W326–W330 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki437

Metabolic PathFinding: inferring relevant pathways in biochemical networks

Didier Croes, Fabian Couche, Shoshana J. Wodak and Jacques van Helden*

SCMBB, Université Libre de Bruxelles, Campus Plaine, CP 263, Boulevard du Triomphe, B-1050 Bruxelles, Belgium

Received February 14, 2005; Revised and Accepted March 25, 2005

ABSTRACT

Our knowledge of metabolism can be represented as a network comprising several thousands of nodes (compounds and reactions). Several groups applied graph theory to analyse the topological properties of this network and to infer metabolic pathways by path finding. This is, however, not straightforward, with a major problem caused by traversing irrelevant shortcuts through highly connected nodes, which correspond to pool metabolites and co-factors (e.g. H_2O, NADP and H^+). In this study, we present a web server implementing two simple approaches, which circumvent this problem, thereby improving the relevance of the inferred pathways. In the simplest approach, the shortest path is computed, while filtering out the selection of highly connected compounds. In the second approach, the shortest path is computed on the weighted metabolic graph where each compound is assigned a weight equal to its connectivity in the network. This approach significantly increases the accuracy of the inferred pathways, enabling the correct inference of relatively long pathways (e.g. with as many as eight intermediate reactions). Available options include the calculation of the *k*-shortest paths between two specified seed nodes (either compounds or reactions). Multiple requests can be submitted in a queue. Results are returned by email, in textual as well as graphical formats (available in http://www.scmbb.ulb.ac.be/pathfinding/).

INTRODUCTION

Biochemical databases contain information about thousands of metabolites and chemical reactions involved in small molecule metabolism (SMM). These data can be represented as a graph, representing all the possible ways in which molecules are converted into one another in the SMM network. We define hereafter as 'raw graph' the graph built by connecting all the annotated reactions to their substrates and products. Several attempts have been made to use path finding algorithms in this graph in order to infer putative metabolic pathways. This, however, yielded disappointing results, for understandable reasons. For example, Kuffner *et al*. (1) found that more than 500 000 paths could be found between glucose and pyruvate. Most of these paths have clearly nothing to do with the known glycolysis pathway, and they most probably reflect artifacts rather than the existence of half-a-million alternative pathways for glucose utilization. Additional information can be used to select the most relevant among these candidate pathways, for instance the stoichiometry of the reactions (2,3), the chemistry of the compounds (4,5) or transcriptome data (6–8).

A recurrent problem is that computed paths in the SMM network tend to traverse shortcuts through highly connected molecules, such as H_2O, ATP and NADP. These molecules are involved as substrate or product in hundreds of reactions, but they generally appear as side metabolites or co-factors, and cannot be considered as valid intermediates between two reactions. One way to circumvent this problem has been to 'filter' the raw graph, by excluding a selection of highly connected compounds (7–12). However, the choice of the compounds to exclude is not obvious, as even the most connected compounds are occasionally used as intermediate metabolites in pathways. For instance, ATP commonly acts as side substrate in reactions where it serves as an energy carrier, but it is also used as a main intermediate between reactions involved in nucleotide metabolism. Recently, we developed an alternative approach based on a weighted graph representation, in which all the compounds are included in the graph, but a weight (cost) is associated with each compound equaling its connectivity in the entire metabolic network (i.e. the number of reactions in which it participates as substrate or product). When searching for the shortest path (the path of minimum weight) the algorithm tends to avoid highly connected compounds whenever possible. An extensive validation performed against 56 annotated pathways showed that this approach allows inferring relevant metabolic pathways when the first and last reactions of the annotated pathway were provided as seeds. The tool is

*To whom correspondence should be addressed. Tel: +32 2 650 5466; Fax: +32 2 650 5425; Email: jvanheld@scmbb.ulb.ac.be

available for academic use through a web interface (http://www.scmbb.ulb.ac.be/pathfinding/).

METHODS

Metabolic graphs

A graph representing SMM network was built with the 5985 reactions and 5082 compounds from the LIGAND database (http://www.genome.jp/kegg/ligand.html). This graph is bipartite: two separate types of nodes are used to represent reactions and compound, respectively, and arcs always connect nodes of different types. Directed arcs are used to represent substrate (compounds → reaction) and product (reaction → compound) relationships. All reactions are considered as potentially reversible. Indeed, even though chemically, some reactions have a strong directionality, under physiological conditions they can be forced in the reverse direction by mass action. We, thus, instantiate two separate nodes per reaction: one for the forward and one for the reverse direction, respectively.

