4-21-2008

Mapping and Filling Metabolic Pathway Holes

Dipendra Kaur

MAPPING AND FILLING METABOLIC PATHWAY HOLES

by

DIPENDRA KAUR

Under the Direction of Dr. Alexander Zelikovsky & Dr. Robert Harrison

The network-mapping tool integrated with protein database search can be used for filling pathway holes. A metabolic pathway under consideration (pattern) is mapped into a known metabolic pathway (text), to find pathway holes. Enzymes that do not show up in the pattern may be a hole in the pattern pathway or an indication of alternative pattern pathway. We present a data-mining framework for filling holes in the pattern metabolic pathway based on protein function, prosite scan and protein sequence homology. Using this framework we suggest several fillings found with the same EC notation, with group neighbors (enzymes with same EC number in first three positions, different in the fourth position), and instances where the function of an enzyme has been taken up by the left or right neighboring enzyme in the pathway. The percentile scores are better when closely related organisms are mapped as compared to mapping distantly related organisms.

INDEX WORDS: Network Mapping, Filling Metabolic Pathway Holes, Dipendra Kaur Thesis.

MAPPING AND FILLING METABOLIC PATHWAY HOLES

by

DIPENDRA KAUR

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science
in the College of Arts and Sciences
Georgia State University

2008

MAPPING AND FILLING METABOLIC PATHWAY HOLES

by

DIPENDRA KAUR

Committee Chair: Dr. Alexander Zelikovsky
Dr. Robert Harrison

Committee: Dr. Irene Weber

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
May 2008

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my thesis advisors, Dr. Alex Zelikovsky and Dr. Robert Harrison for their patient guidance, insightful suggestions for my thesis and making my graduate studies a rewarding experience. It has been a great honor for me to participate in this challenging research in Bioinformatics. Appreciation is also extended to my thesis committee member Dr. Irene Weber for her help in improving the content.

During the course of my studies, I had a tremendous opportunity to interact with a number of individuals in the disciplines of computer science, biology and bioinformatics who have influenced this work and me personally: I would like to thank my graduate advisor Dr. Walter (Bill) Walthall (Biology Department, GSU), and Dr. Raj Sunderraman (CS Department, GSU), for helping me in this project. I would also like to thank Qoing Cheng and Kelly Westbrooks for their support in my thesis.

Finally, I would like to express my sincere thanks to my family members for their patience and support during my graduate studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Most proteins act as part of complex network of reactions taking place in the cells. Therefore, they are also used as drug targets affecting the ability of drugs to enable or disable the target proteins. Proteins also influence signal processing which is a method of communication between various life processes. An understanding of the cell networks requires analyzing these complex interactions as a system. Several proteins show certain conserved structural domains, which may be used in a number of different pathways. While there are several protein homologues conserved in many different organisms, some proteins are unique to a single organism. As more genomes and proteomes are characterized, comparison between genomes and proteomes will allow us to better understand the evolutionary history of these organisms.

Enzymes are proteins (long chains of amino-acids linked by peptide bonds). These enzymes control several metabolic processes within the cells by catalyzing the reactions converting nutrients into energy and new molecules. These new molecules are the building blocks of larger molecules and cell organelles like cell membranes, DNA, polysaccharides and other proteins. The enzymes are able to speed up reactions by lowering the activation energy of the reactants, so that the reactions could take place at normal body temperature and pH. In the absence of enzymes, these reactions would be very slow and unable to support life. An understanding of specific enzymes involved in various metabolic pathways might also lead to the discovery of new drugs that target specific pathways for treatment of diseases [1].

Validation of hundreds of potential pathway candidates by conducting experiments in a wet-lab environment is expensive and time-consuming process. Network mapping is useful for comparing and exploring biological pathways. It can be used for predicting unknown and partially known pathways identifying conserved pathways, indicating potential pathway holes and mapping gaps in existing pathways and filling pathway holes and ambiguities [1].

II. WHAT CAUSES PATHWAY HOLES?

Enzyme annotations in genome databases have been used to predict metabolic pathways present in an organism. An open reading frame is used to determine the amino acid sequence encoded by a gene. An error in reading an ORF (open reading frame) may lead to a pathway hole [10]. If a gene to encode an enzyme that is needed to catalyze a reaction in a metabolic pathway is not identified in an organism's genome, it may result in a pathway hole. Sometimes, due to gaps in research, several sequences may not get specific annotations. Specific function of a protein may not be known during annotation. Reactions catalyzed by those proteins, may result in metabolic pathway holes [3][4]. With further research, some of those proteins get specific annotations and pathway descriptions should be updated in pathway/genome databases.

