*W132–W137* *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 5 May 2010*
*doi:10.1093/nar/gkq312*

# MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks

**Ludovic Cottret[1,*], David Wildridge[2], Florence Vinson[1], Michael P. Barrett[2], Hubert Charles[3,4], Marie-France Sagot[3,5] and Fabien Jourdan[1]**

[1]INRA, UMR1089, Xénobiotiques, F-31000 Toulouse, France, [2]Division of Infection and Immunity, Glasgow Biomedical Research Centre, University of Glasgow, Glasgow, UK, [3]Bamboo Team, INRIA Grenoble-Rhône-Alpes, 38330 Montbonnot Saint-Martin, [4]UMR203 Biologie Fonctionnelle Insectes et Interactions (BF2I), INRA, INSA-Lyon, Université de Lyon, F-69621 Villeurbanne and [5]Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France

Received January 29, 2010; Revised March 30, 2010; Accepted April 17, 2010



## ABSTRACT

**High-throughput metabolomic experiments aim at identifying and ultimately quantifying all metabolites present in biological systems. The metabolites are interconnected through metabolic reactions, generally grouped into metabolic pathways. Classical metabolic maps provide a relational context to help interpret metabolomics experiments and a wide range of tools have been developed to help place metabolites within metabolic pathways. However, the representation of metabolites within separate disconnected pathways overlooks most of the connectivity of the metabolome. By definition, reference pathways cannot integrate novel pathways nor show relationships between metabolites that may be linked by common neighbours without being considered as joint members of a classical biochemical pathway. MetExplore is a web server that offers the possibility to link metabolites identified in untargeted metabolomics experiments within the context of genome-scale reconstructed metabolic networks. The analysis pipeline comprises mapping metabolomics data onto the specific metabolic network of an organism, then applying graph-based methods and advanced visualization tools to enhance data analysis. The MetExplore web server is freely accessible at http://metexplore.toulouse.inra.fr.**

## INTRODUCTION

Metabolomics aims at identifying the metabolome, i.e. the full set of metabolites present in a biological system (1). Cellular metabolite concentrations are ultimately a reflection of cell function (i.e. gene expression regulation and protein interactions). They are modulated by genetic or environmental perturbations and thus can be considered as central to the phenotype of an organism. These metabolites are the inputs and outputs of biochemical reactions organized into the complex system commonly termed the metabolic network. To date, techniques that allow the quantitative measurement of all metabolites within a given system are not available and this confounds systems level analysis of output from metabolomics experiments. Methods that permit meaningful connections to be inferred between metabolites thus offer the potential to enhance metabolite analysis from the network perspective.

The availability of complete genome sequences has allowed the construction of predicted metabolic networks for many organisms by using information on the presence of enzymes inferred from the presence of the genes that encode them and reference to known biochemical pathways whose structure was determined through the methods of classical biochemistry (2,3). The main metabolic databases such as KEGG (4) or BioCyc (5) are built on this pathway-oriented model. In metabolomics, these databases are used for the analysis of metabolites in the context of global metabolism. The MassTRIX web server (6), for example, shows candidate metabolites as coloured objects on the KEGG pathway maps (7). If the positive identification of metabolites has already been made, tools such as the Omics Viewer of BioCyc (8) or the pathway projector (9) allow metabolites to be highlighted on a collection of organism-specific metabolic maps.

However, in such models, the same metabolite is duplicated if it is involved in multiple metabolic pathways. Moreover, some paths linking identified

*To whom correspondence should be addressed. Tel: +33 561285720; Fax: +33 561285244; Email: ludovic.cottret@toulouse.inra.fr

metabolites can span several pathways and thus are difficult to detect. Finally, the list of metabolic pathways supposed to be present in the metabolic databases is often predicted in an automatic way and can contain many errors and omissions (10). Furthermore, novel, organism-specific pathways cannot be integrated into such networks since construction is dependent upon comparison to reference metabolic pathways.