Three types of graphs are derived from the SMM data. The graph containing all the compounds and reactions, hereafter referred to as raw graph. A filtered graph is obtained by excluding a selection of 36 compounds among the most connected ones (list available on the website). The weighted graph is derived from the raw graph by assigning a weight to each compound. By default, the weight of a compound is its connectivity, i.e. the number of reactions in which it participates either as a substrate or as a product. The algorithm allows to assign weights to reactions as well, but this option has so far not been exploited.

Path finding algorithm

A backtracking algorithm finds the *k*-lightest paths, i.e. the paths with the lowest weight. The path weight W is defined as the sum of the weights of its nodes. When a reaction is traversed during path elongation, the passage through the reverse reaction is disabled (9).

WEB INTERFACE

Input

The user has to specify two seed nodes (a source and a target), between which the *k*-shortest paths have to be computed, as well as other parameters (maximum weight, maximum path length, *k*). Alternatively, a file with a list of seed pairs can be uploaded in order to automate multiple searches. Seed nodes can be compounds, reactions or Enzyme Commission (EC) numbers. The EC nomenclature defines a systematic identifier (the EC number) for each catalytic activity. The same EC number can therefore be associated with several reactions. When specifying EC numbers as seed nodes, paths are computed between all the different reaction pairs associated with the specified EC numbers, and the shortest paths among all these are returned.

Processing

Submitted tasks are placed in a queue for further processing by a computer cluster of 36 nodes. The results are sent by email. Multiple tasks can be handled without overloading the server. A typical task takes a few seconds on one node of the cluster. Elapsed time thus primarily depends on the cluster job load.

Output

The output is sent by email. It contains a description of the input parameters (seed nodes, maximal weight, etc.) and a textual description of the paths found (molecule names and IDs). A graphical representation is automatically generated, and can be exported in different formats (png, postscript, dot and svg). The png format can be displayed by most web browsers, but its resolution is limited. Postscript and svg are vectorial formats appropriate for high-quality printing. The dot format allows to visualizing the result with the Graphviz software (http://www.research.att.com/sw/tools/graphviz/).

APPLICATION EXAMPLE

To illustrate the use of our tool, we apply it to search for the five shortest paths between the first and the last reactions of the tryptophan biosynthesis pathway. The graphical output is shown in Figure 1 (weighted graph) and in Figures 3 and 4 of the Supplementary Material. The seed nodes are highlighted with shadowed boxes. Green arrows highlight the lightest path, i.e. the path with the lowest weight (W_{min}). Paths with a weight $W \leqslant W_{min} + 10$ are boxed in green and larger paths in red.

The search in the raw graph (Figure 3 in Supplementary Material) produces a trivial and irrelevant result: the target reaction is reached in two steps, using pyruvate and H_2O as intermediates.

The search in the filtered graph (Figure 4 in Supplementary Material) yields a better result than the raw graph. First, the inferred pathways are more biochemically plausible, as highly connected compounds cannot be used as intermediates anymore. The two shortest inferred paths (green arrows) do not match a single reaction of the annotated pathway (except for the seed nodes), but the third path corresponds perfectly to the annotated pathway (Figure 2).

When the search is performed in the weighted graph (Figure 1), the lightest inferred path matches the annotated pathway perfectly (Figure 2).

VALIDATION OF THE METHOD

The example presented here is clearly not sufficient to draw general conclusions. The same analysis as above was applied to 56 annotated pathways from the bacteria *Escherichia coli* and the yeast *Saccharomyces cerevisiae*, stored in the aMAZE database (13). These pathways cover a large fraction of the annotated pathways having at least three reactions. The results of this validation will be detailed elsewhere (D. Croes, F. Couche, S. Wodak and J. van Helden, submitted for publication). In short, they show that the case illustrated here for the tryptophan biosynthesis pathway is quite representative of the general accuracy. The correspondence between the shortest (lightest) inferred paths and the annotated pathways (computed as the mean of sensitivity and specificity), is very low (28.4%) using the raw graph, increases to 65.5% for the filtered graph and reaches 85.9% for the weighted graph. In addition, even when the annotated path did not correspond to the lightest path, it was often found among the top raking

