BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 7 2005, pages 1189–1193
doi:10.1093/bioinformatics/bti116

Systems biology

Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC)

S. A. Rahman, P. Advani, R. Schunk, R. Schrader and Dietmar Schomburg*

Cologne University BioInformatics Center (CUBIC) and Institute of Biochemistry, Zülpicher Strasse 47, 50674 Köln, Germany

Received on August 12, 2004; revised on October 21, 2004; accepted on October 21, 2004
Advance Access publication November 30, 2004

ABSTRACT

Motivation: Pathway Hunter Tool (PHT), is a fast, robust and user-friendly tool to analyse the shortest paths in metabolic pathways. The user can perform shortest path analysis for one or more organisms or can build virtual organisms (networks) using enzymes. Using PHT, the user can also calculate the average shortest path (Jungnickel, 2002 *Graphs, Network and Algorithm*. Springer-Verlag, Berlin), average alternate path and the top 10 hubs in the metabolic network. The comparative study of metabolic connectivity and observing the cross talk between metabolic pathways among various sequenced genomes is possible.

Results: A new algorithm for finding the biochemically valid connectivity between metabolites in a metabolic network was developed and implemented. A predefined manual assignment of side metabolites (like ATP, ADP, water, CO_2 etc.) and main metabolites is not necessary as the new concept uses chemical structure information (global and local similarity) between metabolites for identification of the shortest path.

Availability: PHT is accessible at http://www.pht.uni-koeln.de

Contact: d.schomburg@uni-koeln.de

INTRODUCTION

With the advent of the 'omics' era, more and more system-based approaches to biological functions are being developed. Metabolome analysis and metabolomics are gaining significance, as they help us to understand the complexity of the underlying cellular networks in organisms. The completion of a large number of genomes has made the comparative study of genomes possible at different levels. One way to gain a better understanding of the sequenced genomes is to analyse the underlying metabolic network and its topology in different genomes. Several databases provide information about metabolic pathways. We have used KEGG (Kanehisa *et al*., 2004) as the basic database for our analysis apart from BRENDA (Schomburg *et al*., 2004) and PROSITE (Hulo *et al*., 2004). A global view of the connectivity in metabolic pathways, and the contribution and usage of certain metabolites in these pathways is highly instructive. Shortest path analysis (Arita, 2004) is one of the most comprehensive methods to analyse a graph (Metabolic Pathways) at different levels (Reactions) in terms of local and global connectivity between metabolites. With the Pathway Hunter Tool (PHT) it is also possible to calculate statistical information (Barabasi and Oltvai, 2004) from the topology arising from the interacting molecules in order to capture the nature of connectivity.

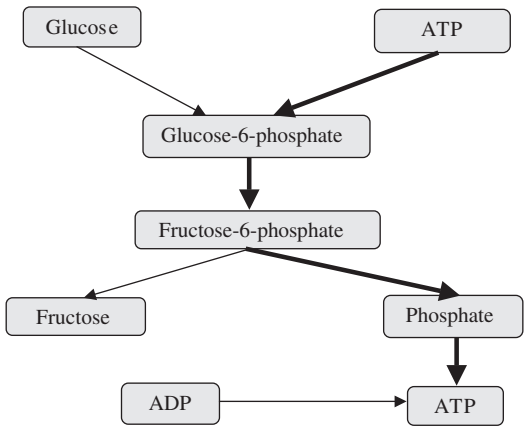

Fig. 1. This figure exemplifies the problem of finding a valid shortest path in the biochemical network.

METHODS

Whereas a number of well-established methods exist for the analysis of shortest paths in graphs, the situation in metabolic networks is a little more complicated. In the example using reactions given in Figure 1, a shortest path algorithm for metabolic pathways is required to follow the path of the thick lines with the result that there exists a path between phosphate and ATP via glucose-6-phosphate (thick line). However, there is no way to produce fructose from phosphate to fructose (thin line). In the third reaction of the scheme, the algorithm has to decide in which direction to go, depending on the starting point being either glucose or phosphate.