III. HOW ARE METABOLIC PATHWAY HOLES FILLED?

Previous research uses nucleotide sequence similarity to known enzyme coding genes [5] and similarity in pathway expression in related organisms to fill pathway holes [6]. A good computational method for predicting the function of proteins is by studying the annotations of similar sequences, because similar sequences usually have common descent, and therefore, similar structure and function [7]. The framework introduced here finds potential enzymes for filling pathway holes by searching the functionally similar proteins using online protein databases. The functionally similar proteins are identified by the type of reaction catalyzed by the enzyme and significant amino acid sequence homology. The type of reaction catalyzed by an enzyme is given by EC notation assigned to the enzyme.

As we know that, an amino acid may be coded by multiple codons (64 codons coding for 20 amino-acids), there may be several different DNA sequences coding for the same protein. In Eukaryotes the presence of introns (non coding regions in a gene), pseudo genes (genes that have lost protein coding ability) and alternate splicing (mechanism where exons can be reconnected in several different combinations to code for different proteins) makes it more complicated for using nucleotide sequence for finding proteins for filling metabolic pathways. Therefore using nucleotide sequence similarity may not be a good approach. The proposed framework for filling pathway holes is based enzyme reaction and amino acid sequence homology and, therefore, should be superior to existing frameworks based on DNA homology [6].

IV. PROBLEM FORMULATION

A metabolic pathway is a series of chemical reactions going on within a cell. These reactions are catalyzed by enzymes. In a metabolic pathway graph the enzymes are represented by vertices and their corresponding reactions are represented by directed edges. The products of one reaction act as substrates for the catalyzing enzymes in the next reaction. If a vertex in text pathway is not matched to a vertex in the pattern pathway, this is identified as a hole or ambiguity. The objective of this framework is to find candidate enzymes for filling pathway holes, based on specific enzyme reaction properties and protein function homology, using computational tools.

V. PROPOSED MODEL TO FILL PATHWAY HOLES

A metabolic pathway can be visualized as directed networks in which vertices correspond to enzymes and edges correspond to reactions. In these networks there is a directed edge from one enzyme to another if the product of the reaction catalyzed by the first enzyme is a substrate of the enzyme catalyzing the second reaction. Mapping metabolic pathways should capture the conserved pathways between different organisms and also identify dissimilarities and ambiguities in the pathways mapped. Mapping of an incomplete metabolic network of a pattern organism into a better known metabolic network can identify possible pathway holes in the pattern as well as suggest possible candidates for filling those pathway holes [2]. The model for proposed framework for filling metabolic pathway holes is shown. Figure 1.

