Tools for the functional interpretation of metabolomic experiments

Monica Chagoyen and Florencio Pazos
Submitted: 31st May 2012; Received (in revised form): 24th July 2012

Abstract
The so-called ‘omics’ approaches used in modern biology aim at massively characterizing the molecular repertoires of living systems at different levels. Metabolomics is one of the last additions to the ‘omics’ family and it deals with the characterization of the set of metabolites in a given biological system. As metabolomic techniques become more massive and allow characterizing larger sets of metabolites, automatic methods for analyzing these sets in order to obtain meaningful biological information are required. Only recently the first tools specifically designed for this task in metabolomics appeared. They are based on approaches previously used in transcriptomics and other ‘omics’, such as annotation enrichment analysis. These, together with generic tools for metabolic analysis and visualization not specifically designed for metabolomics will for sure be in the toolbox of the researches doing metabolomic experiments in the near future.

Keywords: metabolomics; functional analysis; enrichment analysis; metabolic pathways

INTRODUCTION
Metabolomics can be described as the study of the ‘metabolome’, the repertory of small chemical compounds in biological systems, with special emphasis on those forming part of the metabolism (metabolites) [1]. Although based on well-established experimental techniques, metabolomics is one of the last additions to the ‘omics’ wave and generates results that complement those of genomics, transcriptomics and proteomics in deciphering the molecular repertoires of biological systems [2]. As such, metabolomic studies are playing a increasingly important role in different aspects of biomedical research, from basic studies related to Systems Biology [3] to more applied uses related to diagnostics, biomarkers, toxicology, etc. [4]

A typical metabolomics experiment generally has three consecutive phases (Figure 1): (i) Separation of the metabolites within a complex sample using chromatographic techniques such as gas chromatography (GC) or high-performance liquid chromatography (HPLC); (ii) identification of the metabolites separated in the previous phase, mainly by mass spectrometry (MS) or nuclear magnetic resonance (NMR) and (iii) analysis of the data obtained. Metabolomics is a very broad field and many variations of the workflow outlined above exist [2, 5–8]. For example, some particular experiments lack the separation phase and metabolites are identified directly in complex samples, usually by NMR [9]. Additionally, in small-scale experiments or those aimed at detecting a particular metabolite, the third phase is not required and the identified metabolite(s) is the final result of the study. Furthermore, the metabolite identification phase can provide quantitative information associated to the metabolites, such as the concentration or a value associated to the ‘change’ of that metabolite in two conditions/experiments.

In an opinion article, M. Arita [10] suggested three ways in which metabolomics can learn from the more experienced genomics and proteomics to
reach a wider researcher community: (i) collection of publicly free data; (ii) application of a simple data format and (iii) creation of a wiki-based data repository. We could add a fourth item: development of tools for the integration and biological interpretation of experimental results by final users. This is essential as metabolomics techniques continue to improve and provide more high-throughput data.

Indeed, in recent years there have been a number of efforts in all these directions. There is an increase in the number of public databases containing data of interest to metabolomics (for a recent review see [11]). There have been efforts to standardize formats in metabolomics experiments [12, 13], and tools to map and convert compound identifiers, like the Chemical translation service [14], have been developed. Additionally, environments for collaborative data curation, such as Wikipathways [15], are now available.

From the computational analysis point of view, in addition to the crucial steps of metabolite identification and quantification, there is a need to analyze the set or sets of metabolites obtained and put them in a biological context. For this purpose it is essential to understand and interpret metabolite data in terms of the underlying biochemical mechanisms and their phenotypic and physiological consequences [16]. In general, the results of any ‘omics’ technique require a secondary analysis in order to obtain the answers one is looking for. The result of a transcriptomics experiment (e.g. large lists of genes up/downregulated) has to be post-processed in order to be interpreted in biological terms (e.g. pathways or biological processes overrepresented in these genes). The same analysis is required in many metabolomic experiments, e.g. to extract from the raw list of metabolites showing up in the experiment recurrent patterns or enriched ‘features’ (metabolic pathways, physicochemical features, associated diseases...). These features will allow converting these raw lists of molecules into biological results. For small-scale experiments (those involving a small number of compounds) these features can be extracted manually based on expert knowledge, nevertheless, as metabolomic techniques improve and larger lists of metabolites show up from the experiments, more sophisticated and automatic tools are required.

