

2.4.5.2. Combining Molecular Fingerprints and Maximal Common Subgraph

The previously described algorithm (2.4.4,2.4.4.1,2.4.4.2,2.4.4.3) based solely on molecular fingerprint and percentage atomic weight was too stringent to find all the possible molecular interactions. Hence, a dynamic weighted matrix based on the combination of molecular fingerprinting (refer to section 2.4.4.1) and the maximal common subgraph discovery algorithm (refer to section 2.4.5) was used to identify all the connected subgraphs (metabolites) in the reaction process.

2.4.5.3. Mapping Function for the Dynamic Weighted Matrix

Problem: *Given a subgraph G' of reaction and compounds, find all the possible links between the substrate and product metabolites based on the molecular structure similarity.*

Solution: *In order to identify most similar structures a constraint based search approach was conducted. Starting point for the analysis was a matrix M (refer to section 2.4.4.3), where each row M_u and column M_v in the matrix is filled with the similarity $S_{u,v}$ (refer to section 2.4.4.3 and equation (2.15)) for a specified subgraph G' (reaction, metabolites). The search for the most similar structures is conducted by maximizing the value of M_u in mutual dependency of M_v and vice versa or, in other words by minimizing the loss.*

Lemma 1: (Calculation of similarity between two molecules) *Let $H'_{1,2}$ be the MCS (refer to section 2.4.5) of given graphs G_1 (substrate) and G_2 (product). The selected pair (substrate and product) is then mapped using MCS algorithm and the matched sub-structure pattern is removed from the data set (using divide and conquer technique). Hence the resulting graph is $\square H'_{1,2}(G_1)$ and $\square H'_{1,2}(G_2)$.*

The matrix is refilled with the similarity score $S_{u,v}$ between the remaining metabolite structures in the matrix and the process of selection is started iteratively again until no more mapping can be done or the similarity score equals zero (for a perfectly

balanced reaction)(Figure 15) (Figure 16). Here the length of the matrix is dependant on the stoichiometry* of the substrate and products.

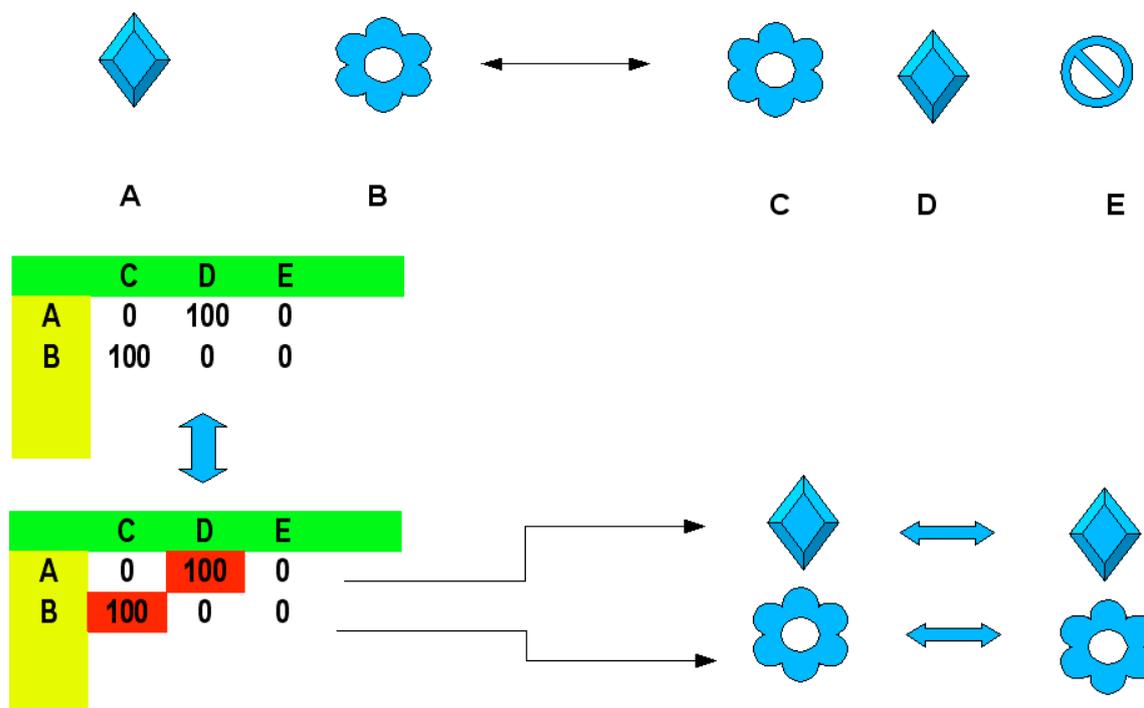


Figure 15. The remaining structures in the data set are mapped using the same concept of MAX-MAX domination of the similarity score (S) in row/column of the matrix (M). The mapping shows A being mapped to D and B being mapped to C. The mapped structures are again removed from the dataset (This is done by mapping them using MCS algorithm. The common set of nodes and edges are deleted from the structure graph). As there is no further data to be mapped in the dataset, the algorithm automatically terminates the mapping search.

* "Die stöchyometrie (Stöchyometria) ist die Wissenschaft die quantitativen oder Massenverhältnisse zu messen, in welchen die chymischen Elemente gegen einander stehen." [Stoichiometry is the science of measuring the quantitative proportions or mass ratios in which chemical elements stand to one another.] ... Jeremias Benjaim Richter, 1792

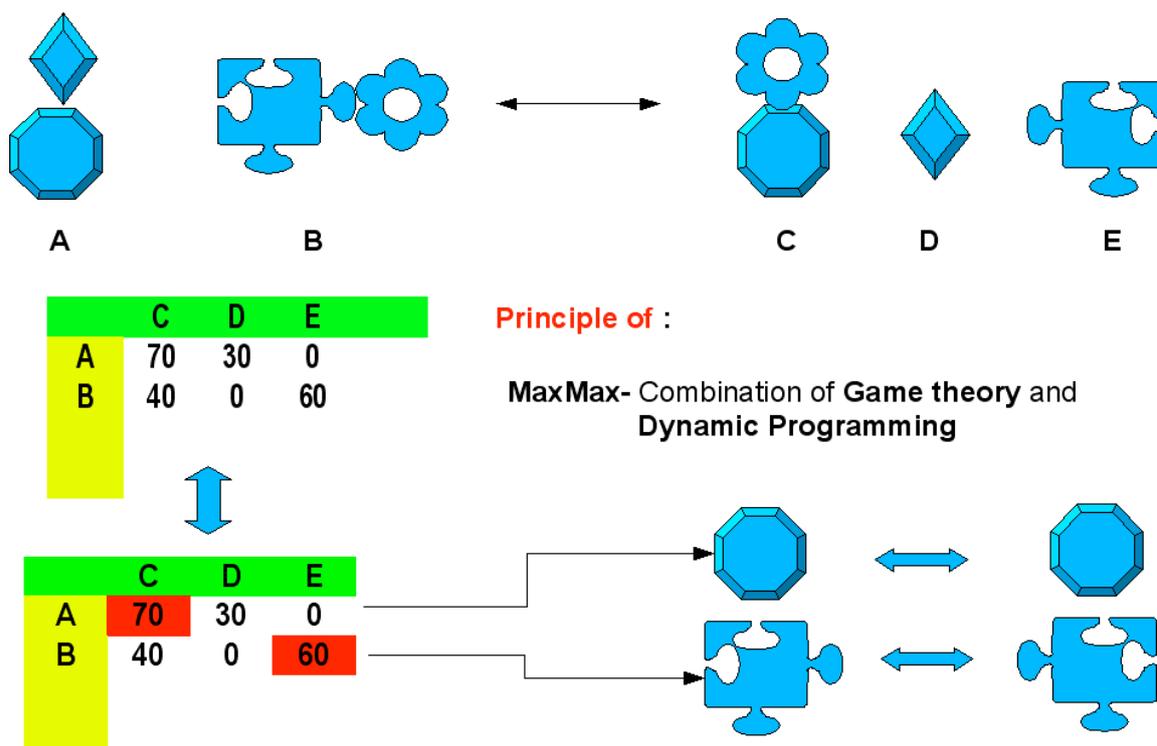


Figure 16. Based on the structural similarity score, structure A is mapped to structure C and structure B is mapped to structure E, as they dominate the scores in rows/columns in the Dynamic Weightage Matrix (M). The mapped part of the structure is then removed from the data set (This is done by mapping them using MCS algorithm. The common set of nodes and edges are deleted from the structure graph).

Thus we were able to find better mapping between substrate and product as each and every pattern, including the sub-structural pattern was mapped (Figure 17). This allows us to understand the conserved functional structural patterns and identify conserved links between different cellular processes.

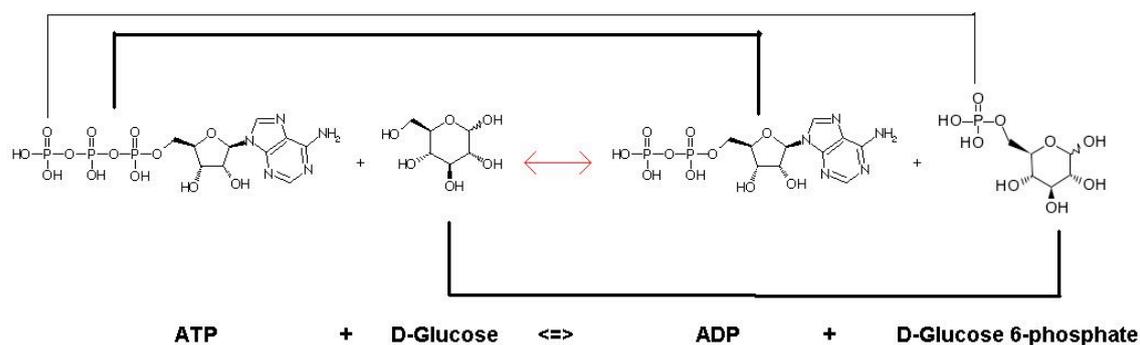


Figure 17. Metabolite mapping obtained from our old algorithm shows that ATP maps to ADP (thick black line) and D-Glucose maps to D-Glucose-6phosphate (thick black line), whereas the improved algorithm can also predict mapping between ATP and D-Glucose-6phosphate (thin black line). The new algorithm is able to match all the possible structure patterns between substrate and product. This algorithm can further be extended for classification of reactions and enzymes apart from path-finding.

2.4.6. Adaptive Breadth First Search Algorithm (BFS) Under Constraints

Problem: Find the k -shortest path between source u and sink v , such that $d_G(u, v)$ of two vertices $u, v \in V$ is the minimum number of edges of shortest paths between u and v in G .

Input: Given a graph pruned network G' of the original $G = (V, E, L)$: $G' \subseteq G$.

Solution: Applying the breadth first search algorithm on the pruned network G' , and find all the possible k -path of same length from a single source u to all the sinks v , via set of edges E

The modified BFS algorithm of Newman (Newman 2001) was further modified by constraining search space in the un-weighted graph \mathbf{G} of the metabolic network. Providing the algorithm with some knowledge about the search space did this.

Definition 18: Given a bipartite graph $G = (V, E, L)$ and its pruned graph $G' = (V', E', L)$ obtained from section 2.4.5.3 apply the BFS algorithm for finding shortest path σ_{uv} in the network. If the given network is a directed bipartite graph $|G|$, then the search space will further reduce by direction constraints ε^* in E' .

Proof:

Condition (a): Given a set of subgraphs $G_i = (V_i, E_i, L)$ of a graph \mathbf{G} . Where each source $u \in V(G_i)$ and $v \in V(G_i)$ sink is associated with reaction $L(v) \subseteq L$ and each edge $uv \in E(G)$ represents the reaction between u and v .

Condition (b): Given a pruned subgraph $G'_i = (V'_i, E'_i, L)$ of a graph \mathbf{G} . Where each source $u' \in V'(G')$ and $v' \in V'(G')$ sink is associated with reaction $L(V') \subseteq L$ and each edge $u'v' \subseteq E'(G)$ represents the reaction between u' and v' . As $|E| \geq |E'|$, therefore $|G'| \leq |G|$.

2.4.7. Tracking Changes in Metabolite Structural Patterns

The metabolic network is a collection of metabolites that share certain common pattern between them to form links. Hence keeping track of the changes in these structure patterns in a defined path will help us understand the intricacies of the network.

The tracing of the metabolite structure patterns can be divided into two parts. While traversing through the metabolic pathway it is possible to set the similarity measure score (**Atom Mapper**) between interacting molecules, and to define minimum amount of structure conserved with respect to this reference molecule at each reaction step (**Atom Tracer**).

2.4.7.1. Local Similarity (Atom Mapper)

Local similarity is defined as the minimum amount of structure change conserved between two interacting metabolites (substrate and product) in the network with respect to the substrate metabolite.

Let u, v be the intermediate metabolites in a given graph G (2.20)

If we define χ_u as the set of all elements χ_{u_i} in vector χ_u whose value is 1 (the "on" bits) and χ_v as the set of all elements χ_{v_i} in vector χ_v whose value is 1. Hence applying equation (2.8) and (2.10), we can deduce the following equation (2.21). Hence the "Local Similarity" L_S between two intermediate metabolites in the network can be defined as

$$L_S = \left(\frac{\chi_u \cap \chi_v}{\chi_u} \right) \times 100 \quad (2.21)$$

2.4.7.2. Global Similarity (Atom Tracer)

Let u'_i, v'_i be the intermediate metabolites in a given subgraph G'_i of G (2.22)

Let u'_{i+1}, v'_{i+1} be the intermediate metabolites in a given subgraph G'_{i+1} of G (2.23)

Let u'_{n-1}, v'_{n-1} be the intermediate metabolites in a given subgraph G'_{n-1} of G (2.24)

Let u'_n, v'_n be the intermediate metabolites in a given subgraph G'_n of G (2.25)

where as $u'_i = u'_{i+1}, \dots, u'_{n-1} = v'_n$ (2.26)

Hence we can trace a common substructure χ_k along a give shortest path σ_{u_i, v_n} path by the following equation based on the equations (2.22), (2.23), (2.24), (2.25), (2.26)

$$\chi_k = u'_i \cap u'_{i+1} \cap u'_{i+2} \cap u'_k \dots \cap u'_{n-1} \cap v'_n \quad (2.27)$$

Thus the “Global Similarity” (2.28) along between two metabolites (source and sink) can be derived by equations (2.27),(2.26)

$$L_G = \left(\frac{\chi_k}{u_i} \right) \times 100 \quad (2.28)$$

2.4.8. Pathway Analysis under Constraints

Predefined exclusion of small metabolites (like ATP, ADP etc) or vertices in the graph 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. Our new algorithm automatically discriminates between side metabolites (like ATP, ADP, Water, CO₂ etc) and main metabolites while finding the shortest path. In order to increase the flexibility of the path finding algorithm, few constraints are allowed on the system.

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** refer to section 2.4.7.1) between interacting molecules and to define the amount of structure change with respect to his reference molecule at each reaction step (**Atom Tracer** refer to section 2.4.7.2).
- By setting the **Minimum path length** (defined by shortest path) and **Maximum path length** (a heuristic measure defined by next shortest path by not including all the previous visited edges in shortest path) between two metabolites in the network, the reported path 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.
- A reaction cannot **repeat itself** in a given path or **no cycle** can be observed in the reported paths, both at level of metabolites and reactions. Also no substrate in a path can become connected product in the next step.
- It is possible set certain on constraints on the path finding algorithm like **via Metabolite** (preference for certain vertices in the graph), **not via Metabolites** (exclusion/deletion of certain vertices in the graph) and **not via Enzymes** (exclusion/deletion of certain edges in the graph).

- It is possible to build a tailored network of metabolites and reactions from the existing network as reference in the database under the option of **Build Virtual Organism**. This is very useful for identification of the missing links in the network.

2.4.9. Variants of the Algorithm

The algorithm in Pathway Hunter Tool (PHT) supports four options.

- Find k-shortest path to convert one metabolite into another in a given network (organism-specific or general metabolic network).
- Find k-shortest paths from a substrate metabolite to all feasible metabolites in a given network (organism-specific or general).
- Find k-shortest path to a product metabolite from all feasible substrate metabolites in a given network (organism-specific or general).
- 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).

2.5. Results

This section highlights the different properties of a metabolic path. All the analyses are carried out using Pathway Hunter Tool (PHT).

2.5.1. Shortest Path Analysis Without Atom Mapper and Atom Trace

The shortest path analysis between Pyruvate and L-Lysine in an amino acid producing bacteria *Corynebacterium glutamicum* results (Figure 18) (Figure 19) in 2 steps (undirected network) and 3 steps (directed network) respectively. Both the reported shortest paths are not valid in biochemical context as they are through a side metabolite or a promiscuous metabolite (highlighted in red). Hence we need an algorithm to skip such promiscuous metabolites while finding shortest path.

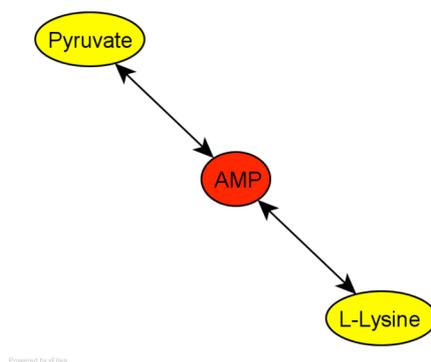
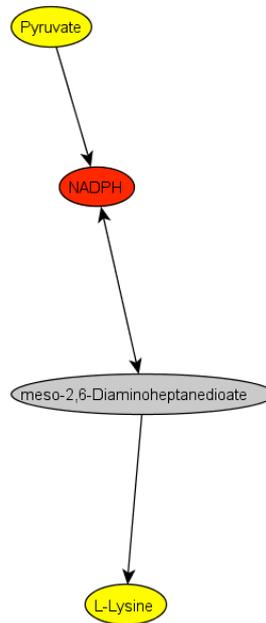


Figure 18. Shortest path between Pyruvate and L-Lysine in *Corynebacterium glutamicum* (cgl) on an undirected metabolic network without mapping information



Powered by yFiles

Figure 19. Shortest path between Pyruvate and L-Lysine in *Corynebacterium glutamicum* (cgl) on a directed metabolic network without mapping information

One option is to delete all such promiscuous metabolites from the network by manually assigning the valid connectivity in the pathway as in the case of KEGG pathway chart. Using such manually curated pathway charts like KEGG may give valid and known biochemical information. Here the shortest path results in 11 reaction steps (Figure 20).



Figure 20. Shortest path between Pyruvate and L-Lysine in *Corynebacterium glutamicum* (*cgl*) on a directed metabolic network with KEGG mapping information

2.5.2. Mapping Examples between Metabolites in a Pathway

One of the shortest paths between beta-D-glucose and 6-Phospho-D-gluconate highlights the structural similarities between the connecting metabolites. The mapping of metabolites in the pathway and their structural similarity based upon binary fingerprint is highlighted in (Figure 21).

2.5.3. Shortest Path Analysis with Atom Mapper and Atom Tracer

The shortest path between pyruvate and L-lysine in *Corynebacterium glutamicum* comprises of 7 reaction steps as per our mapping algorithm (Figure 22). The “Atom Mapper” was set to 15% (refer to 2.4.7.1) and “Atom Tracer” (refer to 2.4.7.2) was set to 5%. The obtained path is shorter than the normal shortest path obtained by KEGG pathway map by 4 reaction steps (refer to section 2.5.1). This is indeed a strong indicator that biochemical mapping would enable the analysis of genome-environment interactions, such as the prediction of new pathways (Figure 23). Such a study can highlight reactions, enzymes and genes that would degrade new environmental compounds.

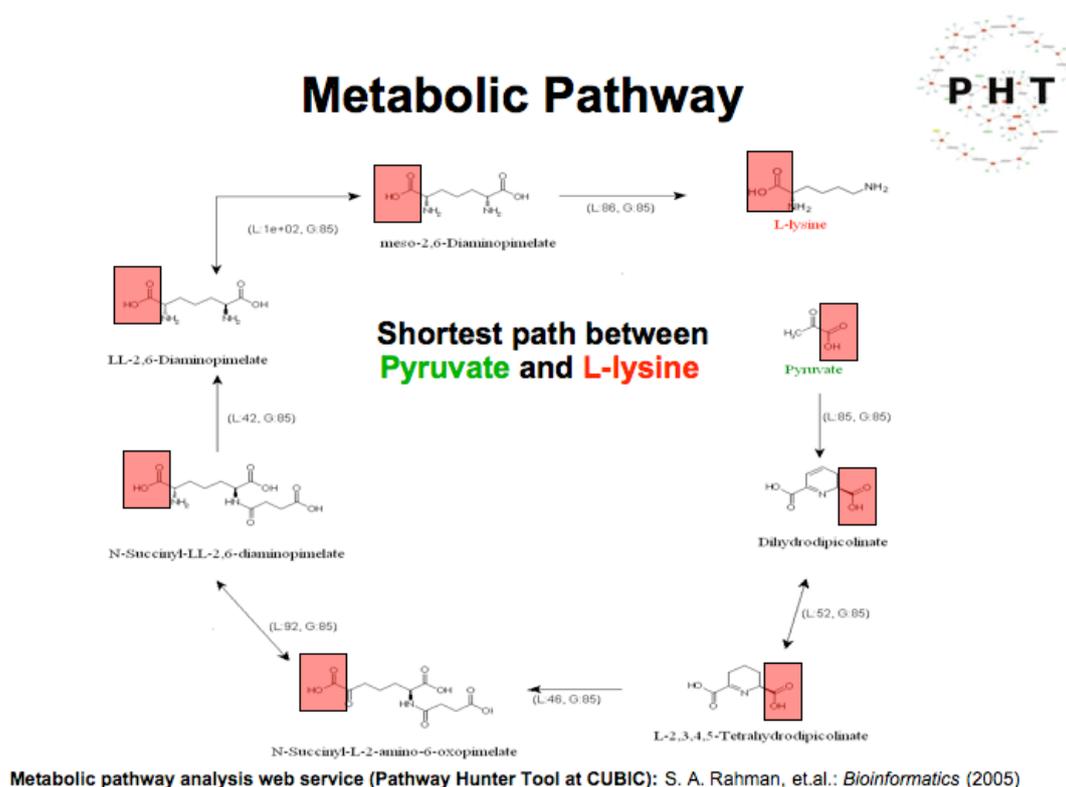
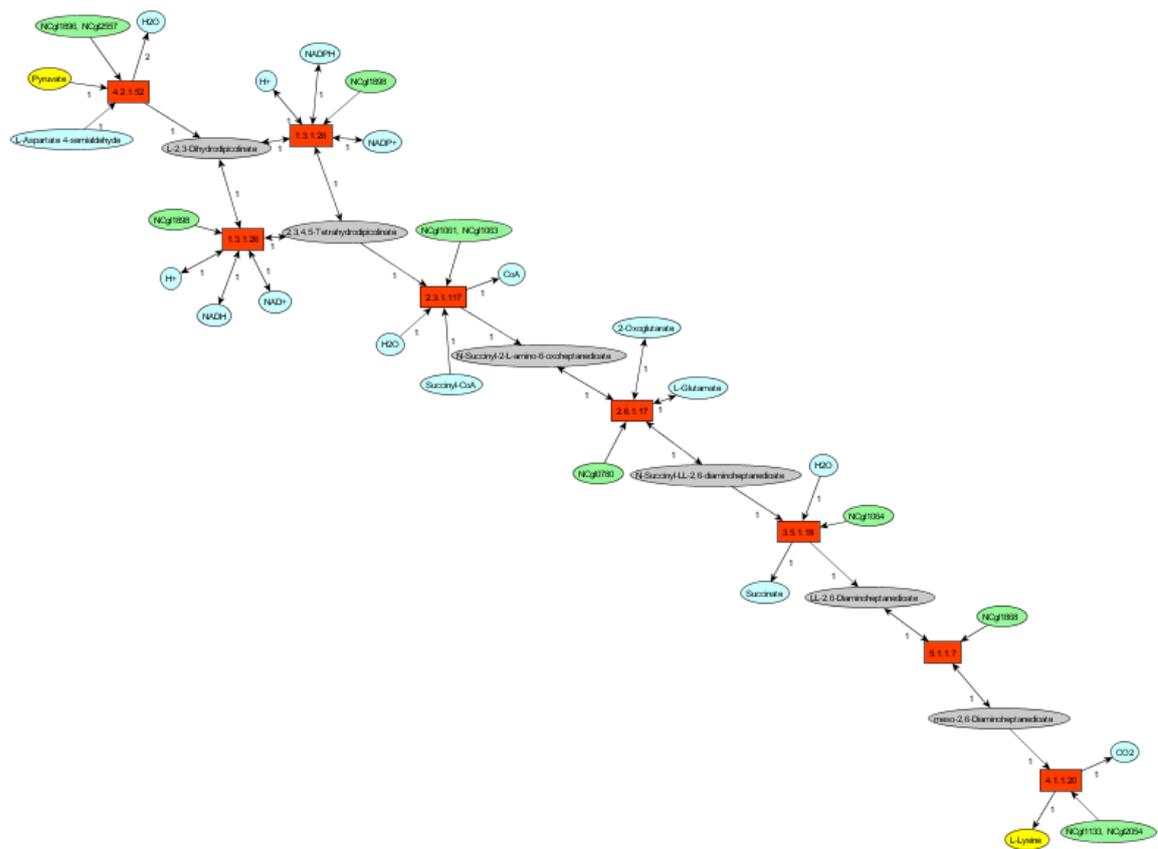


Figure 22. Shortest path between pyruvate and l-lysine in *Corynebacterium glutamicum* comprises of 7 steps as predicted by new mapping algorithm. This path is much shorter than the usual path obtained from hand-drawn maps available through KEGG.