To restore realistic connectivity between metabolites and to overcome the problems of metabolic pathway identification, the metabolism of an organism is best interpreted as a network gathering all pathways into a single structure. The MetExplore web server allows the mapping of data from metabolomic experiments onto genome-scale metabolic networks. Beyond the mapping functionalities, the network can also be analysed using graph-based methods. A graph is a mathematical object made of nodes and edges connecting them. In MetExplore, graphs are used to offer several advanced mining features including choke point analysis (11), computation of biosynthetic capacity (12) and potential precursor determination for any set of identified metabolites.

MetExplore provides novel filters to restrict metabolism to a particular set of pathways and to discard cellular macromolecules such as proteins, RNA and DNA. Filters were also created to remove ubiquitous compounds or generic reactions. Finally, filtered metabolic networks can be exported into SBML files and metabolic graph files. MetExplore provides a unique framework to link metabolomics experiments, metabolic network visualization and modelling.

The fact that many metabolites contribute to multiple pathways means that a network approach can offer advantages to interpretation of metabolomic experiments. A network approach can, for example, be used to show how two metabolites that may each be transformed directly to, or from, a common metabolite, but as members of separate classical pathways, are separated by only a single chemical species within the metabolic network.

## OVERVIEW

MetExplore follows the workflow given in Figure 1. It consists in building tailor-made filtered networks in order to place experimental data into the context of the known, or predicted, metabolism of a selected organism. For each functionality, various outputs are proposed.

The first building block of MetExplore is a relational database (freely available on request) containing the metabolic networks of about 50 organisms. Metabolic information currently comes from BioCyc-like databases (5,13–17). The list of organisms and their source is available in the online documentation of MetExplore. The latter also contains two databases containing metabolic information from multiple organisms: PlantCyc that describes metabolic pathways present in 250 plant species (http://www.plantcyc.org) and MetaCyc that contains 1400 pathways from more than 1800 organisms (5). Analysing the networks as stored in the databases can lead to misinterpretations. For instance, ubiquitous compounds like water or ATP contribute to many reactions and can overload networks in spite of their not representing formal transitional steps in classical biochemical pathways (18). To overcome this limitation, MetExplore proposes various filters that can be applied to the stored networks.

Once a relevant network is built for a given organism, two kinds of function can be used to perform metabolite analysis. First, MetExplore can map to the genome-inferred metabolic networks the list of metabolites generated from metabolomics experiments (using metabolite masses or metabolite identifiers). Second, MetExplore provides computational functions that allow investigation of the network features of a set of metabolites, allowing, for example, searches for potential drug targets.

Each MetExplore function returns a web browsable table and allows visualization of the results through the graph visualization tool, Cytoscape (19). Each filtered