Recently, several software tools have become available for the functional and biological interpretation of metabolomic experiments (Figure 1). These tools are the focus of this review. We classify them in two groups that allow complementary analysis. The first comprises tools that allow mappings and visualizations of a set of metabolites in graph representations of the metabolism (mainly metabolic pathways). The second group comprises tools for the statistical analysis of metabolite annotations, commonly known as enrichment analysis. Tools to perform metabolic modeling (or metabolic simulation) i.e. to study the dynamics of the system are beyond the scope of this review (recent review [17]). The tools and methodologies for the primary analysis of metabolomic data (e.g. identification/ quantification of the molecular species from the original MS or NMR data), as well as those for the concomitant compound detection and statistical assessment of differences (based on PCA treatment of
PATHWAY MAPPING AND VISUALIZATION

Traditionally, information on cell metabolism has been represented as pathways, i.e. a graphical representation of the relationships among enzymes, metabolites and catalyzed reactions. More recently, similar graph-based representations are increasingly being used to describe different types of biological networks (such as signaling pathways and protein–protein interactions).

An obvious first step in the interpretation of metabolic experiments is to map and visualize the identified metabolites and associated experimental measurements in the context of metabolic pathways and other general biological networks. Such visualization can provide a quick overview on the metabolic context of the metabolites showing up in the experiment. For example, it makes possible to assess whether the set of identified metabolites are involved in the same biological pathway or if they are close to each other in the metabolic network. Although locating and visualizing a number of compounds in metabolic charts could look trivial at first sight, automatic and interactive tools are required due to the complexity of the metabolism and its associations with other biological phenomena. Several software applications provide this functionality (Table 1).

Most of these tools are not specifically designed for metabolomic analysis but can be very useful for obtaining a first and quick overview of where in the known metabolism a given set of metabolites is located. We have classified them in two groups, those provided by pathway databases and those developed by third-parties (independent laboratories).

Visualization tools from pathway data providers

Many primary sites with metabolic information provide some features for facilitating its visualization. Among these features is the possibility of highlighting a set of chemical compounds provided by the user. An obvious advantage is that these tools use the most recent version of the metabolic information, since they reside in the providers of such information. A drawback is that, in general, they have fewer features than independent tools specifically designed for that goal, since the characteristics of the former are subordinated to those of the main system. Additionally, these tools are usually restricted to their own data and cannot incorporate those of others in the analysis.

- KEGG’s pathway browser (created and maintained by the Kanehisa Laboratories), includes a functionality to locate and color different entities, including metabolites. Both basic and more advanced coloring functionalities are available. A list of pathways and the number of entities found in each pathway is provided, enabling the user to visualize each pathway one by one. A global view on the metabolism can also be mapped using the KEGG Atlas [18]. This functionality can also be used programmatically through KEGG’s programmatic interface.

- The Pathway Tools ‘cellular overview’ diagram and Omics Viewer of the Biocyc.org resource (created and managed by SRI International) [19] allow mapping and visualization of metabolites. In addition, it is possible to paint data values attached to metabolites, allowing the visualization of experimental quantitative data.

- Reactome [20] is a pathway database managed by the European Bioinformatics Institute. Built around human pathway data, pathways for 20 other species are inferred by orthology. Mapping and visualization of metabolites can be performed using the ‘Map IDs to Pathways’ facility. In addition, Reactome calculates overrepresentation statistics of pathway annotations (see next section).