Powered by yFiles

Figure 23. Shortest path between pyruvate and l-lysine in *Corynebacterium glutamicum* comprises of 7 steps as predicted by new mapping algorithm. This picture highlights the genes, metabolites and enzymes required for conversion (as obtained using Pathway Hunter Tool).

2.5.3.1. Local and Global Similarity Score and its Impact on Path Finding

The local and global similarity score define the minimum amount of structural similarity expected in the reported path. In order to demonstrate the impact of these scoring functions *Escherichia coli K12 MG1655* was chosen as our model organism. The shortest path between “beta-D-glucose” and “pyruvate” was calculated. If the local similarity is increased step by step till no more paths are found, the path length increases with similarity score (Table 3). The impact of global similarity on the path length is very drastic as seen in the following example (Table 4). Hence, number of links in the network is inversely proportion to the local and global similarity score.

Table 3. Local similarity and its impact on the shortest path. Higher the similarity score, longer the path as many links are skipped.

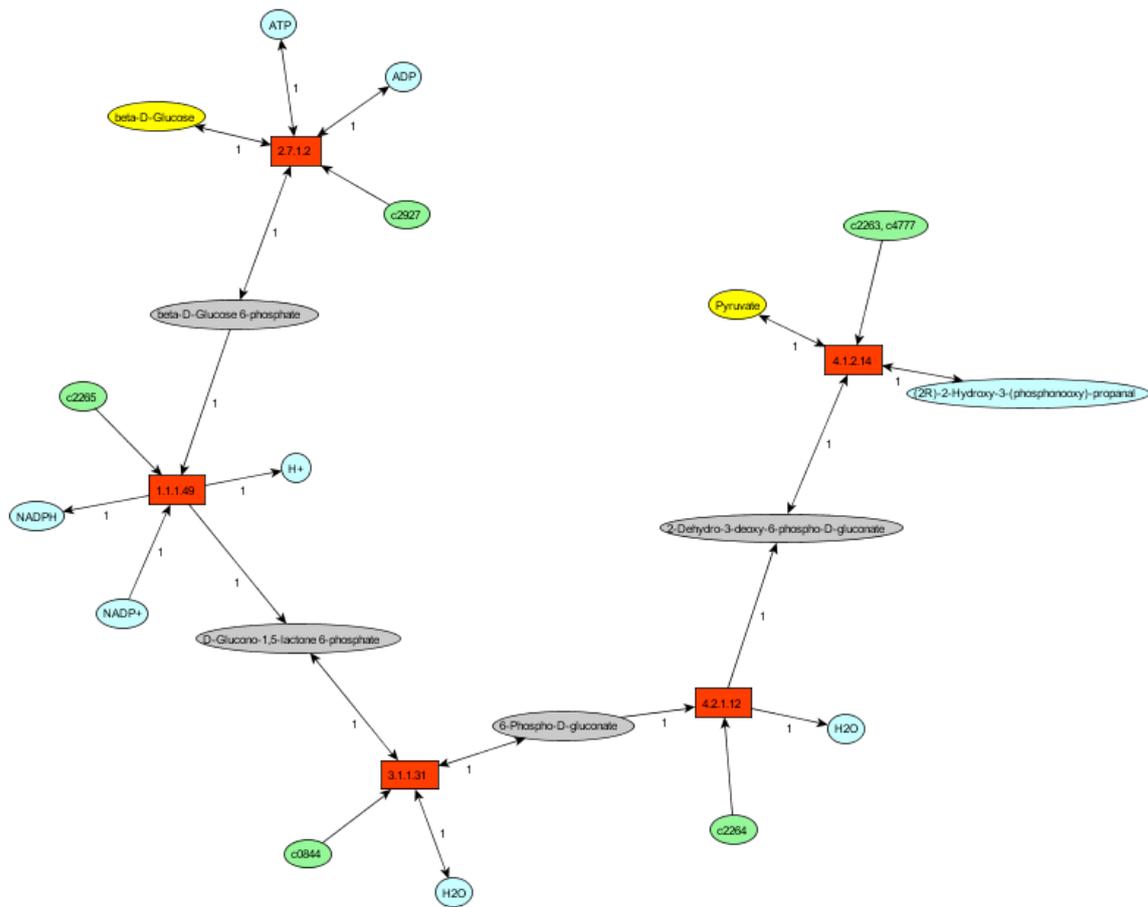
| Atom Mapper | Atom Trace | Path Length | Number of Paths | |
|-------------|------------|-------------|-----------------|--|
| 10 | 5 | 5 | 1 | |
| 15 | 5 | 5 | 1 | |
| 20 | 5 | 6 | 2 | |
| 25 | 5 | 7 | 1 | |
| 30 | 5 | 7 | 1 | |
| 35 | 5 | 7 | 1 | |
| 40 | 5 | 10 | 1 | |
| 45 | 5 | 0 | 0 | |
| 50 | 5 | 0 | 0 | |
| 60 | 5 | 0 | 0 | |

Table 4. Impact of the Global similarity on the length of the reported shortest path.

| Atom Mapper | Atom Trace | Path Length | Number of Paths |
|-------------|------------|-------------|-----------------|
| 10 | 1 | 5 | 1 |
| 10 | 5 | 5 | 1 |
| 10 | 10 | 5 | 0 |
| 10 | 15 | 0 | 0 |
| 15 | 1 | 5 | 1 |
| 15 | 5 | 5 | 1 |
| 15 | 10 | 5 | 1 |
| 15 | 15 | 0 | 0 |
| 15 | 20 | 0 | 0 |
| 20 | 1 | 6 | 2 |
| 20 | 5 | 6 | 2 |
| 20 | 10 | 7 | 1 |
| 20 | 15 | 0 | 0 |

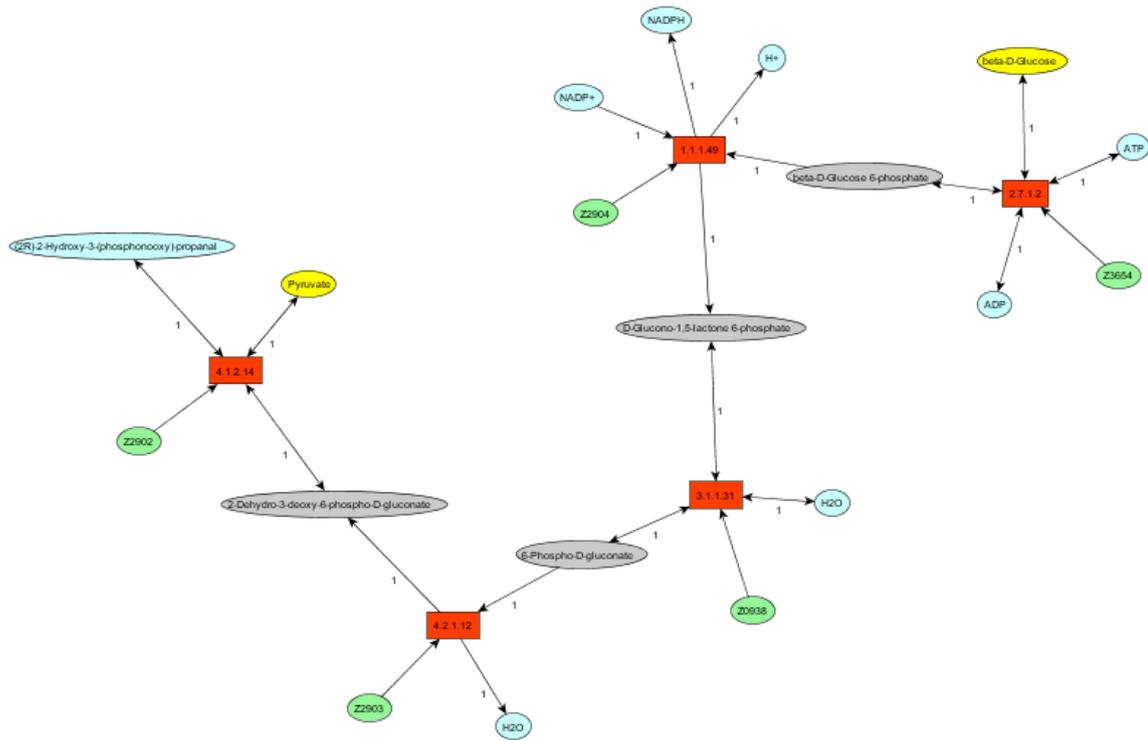
2.5.4. Shortest Path Analysis of the *Escherichia coli*

The shortest path in various *Escherichia coli* strains was calculated using metabolites mapping information (the algorithm explained in sections 2.4.5 and 2.4.7). The shortest path between beta-D-glucose and pyruvate using “Atom Mapper” (refer to 2.4.7.1) as 15% and “Atom Tracer” (refer to 2.4.7.2) at 5% resulted in 5 reaction steps in *Escherichia coli* CFT073 (Figure 24) , *Escherichia coli* O157:H7 EDL933 (Figure 25) , *Escherichia coli* K-12 W3110 (Figure 26), *Escherichia coli* K-12 MG1655 (Figure 27), *Escherichia coli* O157:H7 Sakai (Figure 28).



Powered by yFiles

Figure 24. Shortest path between beta-D-glucose and Pyruvate in *Escherichia coli* CFT073 comprises of 5 steps with local similarity 15% and global similarity 10%.



Powered by yFiles

Figure 25. Shortest path between beta-D-glucose and Pyruvate in *Escherichia coli* H7 EDL933 comprises of 5 steps with local similarity 15% and global similarity 10%.

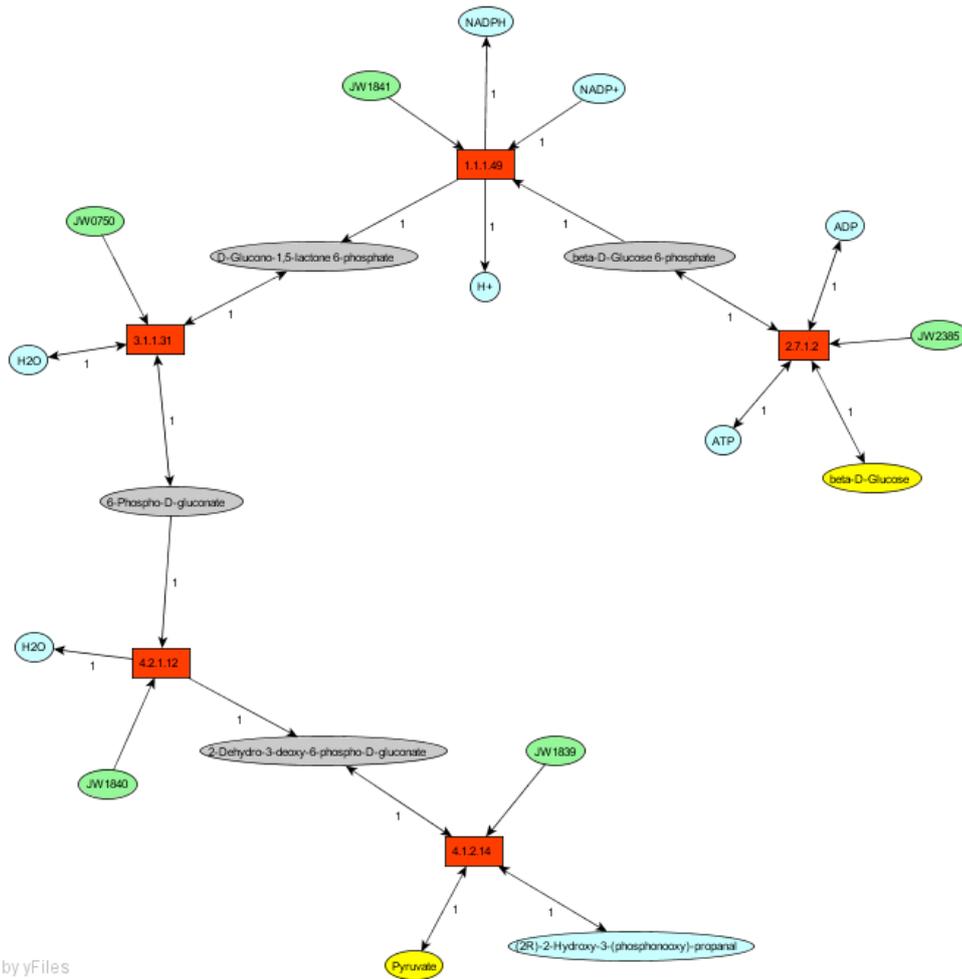


Figure 26. Shortest path between beta-D-glucose and Pyruvate in *Escherichia coli* K12 W3110 comprises of 5 steps with local similarity 15% and global similarity 10%.

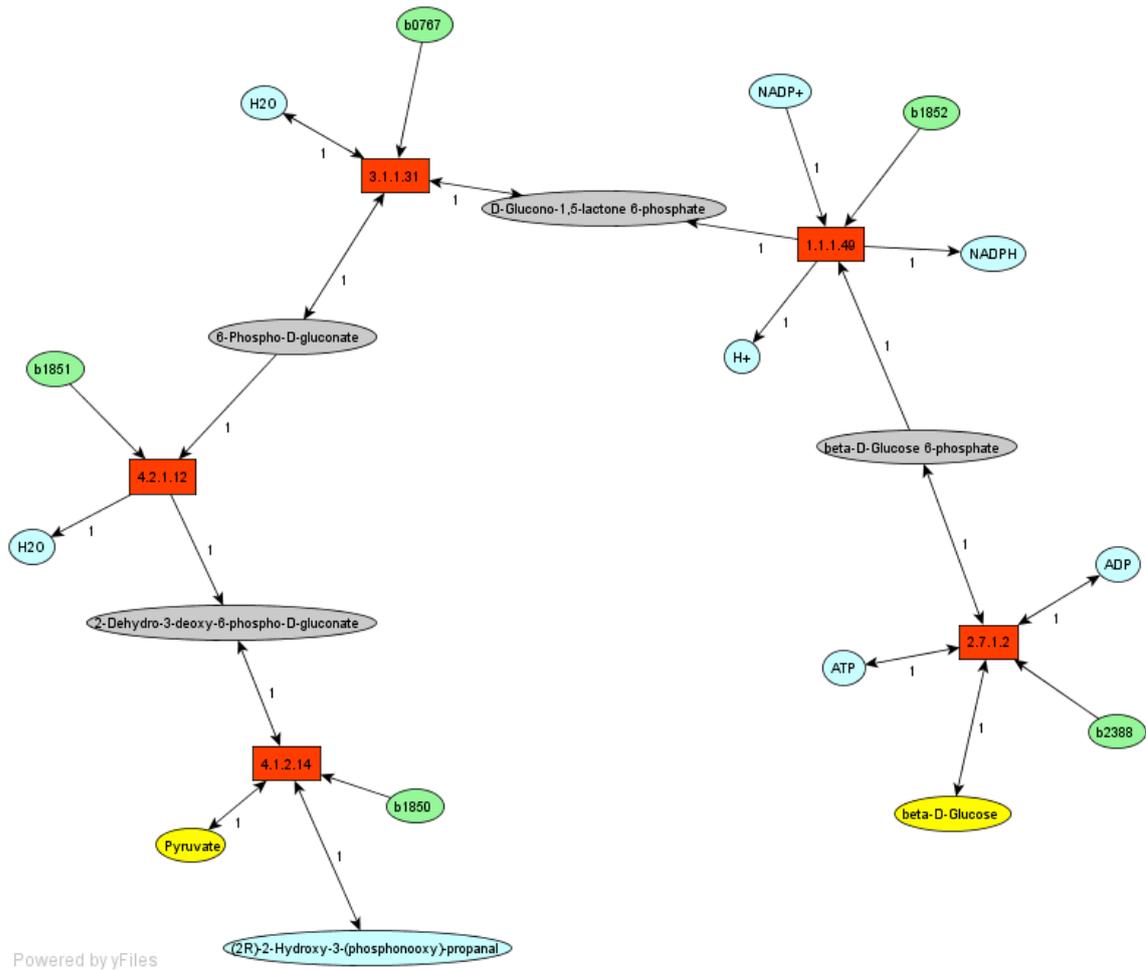


Figure 27. Shortest path between beta-D-glucose and Pyruvate in Escherichia coli K12 MG1655 comprises of 5 steps with local similarity 15% and global similarity 10%.

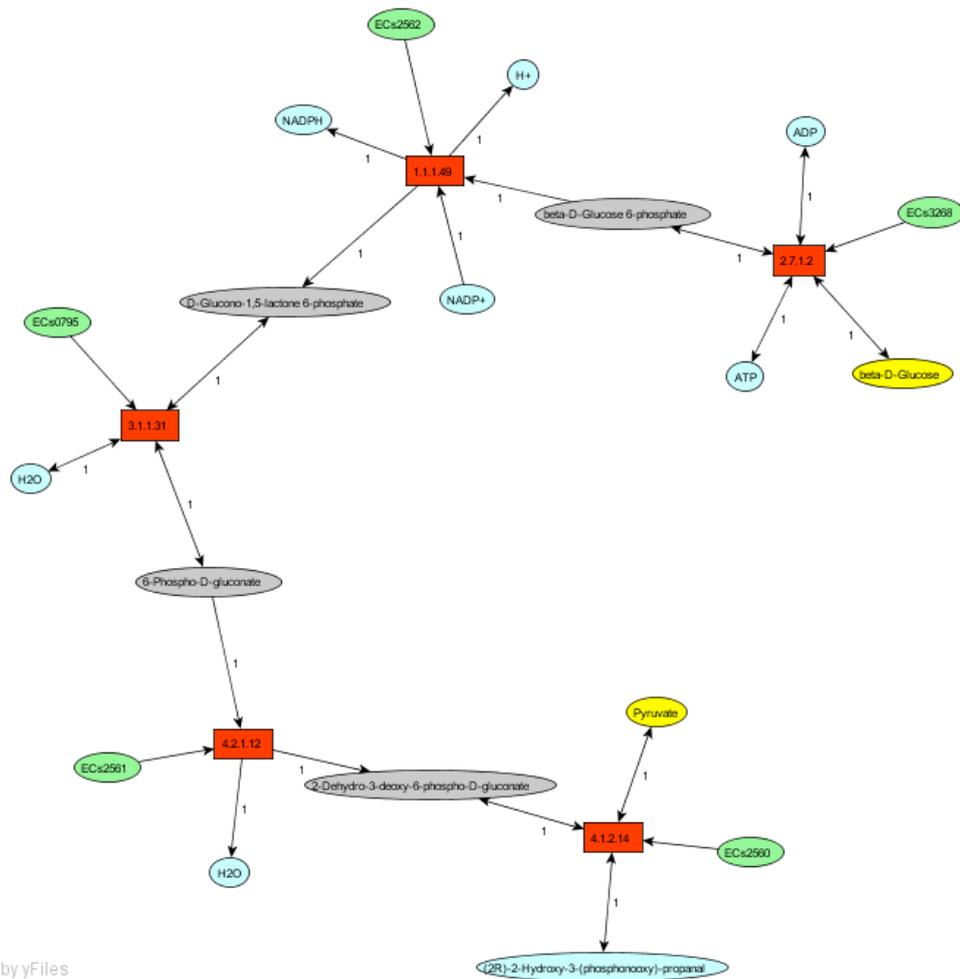


Figure 28. Shortest path between beta-D-glucose and Pyruvate in *Escherichia coli* H7 Sakai comprises of 5 steps with local similarity 15% and global similarity 10%.

2.5.4.1. Network properties

The comparative study (Table 5) of a pathogenic *Escherichia coli* O157:H7 strain and a non-pathogenic bacteria *Corynebacterium glutamicum* is presented in (Table 6) and (Table 7). The results are solely based on the network connectivity.

Table 5. *Escherichia coli* O157:H7, *Corynebacterium glutamicum* and their network features like Degree Distribution (DD) = $2*N/L$, Average Path Length (APL), Average k-Path Length (AKPL)

| Organism | Genes | Enzymes | Reactions | Metabolites | DD | APL | AKPL |
|--|-------|---------|-----------|-------------|------|------|------|
| <i>Escherichia coli</i> O157:H7 Sakai | 5341 | 586 | 1050 | 1044 | 3.52 | 8.06 | 8.52 |
| <i>Corynebacterium glutamicum</i> ATCC 13032 | 3057 | 435 | 899 | 1069 | 2.8 | 8.22 | 8.88 |

Table 6. Top 10 metabolite hubs in *Escherichia coli* O157:H7 Sakai based on connectivity and ranked by incoming degree.

| Metabolite | Incoming degree | Outgoing degree |
|--|-----------------|-----------------|
| Pyruvate | 26 | 24 |
| CoA | 18 | 14 |
| Tetrahydrofolate | 13 | 8 |
| ATP | 12 | 19 |
| 5-Phospho-alpha-D-ribose 1-diphosphate | 11 | 10 |
| L-Glutamate | 11 | 15 |
| Succinate | 11 | 6 |
| Acetate | 10 | 7 |
| beta-D-Fructose 6-phosphate | 10 | 9 |
| D-Galactose | 10 | 5 |

Table 7. Top 10 metabolite hubs based on metabolite connectivity in *Corynebacterium glutamicum* (Ranked by incoming degree).

| Metabolite | Incoming degree | Outgoing degree |
|--|-----------------|-----------------|
| Pyruvate | 14 | 15 |
| CoA | 14 | 14 |
| Tetrahydrofolate | 13 | 8 |
| Acetate | 10 | 6 |
| L-Glutamate | 10 | 13 |
| 5-Phospho-alpha-D-ribose 1-diphosphate | 9 | 10 |
| Succinate | 9 | 6 |
| Acetyl-CoA | 8 | 5 |
| L-Homocysteine | 8 | 4 |
| D-Ribose 5-phosphate | 7 | 6 |
| Glycine | 7 | 6 |
| UMP | 7 | 7 |
| ATP | 7 | 11 |

2.5.4.2. Top 10 Hubs based on k- Shortest Paths

The top 10 metabolite hubs based on the shortest path analysis in a pathogenic *Escherichia coli* O157:H7 strain (Table 8) and a non-pathogenic bacteria *Corynebacterium glutamicum*) (**Error! Reference source not found.**) are presented.