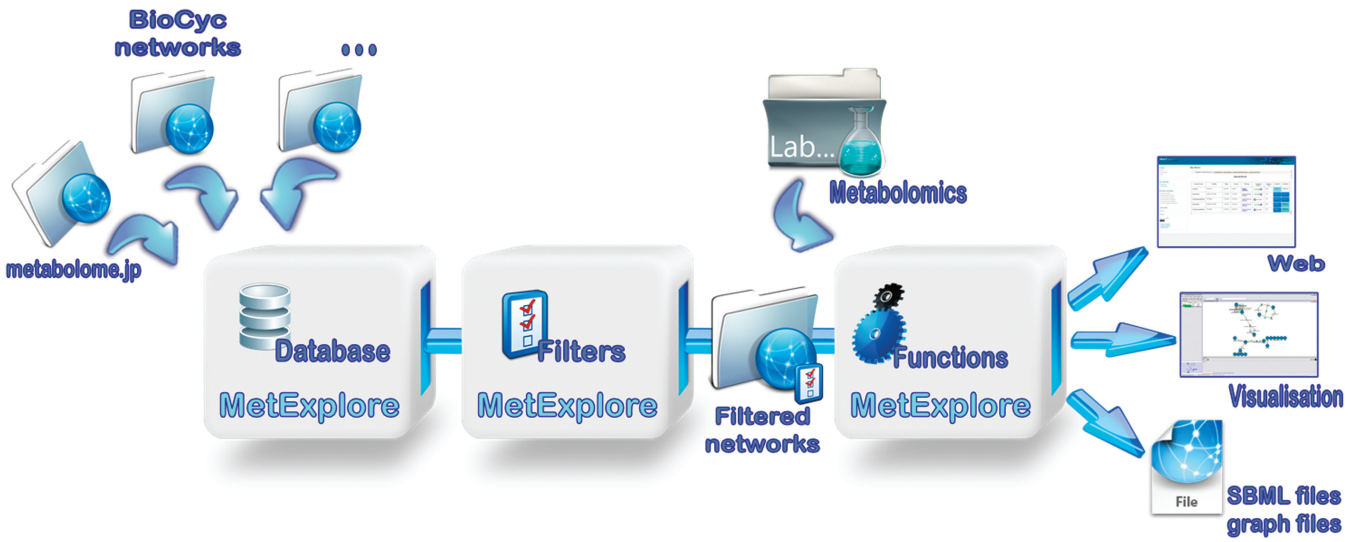


**Figure 1.** MetExplore work flow.

metabolic network can be exported as an SBML file (20) or as a graph file to allow further studies using other modelling tools as described.

## METABOLIC NETWORK FILTERING

The filters available in MetExplore have three functions. First, their implementation can allow investigation of selected subparts of metabolism. Second, they can be used to avoid sources of misinterpretation in metabolic graph analyses and finally, they can aid in providing clearer visualizations of the network, by restricting analyses to only small molecule metabolites excluding cellular macromolecules (proteins, nucleic acids, glycoconjugates, etc.).

A common problem during metabolic network modelling is to deal with the so-called ubiquitous compounds. These compounds, involved in many reactions, can cause artefacts when considered in the same way as components of a linear transformation series as metabolites of classical biochemical pathways. These molecules, e.g. ATP, which provides phosphate to hundreds of reactions, short-circuit the network and confound network analysis (18).

Two ways are proposed in MetExplore to filter out these compounds. First, we can distinguish between main compounds and side compounds in the metabolic pathways stored in the BioCyc-like databases. The main compounds are involved in the backbone of the metabolic pathway, while the side compounds are molecules such as common cofactors that contribute to reactions as chemical donors or recipients without being part of the transformation chain of the metabolic pathway. The MetExplore filter removes metabolites that are annotated as a side compound for given reactions regardless of the metabolic pathway considered. The second MetExplore method to deal with ubiquitous compounds uses a list of 62 cofactor transformations (available in the online documentation). The filter removes these compounds from each reaction in which they participate together. For instance, ATP, ADP and phosphate are removed in each reaction where the transformation 'ATP = ADP + Phosphate' appears. Unlike the previous filter, this function does not use any pathway information and deals with reactions not classified in any metabolic pathway. Filtering ubiquitous compounds also simplifies the visual representation of the metabolic network by removing highly connected nodes.

Finally, a MetExplore user can choose to keep or to remove all of the reactions involved in the pathways identified in the selected organism, but for which no enzymes are assigned [classically called pathway holes (21)].

## METABOLOME MAPPING

### User input

After choosing an organism and tuning the network filters, the required input data for metabolome mapping is a tabulated file. Depending on the selected mode, the first column of the input file corresponds either to measured masses, database metabolite identifiers or metabolite names. In the mass mapping mode, an error limit in p.p.m. (parts per million) has to be defined. The identification of metabolites by their masses is not always satisfactory since, even with high resolution mass spectrometry, the identification may be ambiguous (22), particularly with respect to isomers that by definition are of identical mass. For this reason, MetExplore also uses metabolite identifiers or user-defined names. The names can be described by a regular expression when their syntax is not exactly known. For instance, if the user does not know if the metabolite coenzyme A is stored in the database as co-A, coenzyme-A, or coA, he can use the regular expression co.*a that corresponds to any metabolite name that starts by co, followed by any number of additional characters and ends by a (the case is not sensitive). Additional columns of the input file correspond to numerical values that quantify, for instance, the retention time or the peak intensities.