Table 1: Pathway mapping and visualization software

<table>
<thead>
<tr>
<th>Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioCyc-Omics Viewer</td>
<td><a href="http://biocyc.org">http://biocyc.org</a></td>
</tr>
<tr>
<td>iPPath</td>
<td><a href="http://pathways.embl.de">http://pathways.embl.de</a></td>
</tr>
<tr>
<td>KaPPA-View</td>
<td><a href="http://kpv.kazusa.or.jp/en/">http://kpv.kazusa.or.jp/en/</a></td>
</tr>
<tr>
<td>KEGG</td>
<td><a href="http://www.genome.jp/kegg/pathway.html">http://www.genome.jp/kegg/pathway.html</a></td>
</tr>
<tr>
<td>MapMan</td>
<td><a href="http://mapman.gabipd.org/web/guest/mapman">http://mapman.gabipd.org/web/guest/mapman</a></td>
</tr>
<tr>
<td>MetPA</td>
<td><a href="http://metpa.metabolomics.ca">http://metpa.metabolomics.ca</a></td>
</tr>
<tr>
<td>MGV</td>
<td><a href="http://www.microarray-analysis.org/mayday">http://www.microarray-analysis.org/mayday</a></td>
</tr>
<tr>
<td>Paintomics</td>
<td><a href="http://www.paintomics.org">http://www.paintomics.org</a></td>
</tr>
<tr>
<td>Pathios</td>
<td><a href="http://motif.gia.ac.uk/Pathios/">http://motif.gia.ac.uk/Pathios/</a></td>
</tr>
<tr>
<td>Pathvisio</td>
<td><a href="http://www.pathvisio.org/">http://www.pathvisio.org/</a></td>
</tr>
<tr>
<td>ProMetra</td>
<td><a href="http://www.cebitc.uni-bielefeld.de/groups/bi/software/prometra.info/">http://www.cebitc.uni-bielefeld.de/groups/bi/software/prometra.info/</a></td>
</tr>
<tr>
<td>Reactome</td>
<td><a href="http://www.reactome.org">http://www.reactome.org</a></td>
</tr>
<tr>
<td>VANTED</td>
<td><a href="http://wanted.ipk-gatersleben.de">http://wanted.ipk-gatersleben.de</a></td>
</tr>
</tbody>
</table>

NMR spectra, etc), are not covered here either (recent review [16]).
Independent (or third-party) mapping and visualization tools

In general, these ad hoc tools specifically designed for visualization are richer in features than those offered by the primary sites of metabolic information. Additionally, some of them are able to integrate pathway data from several data sources. On the other hand, in general they do not ensure the user being accessing the newest metabolic information.

• **VANTED (Visualization and Analysis of Networks containing Experimental Data).** VANTED [21] is a stand-alone software application written in Java. It enables mapping experimental data into metabolic graphs. Graphs can be loaded from local files or remotely accessed (e.g. KEGG pathways) and can be further edited. Metabolite mapping is performed individually in each pathway. This tool has the possibility of extending its functionality through plugins. This program might be of interest for users who do not want to use the pathways as they are in the databases but to add/remove reactions from them, thanks to its editing functionality.

• **iPath (Interactive Pathways Explorer).** iPath [22] provides an ‘interactive’ navigation of various biological pathways, including metabolic pathways taken from KEGG. iPath allows to interactively map on top of the pathways any qualitative or quantitative information associated to their entities (reactions, enzymes and metabolites). It is possible, for example, to modify the representation of the metabolites (size, color, etc.) depending on a quantitative value provided by the user, such as their concentration in a sample. These representations are not a priori restricted to a given pathway but the whole metabolic network is shown and the user can zoom it in and out interactively. Nevertheless, KEGG metabolic pathways are still differentiated in these representations with different colors, so that the user can interpret the results in terms of classic pathways. The tool is also intended for performing comparative analysis (e.g. visualizing data from two experiments or conditions). This tool is valuable for users who need to perform complex navigations through the metabolism (movements, zoom in/out, . . .) and to map complex heterogeneous data on them. For performing simple visualizations and/or mapping simple data or for a small number of compounds, maybe other simpler tools are more suitable.

• **MetScape.** MetScape [23] is a plug-in for Cytoscape [24], a generic tool for the interactive visualization and management of biological networks. MetScape provides a graph-based representation of the whole metabolism, in contrast to the pathway-based representation of other tools. Cytoscape’s core functionalities as well as those provided by this plug-in can be used to visualize information associated to the metabolic compounds. This tool is intended for users interested in global and non-redundant views of the metabolic network, instead of focusing on the typical pathways, as well as for those who want to apply Cytoscape’s tools to their data.