Table 8. Shortest path (SP) based top 10 incoming metabolite hubs in Escherichia coli O157:H7.

| Metabolite | SP (IN) | SP (IN) Normalized | SP (OUT) | SP (OUT) Normalized |
|---|---------|--------------------|----------|---------------------|
| Pyruvate | 74061 | 6.08 | 74273 | 6.08 |
| 5-Phospho-alpha-D-ribose 1-diphosphate | 72466 | 6.06 | 72185 | 6.05 |
| (2R)-2-Hydroxy-3-(phosphonoxy)-propanal | 49825 | 5.68 | 50040 | 5.68 |
| D-Ribose 5-phosphate | 43499 | 5.54 | 43509 | 5.54 |
| beta-D-Fructose 6-phosphate | 43283 | 5.54 | 43377 | 5.54 |
| D-Xylulose 5-phosphate | 40630 | 5.48 | 40781 | 5.48 |
| AMP | 38323 | 5.42 | 38049 | 5.41 |
| UMP | 37962 | 5.41 | 37709 | 5.4 |
| Glycerone phosphate | 36096 | 5.36 | 36229 | 5.36 |
| D-Ribulose 5-phosphate | 35610 | 5.34 | 35665 | 5.35 |

The paths were normalized by using the following formula, $\log \left(\frac{\sigma_{sp(m)}}{\sum \sigma_{sp}} \right)$

2.5.4.3. Shortest Path Distribution

The Shortest path distribution in a pathogenic *Escherichia coli* O157:H7 strain and a non-pathogenic bacteria *Corynebacterium glutamicum* are presented in (Table 9). This global similarity was chosen as 5% and Local similarity was chosen as 15%. The percentage of shortest path in each organism is presented in (Table 10).

Table 9. The shortest path distribution between *Escherichia coli* O157:H7 strain and a non-pathogenic bacteria *Corynebacterium glutamicum*, calculated by Pathway Hunter Tool using CUBIC mapping algorithm (Local similarity 15%, Global similarity 5%).

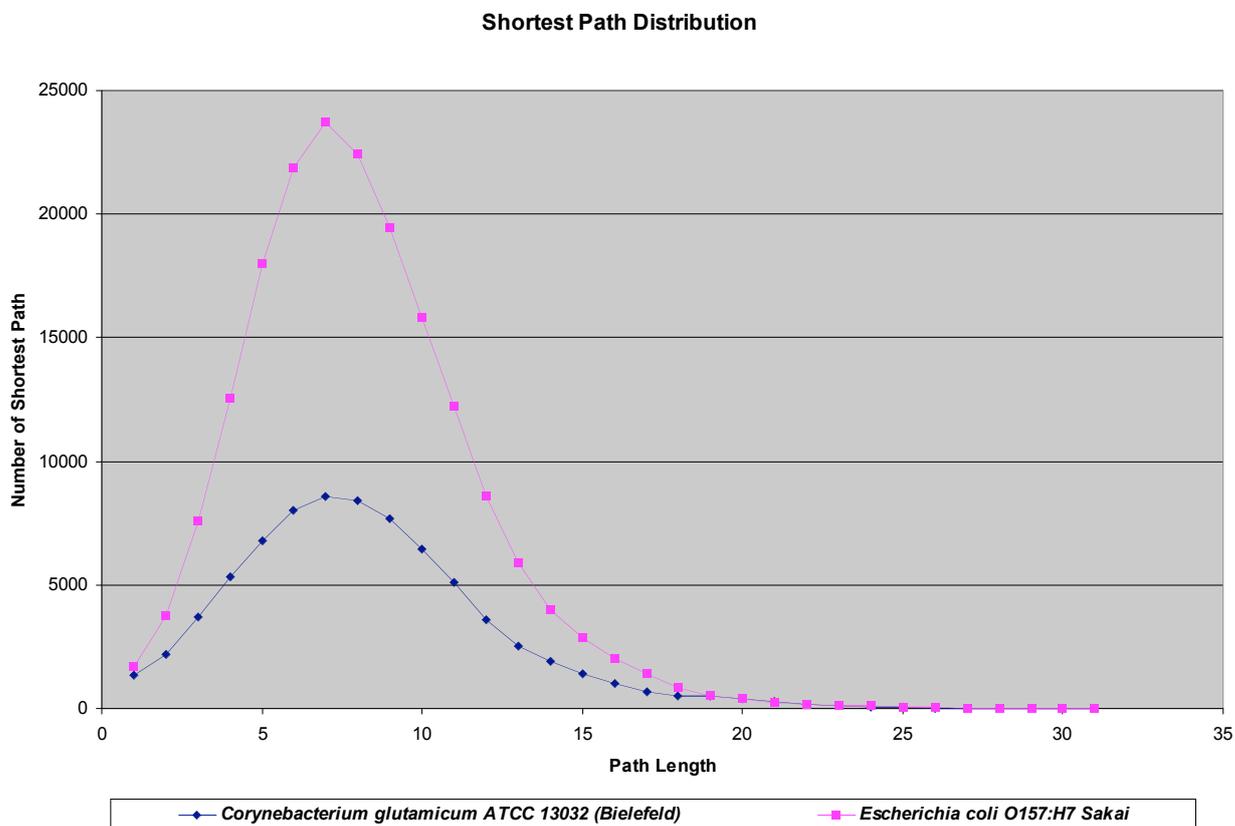
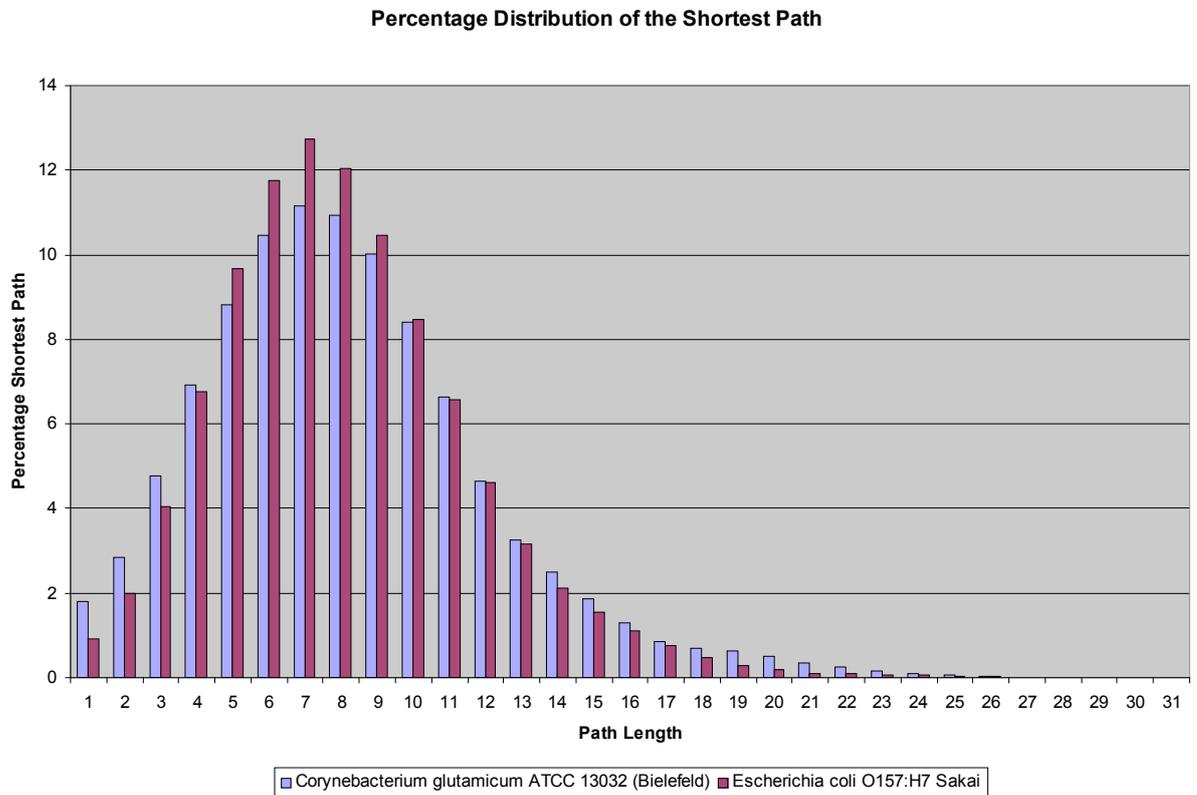


Table 10. Distribution of the shortest path percentage in the pathogenic *Escherichia coli* O157:H7 strain and a non-pathogenic bacteria *Corynebacterium glutamicum*



3 Biochemical Pathway Alignment

3.1 Introduction

A metabolic network can be referred to as a set of interconnected enzymes and small molecules giving rise to anabolic, catabolic and amphibolic pathways. The underlying cellular topology may vary from organism to organism and this may give rise to different functionalities of a pathway (Ravasz, Somera et al. 2002). It is interesting to understand how the connectivity between the networks affects the conversion of metabolites (Hartwell, Hopfield et al. 1999; Csete and Doyle 2002). Some connectivity in the network may be conserved across many genomes (Kelley, Sharan et al. 2003) and vice versa. Thus it is essential to understand the underlying functions of these cellular networks and the cross-talk between them (Chen and Vitkup 2006; Rahman 2006).

Metabolic enzymes represent one of the most important classes of proteins and extensive studies about their sequence, structure and function have been carried out in the past (Rison, Teichmann et al. 2002). Many enzyme families usually catalyze a range of biochemical reactions (Jensen and Gu 1996; Dandekar, Schuster et al. 1999), whereas some of these reactions may also be catalyzed by members of apparently unrelated protein families (Copley and Bork 2000). This fortifies the hypothesis that metabolic enzyme families exhibit complex patterns of divergent and convergent evolution (Hartwell, Hopfield et al. 1999; Albert, Jeong et al. 2000; Chen and Vitkup 2006). The functional assignment by homology (Iliopoulos, Tsoka et al. 2001; Chen and Vitkup 2006) to proteins of known function may overlook issues of evolutionary divergence (Castresana 2001; Chen and Vitkup 2006), which may lead to error propagation in sequence databases (Karp 1998). Understanding the enzyme sequence and function(s), based on the metabolic pathway provides us with a better and deeper understanding of the system (Rison, Teichmann et al. 2002) (Teichmann, Rison et al. 2001; Tsoka and Ouzounis 2001).

In recent years there have been major efforts to understand metabolic pathways based on the concept of pathway alignment (Dandekar, Schuster et al. 1999; Tohsato, Matsuda et al. 2000; Matsuda and Tohsato 2001; Kelley, Sharan et al. 2003; Brandes, Dwyer et al. 2004; Chen and Hofstadt 2004). In one such study the power of a

comparative study of pathways captured the plasticity of the glycolytic pathway (Dandekar, Schuster et al. 1999). These researches highlight the need for robust algorithms to perform pathway reconstruction and pathway alignment for facilitating the discovery of pharmacological targets and complementing biotechnological applications.

3.2 Background

With the stock pile of already sequenced genomes ever increasing, computational reconstruction of metabolic networks becomes a crucial step for metabolic pathway analysis. Robust methods and tools are required to detect and bring out conserved and diverged metabolic pathway connectivity (enzymes/metabolites) in various genomes. One such method is metabolic pathway alignment, which can help us understand the cross-talk between pathways arising from the underlying connectivity at the biochemical level. Alignment tools would find application in various areas, such as genome annotation and would prove valuable to a wide range of users including pharmacologists, metabolic engineers etc.

3.3 Proposed Model

We present a novel method to perform pathway alignment based on the shortest path as implemented in Pathway Hunter Tool (PHT) (Rahman, Advani et al. 2005). The output of the pathway alignment will highlight the convergence/divergence of small molecule(s) coding enzyme-enzyme connectivity in the pathway across genomes. In the first section of this chapter enzymes usage matrix obtained from the *ab-initio* reconstructed network based on the shortest path analysis is presented. In the second section the alignment procedure together with insertions for handling missing links in the network(s) is introduced. The visualization aspect of the alignment, which presents the connectivity information from the obtained matrix, highlighting the intricacies of the pathways in the selected organism is discussed. Thus we combine the top down and bottom up approach to perform the metabolic pathway alignment on the reconstructed metabolic network of the selected genomes from KEGG (Kanehisa, Goto et al. 2004) are combined. The standard Gibbs energy (Mavrovouniotis 1991) change of reactions has been used as an indicator to score/ rank the biochemical flexibility of the pathways. The Gibbs energy-based thermodynamics curve (with an option to add metabolite concentration data for the calculation) of selected ((Mavrovouniotis 1993) and references therein), aligned pathways can then be viewed and compared.

3.4 Method

Exploring the pathway capabilities in different organisms based on connectivity information helps us understand the evolutionary relationship (Forst and Schulten 1999; Forst and Schulten 2001; Teichmann, Rison et al. 2001) as well as alternate pathways (Dandekar, Schuster et al. 1999; Heymans and Singh 2003) available in various species (i.e. Prokaryote, Eukaryote). Using graph theory one can define the metabolic network as a directed graph where $G = (V, E)$ $V \in \text{metabolites}$ and $E \in \text{enzymes}$.

The general procedure used in PAT to perform pathway alignment can be described as:

3.4.3 Data Collection

Collecting the connectivity information from the metabolic network of the genomes by, a) Selecting organisms from the evolutionary tree based on KEGG (Kanehisa, Goto et al. 2004) ontology; and b) Building a consensus network and percentage enzyme occurrence matrix (refer to 3.4.6) from the selected organisms (Figure 29).

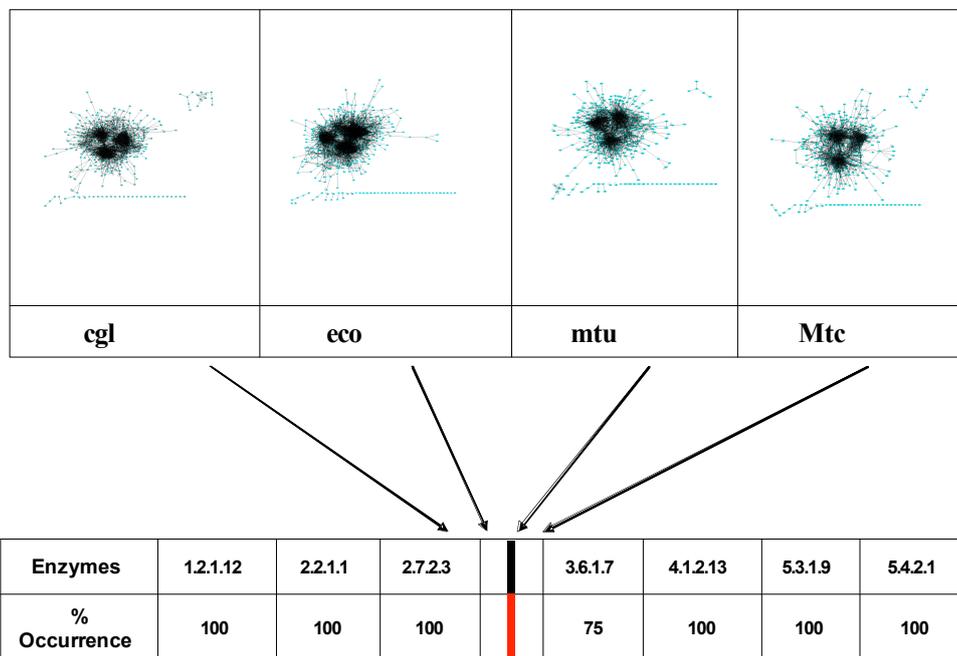


Figure 29. Enzyme-Percentage occurrence matrix resulting from the conversion of beta-d-glucose to pyruvate in the reconstructed network of cgl, eco, mtu, mtc.

3.4.4 Path Finding

Querying a pathway between two metabolites with or without insertion(s) of enzymes in the queried pathway by, a) Finding the shortest path between source and destination metabolites in selected organisms; and b) If possible, finding and extending the shorter shortest path in a genome(s) to match the longer shortest path of other genome(s).

3.4.5 Viewing Aligned Path

The user views the alignment output and saves it as an image (PNG, JPEG, GIF format). Enzymes are coloured (Figure 30) according to their percentage occurrence in the selected reference matrix. A consensus of the alignment is represented by bar charts and they are coloured according to the enzyme preference along the aligned pathway.

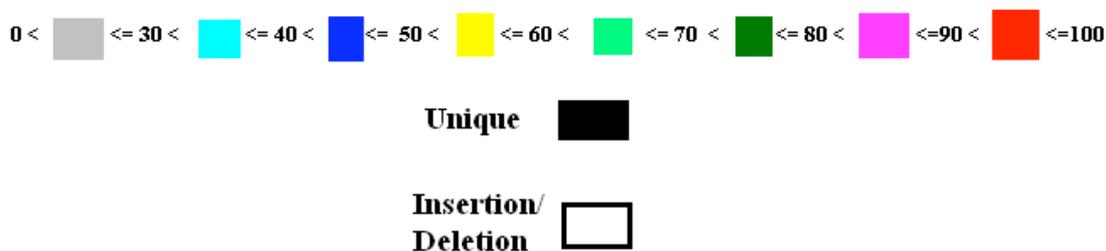


Figure 30. Color code for representing the percentage occurrence of enzymes in the organisms. Insertion(s)/Deletion(s) of enzymes in the alignment is represented by white colors.

3.4.6 Enzyme-Percentage Occurrence Matrix and Enzyme-Enzyme Connectivity Matrix

The enzyme-percent matrices (refer to 3.4.3) were constructed for three classes of organisms - *Bacteria*, *Eukaryote*, *Archae*; representing the percentage occurrence (3.1) of each enzyme $M_{ij} (E)$ occurring in selected organism(s) (“i”th enzyme is “j”th organism).

$$\text{Let } \sigma_j = \sum_{j=1}^m O_j \{ \forall j \in \text{Organisms} \} \quad (3.1)$$

$$M_{ij} (E) = \left(\sum_{j=1}^n E_{i,j} / \sigma_j \right) \times 100 \{ \forall i \in \text{Enzymes} \mid 1 < i < m \} \quad (3.2)$$

The resulting matrix named as *Enzyme-percentage* matrix (3.2) highlights the percentage occurrence of an enzyme across the selected genome. For example the *enzyme-percentage occurrence matrix* for the conversion of beta-D-glucose to pyruvate will appear as shown in (Figure 29). The enzyme 3.6.1.7 is absent in *Mycobacterium tuberculosis* H37Rv (*mtu*) hence its percentage occurrence is 75%.

Enzyme-enzyme connectivity matrix is a sparse matrix, which contains information about the neighbours of each enzyme. An enzyme is connected to another enzyme if and only if it shares a common metabolite between them in a pathway. Finding the shortest path in the network under a certain criterion of local and global similarity score between all the metabolites in the network is used to calculate this matrix.

3.4.7 Pathway Alignment Based on Shortest Path using the Matrices

The basic alignment procedure revolves around the shortest path between a source and destination metabolite. The general alignment procedure can be defined by the following flow chart (Figure 31).

3.4.8 Organizing Consensus Header and Assigning Color Codes

Generating a consensus header from the alignment provides a global view of the alignment. The preference of the organism for certain enzymes at each step of the conversion can be viewed instantly (Figure 31). However, the length of these paths may vary from one organism to another. This may prove to be a hitch for the alignment. To overcome this problem a new method, “*building reference headers*” is implemented. In the “*Building reference headers*” method separates reference headers are built with different path lengths and the paths with identical path lengths are aligned to their corresponding reference header. At the local level, such results indicate whether or not an enzyme is unique to an organism. This is done by examining its percentage occurrence in various organisms which is presented in the *enzyme percentage matrix* where each enzyme is color coded (Figure 30) based on its percentage occurrence.

The resulting output increases the potential for biochemical/metabolic engineering application and has the potential to facilitate the ongoing systems biology research outlook for deciphering the genomes.

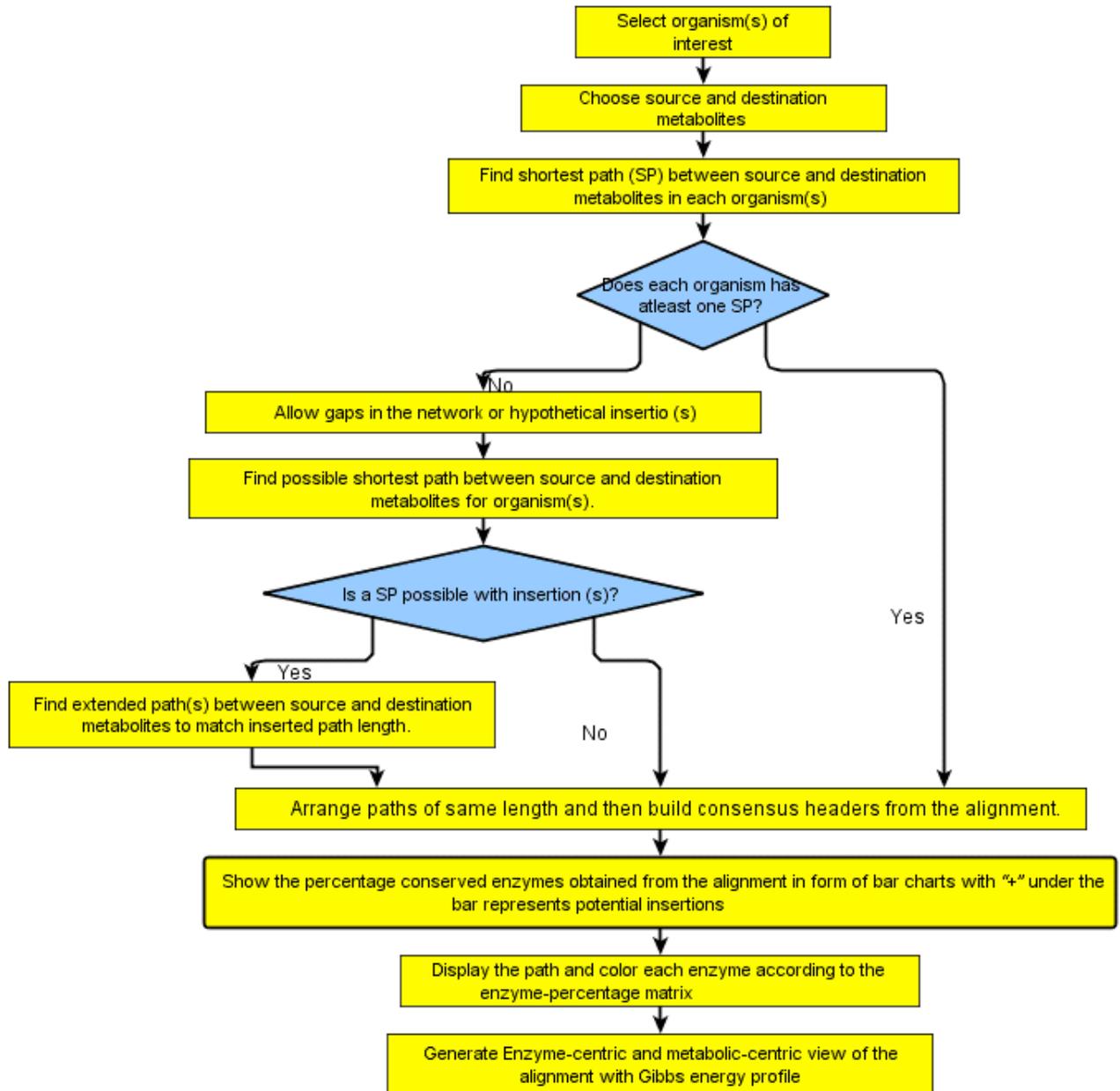


Figure 31. A flowchart for the proposed pathway alignment model.