Therefore it is important to connect two metabolites in a reaction with respect to their structural similarity. We have used the fingerprint algorithm from the Chemistry Development Kit (CDK) (Steinbeck *et al*., 2003) to convert the 2-dimensional chemical structure information to a 1-dimesional binary stream as a fingerprint for faster similarity search (Whittle *et al*., 2003). Using the fingerprints, the similarity between two molecules was calculated using a normalized scoring function obtained by combination of the atomic mass value of the metabolites and the Tanimoto algorithm (Xue *et al*., 2003). This allows avoidance of the false connectivity in the metabolic pathway thus made the path search algorithm more robust.

In order to calculate the shortest path between two metabolites, the breadth-first-search (BFS) algorithm (Jungnickel, 2002) is used in PHT. Higher-order horn logic (HOHL) (Nadathur and Miller, 1990) has been used to satisfy the

*To whom correspondence should be addressed.

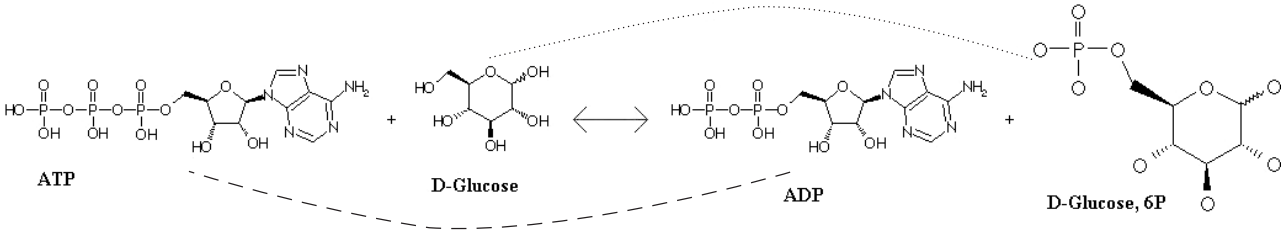

Fig. 2. Metabolite mapping obtained from our new algorithm shows that ATP maps to ADP (dashed line) and D-glucose maps to D-glucose-6phosphate (dotted line).

constraints. Our new algorithm automatically discriminates between side metabolites (like ATP, ADP, water, CO_2, etc.) and main metabolites while finding the shortest path without the need to predefine those. Predefined exclusion of small metabolites in the metabolic pathway may lead to broken links in the network or longer connectivity. This implies that at each reaction step the algorithm should be able to decide which metabolite to choose for further connectivity in the pathway and which to skip.

ALGORITHM

In this section the new algorithm used in PHT to find the shortest path in the biochemical network is described.

1 Definition of the metabolite mapping scoring function

Let A be an educt and B a product metabolite and a and b the number of bits (calculated by the fingerprint algorithm from the CDK, Steinbeck *et al*., 2003) 'on' on A and B metabolites, respectively, c the number of bits 'on' in both A and B, d number of bits 'off' in both A and B. Then we can define the equation in the form of set theory (Jech and Jech, 1997):

$$a = |A|, \quad b = |B|, \quad c = |A \cap B|, \quad d = n - |A \cup B|$$

and

$$a + b - c = |A \cup B|$$

(note: '$|B|$' denotes cardinality of the set), where n is the total number of attributes of an object (e.g. bits in a fingerprint).

Once we are able to formulate the chemical structure (Whittle *et al*., 2003) in terms of set theory the next step was to develop a scoring scheme for the similarity between two metabolites. We have used the Tanimoto Coefficient (Willet *et al*., 1998) for this purpose. The structural similarity between two metabolites A and B can be defined as

$$\text{Tanimoto Coefficient } S_{\text{A,B}} = \frac{|A \cap B|}{|A \cup B|}$$

The percentage atomic mass contribution (PAMC) for two competing educt (A) and product (B) can be defined as hundred times the sum of mass for both the metabolites (A and B) divided by the total mass of the metabolites in that reaction.

$$\text{atomic mass contribution PAMC}_{A,B} = \frac{100 * (M_A + M_B)}{\sum M_R}$$

The mapping scoring function is then defined as the product of similarity score and atomic mass contribution in each reaction between every two competing educt (A) and product (B) metabolites.