Each mass, name or identifier is compared to the information stored in the MetExplore relational database for the selected organism. As queries to external databases are not required, the processing is quite fast: a mapping of 380 masses on the complete metabolic network of MetaCyc, for example, takes ~6 min.

### Results

MetExplore is capable of mapping experimental data uploaded by the user onto user-selected metabolic networks. The output is a list of identified metabolites which is enriched by metabolic network information such as metabolic pathways involving these metabolites.

Whatever the selected mode, a result table as described in Figure 2A is displayed. The name of each identified metabolite is a hyperlink to the source database. The pathways in which the metabolite appears in the filtered metabolic network are also indicated. Each numerical value from the input file is reported and coloured depending on the quartile computed in the whole column to which it belongs. A glyph corresponding to the local topology around the metabolite in the filtered metabolic network is also displayed. Simple visual inspection is thus sufficient to indicate whether the metabolite is a source, an output or a choke point (see definition below) in the filtered metabolic network. To facilitate interpretation of the results, it is possible to visualize identified metabolites on the filtered metabolic network by launching Cytoscape directly from MetExplore. The user does not have to install Cytoscape since it is loaded via Java Web Start (the only requirement is that the browser supports Java 1.5 or higher). A MetExplore visual style is automatically applied and highlights the identified metabolites (Figure 2B). Moreover, metabolic network attributes including Enzyme Commission (EC) numbers, metabolic pathways, masses and chemical formulae are loaded as attributes in Cytoscape. All of these attributes and the SBML file corresponding to the filtered metabolic network are also downloadable from the MetExplore interface.