3.4.9 Finding Alternate Paths based on Metabolites and Enzymes

There can be more than one path for converting a substrate to the final product after a series of reaction steps. These paths were classified on the basis of the metabolites/enzymes (iso-enzymes) involved at each step of conversion.

Two levels of coding - “x.y” were used to classify the pathways. If an alignment between two metabolites results in three paths and if the first two paths evolve with the **same set of connecting metabolites** (but with alternate enzymes doing this conversion). Then this path is shown as “1.1” and “1.2” is represented by “y”. However in the third path, one of the **connecting metabolites** used for conversion has **changed** but the source and end product metabolites of the path remains unaltered. This path is therefore represented as “2.1” which is represented by “x”.

Hence, this implies that the change in convention at the **first level represents** a change in at **least one of the connecting metabolites** during the conversion. If the **second level changes** then there is at **least one change in the connecting enzyme(s)** involved during the conversion. Thus the information about the alternate pathway is coded in the alignment procedure.

3.4.10 Handling gaps in the pathway alignment

It is well-acknowledged that a path between two metabolites might not be discovered due to the missing intermediate conversion step(s) (Osterman and Overbeek 2003; Green and Karp 2004; Brouns, Walther et al. 2006). This may happen due to lack of information about the pathway (i.e. lack of functional data). It would thus be of interest to insert or delete an enzyme in the system and observe the change(s) in the pathway alignment (Ettema, Makarova et al. 2004; Chen and Vitkup 2006). In terms of graph theory, we are interested in finding the shortest path with one or more inserted nodes (missing enzymes). This may change the entire search space and hence the topology of the network may also be affected (especially if the inserted enzyme is highly connected). Hence, this may result in a combination of paths (due to alternate paths) of different lengths thereby increasing the runtime of the alignment. The gap

handling subroutine holds true if and only if no path exists between two metabolites in an organism (Figure 31). In order to avoid combinatory explosion, a maximum of five gaps (hypothetical enzyme(s) to be inserted in the path) is allowed in the alignment.

For the purpose of better visualisation, insertion(s) in our system are reported in “white” color and a “+” sign is inserted below the consensus header. If an enzyme is unique to an organism it is represented in the colour “black”. Handling insertion/deletions in the pathway may help us improve the annotation in the organism. It would also cater to the requirements of pharmacological and biotechnological research (Schilling and Palsson 2000; Papin, Price et al. 2003).

3.4.11 Algorithm Complexity for Handling Alignment

Metabolic pathway alignment for various species may result in paths of different lengths between two metabolites under certain criteria of local and global similarity (refer to section 2.4.7) (Rahman, Advani et al. 2005). Hence it is important to build a consensus header from the alignment by encapsulating path(s) of equal lengths. In order to capture the local preference(s) of the enzymes in the selected organism(s) at various steps of the pathway, a consensus header is generated from the pathway(s) of different path lengths. This increases the quality of information representation without making the alignment visualization too complex.

At the metabolic-centric level, paths of similar lengths were encapsulated and ranked according to the summation of standard Gibbs energy of reaction(s) (refer to section 3.4.12) involved in each pathway. Standard Gibbs energy is an indicator of the tentative biochemical energy (thermodynamics) requirement for a pathway (*i.e.* ΔG of pathway may vary), thereby highlighting the biochemical nature of various pathways (Mavrovouniotis 1993).

This information when combined with genomic data allows us to view the system in an extremely comprehensive manner (a comparative study of pathways between various metabolic networks of the genome(s) will highlight the preference for certain class/type of enzymes at various steps) (Kummel, Panke et al. 2006; von Stockar, Maskow et al. 2006).

Computationally, this means that the Shortest Path σ_{sp} between two metabolites in an organism of interest may result in k steps ($\sigma_{sp}(O) := k \mid k \in O_i$) and in another in “ $k + t$ ” steps ($\sigma_{sp}(O_m) = k+t \mid k+t \in O_{i+m}$). This implies that the path length $k < k+t$ ($\therefore \sigma_{sp}(o) < \sigma'_{sp}$) in the alignment. Hence we will get two consensus headers for each path length k and $k + t$ respectively, whereas the summation of consensus headers will result in length $k + t$.

An alignment between paths of path length of k and $k + t$ in organisms of interest, will require the extension of path length k to the next feasible path of length ' $k+t$ ' in the network. Finding next possible shortest path in a metabolic network is a NP-complete problem due to combinatory explosion arising from the topological search (*i.e.* its difficult to know that a next possible shortest path exists in the network). To find the next possible shortest path in the system, PHT employs heuristics and hence may miss some paths in the alignment while looking for the next possible path.

To overcome this situation, we backtrack (based on consensus header) after each search to find a possible path (if a path was missed, or no shortest path was reported) from the selected *enzyme-enzyme connectivity matrix*¹. Though this may increase the running time of alignment in such cases, it greatly enhances the quality of the reported alignment - a trade-off that will pay high dividends to the potential users.

¹ A sparse matrix containing information about enzyme-enzyme connectivity in the reference network

3.4.12 Gibbs Energy Profile of the Aligned Pathways

Once pathways are aligned, they can also be compared with respect to Gibbs energy changes (refer to section 2.2) of each reaction step. The standard Gibbs energy of formation for each metabolite is computed by a group contribution method (Mavrovouniotis 1991). The basic method was further extended and adapted to cover metabolites containing vinyllog acids and has been implemented in Java. Molecular structures are needed for this procedure and were extracted from the KEGG Ligand database (Kanehisa, Goto et al. 2006). They have been converted to their standard state at pH 7 using the pKa plugin of the ChemAxon software package JChem*. The Gibbs energy prediction method has been applied to the modified structures to give the standard Gibbs energy of formation. By weighted summation of these values for each reaction, the standard Gibbs energy of reaction is obtained[†]. The actual Gibbs energy of reaction may be calculated when concentration values of participating metabolites are at hand (Mavrovouniotis 1993).

* JChem, version 3.1.6, ChemAxon, Budapest, Hungary. www.chemaxon.com/products.html

[†] Please refer Kai Hartmann's PhD Work at CUBIC, Cologne, Germany

3.5 RESULTS

In order to prove the potential of our new algorithm and method to perform metabolic alignment an alignment with the following bacterial genomes was performed.

3.5.3 Comparative Study of Metabolic Network Topology between a Pathogenic and a Non-Pathogenic Bacterium

For conducting a comparative study between *Bacillus subtilis* 168 and *Bacillus anthracis* Sterne metabolic network, we calculated shortest path distribution (Figure 32), the average path length (Table 11) and average alternate paths (Figure 33). It is important to keep track of alternate paths in the metabolic network because this indicates the ability of the organism to survive in adverse conditions.

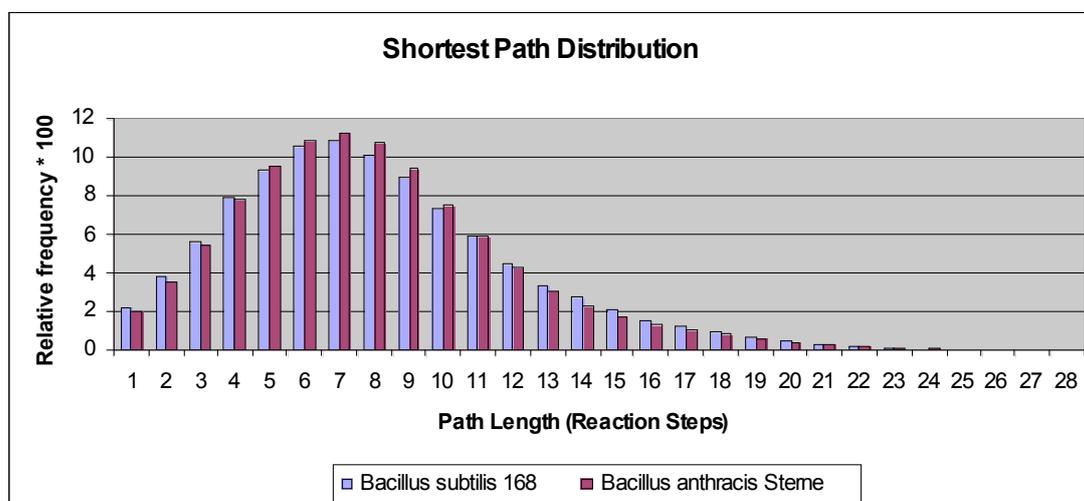


Figure 32. Shortest path distribution in *B. subtilis* 168 and *B. anthracis* Sterne

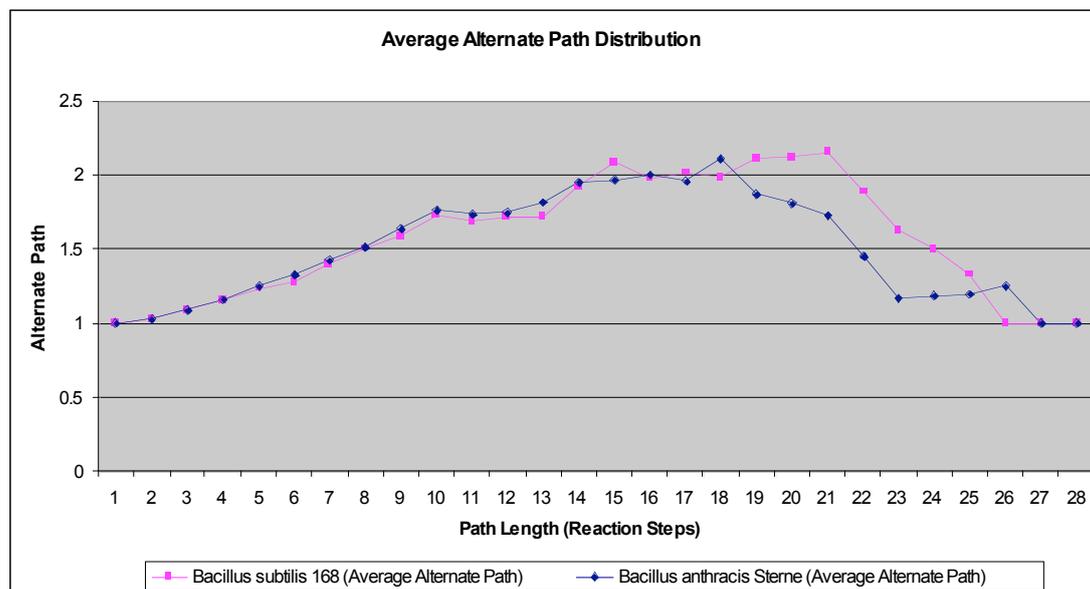


Figure 33. Average alternate path distribution in *B. subtilis* 168 and *B. anthracis* Sterne. The alternate shortest path in the network can be calculated by dividing the total number of shortest paths at each path length (reaction step) by the unique number of shortest path respectively

Table 11. Metabolic network analysis of *B. subtilis* 168 and *B. anthracis* Sterne using shortest path

| Genome | Enzymes | Reactions | Metabolites | Average Path Length (sp/k-sp) | Diameter of the Network | Average Degree |
|----------------------------|---------|-----------|-------------|-------------------------------|-------------------------|----------------|
| <i>B. subtilis</i> 168 | 468 | 974 | 1081 | 7.93/8.59 | 28 | 3.06 |
| <i>B. anthracis</i> Sterne | 455 | 997 | 1127 | 7.98/8.73 | 28 | 3.01 |

Thus, blocking a path may not be lethal as organisms can switch to an alternate path performing similar conversions. Alternate paths can be bio-energetically costlier or longer than the native pathway. Hence though organisms may slow down their metabolic activity, they can still survive.

3.5.4 Performing Pathway Alignment

For demonstrating the pathway alignment algorithm four bacterial pathogens i.e. *Escherichia coli* O157:H7, *Mycobacterium tuberculosis* CDC1551, *Bacillus anthracis* Sterne, *Pseudomonas aeruginosa* and a non-pathogenic bacteria *Bacillus subtilis* was chosen. The local similarity was set to 15% and global similarity was set to 5%. All these alignments were performed using *Enzyme-Enzyme Connectivity Matrix* and *Enzyme-Percentage Occurrence Matrix* built using over 150 bacteria in the PHT database inherited from KEGG.

3.5.4.1 Pathway Plasticity between Bacteria

In order to observe flexibility in the pathways of selected genomes, a comparative analysis based on the pathway alignment was performed. The shortest path was chosen that converts **beta-d-glucose 6-phospate** and **citrate** in selected genomes without gap (Figure 34).

Previous research has demonstrated that pathway plasticity even in the central pathway i.e. glycolysis and pentose pathway is not only often present but seems to be selected by evolution (Forst and Schulten 1999; Schmidt, Sunyaev et al. 2003). This section explains the results obtained from the alignment (Figure 35) and discusses its capabilities in the light of systems biology.

terms of bar charts found above the consensus header of the alignment. For example enzymes 3.1.1.31 (step 2) and 4.1.2.14 (step 4) are found to occur less than 30% time in this alignment.

At each step of the alignment the headers display different preferences (looking at the percentage distribution) of the enzymes at various steps. This implies the existence of other enzymes in the path, which can perform similar tasks. The presence of an iso-enzyme is shown in step 2 and step 3 of (*Bacillus anthracis*) of the alignment where enzyme “2.2.1.1” (section 1.1) is substituted by enzymes “2.7.1.11” and “4.1.2.13” (section 1.2). Hence they perform the same conversion but in the presence of different cofactors thereby highlighting the need to example the metabolites involved in the conversion step. Hence the metabolic centric view (Figure 37) and the Gibbs energy profile can highlight further dynamics of the path and the moieties involved during the conversion process. Using the metabolic profile of the pathway alignment (example between step 3 and step 4) it is clear that there are different metabolites involved in intermediate conversion steps. Such findings will further boost the relevance of comparative network analysis across the genomes and highlights the crosstalk between the pathways.

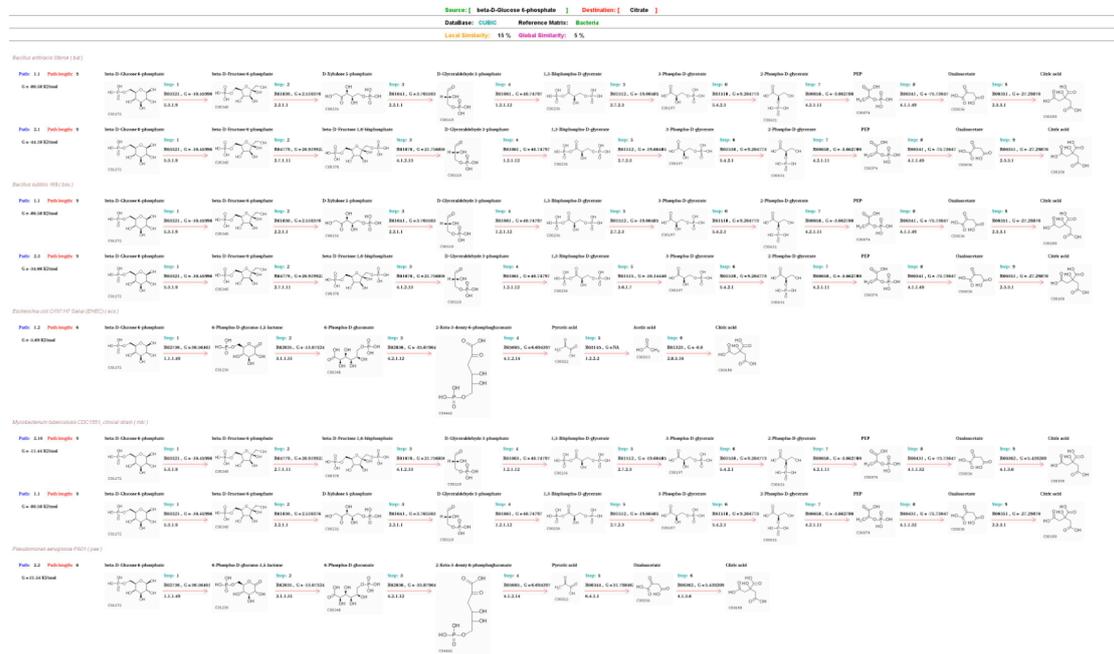


Figure 37. Metabolic centric view of the alignment for the selected paths with deltaG score. Change is deltaG represents change in on of chemical compound constituting the reaction.

This further highlight the fact that selected genomes has a preference for certain enzymes over others (based on functionality or evolutionary past) in select pathways to carry out certain conversions. Hence this method further provides a global and local view to the occurrence of an enzyme in species of interest and its possible relevance in the pathway.

3.5.5 Gibbs Energy Profile of the alignment

The true meaning of the biochemical pathway can be measured quantitatively using Gibbs free energy (refer to 2.2). We have used Gibbs free energy to measure the biochemical cost (energy) of various pathways. Thus combining Gibbs free energy with biochemical pathway alignment is a very natural phenomenon.

3.5.6 Pathway Alignment with Gaps (insertion)

The shortest path in *Helicobacter pylori* for converting “3-phospho-D-Glycerate” to “pyruvate” has 3 reaction steps if an insertion is allowed while performing alignment. The result with an insertion in the alignment of *E. coli* k-12 (Figure 38) shows that a shorter path of length 3 already exists in *E. coli* genome. The inserted enzyme is “2.7.1.40” shown in white color. In the header section the insertion is shown with a “+” sign.

| | |
|--|-----------------------------------|
| Source: [3-Phospho-D-glycerate] | Destination: [Pyruvate] |
| DataBase: CUBIC | Reference Matrix: Bacteria |
| Local Similarity: 15 | Global Similarity: 5 |



Escherichia coli K-12 W3110 (ecj)

Path no: 1.1 **Path length:** 3 5.4.2.1 4.2.1.11 2.7.1.40

Helicobacter pylori J99 (hpi)

Path no: 1.1 **Path length:** 3 5.4.2.1 4.2.1.11 2.7.1.40

Path no: 1.2 **Path length:** 3 5.4.2.4 4.2.1.11 2.7.1.40



Figure 38. The aligned paths between metabolites 3-phospho-D-glycerate and pyruvate in *H. pylori* and *E. coli* K-12. The gaps in the alignment as represented by "+" in the header section.

The metabolic centric view (Figure 39) of the aligned path further highlights the conserved nature of the connecting metabolites between *E.coli* and *H.pylori*.

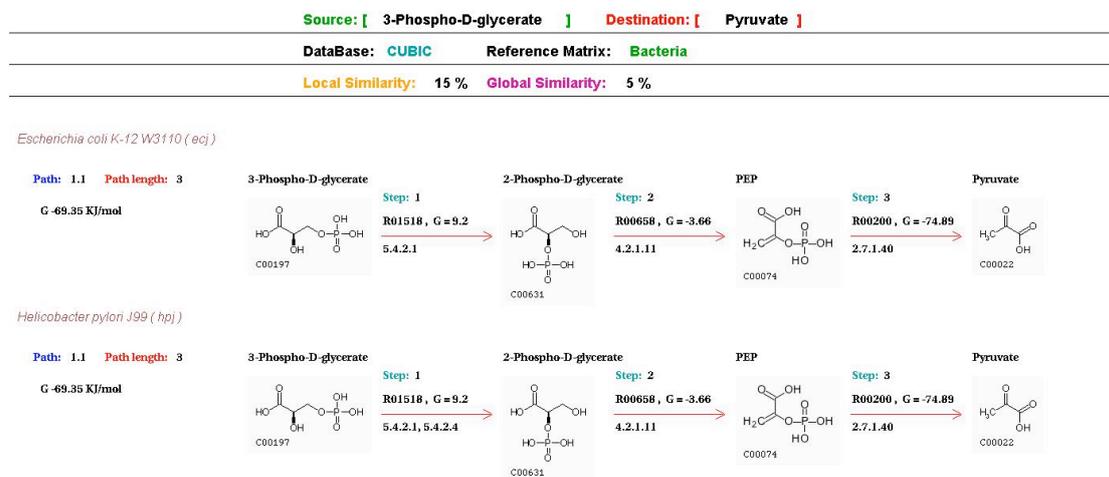


Figure 39. The pathway alignment between 3-phospho-D-glycerate and pyruvate in *H. pylori* and *E. coli* K-12. The metabolic centric view of the pathway is shown here.

The biochemical significance of the insertion can be studied by looking at the Gibbs profile of the aligned pathway. We performed the standard Gibbs energy profile of the aligned path (Figure 40) and found that bio-chemically this insertion is feasible (best results can be calculated by finding the gene coding of this enzyme and metabolic concentration data).

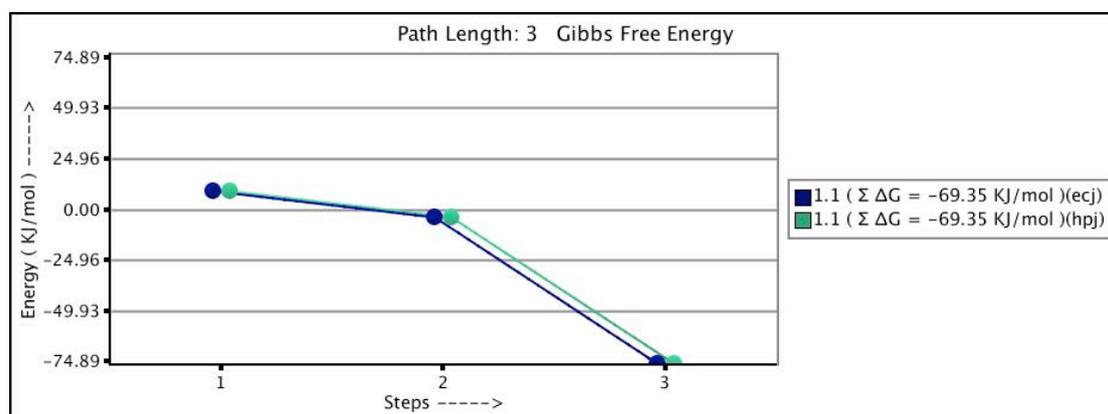


Figure 40. The Gibbs energy profile the alignment pathway between 3-phospho-D-glycerate and pyruvate in *H. pylori* and *E. coli* K-12.

4 Load Point and Choke Point

4.1 Introduction

In the present “omics” era, it becomes increasingly more obvious that network analysis is essential for the analysis of genetic, proteomics and metabolomics data (Hartwell, Hopfield et al. 1999; Grigorov 2005). Large-scale, graph-based mathematical models have been developed to demonstrate the intrinsic hierarchical modularity of metabolic networks (Ravasz, Somera et al. 2002) and their robustness based on the shortest path analysis of the metabolic networks (Arita 2004; Barabasi and Oltvai 2004; Papin, Reed et al. 2004).

A typical metabolic network consists of reactions, metabolites and enzymes, which can be modelled using graph theory (Jeong, Tombor et al. 2000; Schuster, Fell et al. 2000; Girvan and Newman 2002; Oltvai and Barabasi 2002; Steinbeck, Han et al. 2003). These representations lead from a simple graph consisting of edges (reactions) and nodes (metabolites) or vice versa to a complex bipartite graph where two nodes (metabolites) share a common node (reaction/enzymes) (Rahman, Advani et al. 2005). Joining enzymes that share a common metabolite in a path can create enzyme-centric networks. The enzyme-centric view (Horne, Hodgman et al. 2004; Rahman, Advani et al. 2005) simplifies the representation of the metabolic network by removing loose ends in the network (metabolites at the periphery of the network) and forming clusters of interacting enzymes. The gene-centric view has been successfully used in determining co-regulated genes in the metabolic and regulatory networks (Levchenko 2003; Covert, Knight et al. 2004; Luscombe, Babu et al. 2004; Ozbudak, Thattai et al. 2004; Barrett, Herring et al. 2005).