The final score for top competing metabolites can be defined as:

$$\text{Score}_{A,B} = \text{PAMC}_{A,B} * S_{A,B}, \quad 0 \le S_{A,B} \le 1,$$
$$0 \le \text{PAMC}_{A,B} \le 100$$

2 Local mapping metabolites in reactions

The derived scoring function was used to find a suitable mapping between substrate molecules and product molecules. We use a slightly modified form of game theory (http://www.gametheory.net/) in order to map the substrate to the product metabolite. The method consists of construction of a matrix with substrates as rows and products as columns with the score defined above as matrix elements. The score between any substrate or product whose extension is smaller than three bonds is set to zero. A substrate is mapped to a product when the score dominates all other scores in the present row or column respectively. By this procedure we keep track of the maximum structural similarity between two interacting metabolites. Figure 2 illustrates the outcome of our mapping procedure when applied to a reaction.

3 Shortest path between two metabolites

For the calculation of the shortest paths, the two biochemical criteria 'local' and 'global' structural similarity are used, where 'local similarity' is defined as the similarity between two intermediate molecules and 'global similarity' is defined as the amount of conserved structure found between the source metabolite and the destination metabolites after a series of reaction steps (Fig. 3).

The only potential drawback of this method is that not all metabolites in the metabolite databases have structures (e.g. macromolecules like proteins or nucleic acids, or generic molecules like 'an alcohol'). In these cases the user may miss some connectivity due to lack of structural information. It is possible to cross-check this result by switching off the 'Atom Mapper' (local similarity) and 'Atom Tracer' (global similarity) options thereby performing the search on the ligand-number-based mapping obtained from the KEGG reaction database. On the other hand, the power and biochemical relevance of having local similarity and global similarity is very high. In the future we plan to provide non-standard structural information for these metabolites in order to allow the inclusion of such reactions.

Complexity of the algorithm

The shortest path between source and destination metabolite is the minimum number of reaction steps between them (Fig. 4). We consider the metabolic pathway in our system as a directed graph with all edges (reactions) sharing the same cost (here 1). Hence this does not lead us to an NP-complete problem as one can calculate the

