BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 5 2011, pages 730–731
doi:10.1093/bioinformatics/btr001

Systems biology

Advance Access publication on January 5, 2011

MBRole: enrichment analysis of metabolomic data

Monica Chagoyen* and Florencio Pazos

Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), Darwin, 3, 28049 Madrid, Spain

Associate Editor: Olga Troyanskaya

ABSTRACT

Summary: While many tools exist for performing enrichment analysis of transcriptomic and proteomic data in order to interpret them in biological terms, almost no equivalent tools exist for metabolomic data. We present Metabolite Biological Role (MBRole), a web server for carrying out over-representation analysis of biological and chemical annotations in arbitrary sets of metabolites (small chemical compounds) coming from metabolomic data of any organism or sample.

Availability and Implementation: The web server is freely available at http://csbg.cnb.csic.es/mbrole. It was tested in the main web browsers.

Contact: monica.chagoyen@cnb.csic.es

Received on August 11, 2010; revised on October 28, 2010; accepted on December 30, 2010

1 INTRODUCTION

In general, *-omics* data require a secondary analysis in order to be understood in biological terms. In the case of transcriptomics, for example, the primary processing of the raw experimental data provides large lists of hundreds or thousands of genes which are over- or under-expressed in a given condition. These lists are not easily interpretable and they should be further processed in order to get insight into the underlying cause for that change in expression. The most widely used methodology for performing such analysis is termed ‘functional enrichment’ (Khatri and Draghici, 2005). This analysis looks for keywords or descriptors over-represented in the set of genes of interest (e.g. those over-expressed) with respect to a background reference set (e.g. the whole genome or the set of genes printed in the array). Since originally proposed, a hundred of variations and different implementations of enrichment analysis have been proposed (Huang da *et al*., 2009).

While transcriptomic techniques provide information on the mRNA concentrations of the gene repertory of an organism, and proteomics on the respective protein concentrations, metabolomic techniques aim to massively measure the concentrations of the whole metabolite repertory of a given cell (Hollywood *et al*., 2006). Hence, these methodologies complement each other in providing global pictures of the molecular components of living systems. Newer than transcriptomics and proteomics, metabolomic techniques are still facing many technical problems. Maybe due to this, the number of tools for performing secondary analysis (i.e. biological interpretation) of metabolomic data is still scarce, e.g. MetExplore (Cottret *et al*., 2010). As far as we know, only one public tool exists for performing functional enrichment analysis in metabolomic data: MSEA (Xia and Wishart, 2010). That pioneering piece of work has a number of additional interesting features, such as the possibility of using quantitative data (metabolite concentrations). However, that tool is so far restricted to human metabolism. Nevertheless, there exist metabolic databases with multi-organism metabolite annotations whose potential is so far being unexploited for performing this kind of analysis.

With the aim to provide a wide support for analyzing general metabolomic data we have developed MBRole (Metabolite Biological Role). MBRole works with a large number of biological and chemical annotations from the main publicly available metabolite databases and is accessible through a simple and intuitive web interface.

2 METABOLITE ANNOTATIONS

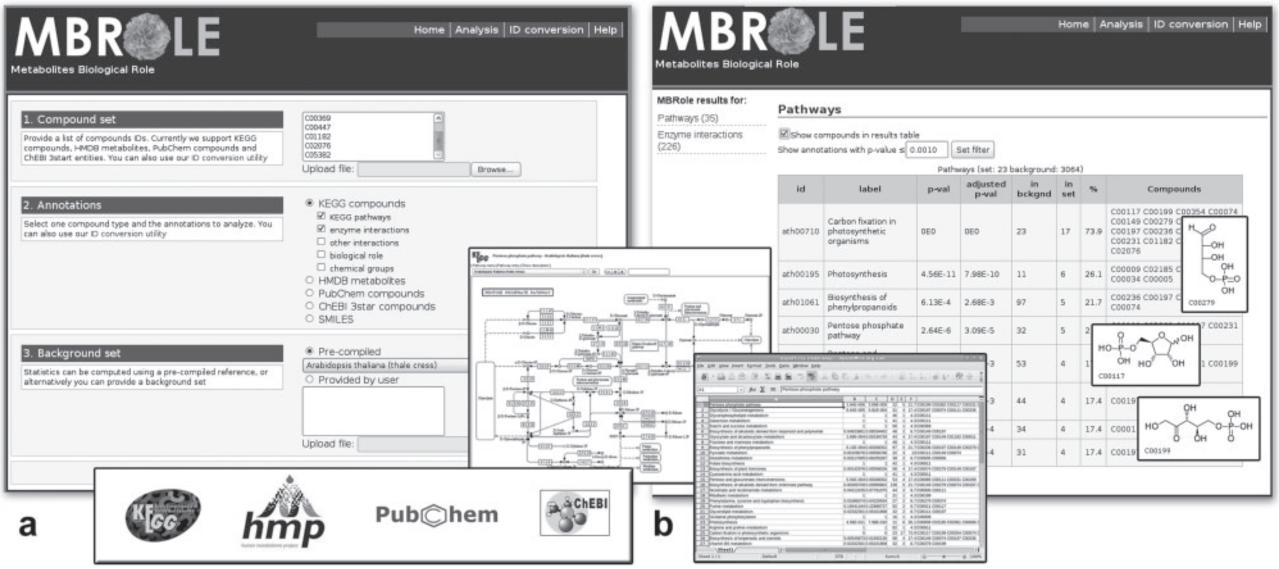
MBRole performs enrichment analysis of a number of annotations of diverse nature coming from different databases. From KEGG (Kanehisa and Goto, 2000) metabolites are annotated with their associated pathways, enzymes and chemical groups, as well as other interactions. From HMDB (Wishart *et al*., 2007), the annotations on pathways, taxonomy, diseases, tissues, biofluids and cellular localizations are gathered. The annotations on pharmacological action are taken from PubChem (pubchem.ncbi.nlm.nih.gov). From ChEBI (Degtyarenko *et al*., 2008), we took the annotations on biological roles, chemical roles and application. Apart from these annotations more related to the biological role of the compounds, the system also uses purely chemical annotations such as the chemical taxonomy of HMDB and the chemical groups detected by the Checkmol software (Heider, 2010).