In addition to the above, a number of other tools are focused on the integrated visualization of –omic data. Among them, ProMetra [25], a web-based multi-omics pathway mapping and visualization tool, where pathways are user defined enriched SVG images. Another similar tool is Paintomics [26], a web-based tool that maps simultaneously quantitative data from both, transcriptomics or proteomics and metabolomics experiments in KEGG pathways. Alternatives to web-based tools exist, such as Pathvisio [27] and MGV [28]. Pathvisio is a powerful stand-alone application for the creation and editing of pathways that can be also be used to map experimental data associated to both, gene/proteins and metabolites. MGV is a plug-in for the transcriptomics analysis workbench Mayday [29] to visualize various ‘omics’ data. There are two additional tools of potential interest for those researchers studying plant metabolomes: KaPPA-View [30] and MapMan [31].

Finally, some pathway analysis and visualization tools are now integrated in general metabolomics data analysis software. This is the case of MetPA [32] which has been recently integrated in the MetaboAnalyst platform [33]. And Pathos [34] that is integrated with MeltDB [35], a web-based platform for the storage and analysis of metabolic experiments, which allows mapping MS data in metabolic pathways.

**ENRICHMENT ANALYSIS**

Since the first appearance of enrichment analysis methods for the functional interpretation of large gene lists in 2002, numerous tools for applying this type of analysis to transcriptomic and proteomic data
are available [36]. These tools are frequently used for calculating and reporting statistical values associated to gene functional annotations provided by several data sources. In general, these tools compare the annotations present in a set of genes of interest (i.e., those up or downregulated in a transcriptomics analysis) with the corresponding annotations in a reference set (i.e., all the genes in the organism or in the array) and report those annotations overrepresented in the interest set, according with a given statistical test. Consequently, these tools are very useful to 'summarize' a long list of genes and to have an idea of the biological phenomenon behind it.

Very recently, a number of publicly available software implementing similar approaches for the analysis of a list of metabolites has been reported (Table 2). These tools, specifically designed for metabolomics, allow the functional interpretation of metabolomic experiments in terms of statistically significant general pathways as well as other biological annotations. One advantage of these approaches is that they can handle heterogeneous and hierarchical vocabularies, which are frequently used for annotating biological entities, including metabolites.

Types of analysis
The tools for enrichment analysis reviewed perform two variations of this kind of analysis: overrepresentation analysis (ORA) and set enrichment analysis (SEA). For both approaches, the user should provide as input a set of metabolites and select the type of annotations to examine. In the case of SEA, a numeric value associated to each metabolite (e.g., its concentration) has to be also provided. The general output is a list of annotations and their associated $P$-value in a tabular format. In addition to ORA and SEA, some tools perform other complementary functions, such as conversion of metabolite identifiers.

### Table 2: Metabolite enrichment analysis software

<table>
<thead>
<tr>
<th>Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSEA</td>
<td><a href="http://www.msea.ca">http://www.msea.ca</a></td>
</tr>
<tr>
<td>MBRole</td>
<td><a href="http://csbg.cnb.csic.es/mbrole">http://csbg.cnb.csic.es/mbrole</a></td>
</tr>
<tr>
<td>MPEA</td>
<td><a href="http://ekhidna.biocenter.helsinki.fi/poxo/mpea/">http://ekhidna.biocenter.helsinki.fi/poxo/mpea/</a></td>
</tr>
<tr>
<td>IMPiLA</td>
<td><a href="http://impala.molgen.mpg.de">http://impala.molgen.mpg.de</a></td>
</tr>
</tbody>
</table>

Reactome and Paintomics, included in the previous section on pathway mapping and visualization, also provide some basic overrepresentation analysis for Reactome and KEGG pathways, respectively.