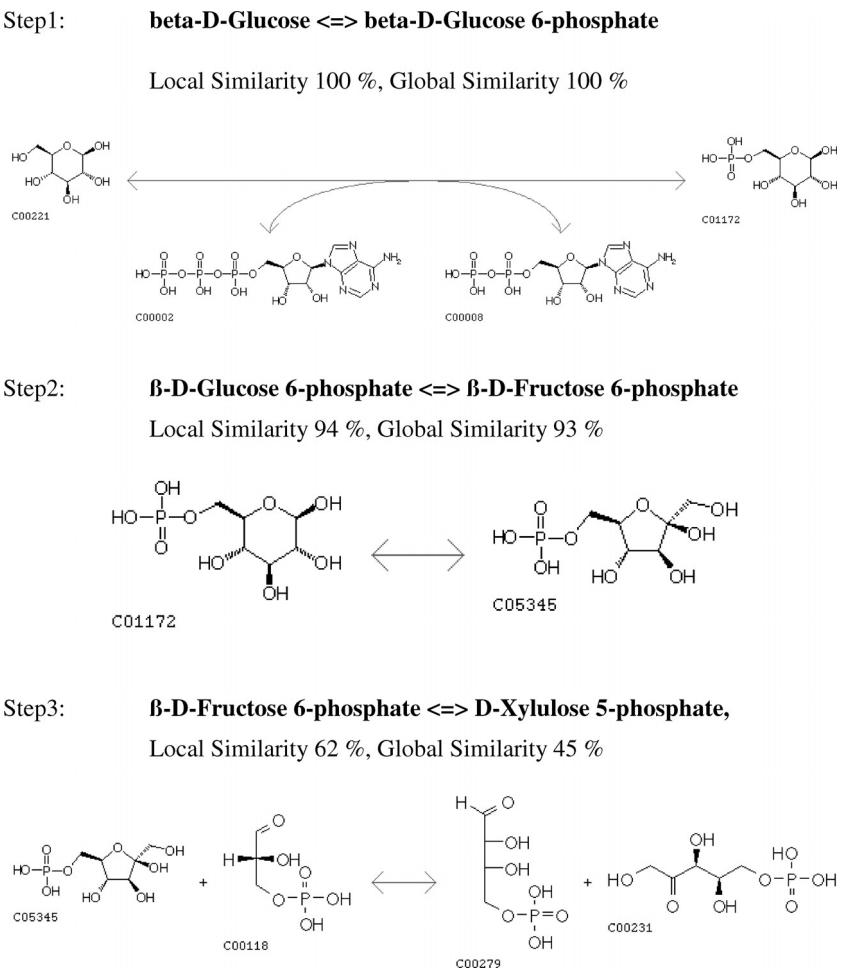

Fig. 3. Shortest path between metabolities β-D-glucose to D-xylulose 5-phosphate is in three steps and only 45% of the structural is common between them globally.

k-shortest path between two metabolites using the BFS (breadth first search) algorithm. HOHL (Nadathur and Miller, 1990) has been used to satisfy the constraints (similarity) with the BFS algorithm in order to calculate *k*-shortest paths between two metabolites (source and destination). This means that the runtime of the tool depends on the metabolites and reactions present in an organism. We are able to generate all possible *k*-shortest paths between two metabolites under given criteria of global and local similarity.

Program options

Presently PHT has four options.

(1) Find k-shortest path to convert one metabolite into another in a given network (organism-specific or general metabolic network).

(2) Find k-shortest paths from a substrate metabolite to all feasible metabolites in a given network (organism-specific or general).

(3) Find k-shortest path to a product metabolite from all feasible substrate metabolites in a given network (organism-specific or general).

(4) Statistical analysis of the metabolic pathways like average path length, diameter of the network, average node connectivity, loose ends in the network, hubs in a given network (organism-specific or general).

User defined constraints

There are sets of user-defined constraints, which can be used for an in-depth network analysis without affecting the biochemical/biological relevance.

- While traversing through the metabolic pathway it is possible to set the similarity measure score (*Atom Mapper*) between interacting molecules and to define the amount of structure change with respect to this reference molecule at each reaction step (*Atom Tracer*).
- By setting the *Minimum path length* and *Maximum path length*, the path between two metabolites in the network can be altered. For example, if the minimum path length is set to 6, then the algorithm will drop paths below it and report the next possible shortest path above or equal to 6, which is the shortest possible path under the given constraint.