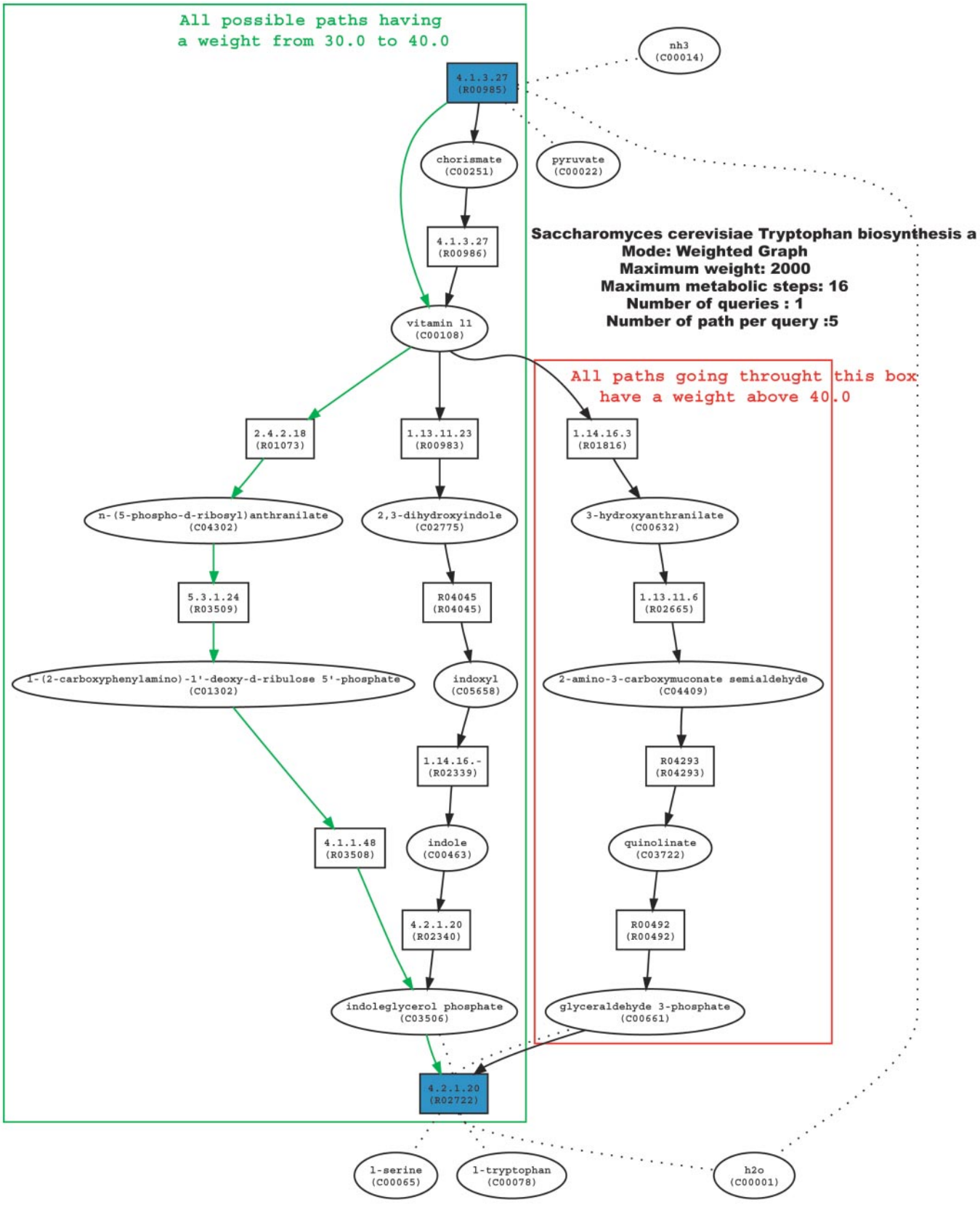

Figure 1. Path finding result: the five lightest pathways found in the weighted graph, when the first and last reactions of the tryptophan biosynthesis are input seeds. Reactions are displayed as rectangles and compounds as ellipses. Seed nodes are highlighted in blue. Green arcs denote the lightest path (or paths in case of ex-aequos). The green box surrounds the path having a weight $W \leq W_{min} + 10$, where W_{min} is the weight of the lightest path. Nodes that are only found in paths heavier than $W_{min} + 10$ are surrounded by using red box. Dashed lines indicate the substrates and products of the seed reactions that are not in the paths.