**ORA**
This analysis performs a statistical test to assess whether a set of metabolites is enriched with a particular annotation (e.g., a pathway) compared to a background set. The statistical test is applied to each annotation found in the user-provided set of metabolites and a $P$-value is calculated. This is also known as singular enrichment analysis.

The aim of ORA is to help the researcher getting insight into the underlying biological mechanisms and functional implications of given metabolite set. Several statistical tests can be used to perform the analysis, e.g., chi-square, Fisher’s exact test, binomial probability and hypergeometric distribution [36].

As multiple annotations are evaluated during the analysis, a correction for multiple testing is usually performed. Therefore, in addition to the ‘raw’ $P$-values, corrected $P$-values are provided.

**SEA**
SEA was first proposed in [37] and later improved [38] for the analysis of gene sets in the context of transcriptomics experiments. The basic idea was to use the whole list of genes in the array (sorted by a score, i.e., fold change of expression), instead of imposing a cut-off value to define a list of overexpressed or underexpressed genes. So, this approach looks for annotations which tend to be associated to extreme scores, but in a sort of ‘continuous’ way [39]. As in ORA, correction for multiple testing is performed.

In its application to the analysis of metabolite sets, it takes into consideration a quantitative measure associated to each metabolite (e.g., concentration). The method evaluates the consistency of each annotation in the top and bottom of the list of metabolites sorted by that quantitative value, compared to a background distribution. Several strategies exist for performing SEA depending, among others, on the statistical test applied and the procedure to create the background distribution.

**Software for enrichment analysis**

**Metabolite set enrichment analysis**
Metabolite Set Enrichment Analysis (MSEA) [40] was the first publicly available tool to perform enrichment analysis of human and mammalian metabolomics experiments. It allows performing overrepresentation analysis and SEA using metabolite concentrations. In addition, MSEA performs single sample profiling (SSP) to assess whether metabolite
concentrations are significantly higher or lower than their normal values. Statistical analysis is performed on the information stored in the Human Metabolome Database (HMDB) [41]. MSEA has been recently integrated into MetaboAnalyst [33], a comprehensive suite of programs for general metabolomics data analysis. This tool is especially suitable for users interested in analyzing human or mammalian metabolomic data.

**Metabolites biological role**
Metabolites Biological Role (MBRole) [42] performs overrepresentation analysis using a wide range of biological and chemical annotations in several organisms. It integrates annotations on metabolites from a number of publicly available databases, such as KEGG, SMPDB [43], HMDB, PubChem [44], ChEBI [45]. These metabolite annotations contain information, among others, on the pathways the metabolite is involved in, enzyme and other protein associations, diseases, pharmacological action, biological or chemical role, as well as chemical taxonomy and chemical groups. The advantage of this tool is the possibility of using many different annotation schemas (not only pathways) for performing enrichment analysis, what is especially interesting for exploratory analyses. Another advantage is that it can handle metabolite sets from many organisms apart from human.

**Metabolite pathway enrichment analysis**
Metabolite Pathway Enrichment Analysis (MPEA) [46] performs overrepresentation analysis of pathway annotations (based on KEGG and SMPDB pathways), as well as SEA of mass-spectrometry tag data. The obvious advantage of this tool is the possibility of using directly the GC–MS information associated to the compounds as quantitative data for performing the analyses.

**Integrated molecular pathway level analysis**
Integrated Molecular Pathway Level Analysis (IMPaLA) [47] performs overrepresentation and SEA using Wilcoxon test simultaneously of genes/proteins and metabolites. This allows the integration of metabolomics as well as transcriptomics or proteomics data in the functional enrichment analysis. Data sources include Reactome, KEGG and Wikipathways. This is the right tool for users who performed metabolomics and transcriptomics/proteomics experiments in parallel for the same sample(s).

**LIMITATIONS AND FUTURE PROSPECTS**
Genomics tries to characterize the repertory of coding genes in a given organism, transcriptomics that of mRNA transcripts and proteomics the corresponding translated proteins. Nevertheless, genomics, transcriptomics and proteomics alone do not provide a complete picture of a living system at the molecular level. Small chemical compounds are crucial for these systems, especially those forming part of the central metabolism, whose transformations provide the mass and energy required by living systems. These small chemical compounds are the target of ‘metabolomics’. In this sense, all these ‘omics’ techniques are complementary to each other in obtaining the molecular repertoires of cellular systems.