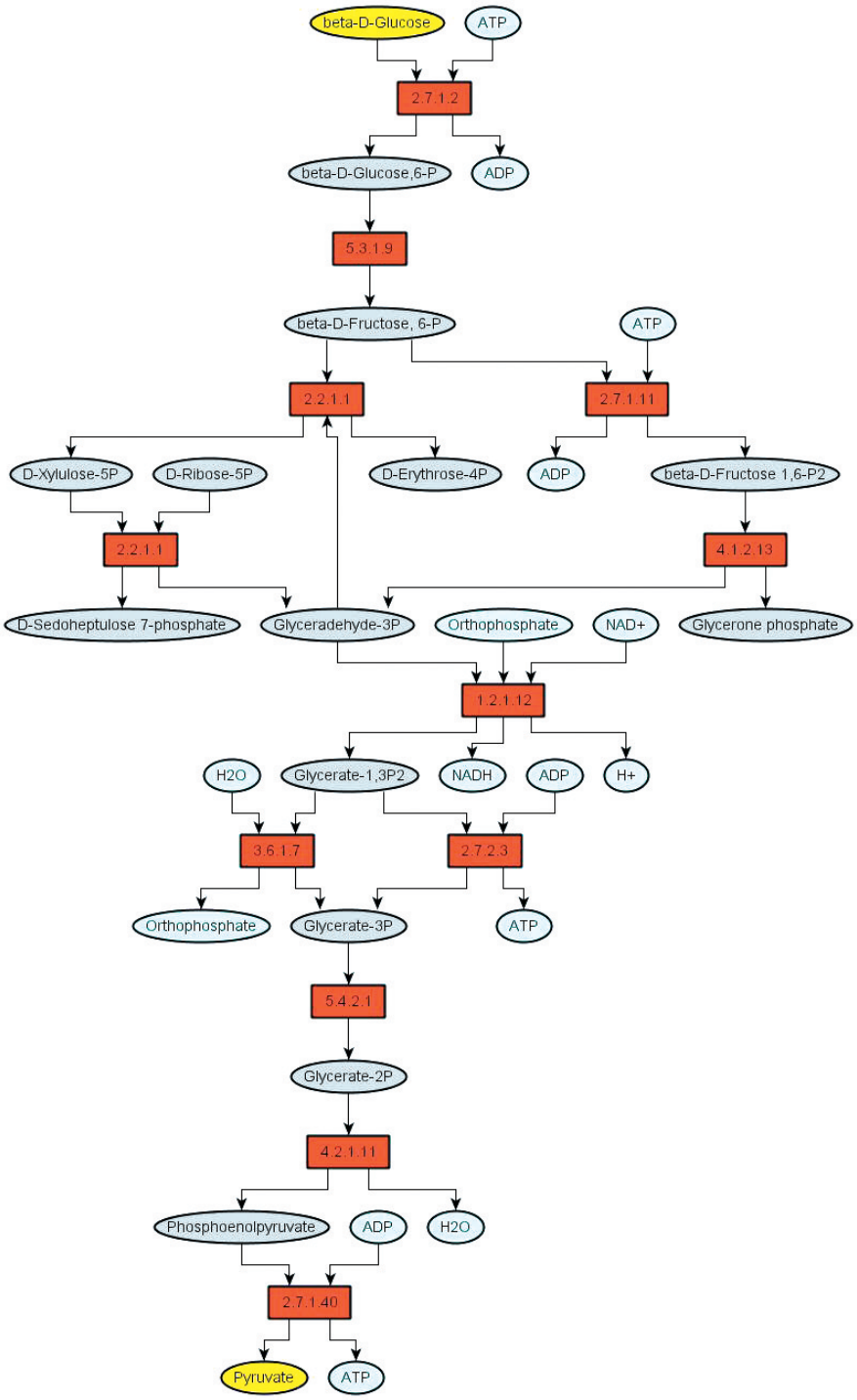

Fig. 4. The shortest path between metabolites β-D-glucose and pyruvate in *E.coli K-12* is nine reaction steps long.

- It is possible to choose *via Metabolite, not via Metabolites* and *not via Enzymes* options for use of a particular set of pathways.
- Under *Build Virtual Organism* it is possible to add one's own set of enzymes and perform further analysis. This is very useful for identification of the missing links in the network.

RESULTS

We performed a shortest path analysis (Fig. 4) in *Escherichia coli* K-12 between beta-D-glucose and pyruvate, which was found to be nine steps long. We considered global similarity and local similarity while traversing the path. The algorithm automatically