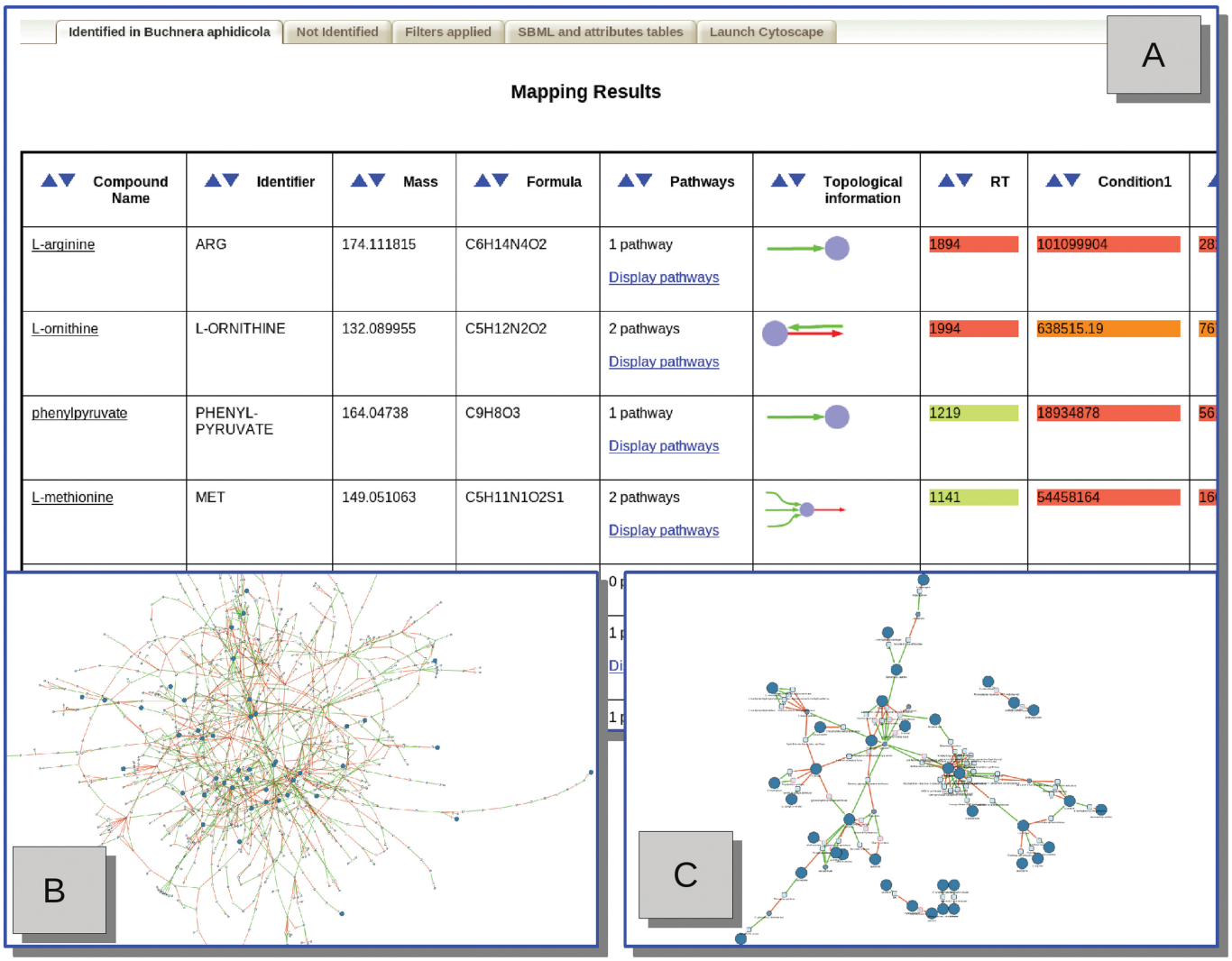


**Figure 2.** Mapping results. (**A**) Table of results (**B**) Visualization of the identified metabolites in the metabolic network. (**C**) Extraction of the subnetwork linking the identified metabolites.

The graphical representation of the entire metabolic network of an organism can be very dense (i.e. with many nodes and overlapping edges). Following reaction paths connecting identified metabolites can be difficult within such an output. To improve visualization in such circumstances, we have developed a Cytoscape plug-in (GapFiller) to compute subnetworks of identified metabolites (23) (Figure 2C). Finally, by creating a MetExplore account, users can create a history of all analyses allowing subsequent consultation without reprocessing.

## METABOLIC GRAPH ANALYSES

In addition to offering the potential to store and visualize data from metabolomics data sets within the context of a virtual metabolome, MetExplore provides several functions based on graph analysis that facilitate understanding of the role of metabolites within the network. For example, these methods can be used for drug target identification (11) or to decipher the biosynthetic links between metabolites (24,12).

### Choke points

One successful and practical use of global metabolic networks has been the introduction of choke point analysis. Choke points are defined as reactions that either uniquely consume a specific substrate or uniquely produce a specific product (11). Choke point analysis, such as used for the malaria parasite *Plasmodium falciparum* has shown current drug targets are far more likely to be choke points within the network than other reactions (11). Furthermore, comparison between choke points in parasites and their hosts identifies parasite-specific choke points as being even better candidate targets.

### Scope

The scope of metabolites allows the identification of the potential biosynthetic capacity of an organism from a

given set of metabolites (12). The scope of a set of metabolites (so-called seeds) is defined as the sum of all metabolites that the seeds are able to produce using the reactions available in an organism. In contrast to the shortest paths computed in simple graphs, the scope concept takes into account the availability of all of the substrates used in a reaction. The scope is computed in an iterative way [referred to as the expansion process (12)]. A table is generated to display information about the metabolites produced during the process.

### Precursors

Precursors are the set of metabolites from which a defined set of target metabolites can be produced. Precursor sets can be calculated by the inverse of the expansion process described in the scope function, starting with a set of given target metabolites, then moving backwards through the dataset until a metabolite is reached that is not produced by any reaction or metabolite already visited during the process. Two outputs can be generated from this computation: a table containing all the metabolites visited during the process and the precursors or a table containing only the precursors. For instance, these precursors might then be considered essential components of any defined culture medium required for growth of a microbial organism or cell type.

When Cytoscape is launched via the scope and precursor functions, the filtered metabolic network and the subnetwork visited during the process are automatically loaded.