In the present work we extend our analysis to the identification of “load points”. The “load points” analysis of metabolites in a metabolic network depends on the ratio of the number of valid k -shortest path passing through the metabolites and its nearest neighbour connectivity. We believe “load points” can complement other existing methods of metabolic network analysis (Schilling, Schuster et al. 1999; Steinbeck, Han et al. 2003; Klamt and Gilles 2004; Croes, Couche et al. 2005). It provides a global view to the metabolic network activity and such information might help in the

analysis of metabolic concentration data obtained from high throughput methods like GC/MS (Fiehn, Kopka et al. 2000; Strelkov, von Elstermann et al. 2004). A global perspective reveals that certain pathways such as the citrate cycle are highly used in the cell. Most of the enzymes/metabolites in the citrate cycle of glycolysis have high “load” values. The load point(s) analysis might help interpret concentration data, or flux data obtained by flux balance analysis or metabolic control analysis. Hence the importance of a metabolite in a metabolic network can be represented and ranked

In chapter 2, an algorithm was developed to identify bio-chemically correct connectivity in the metabolic network (refer to section 2.4.4) by pruning the network based on metabolic structural similarity (Rahman, Advani et al. 2005). The concept of “Global” and “Local” similarity was used to find a valid connectivity in the network (refer to section 2.4.7). The effect of metabolic structural similarity on the reported path and connectivity is very significant as this determines the abstraction level of the sub-network (Hattori, Okuno et al. 2003).

4.2 Formulation of the Critique

In the present work we extend our analysis to the identification of “load points”. The “load points” analysis of metabolites in a metabolic network depends on the ratio of the number of valid k-shortest path passing through the metabolites and its nearest neighbour connectivity. We believe “load points” can complement other existing methods of metabolic network analysis (Schilling, Schuster et al. 1999; Steinbeck, Han et al. 2003; Klamt and Gilles 2004; Croes, Couche et al. 2005). It provides a global view to the metabolic network activity and such information might help in the analysis of metabolic concentration data obtained from high throughput methods like GC/MS (Fiehn, Kopka et al. 2000; Strelkov, von Elstermann et al. 2004). A global perspective reveals that certain pathways such as the citrate cycle are highly used in the cell. Most of the enzymes/metabolites in the citrate cycle of glycolysis have high “load” values. The load point(s) analysis might help interpret concentration data, or flux data obtained by flux balance analysis or metabolic control analysis. Hence the importance of a metabolite in a metabolic network can be represented and ranked by this method.

Choke points are critical points in metabolic networks. Inactivation of choke points may lead to an organism's failure to produce or consume particular metabolites which could cause serious problems for fitness or survival of the organism (Yeh, Hanekamp et al. 2004). We propose a new method to analyse choke points by screening the entire metabolic network of pathogens and report the probable choke points in the network. This extended graph theory model ranks the choke points according to the k-shortest path passing through it and the load (in/out) on it. This ranking has a major advantage as this measure may help determine the biochemical essentiality of a metabolite/enzyme (when a chokepoint enzyme is removed from the network). For example, in *P. falciparum* - a parasite causing malaria in humans, a host cell enzyme 4.2.1.24 (d-aminolevulinate dehydratase; ALAD) involved in heme biosynthesis was suggested as an antimalarial target (Bonday, Dhanasekaran et al. 2000). This enzyme is also a choke point enzyme and identifying such potential targets in the pathogens can accelerate the drug discovery. Also all three clinically validated drug targets for malaria are chokepoint enzymes. A total of 87.5% of proposed drug targets with biological evidence in the literature are chokepoint reactions (Yeh, Hanekamp et al. 2004).

For building the biochemical network we used the LIGAND (Goto, Okuno et al. 2002) database from KEGG (Kanehisa, Goto et al. 2004) as this data model is the backbone for the Pathway Hunter Tool (Rahman, Advani et al. 2005) in addition to BRENDA (Schomburg, Chang et al. 2004). For the predicted choke points in the pathogen we performed a homology search against the human genome using BLAST (Altschul, Madden et al. 1997).

Here we provide a generic framework and model for an automated analysis of metabolic networks by ranking the metabolites on the basis of their load point property. Load points help determine the importance of enzymes and metabolites in the biochemical network. The concept of choke points was used in our study to find potential drug targets in the metabolic network of *Bacillus anthracis* Sterne. The metabolites and enzymes are further ranked on the basis of their loads in the given network. A comparative study was performed between the human metabolic network and pathogen choke points to discriminate human choke points from the pathogenic

bacterial choke points. A homology search was performed against human genome to find non-homologues potential drug targets from the pathogen choke points.

4.3 Method

4.3.1 Data and System

In order to demonstrate the efficiency of our algorithm we chose two microbial organisms, namely *Bacillus subtilis* 168 (Kunst, Ogasawara et al. 1997) and *Bacillus anthracis* Sterne (Read, Salzberg et al. 2002) from the KEGG database. *B. subtilis* is totally innocuous to man and has been widely used in scientific and industrial applications in the past. *B. anthracis* is a pathogen that causes anthrax, which in its pulmonary or digestive form is often lethal to humans. A comparative study of the metabolic networks of these two organisms highlights the analogies and differences between their respective pathways. Of course, there is sometimes more than one reaction for an enzyme in KEGG ligand database. Some of the potential reactions may be irrelevant to the organism, as the organism may not use all the reactions coded by an enzyme.

4.3.2 Data Representation in Graph Theory

Using graph theory we can define the system in terms of a bipartite graph (refer to section 2.4.2), which can be reduced to an enzyme-centric graph and a metabolic-centric graph (refer to section 2.4.1.4). In a bipartite view, two nodes share a common enzyme and the edges define the biological relationship between a set of metabolites and enzymes. In the metabolic-centric view metabolites are nodes and reactions/enzymes are edges whereas in the enzyme-centric view, enzymes are nodes and metabolites are edges.

4.4 ALGORITHM

4.4.1 Load Points

Load points are defined as hot spots in the metabolic network (enzymes/metabolites) based on the ratio of number of k -shortest paths passing through a metabolite/enzyme (in/out) and number of nearest neighbour links (in/out) attached to it, compared to the average load value in the network.

For a given metabolic network, the load l on metabolite m can be defined as (4.1)

$$l_{m_{(in/out)}} = \ln \left(\frac{\left[\frac{\sigma_{sp(m)}}{k(m)} \right]}{\left[\frac{\sum_{i=1}^M \sigma_{sp(i)}}{\sum_{i=1}^M K_i} \right]} \right) \quad (4.1)$$

where $-\infty < l_{m_{(in/out)}} < \infty$, σ_{sp} is the number of shortest paths (in/out) passing through a metabolite m ; k is the number of nearest neighbour links (in/out) for m in the network; $\sigma_{sp(i)}$ is the total number of shortest paths and K is the sum of links in the metabolic network of \mathbf{M} metabolites (where M is the number of metabolites in the network). Use of the logarithm makes the relevant values more distinguishable.

The network model emphasises metabolites participating in the shortest path connectivity, thus minimizing the number of less important links. Since the connectivity is based on the metabolite structural similarity, only metabolites satisfying the similarity constraints are included in the pathway (for example, false links via ATP, ADP etc are excluded by the algorithm). A higher load value will

result if a greater number of shortest paths pass through a node (e.g. maximum number of paths) having a minimum number of nearest neighbour connectivity (e.g. minimum number of edges). In the bio-chemical context, load points can suggest the importance of an enzyme or metabolite in a given static metabolic network of various organisms.

4.4.2 Choke Points

Choke points are those enzymes, which uniquely consume and/or produce a certain metabolite. Choke point enzymes are ranked by the load (in/out) and number of k -shortest paths (in/out) passing through them. Since it is a reasonable assumption that a large number of the biochemical reactions follow the shortest path, we assume that the shortest path count can be a good indicator of the biochemical activity.

In our graph model (Figure 41), node 6 (metabolite) and the unique edges (enzymes) attached to it, all represent choke points. Choke points are bio-chemically essential points in the network. Thus removing a single choke point enzyme (edge between nodes 5 & 6 or 6 & 7) from the network affects the consumption or production of the metabolite(s) (e.g. node 5 or 7) attached to it.

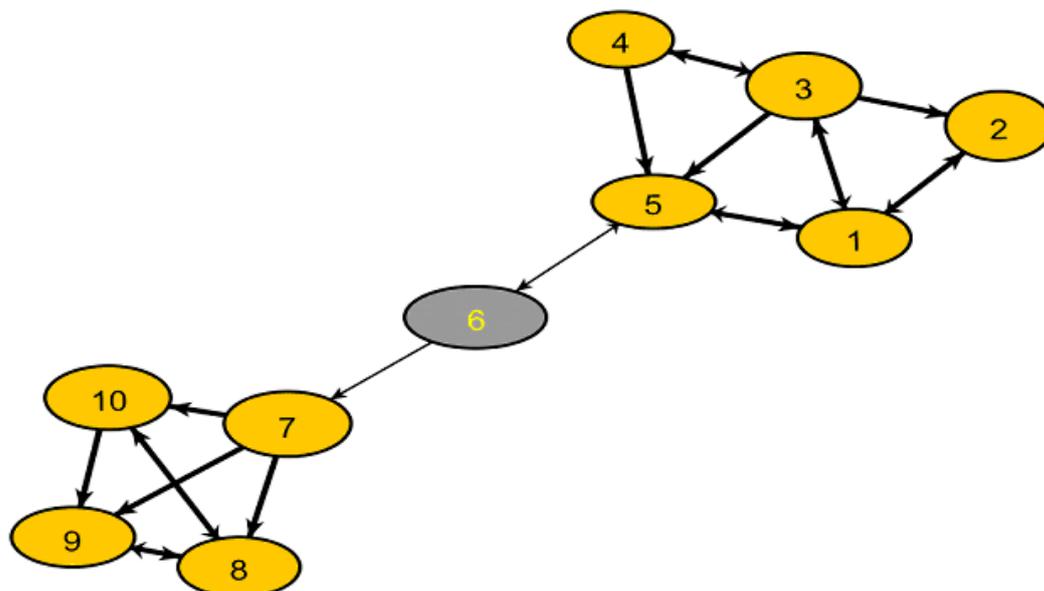


Figure 41. Metabolic-centric view of a graph model. Grey colour node (6) is a choke point (metabolite) and thinner edges adjacent to this node (enzymes) are also choke points. This figure is generated by yEd (<http://www.yworks.com/>).

4.4.3 Hunting and Classification of the Potential Drug Targets

In order to confer biological meaning to the graph-based approach of finding choke points, we proceeded in the following steps.

4.4.4 Reconstruction of the Network

The reconstruction of the metabolic network was done using systems biology knowledge like gene expression data, enzymes and metabolome data for the chosen pathogen (Figure 42).

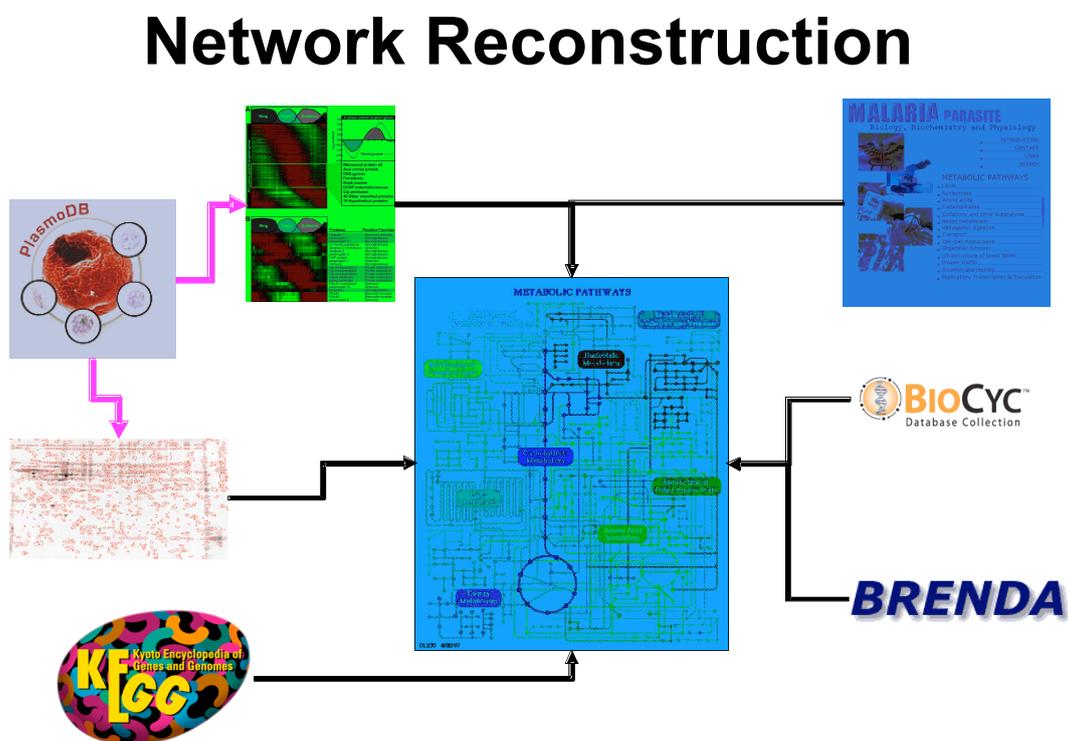


Figure 42. Example of metabolic network reconstruction using systems biology knowledge *i.e.* *Plasmodium falciparum*.

4.4.5 Finding the Choke Points/Non-Choke Points Based Targets

The choke points and the non-choke point enzymes and metabolites were found using PHT. After the calculation of the choke points and non-choke point enzymes in the network, they were ranked on the basis of the number of incoming shortest paths i.e as shown in *Plasmodium falciparum* (Figure 43).

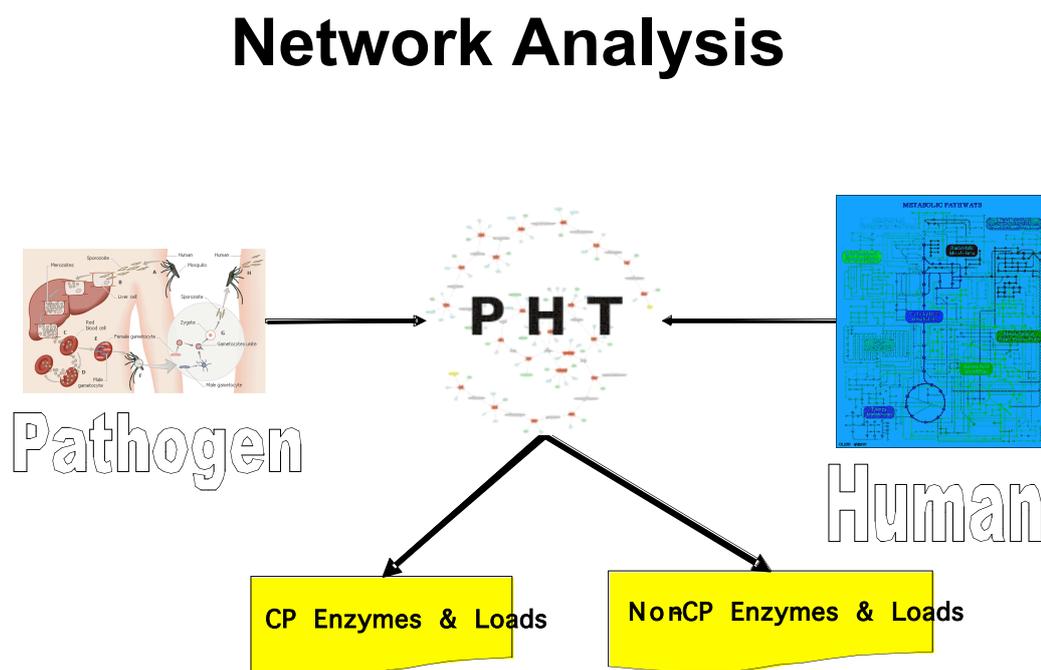


Figure 43. Finding Load and Choke points in the Pathogen and Human metabolic network.

4.4.6 Classification of the Potential Drug Targets

The potential drug targets are divided into four major classes (Figure 44). The Class 1 contains those enzymes which are unique to the pathogen and do not share any homology with human genome. A homology search (Figure 45) was performed between the human and *pathogen's metabolic* network enzymes using BLAST and choke points with a closest homologue with e-values $< 1.0e-02$ were removed. The Class 2 contains those enzymes, which are found in the human genome (based on the

enzyme nomenclature identifier) but they are not homologues to each other and these enzymes are non-choke point enzymes in *Homo sapiens*. The Class 3 enzymes are those enzymes that are common choke points between *Homo sapiens* and pathogen yet they are non-homologous. The Class 4 enzymes are enzymes that are choke points in *Homo sapiens* as well as the pathogen and are also homologous to each other. In order to corroborate the results, existing and potential drug targets from the literatures were also included as a control factor for our prediction. A network based comparative study of the choke points between *pathogen* and *Homo sapiens* was performed using Pathway Hunter Tool (PHT).

Classifying the Targets

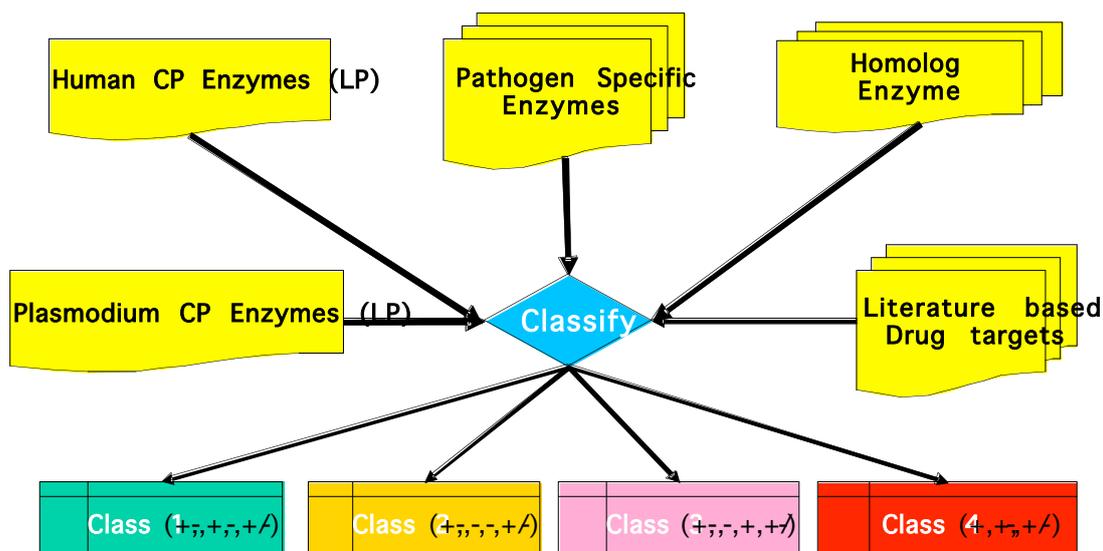


Figure 44. Potential drug targets can be divided in four major categories. Class 1 represents the best and Class 4 the worst. “+” implies that a particular enzyme classification criterion is true (read clockwise).

Comparative Genomics

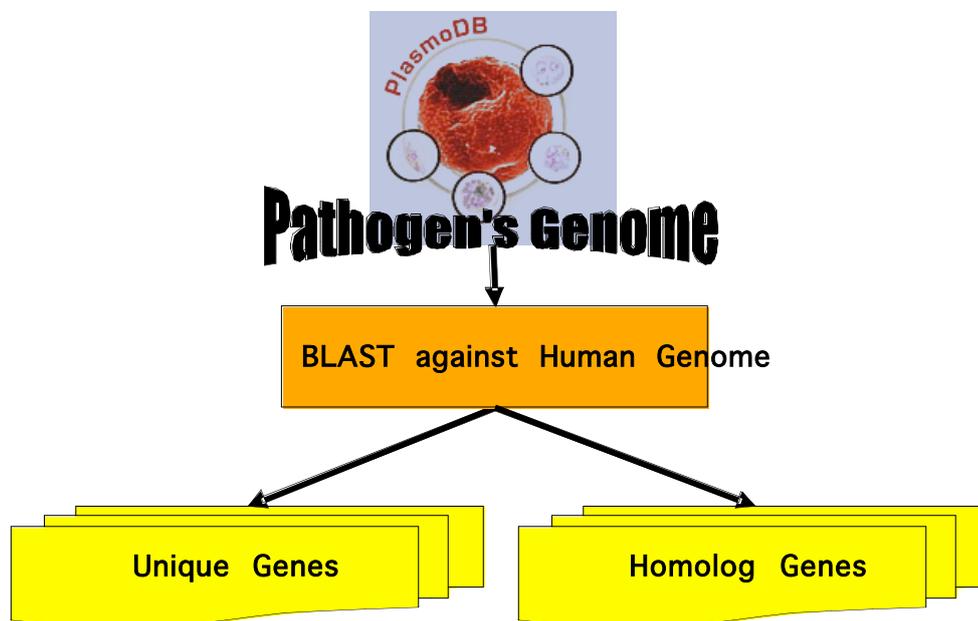


Figure 45. Finding Homolog and Non-Homolog genes between pathogen and Human genome

The similarity measure (refer to 2.4.7) between the two metabolites used in our approach is based on the similarity of their molecular structures as measured by the agreement of their respective 2D molecular fingerprints (Steinbeck, Han et al. 2003). The chosen metabolite similarity criteria (for calculating the shortest pathways) determine the range of network sizes and average degrees of the nodes for the various metabolic networks (Hattori, Okuno et al. 2003; Le, Ho et al. 2004). Thus the higher the global similarity (structural similarity between substrate or source and product or sink in a pathway over a series of intermediate metabolites) and local similarity (the metabolite structural similarity between a pair of consecutive metabolites) cut-off score (Rahman, Advani et al. 2005), the smaller is the network diameter and the average degree of the nodes. In the discussed example the local similarity score was chosen as 15% and global similarity score was chosen as 5%.

4.5 Results

4.5.1 The Choke Point Enzyme Analysis in *Corynebacterium glutamicum*

The choke point enzyme “1.1.1.25” in the Shikimate pathway of *Corynebacterium glutamicum* was knocked out (Figure 46). This resulted in serious repercussion in the growth of the bacteria in the wet lab [≡]. The “load points” also indicated this phenomenon as a major shift in the load graph was observed (Figure 47).

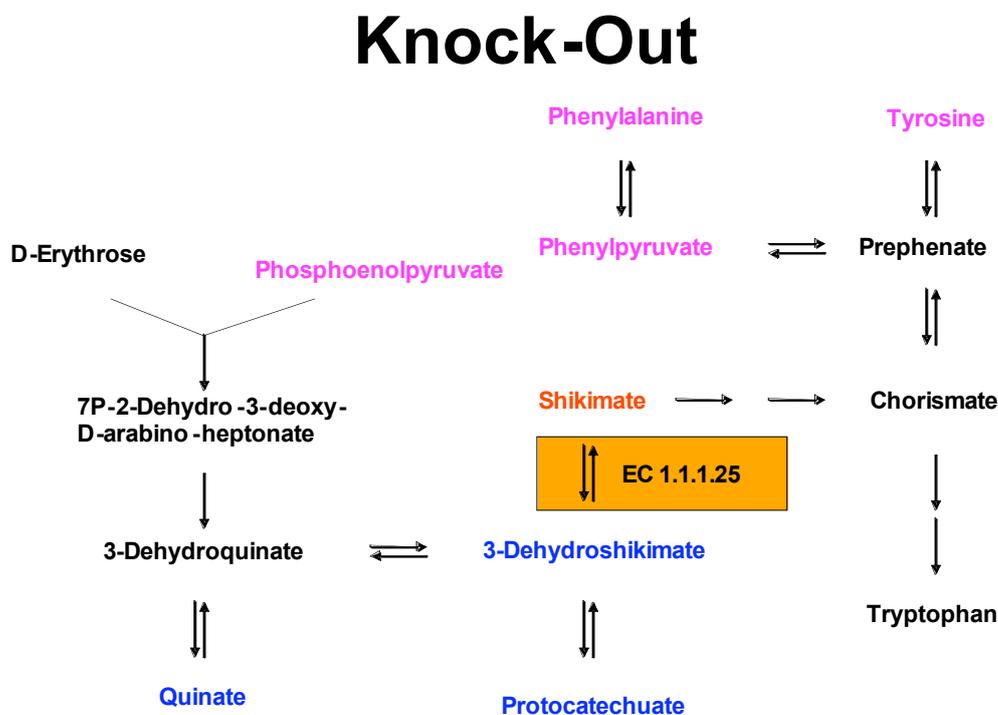


Figure 46. The choke point enzyme 1.1.1.25 in the Shikimate pathway was reported to have high load value.