3 FUNCTIONALITY AND INTERFACE

The main input provided by the user is the set of compounds of interest (e.g. those with significant concentration changes in a given condition). These can be given as identifiers (IDs) of any of the databases mentioned early or as simplified molecular input line entry system (SMILES) strings (www.daylight.com). In the last case, the analysis can be done with purely chemical annotations only. The interface includes a conversion utility which allows the user to perform the analysis using the annotations of a given database even if the IDs are from another database. This conversion utility also accepts CAS registry numbers and metabolite common names.

Additionally, the user has to select the set(s) of annotations (keywords) and the background set of compounds for performing the enrichment analysis (Fig. 1a). The analysis can be performed simultaneously in more than one set of annotations. Annotations

*To whom correspondence should be addressed.



1. Compound set

Provide a list of compounds IDs. Currently we support KEGG compounds, HMDB metabolites, PubChem compounds and ChEBI 3start entities. You can also use our ID conversion utility

C00369
C00447
C01182
C02076
C05382

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 compounds
○ ChEBI 3star compounds
○ 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 cress)
○ Provided by user

Pathways

☑ Show compounds in results table
Show annotations with p-value ≤ 0.0010

Pathways (set: 23 background: 3064)

id	label	p-val	adjusted p-val	in bckgnd	in set	%	Compounds
ath00710	Carbon fixation in photosynthetic organisms	0E0	0E0	23	17	73.9	C00117 C00199 C00354 C00074 C00149 C00279 C00197 C00236 C00231 C01182 C02076
ath00195	Photosynthesis	4.56E-11	7.98E-10	11	6	26.1	C00009 C02185 C00034 C00005
ath01061	Biosynthesis of phenylpropanoids	6.13E-4	2.68E-3	97	5	21.7	C00236 C00197 C00074
ath00030	Pentose phosphate pathway	2.64E-6	3.09E-5	32	5	2	C00231
			3	53	4	1	C00199
			3	44	4	17.4	C0019
			4	34	4	17.4	C0001
			4	31	4	17.4	C0019

Fig. 1. Screenshots of the MBRole web interface. (**a**) Input interface and supported databases; (**b**) results interface and navigation/export capabilities.

for the input set are statistically assessed against the background set. For those databases associating taxonomic information to the compounds (currently only KEGG), the user can select as background set the compounds present in (i.e. metabolized by) his/her organism of interest. For the rest, the background set is the full set of compounds with annotations on each database. Alternatively, the user can provide an arbitrary background set (as a list of compound identifiers). In the case of SMILES the background set must be necessarily provided by the user.

Over-representation analysis is computed using the cumulative hypergeometric distribution. Multiple testing is corrected using the false discovery rate (FDR) (Benjamini and Hochberg, 1995).

The results page (Fig. 1b) contains the list of annotations over-represented in the input set with respect to the background set and their associated *P*-values. The results for the different sets of annotations selected in the previous step can be accessed on the left column. These list(s) of annotations can be directly exported to a spreadsheet program (as .csv comma-separated files) or interactively analyzed. In the interactive list, annotations are hyperlinked to the corresponding databases (when possible). In the case of KEGG pathways, the compounds in the input set are highlighted in the pathway representation. There is a checkbox to include in the table the IDs of the compounds associated to each annotation. These compounds are also hyperlinked to the corresponding database, and their molecular structures (if available in that particular database) are shown when the mouse is over them (Fig. 1b).

The web server includes a help page, a guided tutorial and some pre-computed examples.

4 CONCLUSION

In this work we present MBRole, a tool to perform functional enrichment analysis on metabolic data from any organism or sample using sets of annotations of diverse nature. Only very recently the first tool for performing metabolomic enrichment analysis appeared: MSEA (Xia and Wishart, 2010), which is restricted to human metabolites. Although focused on qualitative annotations, MBRole is general and versatile enough to perform functional enrichment analysis in any metabolomic sample (including additional biological and chemical annotations for human metabolites), and hence complements existing software for the rising field of metabolomics.

ACKNOWLEDGEMENTS

The authors wish to thank the members of the Computational Systems Biology Group (CNB-CSIC), especially Daniel Lopez, for their help with the implementation.

Funding: Spanish Ministry for Science and Innovation (project number BIO2009-11966, in part).

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B*, **57**, 289–300.

Cottret,L. *et al.* (2010) MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res.*, **38**, W132–W137.

Degtyarenko,K. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.

Heider,N. (2010) Functionality pattern matching as an efficient complementary structure/reaction search tool: an open-source approach. *Molecules*, **15**, 5079–5092.

Hollywood,K. *et al.* (2006) Metabolomics: current technologies and future trends. *Proteomics*, **6**, 4716–4723.

Huang da,W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Wishart,D.S. *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res.*, **35**, D521–D526.

Xia,J. and Wishart,D.S. (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.*, **38** (Suppl), W71–W77.