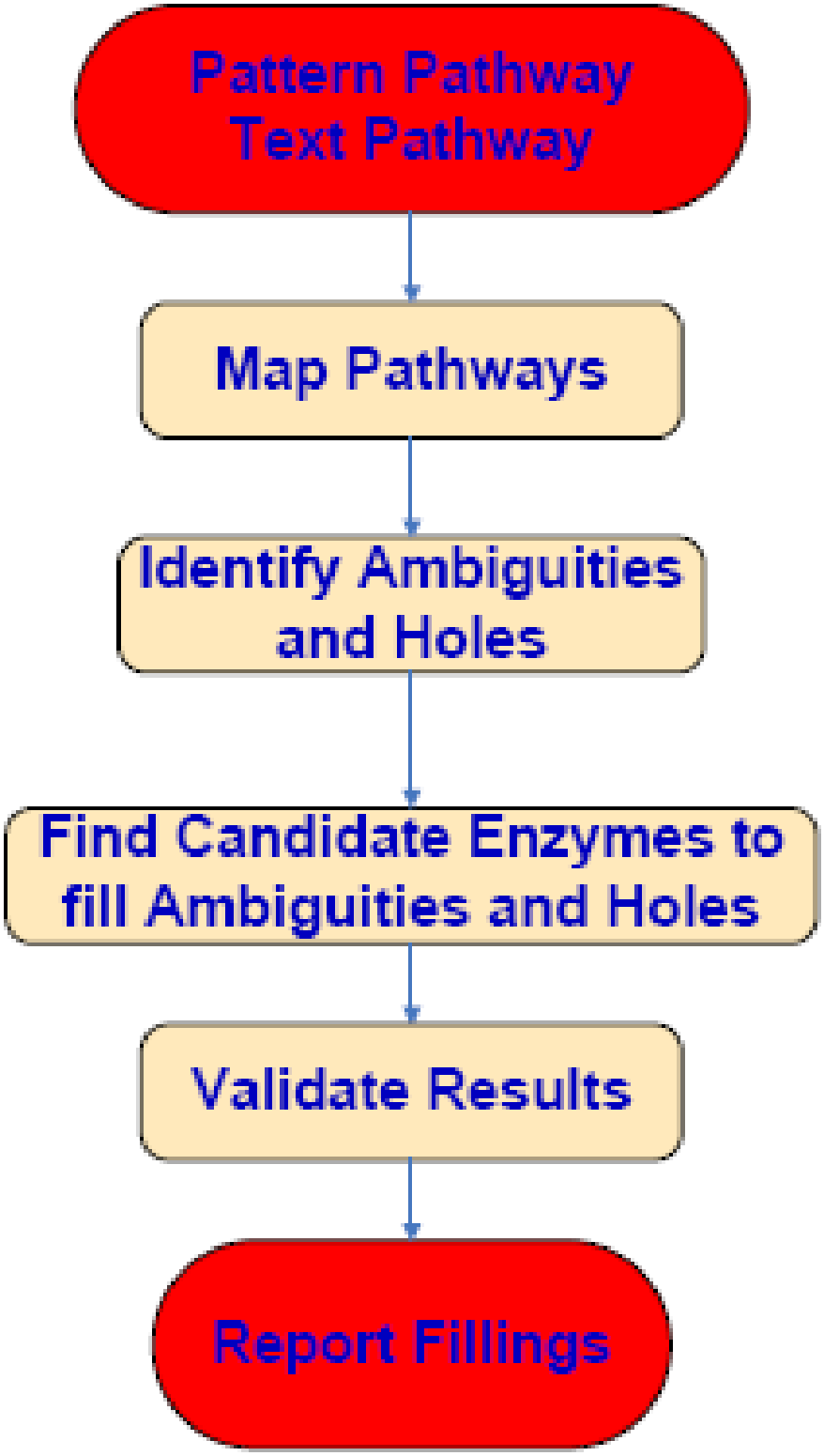

Figure 1: Framework for filling pathway holes

Using EC notation to find functionally similar proteins

Under this system of classification, all enzymes have been classified into 6 classes: EC1-Oxidoreductases EC2 – Transferases, EC 3 – Hydrolases, EC 4 – Lyases, EC 5 – Isomerases and EC 6 - Ligases. Based on the specificity of the enzyme catalyzed, each class is further subdivided into 3 more levels of subclasses. The 4-digit EC number, d1.d2.d3.d4 represents a sub-sub-subclass indication of biochemical reaction. In the mapping approach used, if d1.d2 of two enzymes are different, they are highly dissimilar; if d1 & d2 are same but d3 of two enzymes is different, they are somewhat similar; if d1, d2 & d3 are same but d4 of two enzymes is different, they are relatively more similar. Experimental studies indicate that such a similarity score scheme results in biochemically more relevant pathway matches [2].

Using prosite scan to find functionally similar proteins

PROSITE database contains information about various protein families and domains. Hundreds of proteins involved in various reactions can be grouped in families. Proteins belonging to a family have common functional features and a common ancestor. Certain regions of protein sequences have been well conserved compared to others. These regions are important for the proteins function and 3-D Structure. “By analyzing the constant and variable properties of such groups of similar sequences, it is possible to derive a signature for a protein family or domain, which distinguishes its members from all other unrelated proteins” [14]. These signatures can be used just as finger prints to identify a protein with a specific function. PROSITE database is a

collection of motifs that can be used to identify over a thousand proteins and domains [14]. In cases where the alignment score is close to 50%, we can use prosites to see if the aligned portion of the pattern enzyme has the important prosites found in text enzyme.

Using amino-acid sequence homology to find functionally similar proteins

Several methods have been developed to predict the function of a protein through sequence similarity by comparing the amino-acid sequences of a protein of unknown function with one or more proteins with experimentally or computationally predicted function. It has been observed that the probability of two proteins having similar functions increases with increasing sequence identity and fewer gaps in their alignment. A lower e-value also indicates a stronger functional similarity [12]. But, e-value depends on a number of factors including length of the amino-acid sequence and size of the database searched. Therefore, e-value is not a very reliable score to predict functional similarity. However, sequence similarity alone can only provide a good basis for function prediction; this along with reaction properties of an enzyme can be a strong basis of function prediction.

Finding candidate enzymes to fill ambiguities and holes

The proposed framework follows the following protocol for reporting and recommending fillings for pathway holes. The framework identifies the ambiguous or missing enzymes in the pattern organism from pathway mapping data. These enzymes (EC Notations) are then searched in the pattern organism using Swiss-Prot and

TrEMBL databases. Fasta sequences for reported accession number are extracted (level 1 database search). In some organisms, some proteins may take up multiple functions (protein complexes) [14], therefore fasta sequences for left and right neighboring enzymes in the pattern species are also extracted (level 2 database search). There may be possibility of assigning a wrong EC notation to the enzyme; therefore fasta sequences for the group (EC notation d.d.d.x), in the pattern organism are also extracted (level 3 database search).