[≡] Experiments were conducted in Prof. D. Schomburg’s laboratory of Biochemistry by his metabolome work group.

Understanding Metabolomics Data

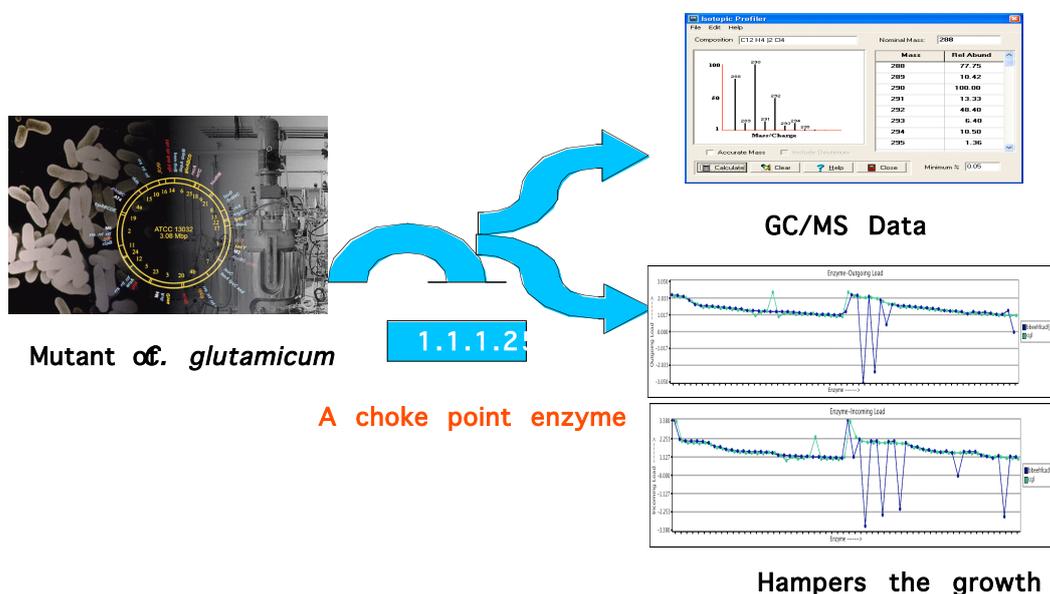


Figure 47. The impact of the knock-out choke point enzyme (gene) 1.1.1.25 as shown in GS/MS peak and the load point analysis.

4.5.1.1 Load Point Analysis of Metabolic Networks between a Pathogenic and a Non-Pathogenic Bacterium

In the pathogenic bacteria *Bacillus anthracis Sterne (bat)* and in non-pathogenic bacteria *Bacillus subtilis 168 (bsu)*, we identified the top 10 metabolite load points (The top '10' load points were chosen purely for convenience – both of presentation and of comprehensibility). The loads on the metabolites differ between the two bacterial networks (incoming load (Table 12) and outgoing load (Table 13)). Most of the metabolites in the list of top 10 load points in pathogenic bacteria *Bacillus anthracis Sterne (bat)* do not match with the load points in non-pathogenic bacteria *Bacillus subtilis 168 (bsu)*.

Table 12. Top 10 Metabolite load points based on incoming load value in *Bacillus subtilis* 168 (bsu) and *Bacillus anthracis* Sterne (bat).

| Metabolites | Rank (bsu) | Rank (bat) | Load bsu | Load bat | Links bsu | Links bat | k-Shortest Path bsu | k-Shortest Path bat |
|--|------------|------------|----------|-----------|-----------|-----------|---------------------|---------------------|
| 2-Phospho -D-glycerate | 1 | 1 | 2.36 | 2.6 | 2 | 2 | 12960 | 14912 |
| GTP | 2 | 2 | 2.29 | 2.39 | 1 | 1 | 6061 | 6060 |
| Formamidopyrimidine nucleoside triphosphate | 3 | 4 | 2.11 | 2.01 | 1 | 1 | 5079 | 4164 |
| 2,5-Diaminopyrimidine nucleoside triphosphate | 4 | 6 | 2.02 | 1.92 | 1 | 1 | 4613 | 3788 |
| Cystathionine | 5 | - | 1.96 | Not Found | 1 | Not Found | 4345 | Not Found |
| 2,5-Diamino-6-(5'-triphosphoryl -3',4'-trihydroxy -2'-oxopentyl) - | 6 | 8 | 1.91 | 1.81 | 1 | 1 | 4145 | 3414 |
| 5-Phospho -alpha -D-ribose 1 -diphosphate | 7 | 7 | 1.82 | 1.82 | 9 | 9 | 34146 | 30861 |
| 2-Amino -4-hydroxy -6-(erythro -1,2,3 -trihydroxypropyl)dihydropteridine | 8 | 10 | 1.79 | 1.09 | 1 | 1 | 3681 | 1653 |
| 3-Phospho -D-glyceroyl phosphate | 9 | 3 | 1.78 | 2.11 | 3 | 3 | 10924 | 13737 |
| Oxaloacetate | 10 | | 1.77 | 1.74 | 6 | 6 | 21594 | 18965 |
| 1-(2-Carboxyphenylamino) -1'-deoxy -D-ribose 5' -phosphate | 11 | 5 | 1.73 | 1.95 | 2 | 2 | 6900 | 7854 |
| Phosphoenolpyruvate | 13 | 6 | 1.59 | 1.83 | 5 | 5 | 15086 | 17363 |
| 3-Phospho -D-glycerate | 14 | 9 | 1.58 | 1.77 | 5 | 5 | 14881 | 16306 |

Table 13. Top 10 Metabolite load points based on outgoing load value in *Bacillus subtilis* 168 (bsu) and *Bacillus anthracis* Sterne (bat).

| Metabolites | Rank (bsu) | Rank (bat) | Load (bsu) | Load (bat) | Links bsu | Links bat | k-Shortest Path bsu | k-Shortest Path bat |
|--|------------|------------|------------|------------|-----------|-----------|---------------------|---------------------|
| 2-Phospho -D-glycerate | 1 | 1 | 2.36 | 2.61 | 2 | 2 | 13028 | 15153 |
| Acetyl adenylate | 2 | 3 | 2.04 | 2.11 | 1 | 1 | 4725 | 4583 |
| Formamidopyrimidine nucleoside triphosphate | 3 | 5 | 2.02 | 1.92 | 1 | 1 | 4613 | 3788 |
| 2,5-Diaminopyrimidine nucleoside triphosphate | 4 | 9 | 1.91 | 1.81 | 1 | 1 | 4145 | 3414 |
| N6-(1,2-Dicarboxyethyl) -AMP | 5 | 13 | 1.81 | 1.74 | 3 | 3 | 11233 | 9548 |
| 3-Phospho -D-glycerate | 5 | 8 | 1.81 | 2.01 | 4 | 4 | 14932 | 16552 |
| 2,5-Diamino -6-(5'-triphosphoryl -3',4' -trihydroxy -2'-oxopentyl) - | 6 | 10 | 1.79 | 1.7 | 1 | 1 | 3681 | 3047 |
| 3-Phospho -D-glyceroyl phosphate | 6 | 2 | 1.79 | 2.13 | 3 | 3 | 10990 | 14035 |
| 1-(2-Carboxyphenylamino) -1'-deoxy -D-ribose 5' -phosphate | 7 | 4 | 1.7 | 1.94 | 2 | 2 | 6732 | 7745 |
| Oxaloacetate | 8 | NR | 1.63 | NR | 7 | NR | 21851 | NR |
| Phosphoenolpyruvate | 9 | 6 | 1.6 | 1.85 | 5 | 5 | 15222 | 17678 |
| Biotinyl -5'-AMP | 10 | 7 | 1.54 | 1.84 | 1 | 1 | 2866 | 3503 |
| Biotin | 14 | 8 | 1.5 | 1.82 | 1 | 1 | 2749 | 3427 |

4.5.1.2 Choke Point Analysis in *B. anthracis* - A Pathogen

An analysis of the top 10 choke points [‡] in *B. anthracis* - a pathogen, revealed a number of possible drug targets against infection of *B. anthracis* are identified (Table 14).

Table 14. A comparative study of top 10 choke points in *Bacillus anthracis* Sterne against the *Homo sapiens* metabolic network. Top 10 Choke point enzymes in *B. anthracis* Sterne ranked by number of shortest paths.

| Enzyme Id | Enzyme Name | Gene Ids | Load (in) | Load (out) | k-SP (in) | k-SP (out) | Human Choke | Top Blast Hit (e-value < 1.0e- | Top Blast Hit (identity) |
|--------------------------|--|-------------------------|-----------|------------|-----------|------------|-------------|--------------------------------|--------------------------|
| 2.4.2.7 | adenine phosphoribosyltransferase | BAS4303 | 1.67 | 1.87 | 31454 | 31454 | + | 7.00E-30 | 42% |
| 2.6.1.1 | aspartate transaminase | BAS1454 | 0.46 | 0.27 | 16600 | 16600 | + | 5.00E-41 | 30% |
| 6.2.1.1 | acetate-CoA ligase | BAS2376 | 0.13 | 0.41 | 15831 | 15831 | + | e-109 | 36% |
| 6.2.1.1 | acetate-CoA ligase | BAS4543 | 0.13 | 0.41 | 15831 | 15831 | + | e-109 | 40% |
| 6.2.1.1 | acetate-CoA ligase | BAS4560 | 0.13 | 0.41 | 15831 | 15831 | + | 5.00E-89 | 34% |
| 3.5.4.16 | GTP cyclohydrolase I | BAS1260 | 2.59 | 2.36 | 14341 | 14341 | + | No Homologue | No Homologue |
| 3.5.4.16 | GTP cyclohydrolase I | BAS1421 | 2.59 | 2.36 | 14341 | 14341 | + | 9.00E-43 | 47% |
| 4.2.1.20 | tryptophan synthase | BAS1161 | 1.67 | 1.26 | 14321 | 14321 | - | No Homologue | No Homologue |
| 4.2.1.20 | tryptophan synthase | BAS1162 | 1.67 | 1.26 | 14321 | 14321 | - | No Homologue | No Homologue |
| 2.7.4.6 | nucleoside-diphosphate kinase | BAS1425 | 0.87 | 0.74 | 14199 | 14199 | + | 7.00E-46 | 57% |
| 4.2.1.11 | phosphopyruvate hydratase | BAS4985 | 1.85 | 2.13 | 13692 | 13692 | + | e-120 | 53% |
| 2.4.2.18 | anthranilate phosphoribosyltransferase | BAS1158 | 1.54 | 1.65 | 12622 | 12622 | - | No Homologue | No Homologue |
| 2.7.4.14 | cytidylate kinase | BAS1407 | 0.82 | 0.92 | 12264 | 12264 | + | No Homologue | No Homologue |
| 2.7.1.40 | pyruvate kinase | BAS3136 | 0.65 | 0.39 | 11919 | 11919 | + | 6.00E-73 | 42% |
| 2.7.1.40 | pyruvate kinase | BAS4492 | 0.65 | 0.39 | 11919 | 11919 | + | e-105 | 44% |

[‡] The top '10' choke points were chosen purely for convenience – both of presentation and of comprehensibility

4.5.2 Potential Drug Targets in *Plasmodium falciparum*

The load point and choke point strategy was implemented on the re-annotated metabolic network of *Plasmodium falciparum*, a malaria-causing agent. Most of the existing/potential drug targets are found by our model (Table 15). We found 31 non-choke points based potential drug targets and 6 of the choke point enzymes targets were found in only one of the malaria life cycle stages. We propose new 29 choke point based anti-malarial potential drug targets. Three proven drugs- fosmidomycin (acting on EC 1.1.1.267), Trimethoprim/ Proguanil/ Pyrimethamine (inhibiting EC 1.5.1.3) and Sulfadiazine/Sulfametopyrazine/ Sulfisoxazole/ Sulfoxone (inhibiting EC 2.5.1.15) are successfully identified by our model. Hence this result, which was used as a control, further highlights the usefulness of our model³.

Table 15. Drug Targets in *Plasmodium falciparum*, and its relevance to our prediction model.

| Targets | CP <i>P.fal</i> | CP <i>H. sapiens</i> | Homology | Drugs/Inhibitors |
|------------|------------------------|----------------------|----------|---|
| 1.1.1.267 | Virtual and 3 stages | - | - | fosmidomycin |
| 1.5.1.3 | + All, - Ring | + | + | Trimethoprim / Proguanil / Pyrimethamine |
| 2.5.1.15 | + All, - Sporozoite | - | - | Dapsone , sulfone /sulfonamide |
| 2.1.1.45 | Merozoite , Sporozoite | - | + | Trimethoprim / Proguanil / Pyrimethamine |
| 1.3.3.1 | Non-CP | + | + | Sulfadiazine/ Sulfametopyrazine / Sulfisoxazole / Sulfoxone |
| 1.14.13.80 | NA | Non-Choke Points | NA | Dapsone |
| 4.1.2.25 | Virtual | NA | NA | Dapsone |
| 2.7.6.3 | + All, - Sporozoite | - | - | Dapsone |
| 1.14.13.67 | NA | Non-Choke Points | NA | Halofantrine / Quinine |
| 3.6.3.14 | Non-CP | - | + | Halofantrine |
| 1.10.2.2 | + All | + | + | Atovaquone |

³ For details please refer the CUBIC project thesis (report) of T. Biru (2005-2006) and J. Padiadpu (2004-2005).

4.5.3 Potential Drug Targets in *Mycobacterium tuberculosis*

The proven potential drug targets, which exist against the tuberculosis producing pathogen *Mycobacterium tuberculosis*, was analysed by our model (Table 16). The results conjure similar prediction rates as observed in the case of the malaria parasite *Plasmodium falciparum*. The non-pathway based drug targets were not picked up by our method as this model is exclusively designed for predicting targets in metabolic networks (Table 17).

Table 16. Classification of the drug targets in *Mycobacterium tuberculosis*

| Enzyme | Genes | Gene ID | Class | Drug/Inhibitor |
|----------|-------|---------|-------|-------------------------------------|
| 2.3.1.41 | fab | MT2306 | I | thiolactomycin |
| 2.1.1.79 | cmaA1 | MT0486 | I | pyrazinamide |
| 1.3.1.9 | inhA | MT1531 | I | Isoniazid , ethoniamide , triclosan |
| 6.3.2.4 | ddlA | MT3059 | II | cycloserine |
| 3.5.2.6 | bla | MT2128 | III | Clavulanic acid |
| 2.1.1.45 | thyA | MT2834 | III | PAS |
| 1.11.1.6 | katG | MT1959 | III | isoniazid |
| 2.3.1.- | aac | MT0275 | + | kanamycin |

Table 17. Classification of the drug targets in *Mycobacterium tuberculosis*

| Enzyme | Genes | Gene ID | Class | Drug / Inhibitor |
|---------------|--------------|----------------|----------------|-------------------------|
| 2.7.7.6 | rpoB | MT0695 | + | Rifampicin |
| 2.7.7.6 | rpoB | MT0695 | + | Rifabutin |
| 2.7.7.6 | rpoC | MT0696 | + | Rifapentene |
| 5.99.1.3 | gyrA | MT0006 | No path | oxfloxacin |
| none | rplJ | MT0680 | No path | Capreomycin |
| none | rpsL | MT0720 | No path | Amikacin |
| none | rpsL | MT0720 | No path | Streptomycin |

5 Discussion

Metabolic network analysis will play a major role in “Systems Biology” in the future. Molecular networks represent the backbone of molecular activity within the cell. Recent studies have taken a comparative approach toward interpreting these networks, contrasting networks of different species and molecular types, and under varying conditions.

5.1 Network Analysis

The shortest path between metabolites can be calculated by various methods (analytical assumption or biochemical knowledge). We proposed a novel method, which uses biochemical information to calculate the shortest path in the metabolic network. In order to achieve such a meticulous approach we have mapped each substrate in the bio-chemical reaction to its corresponding product (refer to section 2.1.5 and 2.1.6). For the calculation of the shortest paths the two biochemical criteria “local” and “global” structural similarity were used, where “local similarity” is defined as the similarity between two intermediate molecules and “global similarity” is defined as the amount of conserved structure (refer to section 2.5.2) found between the source metabolite and the destination metabolites after a series of reaction steps (refer to section 2.4.7). These two criteria helped us to achieve bio-chemically valid shortest path (refer to section 2.5.2 and 2.5.3) in the metabolic network without making any analytical presumptions (elimination of the frequently occurring metabolites in the network).

The atomic mapping algorithm developed by us handles the biochemical mapping in the network, while dynamically finding the paths. Since we use binary fingerprints-a heuristic method (refer to section 2.4.4.1) to find biochemical similarity, we might get some false positive mapping in the network. Though it keeps track of the structural changes molecular fingerprint is not a very sensitive measure. In future it will be a useful to change fingerprints with actual atom counts (of the metabolites) for keeping track of the conserved /matched chemical moieties in the pathway (refer to the MCS method on 2.4.5). Another 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”), as a result of which, the user may miss some connectivity due to lack of structural information.

The shortest path between source and destination metabolite is the minimum number of reaction steps between them. We consider the metabolic pathway in our system to be a directed graph, and all the edges (reactions) share 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 BFS (Breadth First Search) algorithm. The biochemical knowledge about the chemical similarity (refer to section 2.4.2, 2.4.5 and 2.4.6) has been used to satisfy the constraints (similarity) with 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.

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 the future we plan to provide non-standard structural information for these metabolites in order to allow the inclusion of such reactions.

The impact of local similarity is more than global similarity while path finding (refer to 2.5.3.1). This may be because the number of connectivities in the network is directly proportional to the local similarity score. The strength of using shortest path with biochemical knowledge is very well cited in this example (refer to section 2.5.1, 2.5.2). The shortest path between pyruvate and L-lysine resulted in 11 steps in KEGG pathway map whereas by using chemical knowledge we found a shorter path of length 7 (refer to 2.5.3). Thus such missing links can be explored with this algorithm which can be very helpful while interpreting –omics data set.

A comparative study of the network of *E.coli* was performed (refer to 2.5.4). A subset of shortest path (between beta-d-glucose and pyruvate) was found to be similar at the biochemical level. When the metabolic network of *C. glutamicum* was compared with *E.coli* (refer to section 2.5.4.2 and 2.5.4.1) it was evident that they differ both in terms of connectivity and shortest path. It is clear that at the biochemical level the top 10 hubs (in both the organism) based on shortest path may be a better criterion to understand the network rather than just connectivity. On the other hand one can also observe that highly connected metabolites do not occur as top scores in shortest path hubs. While performing the shortest path analysis (distribution of shortest paths refers to section 2.5.4.3) it was also evident that *E.coli* has a greater number of alternate paths^ψ although both the genomes have approximately similar average path length. This further highlights the strength of our algorithm that by using chemical information for calculating pathways, we can achieve biochemical meaningful results.

Thus we are able to calculate a valid shortest path in network. Using Pathway Hunter Tool (PHT) we are able to perform comparative studies between the genomes i.e. hubs, alternate path, paths under constraints (via not via metabolites etc).

^ψ This might also be because of the fact that *E.coli* has been a model organism since long, hence we have better understanding of the genome.

5.2 Pathway Alignment

With the stock pile of already sequenced genomes ever increasing, computational reconstruction of metabolic pathways becomes a crucial step for metabolic pathway analysis (Sharan and Ideker 2006). We present a novel method to perform metabolic pathway alignment from the reconstructed metabolic networks of various genomes. Our method highlights both conserved as well as diverged parts in the metabolic networks of various genomes. This study further can be used to bring out the alternate path, iso-enzymes and evolutionary relationship that may give us a better understanding of the genomes. The *enzyme-percentage matrix* (refer section 3.4.6) and *enzyme-enzyme matrix* can not only provide information about the enzyme preference in the genomes but it can also help us to annotate organisms based on evidence of molecular interactions.

The new tool provides flexibility to the user both at the level of selection of genomes based on the existing phylogenetic profile and pathway alignment based on the chosen source (substrate) and destination (product) metabolite. The user can also build reference matrices from the chosen genomes (network topology) and set the abstraction level of the graph based on the molecular similarity (local and global similarity) as implemented in PHT. Certain enzymes can be skipped while aligning the pathway by the option “not via enzymes” or “not via metabolites” thereby giving the user complete control over the pathway elucidation.

Any pathway alignment without gap (insertion/deletion) may sound very stringent, especially since almost all the annotated metabolic pathways have gaps. To overcome this bottleneck in the pathway alignment we provide the user with a gap insertion option thus giving the user a new insight into the pathways. Further evidences about the existence of the suggested enzymes (insertions/gaps) can be obtained by looking into different annotation database. The “gap filling” or “hole filling” idea in the metabolic alignment may take more computation time but it yields high dividends for the users (refer section 3.4.10 and section 3.5.6).

From the perspective of the algorithm, pathway alignment is an NP-complete problem as finding all the possible paths (of various lengths) in the network is not possible (refer section 3.4.11). Hence we allow a maximum of five gaps in the network as it affects the run time due to increase in the search space. We intend to include an automated annotation module for the suggested insertions (enzymes) while performing the pathway alignment. We expect our tool to be of use in different applications and of value to a wide range of users including pharmacologists, metabolic engineers and in genome annotation.

In the results section a comparative study between two bacilli strains (pathogenic and non pathogenic) were performed. They exhibited similar network topology except that they had alternate paths of different lengths (refer to section 3.5.3). Thus we need a method to amplify these changes/variations in the genomes.

Hence we performed a pathway alignment between four pathogens and a non-pathogenic bacterium (refer to section 3.5.4). The results were promising as one could expect preferences of various enzymes/metabolites in these genomes. The pathogenic strains *Escherichia coli* O157:H7 and *Pseudomonas aeruginosa* have a shorter shortest path between them, while *Mycobacterium tuberculosis* CDC1551, *Bacillus anthracis* Sterne, and a non-pathogenic bacteria *Bacillus subtilis* have a longer shortest path for the same conversion (refer to section 3.5.4, 3.5.4.1). The enzymes preferences between them also vary (refer to section 3.5.4.2). The pathogenic strain of *Bacillus anthracis* at step 5 does not use enzyme EC 3.6.1.1., while the non-pathogenic strain *Bacillus subtilis* uses this enzyme. The colour codes further highlight that this enzyme is only used by 30% to 40 % of bacteria (although this knowledge is limited to the present data). The bar charts above the header further highlight the enzyme preferences at certain steps of conversion between these genomes. Thus the **global information** about enzyme usage in bacteria and **local information** about enzyme preference at various steps of conversion is very well captured by this representation.

Further investigation of the metabolite connectivity in the metabolic-centric view of the alignment makes clear that certain metabolites are conserved and some of them vary (refer to 3.4.9). Invariably, it is the side metabolites (or non-connecting) that vary in the pathway rather than the connecting metabolites. The impact of such variation will change the Gibbs energy profile of the pathway and its dynamics. Thus the Gibbs energy perspective gives us an idea of the thermodynamic feasibility of the reactions.

Since the present network suffers with gaps, pathway alignment that allows filling of these gaps is very beneficial. One such example is demonstrated in the section 3.4.10. The pathway alignment between *E.coli* and *H.pylori* was performed with one gap. We were successfully able to predict a potential missing enzyme in this pathway (refer section 3.5.6). It was evident from the alignment that if we insert an enzyme EC 2.7.1.40 then we will be able to complete the path. Since this enzyme is used by 80% of the genome, it was highly probable that this was a good hit. On the other hand the Gibbs energy profile was also favourable, as the connecting metabolites do not change. Another extension can be the search for the homologous gene in *H.pylori*, which may code for this enzyme.