## METABOLIC NETWORK EXPORT

While MetExplore, especially used in conjunction with visualization environments such as Cytoscape, offers multiple functions regarding presentation and analysis of metabolomic data sets, clearly there are a multitude of additional software options that may be useful in further analysis of MetExplore file outputs. SBML is an XML-based format dedicated to the description of systems biology type data sets including metabolic networks (20). Filtered networks from each MetExplore function described above and the corresponding attributes can be exported directly to SBML. Moreover, the extended SBML that we created stores information about masses, EC numbers, links between genes, reactions and pathways not present in a classical SBML file. This format is described in the online documentation.

From the filtered metabolic networks, MetExplore is able to build three kinds of graphs: the compound graph, the reaction graph and the bipartite graph [for definitions see (3)]. Each one is available in edge-list format and can be visualized in Cytoscape or used to perform other graph analyses.

## DISCUSSION AND CONCLUSION

Here, we describe novel software to help visualize, navigate, mine and draw biological inference from metabolomics experiments set within the context of global metabolic networks. MetExplore gathers a set of original functionalities that can be easily and freely accessed through its web server. The mapping function enables identification of metabolites predicted within the metabolism of a given organism, using either mass information, database identifiers or their simple names. Since metabolite names are frequently poorly formatted, MetExplore allows the inclusion of regular expressions to describe them. As MetExplore does not require access to other web servers, the processing involved in mapping is very fast as it queries only the MetExplore database. Rather than mapping the identified metabolites to individual pathways or to a collection of pathways, MetExplore maps them onto a single metabolic network. This yields topological information about the identified metabolites. Another advantage of MetExplore resides in its ability to highlight identified metabolites within the metabolic network by directly launching Cytoscape from the MetExplore interface. It is thus possible to visualize output from metabolomic experiments with the attributes of the metabolic network also loaded. To generate a clearer visualization or to study only a subsection of the metabolic network, MetExplore contains various filters to be applied to networks prior to mapping. Moreover, the GapFiller plug-in installed in the Cytoscape version loaded from MetExplore further facilitates analysis by computing subnetworks that link identified metabolites.

MetExplore also provides additional high-level functions that allow further inference from the mapping results. By computing the choke points in a filtered metabolic network, it is possible in MetExplore to detect weak points in the metabolic network, which offer potential drug targets. Furthermore, the scope function implemented in MetExplore helps to decipher which metabolites and reactions can be affected by the changes in concentration of a set of metabolites. In a converse function, since metabolomics generally measures the outputs of the metabolism, MetExplore also computes which metabolites are necessary to produce those observed experimentally. The results of these functions can also be directly visualized by launching Cytoscape. Finally, the possibility to save the filtered metabolic networks into SBML and several graph formats facilitates the first steps of the metabolic network modelling.

We are currently working on further improvements for MetExplore. The database currently contains only about 50 organism-specific data sets. We plan to increase this number, especially by including data sets from other databases such as KEGG (4) or HMDB (25). Furthermore, we are working on the possibility of applying the MetExplore functions to a metabolic network uploaded by the user.

Other graph analysis tools will be included in future versions of MetExplore. For instance, some topological graph measures including betweeness centrality (26) should facilitate an appreciation of the importance of identified metabolites in a metabolic network. Inclusion of data from other sources (e.g. proteomics and transcriptomics) is also under consideration for enhanced analysis through MetExplore.

