole, which is underway.)

## References

1. Hollywood K, Brison DR, Goodacre R (2006) Metabolomics: current technologies and future trends. *Proteomics* 6(17):4716–4723. doi:10.1002/pmic.200600106
2. da Huang W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37(1):1–13. doi:10.1093/nar/gkn923
3. Chagoyen M, Pazos F (2013) Tools for the functional interpretation of metabolomic experiments. *Brief Bioinform* 14(6):737–744. doi:10.1093/bib/bbs055
4. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P (2011) iPath2.0: interactive pathway explorer. *Nucleic Acids Res* 39(Web Server Issue):W412–W415. doi:10.1093/nar/gkr313
5. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42(Database issue):D199–D205. doi:10.1093/nar/gkt1076
6. Chagoyen M, Pazos F (2011) MBRole: enrichment analysis of metabolomic data. *Bioinformatics* 27(5):730–731. doi:10.1093/bioinformatics/btr001
7. Xia J, Wishart DS (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res* 38(Web Server Issue):W71–W77. doi:10.1093/nar/gkq329
8. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendrakar T, Williams M, Neumann S, Rocca-Serra P, Maguire E, Gonzalez-Beltran A, Sansone SA, Griffin JL, Steinbeck C (2013) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res* 41(Database Issue):D781–D786. doi:10.1093/nar/gks1004