Thus we define a novel method examining metabolic capabilities of various genomes in the light of systems biology to bring out topological complexity and the crosstalk between the connectivity (Rahman, Jonnalagadda et al. 2005). Along with highlighting the conserved/diverged enzymes in the pathways, our study also provides a local and global outlook to metabolic pathway alignment. We are able to predict potentially missing enzymes in the pathway. The metabolic pathway flexibility and their cross-talk will highlight the adaptation potential in the genomes and allow us to tap organism-specific enzymes for potential drug targeting (especially in drug-resistant strains of parasites).

5.3 Potential Drug Targeting using Load Point and Choke Point

In certain common diseases, e.g., diabetes and obesity, metabolic dysfunction is a core aspect of the patho-physiology. In others, such as cancer, it is secondary, but nevertheless required for disease progression, with, for example, malignant cells requiring increased glycolytic metabolism and nucleic acid synthesis to be able to divide rapidly in an oxygen-deficient tumor environment. Notably, many important drugs (e.g., certain antibiotic and anticancer agents and the leading cholesterol lowering agents) target specific metabolic reactions. Hence, an improved understanding of metabolism is likely to have great value in developing better drugs and disease treatments.

Drug target identification based on “omics” networks (Giaever, Flaherty et al. 2004; Holzhutter and Holzhutter 2004; Yeh, Hanekamp et al. 2004; di Bernardo, Thompson et al. 2005) is a very promising approach that has only recently become possible. The concept of choke points (Dawson and Elliott 1980) in a given network contributes effectively in the identification of the lethality/bottleneck (here potential drug targets) in a network. Since a high load on a certain enzyme means that a large number of shortest paths go through it, therefore indicating a position in the central metabolism, we assume that ranking choke points on the basis of load will move enzymes with a higher probability of biochemical lethality to the top of the candidate list. A comparative study of choke points with the human metabolic network is essential to identify possible interference of the drugs with the human metabolism which might lead to side effects. It has to be kept in mind though, that presently a large number of genes have unidentified functions which could lead to erroneous prediction of choke points. For example, often drug targets are identified by a unique pathogen-specific metabolic activity, as in the case of reverse transcriptase in the case of HIV (Imamichi 2004). However, the screening of the entire metabolic network of the pathogen to find choke point-based potential drug candidates followed by a comparative study with human metabolic network provides additional targets. Examples are the anti-malarial drugs (Sixsmith, Watkins et al. 1984) pyrimethamine and cycloguanil, targeting a choke point enzyme dihydrofolate reductase (1.5.1.3) (also a human homologue) in *Plasmodium falciparum* with some side effects on humans but lethal to the parasite.

Using better pathogen genome annotation (refer to section 4.4.4) and the robust algorithm in the Pathway Hunter Tool (Rahman and Schomburg 2006), we provide a strong foundation for the identification of Choke Point enzymes by metabolic network analysis (refer to section 4.4.6). This has resulted in the prediction and classification of potential drug targets in few of the most lethal disease-causing organisms. As already proved by previous studies, identifying Choke Point enzymes is one of the systematic methods of identifying potential metabolic drug targets (Yeh, Hanekamp et al. 2004) (Rahman and Schomburg 2006). This robust method enables the prediction of potential targets in metabolic pathways of a pathogenic organism (refer to section 4.5.3 and 4.5.2). The lack in functional annotations poses the major drawback in any analysis of a metabolic network due to the lack of connecting enzymes in the pathways. For example, in *M. tuberculosis* we have many hypothetical proteins belonging to the unique PE (Pro-Glu) and PPE (Pro-Pro-Glu) protein families. We may have additional Choke Points due to this lack of connectivity (holes in the network), making it a unique enzyme in the pathway. We may also lack few critical enzymes due to gaps in the functional annotations. An important and better annotation of the pathogen genome would greatly reduce the false Choke Point enzymes in the metabolic network.

There are few gaps in the pathways for which no enzyme(s) has been detected in the genome. This implies that either, a) there are enzymes in the genome that have not been identified, or b) enzymatic functions have not been assigned to the identified proteins, or c) an alternate pathway in the organism exists that does not involve the reaction, or d) there is parasite importation of the enzymatic activity from the host. If the organism produces these missing enzymes, they must reside in the un-annotated (functional) regions of the genome. If the organism does not produce them, they or their products may be imported from the host, a variant pathway that does not use the reaction may exist (refer to 3.4.10 for hole filling), or the pathway may not exist at all. Such discoveries are very important in drug discovery (Morett, Korbel et al. 2003). This further fortifies our study and makes it valuable to the immunological and molecular research community as alternate paths are promising targets of drug discovery.

In addition, we found that lack of sequence similarity has a predictive value for considering a choke point enzyme as a biologically validated drug target in pathogens. Although we have more than hundred choke point enzymes in each studied pathogens, the classification based on the relevance with the human pathway narrowed down the counts to only few critical enzymes with very few exceptions. Enzymes with no significant sequence similarity to any known human enzyme or protein make these predicted choke point enzymes very important as targets for drug designers (refer to section 4.4.5, 4.4.6, and 4.4.3). Even enzymes with some sequence similarity (low e-value) can be considered as drug targets; given the fact that many such enzymes are already targets for currently used drugs (Rahman 2006) (Table 15), (Table 16) and (Table 17).

We were successful in finding new potential drug targets in pathogen *M.tb* and malarial parasite (*P.fal*). Apart from this we also found few choke point drug targets, which already exist as successful drugs thus complimenting our method^{*}. Most of the proposed choke points found support in the existing literature. This fortifies our concept and brings out the effectiveness and robustness of our prediction method. The results obtained from our *in silico* method illustrate that many proposed drug targets in the literature are predicted Choke Point enzymes (Table 15). The effectiveness of our Choke Point analysis for target prediction can be improved with refinement of the underlying metabolic network. The structure of the metabolic network will improve with further annotation of the metabolic functions in the organism, as well as the incorporation of additional types of information, expression at different cellular growth phases and cellular localization of specific enzymes (refer to section 4.4.4).

^{*} For Further details please refer CUBIC project report of T. Biru (2005-2006) and J. Padiadpu (2005-2006).

An analysis of the top 10 choke points (The top ‘10’ choke points were chosen purely for convenience – both of presentation and of comprehensibility) in *B. anthracis* - a pathogen, is presented (Table 14). In (Table 14), a number of possible drug targets against infection of *B. anthracis* are identified. We found that the enzymes tryptophan synthase (EC: 4.2.1.20) and anthranilate phosphoribosyltransferase (EC: 2.4.2.18) could be effective potential drug targets (refer to section 4.5.1.2). Neither of these enzymes are chokepoints in the human metabolic network nor do they share a significant homology with the human genome (Table 14). This means that blocking these enzymes might affect the pathogen but not the human as there exists an alternate pathway.

This approach may contribute to the first identification of potential target enzymes for rational drug design. However, it must be noted that the absence of complete pathway information may lead to false identification of choke points. Additional computational, biological and/or experimental methods or data will further narrow down the list of potential drug targets.

To summarize, our analysis is based on the, a) sound biochemical significance obtained from experimental facts as stated in the literature, b) a comparative study of human and pathogen metabolic network and c) sequence similarity (homologue) search of functional assignment.

Although we have developed a robust method to predict the potential targets in metabolic pathways of organisms in general, there is always room for further improvement. Providing a complete and better annotation *in vivo* (thereby reducing the identification of false Choke Point enzymes and providing previously unreported Choke Point enzymes in the metabolic network) is one of the first steps in this direction. The current analysis includes only the completely annotated enzymes in each organism. Including all the available enzymes for the organisms, such as putative enzymes, may complete the analysis of the metabolic network. The enzymes obtained as Choke Point enzymes in the analysis can be considered as potential drug targets although many other issues need to be taken into consideration (e.g. Non-Choke Point enzymes and their impact). Another approach could be combining Choke Point analysis with chemo-genomic profiling (micro-array data) (di Bernardo, Thompson et

al. 2005) to observe functional responses on the targets. Our provisional targets need to be examined further, both computationally and experimentally for these additional features. The rapid emergence of multi-drug resistant strains of these potentially lethal pathogens calls for the identification of new targets. The discovery of new targets may lead to a drug formulation that would be able to counteract the resurgence of these diseases.

Our results highlight the local and global properties of complex biological metabolic networks. Thus the load (in/out) of metabolites is a more global indicator of their importance, as compared to mere connectivity information. The network model and algorithm presented can process the information contained in the topology of the metabolic network and extract knowledge about the function, role and importance of the metabolites in a network. The extended graph-based choke point concept can facilitate drug discovery and ranking choke points based on their load values may be a likely pointer to the lethality level of such potential drug targets in the network. Further study and comparative analysis of various metabolic networks based on our network model can be beneficial for *in vivo* and *in vitro* studies (refer to section 4.5.1). As a note of caution we would like to add that presently such an analysis is limited by the limited accuracy and completeness of pathway annotations and by the lack of knowledge of the proteins actually present in a certain state of the cell.

The algorithm described has been implemented in the Pathway Hunter Tool (PHT) with the aim of identifying enzymes for potential drug targets and designing synthetic networks with highly specialised metabolic functions.

6 Outlook

Metabolic network analysis will play a major role in “Systems Biology” in the future. Molecular networks represent the backbone of molecular activity within the cell. Recent studies have taken a comparative approach toward interpreting these networks, contrasting networks of different species and molecular types, and under varying conditions. The knowledge about metabolites involved in the cellular process will improve the connectivity information. We need to substitute the molecular fingerprint score while tracking the changes in the metabolite structure by the actual count of the atoms. Apart from finding shortest path, a robust algorithm is needed to find all the paths (path of variable lengths between fixed source and destination) and their impact on the Gibbs energy profile. This will also improve pathway alignment profiles. We can further investigate the role of small molecules and their impact on the enzymatic activity by understanding the reaction mechanism and evolutionary stature of the enzymes. Finding choke point(s), which is lethal to multiple organisms, will be very helpful. We can automate the drug identification process by integrating text mining, micro-array, and proteome and metabolome data for network analysis.

List of Figures

- Figure 1. Edges in a directed graph point one way, such as a graph of metabolic network where edges implies relationship from one metabolite (substrate) to another (product) 2
- Figure 2. Undirected graphs have edges pointing both ways, such as a enzyme-enzyme graph. These represent self-graphs. 3
- Figure 3. A bipartite graph has edges connecting two different sets of nodes, such as reactions and the metabolites coding it. A weighted graph has weights for each edge, such as the stoichiometric weight of the reaction. 3
- Figure 4. An example of an undirected graph. The shortest distance between node B to A is one step. The incoming/outgoing degree of node A is 5. 6
- Figure 5. An example of a directed graph. The shortest path between node B to A is one step and between A to B is three steps. The incoming degree of node A is 4, whereas outgoing degree is 1. 7
- Figure 6. Three models had a direct impact on our understanding of biological networks. They were Random, Scale-free and Hierarchical networks 12
- Figure 7. Bipartite view of one of the shortest paths between pyruvate and citrate in *Bacillus subtilis* 168 (A). Metabolic-centric view (B-top-right) and enzyme-centric view (C-bottom-right) of the above mentioned path. 15
- Figure 8. Map of protein-protein interactions in yeast. Each point represents a different protein and each line indicates that the two proteins are capable of binding to one another. Only the largest cluster, which contains ~78% of all proteins, is shown. The colour of a node signifies the phenotypic effect of removing the corresponding protein (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown) 21
- Figure 9. General Metabolic Pathway. Adapted from <http://www.genome.jp/kegg/pathway/map/map01100.html> 22
- Figure 10. Valine, Leucine and Isoleucine in *Corynebacterium glutamicum*. Adapted from KEGG <http://www.genome.jp> 23
- Figure 11. Shortest Path analysis for the Glycolysis pathway: The green lines represent one of the valid paths between 'alpha-D-glucose' and 'pyruvate', while the pink line represents a bio-chemically invalid shortest path via ADP 26
- Figure 12. Adapted from Arita's (Arita 2004) paper. Two ways to represent the reaction of EC 2.3.1.35. In this reaction, the acetyl moiety of N-acetyl L-ornithine is transferred to L-glutamate to form N-acetyl L-glutamate. (Lower Left) In the scheme of Jeong *et al.* (Jeong, Tombor *et al.* 2000), its two substrates and two products are equally linked to the object representing the EC number, irrespective of their structural changes. (Lower Right) In Arita's (Arita 2004) scheme, conserved sub-structural moieties, coded in different colors, are computationally detected and each link is associated with the information of which atom goes where. 30
- Figure 13. Enzyme class and its reaction mechanism can be solved by mapping the respective metabolites in the reaction dataset. 38
- Figure 14. Metabolite mapping obtained from our algorithm shows that ATP maps to ADP (green line) and D-Glucose maps to D-Glucose-6phosphate (red line). 48
- Figure 15. The remaining structures in the data set are mapped using the same concept of MAX-MAX domination of the similarity score (S) in row/column of the matrix (M). The mapping shows A being mapped to D and B being mapped to C. The mapped structures are again removed from the dataset (This is done by mapping them using MCS algorithm. The common set of nodes and edges are deleted from the structure graph). As there is no further data to be mapped in the dataset, the algorithm automatically terminates the mapping search. 52
- Figure 16. Based on the structural similarity score, structure A is mapped to structure C and structure B is mapped to structure E, as they dominate the scores in rows/columns in the Dynamic Weightage Matrix (M). The mapped part of the structure is then removed from the data set (This is done by mapping them using MCS algorithm. The common set of nodes and edges are deleted from the structure graph). 53
- Figure 17. Metabolite mapping obtained from our old algorithm shows that ATP maps to ADP (thick black line) and D-Glucose maps to D-Glucose-6phosphate (thick black line), whereas the improved algorithm can also predict mapping between ATP and D-

| | |
|---|-----|
| Glucose-6-phosphate (thin black line). The new algorithm is able to match all the possible structure patterns between substrate and product. This algorithm can further be extended for classification of reactions and enzymes apart from path-finding. | 54 |
| Figure 18. Shortest path between Pyruvate and L-Lysine in <i>Corynebacterium glutamicum</i> (cgl) on an undirected metabolic network without mapping information | 61 |
| Figure 19. Shortest path between Pyruvate and L-Lysine in <i>Corynebacterium glutamicum</i> (cgl) on a directed metabolic network without mapping information | 62 |
| Figure 20. Shortest path between Pyruvate and L-Lysine in <i>Corynebacterium glutamicum</i> (cgl) on a directed metabolic network with KEGG mapping information | 63 |
| Figure 21. Tracing and mapping metabolites in reaction pairs found in a biochemical conversion pathway. | 64 |
| Figure 22. Shortest path between pyruvate and l-lysine in <i>Corynebacterium glutamicum</i> comprises of 7 steps as predicted by new mapping algorithm. This path is much shorter than the usual path obtained from hand-drawn maps available through KEGG. | 65 |
| Figure 23. Shortest path between pyruvate and l-lysine in <i>Corynebacterium glutamicum</i> comprises of 7 steps as predicted by new mapping algorithm. This picture highlights the genes, metabolites and enzymes required for conversion (as obtained using Pathway Hunter Tool). | 66 |
| Figure 24. Shortest path between beta-D-glucose and Pyruvate in <i>Escherichia coli</i> CFT073 comprises of 5 steps with local similarity 15% and global similarity 10%. | 69 |
| Figure 25. Shortest path between beta-D-glucose and Pyruvate in <i>Escherichia coli</i> H7 EDL933 comprises of 5 steps with local similarity 15% and global similarity 10%. | 70 |
| Figure 26. Shortest path between beta-D-glucose and Pyruvate in <i>Escherichia coli</i> K12 W3110 comprises of 5 steps with local similarity 15% and global similarity 10%. | 71 |
| Figure 27. Shortest path between beta-D-glucose and Pyruvate in <i>Escherichia coli</i> K12 MG1655 comprises of 5 steps with local similarity 15% and global similarity 10%. | 72 |
| Figure 28. Shortest path between beta-D-glucose and Pyruvate in <i>Escherichia coli</i> H7 Sakai comprises of 5 steps with local similarity 15% and global similarity 10%. | 73 |
| Figure 29. Enzyme-Percentage occurrence matrix resulting from the conversion of beta-d-glucose to pyruvate in the reconstructed network of cgl, eco, mtu, mtc. | 83 |
| Figure 30. Color code for representing the percentage occurrence of enzymes in the organisms. Insertion(s)/Deletion(s) of enzymes in the alignment is represented by white colors. | 84 |
| Figure 31. A flowchart for the proposed pathway alignment model. | 87 |
| Figure 32. Shortest path distribution in <i>B. subtilis</i> 168 and <i>B. anthracis</i> Sterne | 93 |
| Figure 33. Average alternate path distribution in <i>B. subtilis</i> 168 and <i>B. anthracis</i> Sterne. The alternate shortest path in the network can be calculated by dividing the total number of shortest paths at each path length (reaction step) by the unique number of shortest path respectively | 94 |
| Figure 34. The pathway alignment between beta-D-Glucose 6-P and Citrate in four bacterial pathogens and one non-pathogenic bacterium. The assigned local similarity is 15% and global similarity is 5% | 96 |
| Figure 35. Progressive pathway alignment highlighting the enzymes preferred at various steps of the alignment | 96 |
| Figure 36. Progressive pathway alignment highlighting the enzymes preferred at various steps of the alignment. Consensus of the alignment is shown as bar charts and the percentage occurrence of each enzyme in bacteria is shown by the respective colors. | 98 |
| Figure 37. Metabolic centric view of the alignment for the selected paths with deltaG score. Change is deltaG represents change in on of chemical compound constituting the reaction. | 100 |
| Figure 38. The aligned paths between metabolites 3-phospho-D-glycerate and pyruvate in <i>H. pylori</i> and <i>E. coli</i> K-12. The gaps in the alignment as represented by "+" in the header section. | 102 |
| Figure 39. The pathway alignment between 3-phospho-D-glycerate and pyruvate in <i>H. pylori</i> and <i>E. coli</i> K-12. The metabolic centric view of the pathway is shown here. | 103 |
| Figure 40. The Gibbs energy profile the alignment pathway between 3-phospho-D-glycerate and pyruvate in <i>H. pylori</i> and <i>E. coli</i> K-12. | 103 |
| Figure 41. Metabolic-centric view of a graph model. Grey colour node (6) is a choke point (metabolite) and thinner edges adjacent to this node (enzymes) are also choke points. This figure is generated by yEd (http://www.yworks.com/). | 112 |

| | |
|---|-----|
| Figure 42. Example of metabolic network reconstruction using systems biology knowledge <i>i.e. Plasmodium falciparum.</i> | 113 |
| Figure 43. Finding Load and Choke points in the Pathogen and Human metabolic network. | 114 |
| Figure 44. Potential drug targets can be divided in four major categories. Class 1 represents the best and Class 4 the worst. “+” implies that a particular enzyme classification criterion is true (read clockwise). | 115 |
| Figure 45. Finding Homolog and Non-Homolog genes between pathogen and Human genome | 116 |
| Figure 46. The choke point enzyme 1.1.1.25 in the Shikimate pathway was reported to have high load value. | 117 |
| Figure 47. The impact of the knock-out choke point enzyme (gene) 1.1.1.25 as shown in GS/MS peak and the load point analysis. | 118 |

List of Tables

| | |
|---|-----|
| Table 1. Basic Graph Theory Notations | 5 |
| Table 2. Comparison of four <i>E. coli</i> network analyses. The top 10 hub metabolites and ALs reported in each study. Wagner and Fell (Wagner and Fell 2001) computed several versions of the network. The one shown here is the substrate-based network where ATP, ADP, NAD, NADP, NADH, NADPH, carbon dioxide, ammonia, sulfate, thioredoxin, (ortho) phosphate (P), and pyrophosphate (PP) are removed. | 31 |
| Table 3. Local similarity and its impact on the shortest path. Higher the similarity score, longer the path as many links are skipped. | 67 |
| Table 4. Impact of the Global similarity on the length of the reported shortest path. | 68 |
| Table 5. <i>Escherichia coli</i> O157:H7, <i>Corynebacterium glutamicum</i> and their network features like Degree Distribution (DD) = $2*N/L$, Average Path Length (APL), Average k-Path Length (AKPL) | 74 |
| Table 6. Top 10 metabolite hubs in <i>Escherichia coli</i> O157:H7 Sakai based on connectivity and ranked by incoming degree. | 74 |
| Table 7. Top 10 metabolite hubs based on metabolite connectivity in <i>Corynebacterium glutamicum</i> (Ranked by incoming degree). | 74 |
| Table 8. Shortest path (SP) based top 10 incoming metabolite hubs in <i>Escherichia coli</i> O157:H7. | 75 |
| Table 9. The shortest path distribution between <i>Escherichia coli</i> O157:H7 strain and a non-pathogenic bacteria <i>Corynebacterium glutamicum</i> , calculated by Pathway Hunter Tool using CUBIC mapping algorithm (Local similarity 15%, Global similarity 5%). | 76 |
| Table 10. Distribution of the shortest path percentage in the pathogenic <i>Escherichia coli</i> O157:H7 strain and a non-pathogenic bacteria <i>Corynebacterium glutamicum</i> | 77 |
| Table 11. Metabolic network analysis of <i>B. subtilis</i> 168 and <i>B. anthracis</i> Sterne using shortest path..... | 94 |
| Table 12. Top 10 Metabolite load points based on incoming load value in <i>Bacillus subtilis</i> 168 (bsu) and <i>Bacillus anthracis</i> Sterne (bat)..... | 119 |
| Table 13. Top 10 Metabolite load points based on outgoing load value in <i>Bacillus subtilis</i> 168 (bsu) and <i>Bacillus anthracis</i> Sterne (bat)..... | 120 |
| Table 14. A comparative study of top 10 choke points in <i>Bacillus anthracis</i> Sterne against the <i>Homo sapiens</i> metabolic network. Top 10 Choke point enzymes in <i>B. anthracis</i> Sterne ranked by number of shortest paths. | 121 |
| Table 15. Drug Targets in <i>Plasmodium falciparum</i> , and its relevance to our prediction model. | 122 |
| Table 16. Classification of the drug targets in <i>Mycobacterium tuberculosis</i> | 123 |
| Table 17. Classification of the drug targets in <i>Mycobacterium tuberculosis</i> | 124 |

6. Literatures

- Adamic, L. A., Huberman, et al. (2000). "Power-Law Distribution of the World Wide Web." Science **287**(5461): 2115a-.
- Akutsu, T. (2004). "Efficient extraction of mapping rules of atoms from enzymatic reaction data." J Comput Biol **11**(2-3): 449-62.
- Albert, R. and A.-L. Barabasi (2002). "Statistical mechanics of complex networks." Reviews of Modern Physics **74**: 47.
- Albert, R. and A. L. Barabasi (2000). "Topology of evolving networks: local events and universality." Phys Rev Lett **85**(24): 5234-7.
- Albert, R., H. Jeong, et al. (2000). "Error and attack tolerance of complex networks." Nature **406**(6794): 378-82.
- Alon, U. (2003). "Biological networks: the tinkerer as an engineer." Science **301**(5641): 1866-7.
- Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.
- Anderson, L. and J. Seilhamer (1997). "A comparison of selected mRNA and protein abundances in human liver." Electrophoresis **18**(3-4): 533-7.
- Arita, M. (2004). "The metabolic world of Escherichia coli is not small." Proc Natl Acad Sci U S A **101**(6): 1543-7.
- Banavar, J. R., A. Maritan, et al. (1999). "Size and form in efficient transportation networks." Nature **399**(6732): 130-2.
- Barabasi, A.-L. (2002). Linked: The New Science of Networks., Perseus, Cambridge, MA.