## FUNDING

French projects (ANR REGLIS NT05-3_45205 and ANR MIRI BLAN08-1335497); French-UK projects (ANR-BBSRC MetNet4SysBio ANR-07-BSYS 003 02 and ANR-BBSRC SysTryp). Funding for open access charge: University of Glasgow.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Fiehn,O. (2001) Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Funct. Genomics*, **2**, 155–168.
2. Francke,C., Siezen,R.J. and Teusink,B. (2005) Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol.*, **13**, 550–558.
3. Lacroix,V., Cottret,L., Thébault,P. and Sagot,M.-F. (2008) An introduction to metabolic networks and their structural analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 594–617.
4. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
5. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M. *et al.* (2010) The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–D479.
6. Suhre,K. and Schmitt-Kopplin,P. (2008) MassTRIX: mass translator into pathways. *Nucleic Acids Res.*, **36**, W481–W484.
7. Okuda,S., Yamada,T., Hamajima,M., Itoh,M., Katayama,T., Bork,P., Goto,S. and Kanehisa,M. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.*, **36**, W423–W426.
8. Paley,S.M. and Karp,P.D. (2006) The pathway tools cellular overview diagram and omics viewer. *Nucleic Acids Res.*, **34**, 3771–3778.
9. Kono,N., Arakawa,K., Ogawa,R., Kido,N., Oshita,K., Ikegami,K., Tamaki,S. and Tomita,M. (2009) Pathway projector: web-based zoomable pathway browser using kegg atlas and google maps api. *PLoS ONE*, **4**, e7710 [11 November 2009, Epub ahead of print].
10. Ginsburg,H. (2009) Caveat emptor: limitations of the automated reconstruction of metabolic pathways in Plasmodium. *Trends Parasitol.*, **25**, 37–43.
11. Yeh,I., Hanekamp,T., Tsoka,S., Karp,P.D. and Altman,R.B. (2004) Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome Res.*, **14**, 917–924.
12. Handorf,T., Ebenhöh,O. and Heinrich,R. (2005) Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J. Mol. Evol.*, **61**, 498–512.
13. Evsikov,A.V., Dolan,M.E., Genrich,M.P., Patek,E. and Bult,C.J. (2009) Mousecyc: a curated biochemical pathways database for the laboratory mouse. *Genome Biol.*, **10**, R84.
14. Keseler,I.M., Bonavides-Martinez,C., Collado-Vides,J., Gama-Castro,S., Gunsalus,R.P., Johnson,D.A., Krummenacker,M., Nolan,L.M., Paley,S., Paulsen,I.T. *et al.* (2009) Ecocyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res.*, **37**, D464–D470.
15. Romero,P., Wagg,J., Green,M.L., Kaiser,D., Krummenacker,M. and Karp,P.D. (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, **6**, R2.
16. Vallenet,D., Labarre,L., Rouy,Z., Barbe,V., Bocs,S., Cruveiller,S., Lajus,A., Pascal,G., Scarpelli,C. and Médigue,C. (2006) Mage: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res.*, **34**, 53–65.
17. Zhang,P., Foerster,H., Tissier,C.P., Mueller,L., Paley,S., Karp,P.D. and Rhee,S.Y. (2005) Metacyc and aracyc. metabolic pathway databases for plant research. *Plant Physiol.*, **138**, 27–37.
18. Arita,M. (2004) The metabolic world of Escherichia coli is not small. *Proc. Natl Acad. Sci. USA*, **101**, 1543–1547.
19. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
20. Finney,A. and Hucka,M. (2003) Systems biology markup language: level 2 and beyond. *Biochem. Soc. Trans.*, **31(Pt 6)**, 1472–1473.
21. Green,M.L. and Karp,P.D. (2004) A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, **5**, 76.
22. Kind,T. and Fiehn,O. (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, **7**, 234.
23. Jourdan,F., Cottret,L., Wildridge,D., Scheltema,R., Hillenweck,A., Barrett,M.P., Zalko,D., Watson,D.G. and Debrauwer,L. (2010) Use of reconstituted metabolic networks to assist in metabolomic data visualization and mining. *Metabolomics*, doi:10.1007/s11306-009-0196-9.
24. Handorf,T., Christian,N., Ebenhöh,O. and Kahn,D. (2007) An environmental perspective on metabolism. *J. Theor. Biol.*, **252**, 530–537.
25. Wishart,D.S., Knox,C., Guo,A., Eisner,R., Young,N., Gautam,B., Hau,D.D., Psychogios,N., Dong,E., Bouatra,S. *et al.* (2009) Hmdb: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**, D603–D610.
26. Liu,W.-C., Lin,W.-H., Davis,A.J., Jordàn,F., Yang,H.-T. and Hwang,M.-J. (2007) A network perspective on the topological importance of enzymes and their phylogenetic conservation. *BMC Bioinformatics*, **8**, 121.