Metabolic pathway analysis web service

(Pathway Hunter Tool at CUBIC)

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

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

Motivation: Pathway Hunter Tool (PHT) (Syed Asad Rahman et al., 2004) 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), average alternate path and the top 10 hubs in the metabolic network. The comparative study of metabolic connectivity and the cross talk between metabolic pathways between 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, CO2 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: Pathway Hunter Tool (PHT) is accessible at http://www.pht.uni-koeln.de.

Contact: asad.rahman@uni-koeln.de, D.Schomburg@uni-koeln.de

Keywords: Pathway Hunter Tool, PHT, Shortest Path, Metabolic Pathways, Metabolic Networks, Tools, Pathway Database.
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 higher attention and help 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 can be achieved by analysis of 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 pathway, the contribution and usage of certain metabolite in these pathways is highly instructive. Shortest path analysis (Arita, 2004) is one of the best-defined methods to analyse a graph (Metabolic Pathways) at different levels in terms of local and global connectivity. With 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.

METHOD

Whereas a number of long-established methods exist for the analysis of shortest paths in graphs the situation in metabolic networks is a little more complicated. In the example of 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) but that there is no way to produce fructose from phosphate to fructose (thin line). In the third reaction of the scheme the algorithm has to distinguish 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-dimensional 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 allowed to avoid the false connectivity in the metabolic pathway and made the path search algorithm more robust.

In order to calculate the shortest path between two metabolites, the depth first search (DFS) algorithm (Jungnickel, 2002) is used in PHT. Higher-Order Horn Logic (HOHL) (Nadathur and Miller, 1990) has been used to satisfy the constraints. Our new algorithm automatically discriminates between side metabolites (like ATP, ADP, Water, CO2 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 means 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 Pathway Hunter Tool (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 \( a, b \) the number of bits (calculated by the fingerprint algorithm from the Chemistry Development Kit (CDK) (Steinbeck et al., 2003)) “on” on \( A \) or \( B \) metabolites, respectively, \( c = \) the number of bits “on” in both \( A \& B \), \( d = \) number of bits “off” in both \( A \& B \), then we can define the equation in form of set theory (Jech and Jech, 1997).

\[
\begin{align*}
  a &= |A|, \\
  b &= |B|, \\
  c &= |A \cap B|, \\
  d &= n - |A \cup B| \\
\end{align*}
\]

and \( a + b - c = |A \cup B| \) (Note: ‘\( |B| \’ denotes cardinality of the set)

where \( n \) = 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 the development of a scoring scheme for the similarity between two metabolites. We have used the Tanimoto Coefficient (Willet et al., 1998)
for this purpose, i.e. the structural similarity between two metabolites A and B can be defined as

- Tanimoto Coefficient \( S_{A,B} = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \)

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 & B) divided by the total mass of the metabolites in that reaction.

- Atomic Mass Contribution \( \text{PAMC}_{A,B} = 100 \times \frac{(M_A + M_B)}{\sum M_R} \)

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.

Final score for top competing metabolites can be defined as

\[ \text{Score}_{A,B} = \text{PAMC}_{A,B} \times S_{A,B} \]

Where \( 0 \leq S_{A,B} \leq 1 \)

and \( 0 \leq \text{PAMC}_{A,B} \leq 100 \)

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 of 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 either 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. Fig 2 illustrates the outcome 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 given by the fact 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. In order to cross check this result it is possible to switch 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 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 NP-complete problem as one can calculate the k-shortest path between two metabolites using the BFS (Breadth First Search) algorithm. Higher-Order Horn Logic (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 Pathway Hunter Tool (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 his 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 six, then the algorithm will drop paths below it and report the next possible shortest path above or equal to six, which is the shortest possible path under the given constraint.
• 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 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 turned out to be nine steps long. We considered global similarity and local similarity while traversing the path. The algorithm automatically 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, Escherichia coli K-12 and Mycobacterium tuberculosis (fig. 5). We were interested in finding the shortest path between “D-Erythrose 4-phosphate” and “Chorismate”, which turned out to be in 7 reaction steps in all these cases. Looking closely into fig. 5 it is clear that different pathways are possible to convert “D-Erythrose 4-phosphate” to “Chorismate” in the reference map, or in Corynebacterium glutamicum, Escherichia coli K-12 and Mycobacterium 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

Pathway Hunter Tool (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://infosun.fmi.uni-passau.de/Graphlet/GML/) files. These portable files can be saved on the clients system and can be viewed later in any dynamic
layout software that read the GML format (e.g. the yEd graphical editor).

- Pathway Hunter Tool (PHT) also generates “Enzyme-Enzyme” connectivity matrix, which can be used for pathway alignment and other studies. The “Reaction-Organism Matrix” highlights the presence of reaction 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


FIGURES:

![Biochemical Network Diagram]

**Figure 1**: This figure exemplifies the problem of finding a valid shortest path in the biochemical network.
Figure 2: Metabolite mapping obtained from our new algorithm shows that ATP maps to ADP (green line) and D-Glucose maps to D-Glucose-6-phosphate (red line).
Step 1: beta-D-Glucose $\leftrightarrow$ beta-D-Glucose 6-phosphate

Local Similarity 100 %, Global Similarity 100 %

Step 2: beta-D-Glucose 6-phosphate $\leftrightarrow$ beta-D-Fructose 6-phosphate

Local Similarity 94 %, Global Similarity 93 %

Step 3: beta-D-Fructose 6-phosphate $\leftrightarrow$ D-Xylulose 5-phosphate,

Local Similarity 62 %, Global Similarity 45 %

Figure 3: Shortest path between metabolites beta-D-Glucose to D-Xylulose 5-phosphate is in 3 steps and only 45% of the structural is common between them globally.
Figure 4: The shortest path between metabolites beta-D-Glucose and Pyruvate in *Escherichia coli* K-12 is 9 reaction steps long.
Figure 5: Enzyme-Enzyme connectivity map highlights the shortest path (7 reaction steps) between “D-Erythrose 4-phosphate” and “Chorismate” in the KEGG reference map and Corynebacterium glutamicum, Escherichia coli K-12 and Mycobacterium 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).