Pair-wise sequence alignment is done between text and pattern enzymes. If any significant alignments are found in level 1 search, they are reported as fillings for pathway holes. Significant alignments are selected as those having same EC notation as text enzyme in the mapped location, and greater than 50% sequence identity. In case no significant alignments are found in level 1 database search, it is assumed that the function of this enzyme has been taken up left or right neighbors in the pathway or there may be an error in assigning the EC notation to the enzymes. Therefore, pair-wise alignment is done for left/right neighbors in the pattern organism with the text enzyme; pair-wise alignment is also done between group neighbors with the text enzyme. The corresponding alignment scores for all these alignments are reported. The alignments with alignment score in the range of 50% are checked for prosites. The recommended fillings would provide a good basis for experimental verification. Finally, keyword search is done to report Pubmed links for fillings or recommended fillings from published literature. Figure 2

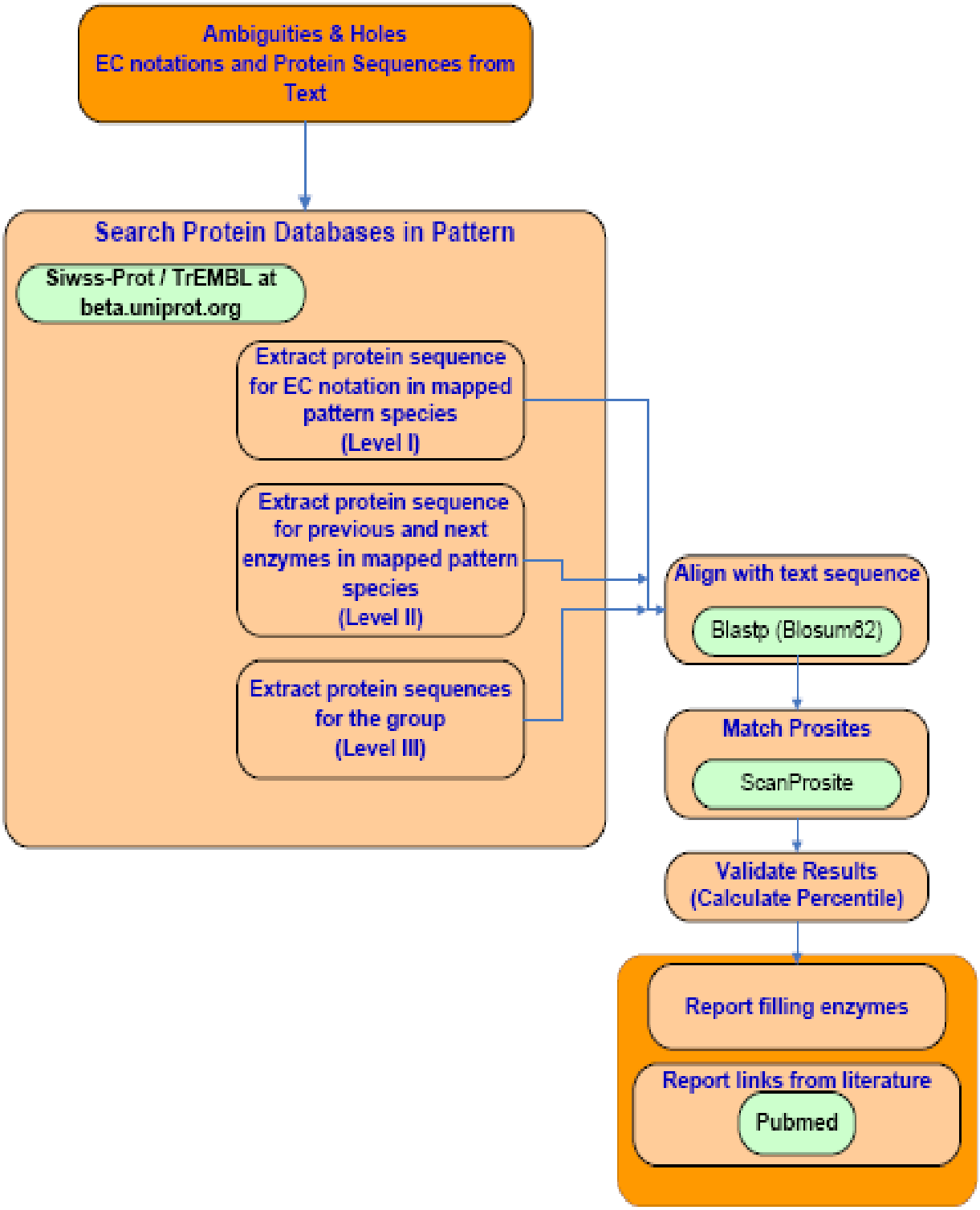

Figure 2: Framework for finding candidate enzymes to fill ambiguities and pathway holes

Framework for filling metabolic pathway holes

Metabolic pathway mapping data is provided as an input to the framework for filling pathway holes. The mapping data comprises of the following information in that order:

- P-Species : pattern species name
- P-PW : pattern pathway name
- T-Species : text species name
- T-PW : text pathway name
- GAPLIST : text start enzyme[pattern start enzyme];GAPS IN TEXT; text end enzyme[pattern end enzyme]
- GAPS IN TEXT : gaps separated by semicolon

The framework for filling pathway holes identifies the pattern organism, text organism, holes in the pattern pathway, and left and right neighboring enzymes. The holes are identified as the enzymes (EC notations) present in the text organism, but, absent or ambiguous in the pattern organism. The framework first extracts the text organism's fasta sequences for enzymes appearing as holes in the pattern. A first level search is conducted by querying the database for the enzymes that appear as holes in the pattern organism, and their fasta sequences are extracted. Pair-wise alignment is done between the fasta sequences for pattern and text.