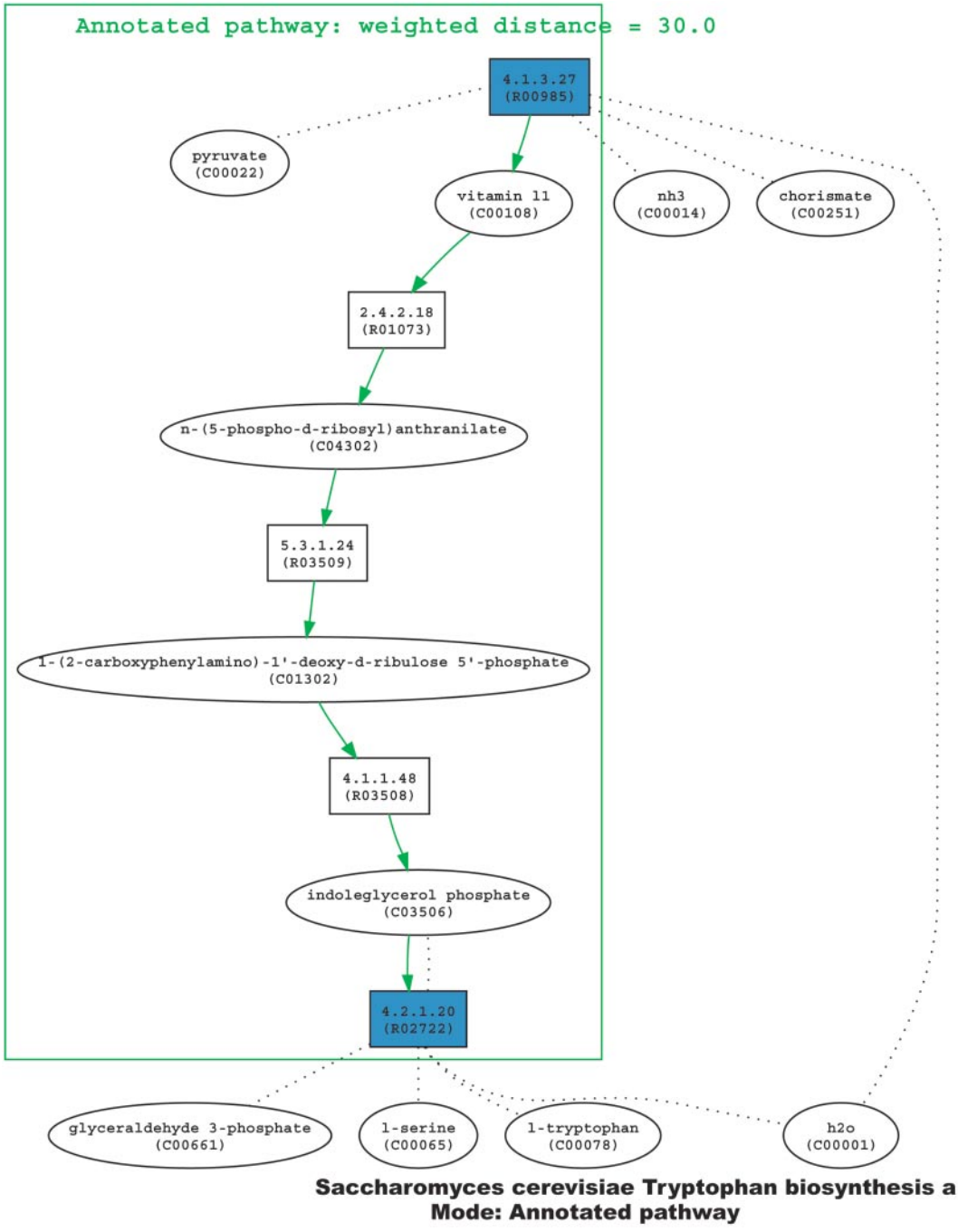

Figure 2. Pathway annotated in aMAZE (source: http://www.amaze.ulb.ac.be/) for the tryptophan biosynthesis in yeast.

paths: the correspondence between the annotated pathways and the best matching path among the five shortest (lightest) ones was on average as high as 90.1%.