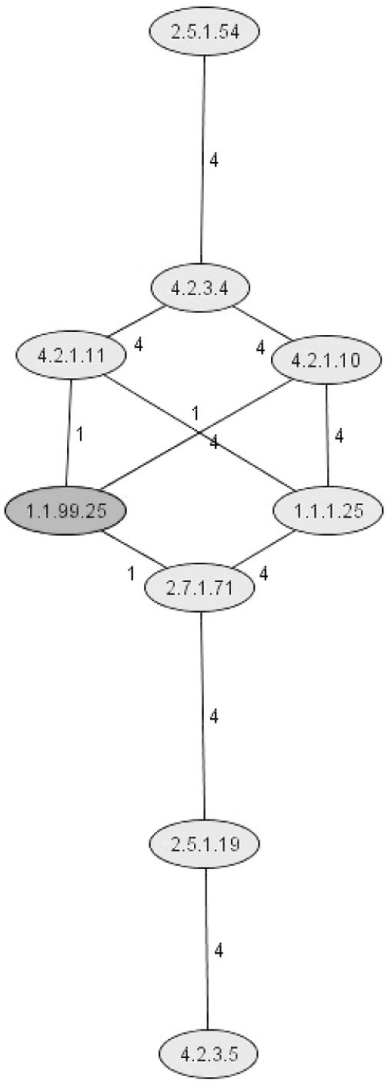

Fig. 5. Enzyme–enzyme connectivity map highlights the shortest path (seven reaction steps) between 'D-erythrose 4-phosphate' and 'chorismate' in the KEGG reference map and *C.glutamicum, E.coli K-12* and *M.tuberculosis*. The weights given at the connections reflect the number of occurrences of this step in the queried pathways. 1.1.99.25 is found only in the reference map (originating from *Acinetobacter calcoaceticus*).

identifies the correct connectivity between the metabolites at each reaction step.

We also performed a comparative study between the KEGG reaction reference map, *Corynebacterium glutamicum, E.coli* K-12 and *Mycobacterium tuberculosis* (Fig. 5). We were interested in finding the shortest path between 'D-erythrose 4-phosphate' and 'chorismate', which was found to be in seven reaction steps in all these cases.

Looking closely at Figure 5, it is clear that different pathways are possible to convert 'D-erythrose 4-phosphate' to 'chorismate' in the reference map, or in *C.glutamicum, E.coli* K-12 and *M.tuberculosis* (score 4 on the edge). Some organisms may use enzyme '1.1.99.25' (blue colour) to perform the same conversion (score 1 on the edge).

OUTPUT FORMAT

PHT generates three kinds of output:

- A *Text*-based output can be viewed immediately in the browsers and is supplied with hyperlinks to other database, like BRENDA, KEGG and PROSITE.
- A *Graphical* view of the output is generated for 'Metabolic Pathways' and 'Enzyme' connectivity as graph modeling language (GML) (http://www.infosun.fmi.uni-passau.de/Graphlet/GML/) files. These portable files can be saved on the client's system and can be viewed later in any dynamic layout software that can read the GML format (e.g. the yEd (http://www.yworks.com/products/yed/ graphical editor).
- An 'enzyme–enzyme' connectivity matrix, which can be used for pathway alignment and other studies. The 'reaction-organism matrix' highlights the presence of reactions in organisms by binary 1 and 0 for absence.

ACKNOWLEDGEMENTS

The authors are grateful for financial support by the German Federal Ministry for Education and Research (BMBF).

REFERENCES

Arita,M. (2004) The metabolic world of Escherichia coli is not small. *Proc. Natl Acad. Sci. USA*, **101**, 1543–1547.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32** (Database issue), D134–137.

Jech,T.J. and Jech,T. (1997) *Set Theory*. Springer-Verlag, Berlin.

Jungnickel,D. (2002) *Graphs, Network and Algorithm*. Springer Verlag, Berlin.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32** (Database issue), D277–280.

Nadathur,G. and Miller,D. (1990) Higher-order Horn clauses. *JACM*, **37**, 777–814.

Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32** (Database issue), D431–433.

Steinbeck,C., Han,Y., Kuhn,S., Horlacher,O., Luttmann,E. and Willighagen,E. (2003) The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500.

Whittle,M., Willett,P., Klaffke,W. and van Noort,P. (2003) Evaluation of similarity measures for searching the dictionary of natural products database. *J. Chem. Inf. Comput. Sci.*, **43**, 449–457.

Willet,P., Barnard,J.M. and Downs,G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, **38**, 938–996.

Xue,L., Godden,J.W., Stahura,F.L. and Bajorath,J. (2003) Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.*, **43**, 1151–1157.