At second level, fasta sequences for the left and right neighbors in the pattern organism are extracted. Pair-wise alignment is conducted between text sequence for the hole and the left and right neighboring enzymes in the pattern pathway.

At level three, fasta sequences for the group neighbors (EC notation d.d.d.x) for the pattern organism are extracted. Pair-wise alignment of the group neighbors is done with text sequence and significant alignments are reported as alternative pathways.

An alignment score of greater than 50% is considered to be significant. Prosite matching is done between the text and pattern enzyme if the alignment score is in the range of 50%. Proteins showing highest alignment and prosite matching score are selected as candidate fillings.

Finally, statistics are collected to see the number fillings found at each level. The output for filling pathway holes and statistics for the batch are reported in two separate text files.

Examples

We distinguish two types of pathway holes identified as a result of pathway mapping.

A hole representing an enzyme with partially or completely unknown EC notation (e.g., EC 1.2.4.- or -.-.-.-) in the currently available pathway description. This type of holes is caused by ambiguity in identifying a gene and its product in an organism Example: Filling pathway hole in the mapping of glutamate degradation VII pathway in B.subtilis to glutamate degradation VII pathway in T.thermophilus.

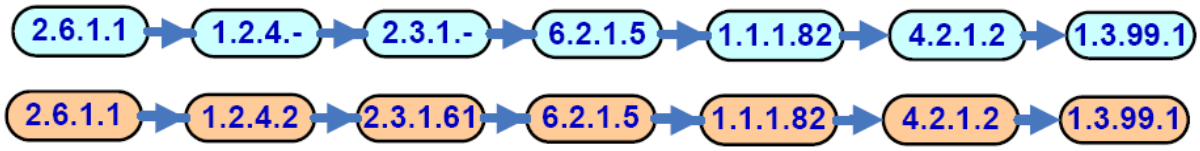

Figure 3: Mapping of glutamate degradation VII pathways from B.subtilis to T.thermophilus

The pattern in the above example contains two pathway holes (the corresponding enzymes are shaded). The mapping results indicate that similar corresponding enzymes 2.3.1.61 and 1.2.4.2 with similar functions can be found in *T.thermophilus*. The proposed tool queries the Swiss-Prot and TrEMBL databases to see if enzyme 2.3.1.61 and 1.2.4.2 have been reported for *B. subtilis*, and also does a pair-wise sequence alignment for all the pattern accession numbers reported, with text accession numbers. We found that these two enzymes have been reported in Swiss-Prot database for *B.subtilis* as P16263 and P23129 respectively, and their sequence homology was found to be greater than 50%. Therefore we recommend filling these pathway holes with enzymes 2.3.1.61 and 1.2.4.2.

1) A hole representing an enzyme that is completely missing from the currently available pathway description. This type of holes occurs when the gene encoding an enzyme is not identified in an organism's genome. Figure 4.

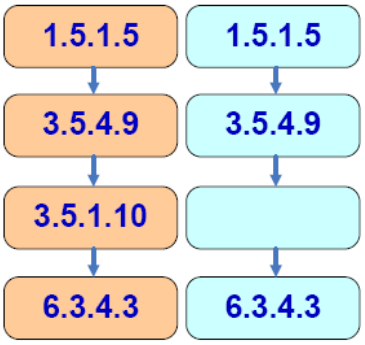

Figure 4: Mapping of formaldehyde oxidation V pathway in B.subtilis to formy1THF biosynthesis pathway in E.coli

The proposed framework to fill pathway holes makes full use of EC encoding which classifies enzymes on the basis of their reaction, and amino acid sequence similarity, which reflects their common origin. Example: Mapping of formaldehyde oxidation V pathway in *B.subtilis* to formy1THF biosynthesis pathway in E. coli.

In this case the enzyme 3.5.1.10 is present between 3.5.4.9 and 6.3.4.3 in E. coli, but absent in the pathway description for B.subtilis. The Swiss-Prot and TrEMBL database search shows that this enzyme is completely missing from B. subtilis and therefore this hole does not allow an easy fix. There is a possibility that this enzyme has not yet been included in the database but has been already identified. We then search for this enzyme in closely related organisms. The Swiss-Prot database search returns the accession number Q5WE95 in *B.clausii, which* is very close to *B. subtilis*. The amino-acid sequence for Q5WE95 is then aligned with P37051 and P0A440, the accession number for enzyme 3.5.1.10 in *E.coli*. The sequence alignment was found to be greater than 50%. Therefore we recommend filling this pathway hole with enzymes 3.5.1.10.