- Barabasi, A. L. and R. Albert (1999). "Emergence of scaling in random networks." Science **286**(5439): 509-12.
- Barabasi, A. L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nat Rev Genet **5**(2): 101-13.
- Barrett, C. L., C. D. Herring, et al. (2005). "The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states." Proc Natl Acad Sci U S A **102**(52): 19103-8.
- Batagelj, V. and A. Mrvar (1998). "Pajek—program for large network analysis." Connections **21**: 47-57.
- Bender, A., H. Y. Mussa, et al. (2004). "Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance." J Chem Inf Comput Sci **44**(5): 1708-18.
- Bonday, Z. Q., S. Dhanasekaran, et al. (2000). "Import of host delta-aminolevulinate dehydratase into the malarial parasite: identification of a new drug target." Nat Med **6**(8): 898-903.
- Brandes, U., T. Dwyer, et al. (2004). "Visual Understanding of Metabolic Pathways Across Organisms Using Layout in Two and a Half Dimensions." Journal of Integrative Bioinformatics **0002**.
- Bray, D. (2003). "Molecular networks: the top-down view." Science **301**(5641): 1864-5.
- Brouns, S. J., J. Walther, et al. (2006). "Identification of the missing links in prokaryotic pentose oxidation pathways: Evidence for enzyme recruitment." J Biol Chem.
- Cakir, T., K. R. Patil, et al. (2006). "Integration of metabolome data with metabolic networks reveals reporter reactions." Mol Syst Biol **2**: 50.
- Castresana, J. (2001). "Comparative genomics and bioenergetics." Biochim Biophys Acta **1506**(3): 147-62.

- Catchpole, G. S., M. Beckmann, et al. (2005). "Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops." Proc Natl Acad Sci U S A **102**(40): 14458-62.
- Chen, L. and D. Vitkup (2006). "Predicting genes for orphan metabolic activities using phylogenetic profiles." Genome Biol **7**(2): R17.
- Chen, M. and R. Hofstadt (2004). "PathAligner : Metabolic Pathway Retrieval and Alignment." Appl Bioinformatics **3**(4): 241-52.
- Chung, F. and L. Lu (2002). "The average distances in random graphs with given expected degrees." Proc Natl Acad Sci U S A **99**(25): 15879-82.
- Cohen, R. and S. Havlin (2003). "Scale-free networks are ultrasmall." Phys Rev Lett **90**(5): 058701.
- Copley, R. R. and P. Bork (2000). "Homology among (betaalpha)₈ barrels: implications for the evolution of metabolic pathways." J Mol Biol **303**(4): 627-41.
- Cormen, T. H., C. E. Leiserson, et al. (2001). Introduction to Algorithms, MIT Press and McGraw-Hill.
- Covert, M. W., E. M. Knight, et al. (2004). "Integrating high-throughput and computational data elucidates bacterial networks." Nature **429**(6987): 92-6.
- Croes, D., F. Couche, et al. (2005). "Metabolic PathFinding: inferring relevant pathways in biochemical networks." Nucleic Acids Res **33**(Web Server issue): W326-30.
- Croes, D., F. Couche, et al. (2006). "Inferring meaningful pathways in weighted metabolic networks." J Mol Biol **356**(1): 222-36.
- Csete, M. E. and J. C. Doyle (2002). "Reverse engineering of biological complexity." Science **295**(5560): 1664-9.

- Dandekar, T., S. Schuster, et al. (1999). "Pathway alignment: application to the comparative analysis of glycolytic enzymes." Biochem J **343 Pt 1**: 115-24.
- Dawson, S. V. and E. A. Elliott (1980). "Use of the choke point in the prediction of flow limitation in elastic tubes." Fed Proc **39(10)**: 2765-70.
- di Bernardo, D., M. J. Thompson, et al. (2005). "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks." Nat Biotechnol **23(3)**: 377-83.
- Erdős, P. and A. Rényi (1960). "On the evolution of random graphs." Publ. Math. Inst. Hung. Acad. Sci. **5**: 17-61.
- Ettema, T. J., K. S. Makarova, et al. (2004). "Identification and functional verification of archaeal-type phosphoenolpyruvate carboxylase, a missing link in archaeal central carbohydrate metabolism." J Bacteriol **186(22)**: 7754-62.
- Fell, D. A. and A. Wagner (2000). "The small world of metabolism." Nat Biotechnol **18(11)**: 1121-2.
- Fiehn, O., J. Kopka, et al. (2000). "Metabolite profiling for plant functional genomics." Nat Biotechnol **18(11)**: 1157-61.
- Flower, D. R. (1998). "On the Properties of Bit String-Based Measures of Chemical Similarity." J. Chem. Inf. Model. **38(3)**: 379-386.
- Forst, C. V. and K. Schulten (1999). "Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information." J Comput Biol **6(3-4)**: 343-60.
- Forst, C. V. and K. Schulten (2001). "Phylogenetic analysis of metabolic pathways." J Mol Evol **52(6)**: 471-89.
- Fraser, H. B., A. E. Hirsh, et al. (2002). "Evolutionary rate in the protein interaction network." Science **296(5568)**: 750-2.
- Giaever, G., A. M. Chu, et al. (2002). "Functional profiling of the *Saccharomyces cerevisiae* genome." Nature **418(6896)**: 387-91.

- Giaever, G., P. Flaherty, et al. (2004). "Chemogenomic profiling: identifying the functional interactions of small molecules in yeast." Proc Natl Acad Sci U S A **101**(3): 793-8.
- Girvan, M. and M. E. Newman (2002). "Community structure in social and biological networks." Proc Natl Acad Sci U S A **99**(12): 7821-6.
- Goto, S., Y. Okuno, et al. (2002). "LIGAND: database of chemical compounds and reactions in biological pathways." Nucleic Acids Res **30**(1): 402-4.
- Green, M. L. and P. D. Karp (2004). "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases." BMC Bioinformatics **5**: 76.
- Grigorov, M. G. (2005). "Global properties of biological networks." Drug Discov Today **10**(5): 365-72.
- Gu, Z., L. M. Steinmetz, et al. (2003). "Role of duplicate genes in genetic robustness against null mutations." Nature **421**(6918): 63-6.
- H. Jeong, S. P. M., A.-L. Barabási & Z. N. Oltvai (2001). Map of protein-protein interactions in yeast. PPI_Network.
- Han, J. D., N. Bertin, et al. (2004). "Evidence for dynamically organized modularity in the yeast protein-protein interaction network." Nature **430**(6995): 88-93.
- Hartwell, L. H., J. J. Hopfield, et al. (1999). "From molecular to modular cell biology." Nature **402**(6761 Suppl): C47-52.
- Hasty, J., D. McMillen, et al. (2002). "Engineered gene circuits." Nature **420**(6912): 224-30.
- Hattori, M., Y. Okuno, et al. (2003). "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways." J Am Chem Soc **125**(39): 11853-65.
- Hattori, M., Y. Okuno, et al. (2003). "Heuristics for chemical compound matching." Genome Inform **14**: 144-53.

- Heymans, M. and A. K. Singh (2003). "Deriving phylogenetic trees from the similarity analysis of metabolic pathways." Bioinformatics **19 Suppl 1**: i138-46.
- Holliday, J. D., S. S. Ranade, et al. (1995). "A Fast Algorithm For Selecting Sets Of Dissimilar Molecules From Large Chemical Databases." Quantitative Structure-Activity Relationships **14(6)**: 501-506.
- Holzhutter, S. and H. G. Holzhutter (2004). "Computational design of reduced metabolic networks." Chembiochem **5(10)**: 1401-22.
- Horne, A. B., T. C. Hodgman, et al. (2004). "Constructing an enzyme-centric view of metabolism." Bioinformatics **20(13)**: 2050-5.
- Hubalek, Z. (1982). "Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. ." Biological Reviews of the Cambridge Philosophical Society **57(4)**: 669-689.
- Hulo, N., C. J. Sigrist, et al. (2004). "Recent improvements to the PROSITE database." Nucleic Acids Res **32(Database issue)**: D134-7.
- Ideker, T. (2004). "A systems approach to discovering signaling and regulatory pathways--or, how to digest large interaction networks into relevant pieces." Adv Exp Med Biol **547**: 21-30.
- Iliopoulos, I., S. Tsoka, et al. (2001). "Genome sequences and great expectations." Genome Biol **2(1)**: INTERACTIONS0001.
- Imamichi, T. (2004). "Action of anti-HIV drugs and resistance: reverse transcriptase inhibitors and protease inhibitors." Curr Pharm Des **10(32)**: 4039-53.
- Jaccard, P. (1912). "The distribution of the flora of the alpine zone." New Phytologist **11**: 37-50.
- Jensen, R. A. and W. Gu (1996). "Evolutionary recruitment of biochemically specialized subdivisions of Family I within the protein superfamily of aminotransferases." J Bacteriol **178(8)**: 2161-71.

- Jeong, H., S. P. Mason, et al. (2001). "Lethality and centrality in protein networks." Nature **411**(6833): 41-2.
- Jeong, H., B. Tombor, et al. (2000). "The large-scale organization of metabolic networks." Nature **407**(6804): 651-4.
- Jungnickel, D. (2002). Graphs, Networks and Algorithm, Springer-Verlag.
- Kanehisa, M., S. Goto, et al. (2006). "From genomics to chemical genomics: new developments in KEGG." Nucleic Acids Res **34**(Database issue): D354-7.
- Kanehisa, M., S. Goto, et al. (2004). "The KEGG resource for deciphering the genome." Nucleic Acids Res **32**(Database issue): D277-80.
- Kann, V. (1992). On the approximability of the maximum common subgraph problem. Proc. 9th Annual Symposium on Theoretical Aspects of Computer Science.
- Karp, P. D. (1998). "What we do not know about sequence analysis and sequence databases." Bioinformatics **14**(9): 753-4.
- Karp, P. D., C. A. Ouzounis, et al. (2005). "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes." Nucleic Acids Res **33**(19): 6083-9.
- Kelley, B. P., R. Sharan, et al. (2003). "Conserved pathways within bacteria and yeast as revealed by global protein network alignment." Proc Natl Acad Sci U S A **100**(20): 11394-9.
- Kitano, H. (2002). "Computational systems biology." Nature **420**(6912): 206-10.
- Klamt, S. and E. D. Gilles (2004). "Minimal cut sets in biochemical reaction networks." Bioinformatics **20**(2): 226-34.
- Koonin, E. V., Y. I. Wolf, et al. (2002). "The structure of the protein universe and genome evolution." Nature **420**(6912): 218-23.
- Krapivsky, P. L., G. J. Rodgers, et al. (2001). "Degree distributions of growing networks." Phys Rev Lett **86**(23): 5401-4.

- Kromer, J. O., O. Sorgenfrei, et al. (2004). "In-depth profiling of lysine-producing *Corynebacterium glutamicum* by combined analysis of the transcriptome, metabolome, and fluxome." J Bacteriol **186**(6): 1769-84.
- Kummel, A., S. Panke, et al. (2006). "Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data." Mol Syst Biol **2**: 2006 0034.
- Kunst, F., N. Ogasawara, et al. (1997). "The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*." Nature **390**(6657): 249-56.
- Lander, E. S. (1996). "The new genomics: global views of biology." Science **274**(5287): 536-9.
- Lappe, M. and L. Holm (2004). "Unraveling protein interaction networks with near-optimal efficiency." Nat Biotechnol **22**(1): 98-103.
- Le, S. Q., T. B. Ho, et al. (2004). "A novel graph-based similarity measure for 2D chemical structures." Genome Inform **15**(2): 82-91.
- Le, S. Q., T. B. Ho, et al. (2004). "A Novel Graph-Based Similarity Measure for 2D Chemical Structures." Genome Inform Ser Workshop Genome Inform **15**(2): 82-91.
- Levchenko, A. (2003). "Dynamical and integrative cell signaling: challenges for the new biology." Biotechnol Bioeng **84**(7): 773-82.
- Li, S., C. M. Armstrong, et al. (2004). "A map of the interactome network of the metazoan *C. elegans*." Science **303**(5657): 540-3.
- Lieberman, E., C. Hauert, et al. (2005). "Evolutionary dynamics on graphs." Nature **433**(7023): 312-6.
- Liljeros, F., C. R. Edling, et al. (2001). "The web of human sexual contacts." Nature **411**(6840): 907-8.

- Luscombe, N. M., M. M. Babu, et al. (2004). "Genomic analysis of regulatory network dynamics reveals large topological changes." Nature **431**(7006): 308-12.
- Ma, H. and A. P. Zeng (2003). "Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms." Bioinformatics **19**(2): 270-7.
- Ma, H. W. and A. P. Zeng (2003). "The connectivity structure, giant strong component and centrality of metabolic networks." Bioinformatics **19**(11): 1423-30.
- Ma'ayan, A., A. Lipshtat, et al. (2006). "Topology of resultant networks shaped by evolutionary pressure." Phys Rev E Stat Nonlin Soft Matter Phys **73**(6 Pt 1): 061912.
- Matsuda, H. and Y. Tohsato (2001). "[Detection of similar reaction patterns by using metabolic pathway alignment]." Tanpakushitsu Kakusan Koso **46**(16 Suppl): 2550-4.
- Mavrouniotis, M. L. (1991). "Estimation of standard Gibbs energy changes of biotransformations." J Biol Chem **266**(22): 14440-5.
- Mavrouniotis, M. L. (1993). "Identification of localized and distributed bottlenecks in metabolic pathways." Proc Int Conf Intell Syst Mol Biol **1**: 275-83.
- McGregor., J. J. (1982). "Backtrack search algorithms and the maximal common subgraph problem." Software -Practice and Experience **12**: 23-24.
- Milgram, S. (1967). "The small-world problem." Psychology Today **1**: 60-67.
- Morett, E., J. O. Korb, et al. (2003). "Systematic discovery of analogous enzymes in thiamin biosynthesis." Nat Biotechnol **21**(7): 790-5.
- Newman, M. E. (2001). "Clustering and preferential attachment in growing networks." Phys Rev E Stat Nonlin Soft Matter Phys **64**(2 Pt 2): 025102.

- Newman, M. E. (2001). "Scientific collaboration networks. I. Network construction and fundamental results." Phys Rev E Stat Nonlin Soft Matter Phys **64**(1 Pt 2): 016131.
- Newman, M. E. (2001). "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality." Phys Rev E Stat Nonlin Soft Matter Phys **64**(1 Pt 2): 016132.
- Newman, M. E. (2001). "The structure of scientific collaboration networks." Proc Natl Acad Sci U S A **98**(2): 404-9.
- Newman, M. E., S. H. Strogatz, et al. (2001). "Random graphs with arbitrary degree distributions and their applications." Phys Rev E Stat Nonlin Soft Matter Phys **64**(2 Pt 2): 026118.
- Newman, M. E., D. J. Watts, et al. (2002). "Random graph models of social networks." Proc Natl Acad Sci U S A **99** **Suppl 1**: 2566-72.
- Newman, M. E. J. (2003). "The structure and function of complex networks." SIAM Review **45**(2): 167–256.
- Oliver, S. G. (2006). "From genomes to systems: the path with yeast." Philos Trans R Soc Lond B Biol Sci **361**(1467): 477-82.
- Oltvai, Z. N. and A. L. Barabasi (2002). "Systems biology. Life's complexity pyramid." Science **298**(5594): 763-4.
- Osterman, A. and R. Overbeek (2003). "Missing genes in metabolic pathways: a comparative genomics approach." Curr Opin Chem Biol **7**(2): 238-51.
- Ozbudak, E. M., M. Thattai, et al. (2004). "Multistability in the lactose utilization network of Escherichia coli." Nature **427**(6976): 737-40.
- Papin, J. A., N. D. Price, et al. (2003). "Metabolic pathways in the post-genome era." Trends Biochem Sci **28**(5): 250-8.

- Papin, J. A., J. L. Reed, et al. (2004). "Hierarchical thinking in network biology: the unbiased modularization of biochemical networks." Trends Biochem Sci **29**(12): 641-7.
- Papp, B., C. Pal, et al. (2004). "Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast." Nature **429**(6992): 661-4.
- Podani, J., Z. N. Oltvai, et al. (2001). "Comparable system-level organization of Archaea and Eukaryotes." Nat Genet **29**(1): 54-6.
- Rahman, S. A. (2006). Malaria Drug Targets.
- Rahman, S. A. (2006). SP based Hub Incoming CGL.
- Rahman, S. A., P. Advani, et al. (2005). "Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC)." Bioinformatics **21**(7): 1189-93.
- Rahman, S. A., P. S. Jonnalagadda, et al. (2005). "Metabolic Network Analysis: Implication And Application." BMC Bioinformatics **6**(Suppl 3)(S12).
- Rahman, S. A. and D. Schomburg (2006). "Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks." Bioinformatics **22**(14): 1767-74.
- Ravasz, E., A. L. Somera, et al. (2002). "Hierarchical organization of modularity in metabolic networks." Science **297**(5586): 1551-5.
- Raymond, J. W., E. J. Gardiner, et al. (2002). "Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm." J Chem Inf Comput Sci **42**(2): 305-16.
- Raymond, J. W. and P. Willett (2002). "Maximum common subgraph isomorphism algorithms for the matching of chemical structures." J Comput Aided Mol Des **16**(7): 521-33.
- Read, T. D., S. L. Salzberg, et al. (2002). "Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis." Science **296**(5575): 2028-33.

- Rison, S. C., S. A. Teichmann, et al. (2002). "Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*." J Mol Biol **318**(3): 911-32.
- Schilling, C. H. and B. O. Palsson (2000). "Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis." J Theor Biol **203**(3): 249-83.
- Schilling, C. H., S. Schuster, et al. (1999). "Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era." Biotechnol Prog **15**(3): 296-303.
- Schmidt, S., S. Sunyaev, et al. (2003). "Metabolites: a helping hand for pathway evolution?" Trends Biochem Sci **28**(6): 336-41.
- Schomburg, I., A. Chang, et al. (2004). "BRENDA, the enzyme database: updates and major new developments." Nucleic Acids Res **32**(Database issue): D431-3.
- Schuffenhauer, A., V. J. Gillet, et al. (2000). "Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors." J Chem Inf Comput Sci **40**(2): 295-307.
- Schuster, S., D. A. Fell, et al. (2000). "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks." Nat Biotechnol **18**(3): 326-32.
- Schwikowski, B., P. Uetz, et al. (2000). "A network of protein-protein interactions in yeast." Nat Biotechnol **18**(12): 1257-61.
- Sharan, R. and T. Ideker (2006). "Modeling cellular machinery through biological network comparison." Nat Biotechnol **24**(4): 427-33.
- Simeonidis, E., S. C. Rison, et al. (2003). "Analysis of metabolic networks using a pathway distance metric through linear programming." Metab Eng **5**(3): 211-9.

- Sixsmith, D. G., W. M. Watkins, et al. (1984). "In vitro antimalarial activity of tetrahydrofolate dehydrogenase inhibitors." Am J Trop Med Hyg **33**(5): 772-6.
- Steinbeck, C., Y. Han, et al. (2003). "The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics." J Chem Inf Comput Sci **43**(2): 493-500.
- Strelkov, S., M. von Elstermann, et al. (2004). "Comprehensive analysis of metabolites in *Corynebacterium glutamicum* by gas chromatography/mass spectrometry." Biol Chem **385**(9): 853-61.
- Strogatz, S. H. (2001). "Exploring complex networks." Nature **410**(6825): 268-76.
- Teichmann, S. A., S. C. Rison, et al. (2001). "The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*." J Mol Biol **311**(4): 693-708.
- Tohsato, Y., H. Matsuda, et al. (2000). "A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy." Proc Int Conf Intell Syst Mol Biol **8**: 376-83.
- Tsoka, S. and C. A. Ouzounis (2001). "Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*." Genome Res **11**(9): 1503-10.
- Verhoeckx, K. C., S. Bijlsma, et al. (2004). "Characterization of anti-inflammatory compounds using transcriptomics, proteomics, and metabolomics in combination with multivariate data analysis." Int Immunopharmacol **4**(12): 1499-514.
- von Stockar, U., T. Maskow, et al. (2006). "Thermodynamics of microbial growth and metabolism: an analysis of the current situation." J Biotechnol **121**(4): 517-33.
- Wagner, A. and D. A. Fell (2001). "The small world inside large metabolic networks." Proc Biol Sci **268**(1478): 1803-10.
- Wall, M. E., W. S. Hlavacek, et al. (2004). "Design of gene circuits: lessons from bacteria." Nat Rev Genet **5**(1): 34-42.

- Wasserman, S. and K. Faust (1994). Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press.
- Watts, D. J. (1999). Small Worlds. Princeton, Princeton University Press.
- Watts, D. J. (2003). Six Degrees: The Science of a Connected Age., Norton, New York.
- Watts, D. J. and S. H. Strogatz (1998). "Collective dynamics of 'small-world' networks." Nature **393**(6684): 440-2.
- Whittle, M., P. Willett, et al. (2003). "Evaluation of similarity measures for searching the dictionary of natural products database." J Chem Inf Comput Sci **43**(2): 449-57.
- Willett, P. (2003). "Similarity-based approaches to virtual screening." Biochemical Society Transactions **31**: 603-606.
- Willett, P., J. M. Barnard, et al. (1998). "Chemical Similarity Searching." J. Chem. Inf. Model. **38**(6): 983-996.
- Winzler, E. A., D. D. Shoemaker, et al. (1999). "Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis." Science **285**(5429): 901-6.
- Xia, Y., H. Yu, et al. (2004). "Analyzing cellular biochemistry in terms of molecular networks." Annu Rev Biochem **73**: 1051-87.
- Yeh, I., T. Hanekamp, et al. (2004). "Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery." Genome Res **14**(5): 917-24.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. Rainer Schrader betreut worden.

Teilpublikationen:

- I. **Observing local and global properties of metabolic pathways: "Load points" and "Choke points" in the metabolic network:** Syed Asad Rahman and Dietmar Schomburg: *Bioinformatics* (2006) 22: 1767-1774.
- II. **Metabolic Network Analysis: Implication and Application:** S. A. Rahman et. al.: *BMC Bioinformatics* (2005): 471-2106-5-6-S3-S12
- III. **Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC):** S. A. Rahman, P. Advani, R. Schunk, R. Schrader, Dietmar Schomburg: *Bioinformatics* (2005) 21: 1189-1193.

Syed Asad Rahman

1. Referent: Prof Dr. R. Schrader
2. Referent: Prof. Dr. D. Schomburg

Eingereicht: 12th Dec 2006
Tag der muendlichen Pruefung: 14th Feb.2007

SYED ASAD RAHMAN

Geachlecht: männlich, **Nationalität:** Indisch,
Geburtsdatum: 9. September 1976,
Verheiratet, keine Kinder

17 Loris Court, Cambridge, CB1 9GF, UK

| | |
|---|---|
| <u>Ausbildung</u> <ul style="list-style-type: none">• Doktorand (Informatik) CUBIC, Universität zu Köln, Köln, Deutschland• Postgraduierten-Zertifikat in Bioinformatik CUBIC, Universität zu Köln, Köln, Deutschland• Bachelor of Engineering (Computer Science) BVBCET, Karnatak University, Karnataka, Indien | April 2003- heute März 2002-März 2003 Mai 1995-Okt 2000 |
| <u>Auszeichnungen und Stipendien</u> <ul style="list-style-type: none">• European Science Foundation (ESF) finanzierte für "Modelling Metabolism and Signal Transduction", St. Hugh's College, University of Oxford, UK.• European Science Foundation (ESF) finanzierte für "Advanced Data Mining and Visualisation Approaches to Systems Biology", University of Ulster, UK.• Federal Ministry of Education and Research (BMBF) finanzierte "Stipendium für hochqualifizierte Studenten" des postgraduierten-Ausbautudiums in Bioinformatik am CUBIC, Universität zu Köln, Köln, Deutschland. | Sep 2004 Nov 2003 2002-2003 |