In spite of these parallelisms, there are fundamental differences between metabolomics and other ‘omics’. Maybe the most important one is the degree of coverage. While the other ‘omics’, at least in theory, are intended for characterizing the whole repertory of genes, transcripts and proteins, current metabolomics approaches usually cover a relatively small fraction of a sample’s metabolome (estimated to be \( \sim 10\% \) [1]). The bottleneck is mainly in the metabolite identification step, and hence it is supposed to be alleviated as the technologies used for this identification, especially MS, are improved [5]. Another important difference is that metabolomic techniques are not restricted to a pure sample/culture and can be applied to any sample (complex fluids such as blood, environmental samples, etc.) This is also changing now for other ‘omics’ as, for example, DNA sequencing is being applied to environmental samples [48], and the so-called ‘next-generation sequencing’ methods can also be applied to complex samples.

Most computational tools designed for metabolomics are intended for the metabolite identification step (analysis of MS and NMR spectra, etc.). Nevertheless, as metabolomics experiments identify more and more metabolites, a functional analysis of these sets is required in order to gain some biological insight.

A number of tools for the graphical representation of metabolic pathways and other biological networks can be used to map and visualize metabolomics results. As exploratory tools they allow researchers to examine experimental results in a biological context, navigating throughout various views of the metabolism and other relevant biological processes.
Recently other computational tools based on ‘enrichment analysis’, an approach widely used in transcriptomics, started to appear. When enrichment analysis is applied in metabolomics, there are two main difficulties, as compared to transcriptomics and proteomics analysis. The first is the lack of functional annotations of metabolites, compared to genes/proteins for which large annotation efforts exist [49]. The second is the difficulty in establishing an optimal background set as, in contrast to the genome, the metabolome of a given organism is unknown. In practical terms, a partial metabolome can be inferred from a metabolic reconstruction or from a set of partial metabolic pathways. Enrichment analysis tools can be applied with any annotation associated to the metabolites. When used for the analysis of metabolic pathways it suffers from the fact that pathways are human abstractions. To handle this problem, alternative methods are emerging, which use a different definition of ‘pathway’, based on the topological properties of metabolic networks [50].

To better interpret metabolite sets there is a need to create more resources with annotations on the biological activities of metabolites, beyond metabolic pathways. In this line of work there are organism-centric compilations such as the Human Metabolome Database [41] or the more recent Yeast Metabolome Database [51] that provide several types of functional annotations. Metabolite annotations can also be inferred from other resources not specifically designed for that purpose, e.g. by using the relationships between a given metabolite and a number of proteins as annotations for that metabolite. For example, the integration of various type of data has enabled to build a network of compound–protein interactions in STITCH [52]. Similarly, automatic analysis of the literature by the Metab2MeSH [53] method recently compiled associations between chemical compounds and MeSH terms. MeSH (Medical Subject Headings) is the controlled vocabulary used for indexing articles in MEDLINE [54].

It is expectable that in the future metabolomic experiments become more massive, and more resources aimed at annotating metabolites (directly or indirectly) are created in order to feed enrichment analysis and other methods. In that stage, metabolomic techniques will be routinely used as other ‘omics’ technologies and their results can be therefore combined.

Key Points

- As metabolomic techniques develop, automatic methods for analyzing metabolite sets in order to obtain meaningful biological information are required.
- Recently several software tools have become available for the functional and biological interpretation of metabolomic experiments, which are review here.
- Some tools map and display metabolites and associated experimental data in visual representations of the metabolism.
- Others perform statistical analysis of metabolite annotations, commonly known as enrichment analysis.

Acknowledgements

The authors want to thank the members of the Computational Systems Biology Group (CNB-CSIC) for interesting discussions and support.

FUNDING

This work was partially supported by the Spanish Ministry for Economy and Competitiveness [project number BIO2010–22109].

References