If such a search would not return any hits in close relatives, then we would investigate if the function of this enzyme has been taken up by one of the adjoining enzymes (in this case 3.5.4.9 or 6.3.4.3). Alternatively, there may be another pathway existing for this function which can be verified by searching the group neighbors i.e. by finding sequence homology in 3.5.1.6, 3.5.1.7, 3.5.1.8…3.5.1.15 etc.

VI. IMPLEMENTATION OF THE FRAMEWORK

Requirement Analysis

- Gather mapping data for the four organisms: *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Halobacterium sp.*
- Build mapping data parsing tool.
- Build Uniprot Client that gets accession numbers for a given EC notation, form beta.uniprot.org
- Build Prosite client that gets prosite ids for a given accession number.
- Build tool to download Fasta file for given accession number.
- Build BlastP client for pairwise alignment of fasta files.
- Build filling finder.
- Build tools to gather statistics.

Tools Used

Network mapping tool designed by Cheng Q., Harrison R., Zelikovsky A. [2] has been used for pathway mapping. Swiss-Prot and TrEMBL databases at http://beta.uniprot.org are used for searching functionally similar enzymes and

extracting their fasta sequences. Blastp tool using Blosum62 matrix has been used for protein sequence alignment. Keyword search for links to literature is done using www.pubmedcentral.nih.gov. The framework has been tested on metabolic network data drawn from BioCyc [8], EcoCyc, a model organism database for *E.coli* [9], and MetaCyc, a collection of metabolic pathways and enzymes from a variety of organisms, primarily microorganisms [9].

The components for database query, and analysis have been developed using java version 1.6, so that they are compatible with both Windows and Unix based operating systems. The online database at http://beta.uniprot.org serves xhtml and xml pages, therefore javax.xml.xpath libraries have been used for database query.

Class Diagram

Table 1: Class Diagram for Mapping and Filling Pathways

FillPathWayHoles Fields ➢ patternPathway ➢ patternSps ➢ textPathway ➢ textsps ➢ queryString Conuctructors ➢ FillPathwayHoles(String) Methods ➢ Main()	**Uniprot Client** Fields ➢ uniprot_URL ➢ URL_Encoding ➢ xPath_Query Methods ➢ query ➢ searchUniprot
BlasrPClient Fields ➢ patternFastaFileName ➢ textFastsFileName Constructors ➢ BlastPClient(String,String) Methods ➢ alignSequences ➢ getScore	**GetProsite** Fields ➢ uniprot_URL_Prefix ➢ uniprot_URL_Suffix ➢ uRL_Encoding ➢ xPath_Query Methods ➢ query ➢ SearchUniprot

DownLoadFasta Fields ➢ accessionNumber ➢ fastaFileName Conuctructors ➢ DownLoadFasta(String) Methods ➢ getFasta()	**Filling Finder** Fields ➢ scoreNFileName Conuctructors ➢ FillingFinder(String) Methods ➢ reverseSort ➢ getFillings
PairWiseAlignment Fields ➢ pAccessionNumbers ➢ tAccessionNumbers Conuctructors ➢ PairWiseAlignment(String, String) Methods ➢ alignAll()	**GroupNeighborFinder** Fields ➢ scoreNfileName Constructors ➢ GroupNeighborFinder(String, String, String, int) Methods ➢ reverseSort ➢ getFillings
GetStatistics Fields ➢ inputFileName ➢ outputFileName ➢ Read Conuctructors ➢ GetStatistics(String) Methods ➢ getStatistics ➢ main()	**BlastP executable used** Bl2seq.exe

VII. MAPPING EXPERIMENTS

Four organisms selected to perform pathway mapping cover the biological diversity of all three domains of life: Archea, Bacteria and Eukaryota. Archea and Bacteria belong to prokaryotes group of cell types, while yeast belongs to the eukaryotic domain. *E.coli* being the most studied organism has been compared with three other organisms. The four organisms are:

- *Escherichia coli (Bacteria, Prokaryota)*
- *Bacillus subtilis (Bacteria, Prokaryota)*
- *Saccharomyces cerevisiae (Yeast, Eukaryota)*
- *Halobacterium sp (Archea, Prokaryota)*

While different enzymes catalyze different reactions in a metabolic pathway, some enzymes participate in similar reactions in several different pathways in the same organism [2]. Therefore, mapping experiments performed for the purpose of identifying and filling pathway holes are:

- *Escherichia coli* into *Saccharomyces cerevisiae, Bacillus subtilis*
- *Bacillus subtilis* into *Saccharomyces cerevisiae, Escherichia coli*
- Self mapping of *Escherichia coli*
- Self mapping of *Saccharomyces cerevisiae*
- Self mapping of *Bacillus subtilis*
- *Halobacterium* sp into *E.coli*, *Saccharomyces cerevisiae, Bacillus subtilis*
- *E.coli*, *Saccharomyces cerevisiae, Bacillus* substilis into *Halobacterium sp.*
- Self mapping of *Halobacterium sp.*

VIII. VALIDATION OF RESULTS

Cross Validation

A node in the pattern pathway was deleted. The framework was applied on this data to see :

1. The artificially created pathway hole is identified
2. The filling candidate suggested by the framework is the same as the one deleted.