POTENTIAL APPLICATIONS

The basic functionality provided by our Metabolic Path-Finding tool is to find one or several pathways between two seed nodes (reactions or compounds). When this search is performed in the weighted graph, the inferred paths generally correspond to biochemically valid pathways. The inference of metabolic pathways between enzymes can be useful for probing the functional relationships between pairs of enzyme coding genes believed to be associated in various ways (e.g. genes whose homologs are involved in fusion events, synteny pairs, genes belonging to the same operon, genes having correlated expression profiles and so on), or between pairs of interacting enzymes characterized by two-hybrid screens or part of the same protein complexes identified by pull-down experiments).

An alternative approach would be to map such pairs of related enzymes directly onto known pathways stored in a

database such as the Kyoto Encyclopaedia of Gene and Genomes (KEGG) (14,15). One drawback to this approach is that pathways represent a somewhat arbitrary segmentation of metabolism (as witnessed by the differences in the definition of pathways used in the different databases, such as EcoCyc, KEGG and WIT). The path finding procedure described here ignores such segmentation, enabling links between reactions annotated in distinct pathways.

A second limitation of pathway mapping resides in the fact that it limits us to the current state of pathway annotation, which usually lags behind that of individual reactions and enzymes. Indeed, our current knowledge of metabolism is mainly based on a very few model organisms (*Escherichia*, *Salmonella*, *Saccharomyces* and a few mammals). Moreover, the level of annotation is uneven, ranging between some very well-characterized pathways involved in primary metabolism, and secondary pathways involved, for example, in detoxification. In addition, even in the central metabolism, alternative pathways can be used by different organisms, or as a result of gene knockouts. Some of these have been characterized experimentally, but many of them remain to be discovered. With the availability of hundreds of sequenced genomes (16), the Metabolic PathFinding presented here should be a valuable tool for analyzing the evolution of metabolic pathways.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

KEGG is gratefully acknowledged for free use of the data curated in the LIGAND database. This work was partially supported by funds from the Government of the Brussels Region, and the BioSapiens Network of Excellence funded under the 6th Framework programme of the European Union (LSHG-CT-2003-503265). Funding to pay the Open Access publication charges for this article was provided by the 6th Framework programme of the European Union (LSHG-CT-2003-503265.

Conflict of interest statement. None declared.

REFERENCES

1. Kuffner,R., Zimmer,R. and Lengauer,T. (2000) Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, **16**, 825–836.
2. Schuster,S. and Hilgetag,C. (1994) On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, **2**, 165–182.
3. Schuster,S., Dandekar,T. and Fell,D.A. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.
4. Arita,M. (2000) Metabolic reconstruction using shortest paths. *Simulat. Pract. Theory*, **8**, 109–125.
5. Arita,M. (2004) The metabolic world of *Escherichia coli* is not small. *Proc. Natl Acad. Sci. USA*, **101**, 1543–1547.
6. Zien,A., Kuffner,R., Zimmer,R. and Lengauer,T. (2000) Analysis of gene expression data with pathway scores. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 407–417.
7. van Helden,J., Gilbert,D., Wernisch,L., Schroeder,M. and Wodak,S. (2001) Applications of regulatory sequence analysis and metabolic network analysis to the interpretation of gene expression data. In Gascuel, O. and Sagot, M.F. (eds), *Computational Biology: First International Conference on Biology, Informatics, and Mathematics, JOBIM 2000*, volume LNCS 2066. Springer, pp. 155–172.
8. van Helden,J., Wernisch,L., Gilbert,D. and Wodak,S.J. (2002) Graph-based analysis of metabolic networks. In Mewes, H.W. *et al.* (eds), Bioinformatics and Genome Analysis. Springer-Verlag, pp. 245–274.
9. Fell,D.A. and Wagner,A. (2000) The small world of metabolism. *Nat. Biotechnol.*, **18**, 1121–1122.
10. Wagner,A. and Fell,D.A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. Lond. B Biol. Sci.*, **268**, 1803–1810.
11. Rison,S.C., Teichmann,S.A. and Thornton,J.M. (2002) Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*. *J. Mol. Biol.*, **318**, 911–932.
12. Simeonidis,E., Rison,S.C., Thornton,J.M., Bogle,I.D. and Papageorgiou,L.G. (2003) Analysis of metabolic networks using a pathway distance metric through linear programming. *Metab. Eng.*, **5**, 211–219.
13. Lemer,C., Antezana,E., Couche,F., Fays,F., Santolaria,X., Janky,R., Deville,Y., Richelle,J. and Wodak,S.J. (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res.*, **32**, D443–D448.
14. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.
15. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
16. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.