The framework correctly identified the holes and suggested correct enzymes as fillings.

Algorithm for percentile calculation

Fillings found within the same EC group, i.e., with same EC number in first three positions and different in the fourth position, were compared to all the other pattern text alignments in that group. We need to decide if the enzyme within the corresponding EC group having the best match (the maximum similarity) with the corresponding text enzyme is sufficiently good candidate for filling. Our decision is based on the alignment score. If this score is sufficiently high then we report the candidate, otherwise, we would not be confident in such candidate. We say that the alignment score is high if it is within 25% of best alignment scores for proteins in the same EC group.

Fillings found within the same EC group were evaluated to see where they stand, compared to all the other pattern text alignments in that group. All enzyme EC

notations for the pattern and text were downloaded from Swisprot database. A subset of all enzymes having Text enzymes minus Pattern enzymes was created. Another subset of all enzymes in Pattern minus text enzymes was created. The enzymes in these two subsets were then aligned within the same EC group (x.y.z.any) and the alignment scores recorded in a table. Now the alignment score of the candidate enzyme was compared to see where it stands in this table. Following algorithm was used to calculate a percentile score:

1. Let P be all the EC numbers for pattern (sorted in descending order).
2. Let T be all the EC numbers for text (sorted in descending order).
3. For each enzyme in T-P find the maximum similarity score S with proteins X.Y.Z.any in P-T
4. R=sorted list of all S in decreasing order (as shown in Table 1 column 2).
5. Get the alignment score for the filling alignment and its order O in R. O denotes the number of alignments with same similarity score or higher (as shown in Table 1 column 1).
6. Percentile=O/total number of alignment scores in R (Example shown in Table 1).

Table 2: Showing Raw Score calculation for pattern Bacillus subtilis and Text Escherichia coli.

Order	A-Result
→ 1.	1.00
2.	0.91
3.	0.90
4.	0.88
5.	0.87
6.	0.85
7.	0.83
8.	0.81
→ 9.	0.80
10.	0.78
11.	0.77
12.	0.76
13.	0.75
14.	0.73
15.	0.72
16.	0.71
17.	0.70
18.	0.69
19.	0.68
20.	0.66
21.	0.64
22.	0.63
23.	0.62
24.	0.61
25.	0.60
26.	0.59
27.	0.58
28.	0.57
29.	0.56
30.	0.55
31.	0.54
32.	0.53
33.	0.52
34.	0.51
35.	0.50

1/35=0.02

IX. RESULTS AND DISCUSSION

The metabolic pathway data in this study was gathered from BioCyc [8], EcoCyc, a model organism database for *E.coli* [9], and MetaCyc, a collection of metabolic pathways and enzymes from a variety of organisms, primarily microorganisms [9]. We selected four organisms *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae* and *Halobacterium species*. *E.coli* and *B.subtillis* belong to prokaryotic group of organisms, and to Domain Bacteria. *Saccharomyces cerevisiae* belongs to eukaryotic group of organisms and Fungi Kingdom. It is the most studied

eukaryote. *Halobacterium.sp* also belongs to prokaryotic group of organisms, and the domain is Archaea.

We mapped metabolic 105 pathways from *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae* and *Halobacterium sp*. Twenty five pathway holes and ambiguities were found using network mapping. These holes and ambiguities were processed using the data mining framework described. We found eleven fillings where the enzymes existed in the pattern organism with the same EC notation as in text organism. These pathway holes are therefore mistakes in the pathway data that needs to be fixed. Five fillings were found with left or right neighboring enzymes in the pathway. These fillings show that the left or right neighbors in the pathway have taken multiple roles. Nine fillings were found with the enzymes that have same numbers in the first three positions in the EC notations and different number in the fourth position in the EC notation. These fillings also suggest that some closely related enzymes have taken up multiple roles. These fillings can be used as the candidate enzymes to conduct wet lab experiments to find out their actual functions. From the percentile calculations it was found that we get better results by mapping of closely related organisms like *B.subtillis and E.coli*, both of which belong to domain Bacteria, and percentile values falling in the top quarter. The percentile scores between *Halobacterium - E.coli and Halobacterium - B.subtillis* are also good as they both belong to prokaryotic group of organisms. The percentile scores between Halobacterium and Yeast, which belong to two different groups of organisms, prokaryota and eukaryota respectively, fall in the 3rd and 4th quartile. Tables 3 to 12 show the fillings found using this framework.

Table 3: Pattern: Bacillus subtilis. Text: Escherichia coli. Fillings found with same EC number

N	P: Pattern Pathway T: Text Pathway Mapping of Pattern Enzymes with Text Enzymes	Missing Enzyme in Pattern	Similarity Score & Acc Num, EC Num
1	P: Gamma glutamyl cycle T: Superpathway of glycolysis pyruvate dehydrogenase TCA and glyoxylate bypass P: 2.3.2.2; _____; 2.3.2.4 T: 2.3.3.9; 1.1.1.37; 2.3.3.1	EC 1.1.1.37	0.51 P49814 EC 1.1.1.37
2	P: Acetyl CoA fermentation to butyrate T: Superpathway of glycolysis pyruvate dehydrogenase TCA and glyoxylate bypass P: 1.1.1.35; _____; 4.2.1.55 T: 1.1.1.37; 2.3.3.1; 4.2.1.3	EC 2.3.3.1	0.53 P39120 EC 2.3.3.1
3	P: Alanine biosynthesis I T: Superpathway of lysine threonine methionine and S-adenosyl L-methionine biosynthesis P: 2.6.1.66;_____; _____; _____; _____; _____; _____;_____; 5.1.1.1 T: 2.6.1.1; 2.7.2.4;1.2.1.11; 4.2.1.52;1.3.1.26; 2.3.1.117; 2.6.1.17; 3.5.1.18; 5.1.1.7	EC 2.7.2.4	0.75 Q04795 EC 2.7.2.4
4	P: Alanine biosynthesis I T: Superpathway of lysine threonine methionine and S-adenosyl-L-methionine biosynthesis P: 2.6.1.66;_____; _____; _____; _____; _____; _____; _____; 5.1.1.1 T: 2.6.1.1; 2.7.2.4; 1.2.1.11; 4.2.1.52;1.3.1.26; 2.3.1.117; 2.6.1.17; 3.5.1.18; 5.1.1.7	EC 4.2.1.52	0.66 Q04796 EC 4.2.1.52

Table 4: Pattern: Bacillus subtilis. Text: Escherichia coli. Fillings found with Left/Right Neighbors.

N	P: Pattern Pathway T: Text Pathway Mapping of Pattern Enzymes with Text Enzymes	Missing Enzyme in Pattern	Similarity Score Acc Num & Acc Num	Percentile
1	P: Alanine biosynthesis I T: Superpathway of lysine threonine methionine and S-adenosyl-L-methionine biosynthesis P: 2.6.1.66; _____; _____; _____; _____; _____; _____; _____; 5.1.1.1 T: 2.6.1.1; 2.7.2.4 ; 1.2.1.11; 4.2.1.52; 1.3.1.26; 2.3.1.117; 2.6.1.17; 3.5.1.18; 5.1.1.7	EC 3.5.1.18	1.0 P37112 EC 3.5.1.14	0.02
2	P: Alanine biosynthesis I T: Superpathway of lysine threonine methionine and S-adenosyl-L-methionine biosynthesis P: 2.6.1.66; _____; _____; _____; _____; _____; _____; _____; 5.1.1.1 T: 2.6.1.1; 2.7.2.4; 1.2.1.11; 4.2.1.52; 1.3.1.26; 2.3.1.117; 2.6.1.17; 3.5.1.18; 5.1.1.7	EC 2.6.1.17	0.8 P39754 EC 2.6.1.16	0.26

Table 5: Pattern: Bacillus subtilis. Text: Escherichia coli. Fillings found with group neighbors (same EC number in first three positions, different in the fourth position) Raw Score calculation shown in Table 1.

N	P: Pattern Pathway T: Text Pathway Mapping of Pattern Enzymes with Text Enzymes	Missing Enzyme in Pattern	Similarity Score, Acc Num & EC Num
1	P: Alanine biosynthesis I T: Superpathway of lysine threonine methionine and S-adenosyl-L-methionine biosynthesis P: 2.6.1.66; _____; _____; _____; _____; _____; _____; _____; 5.1.1.1 T: 2.6.1.1; 2.7.2.4; 1.2.1.11; 4.2.1.52; 1.3.1.26; 2.3.1.117; 2.6.1.17; 3.5.1.18; 5.1.1.7	EC 1.2.1.11	0.7 P10725 EC 5.1.1.1
2	P: Alanine biosynthesis I T: Superpathway of lysine threonine methionine and S-adenosyl-L-methionine biosynthesis P: 2.6.1.66; _____; _____; _____; _____; _____; _____; _____; 5.1.1.1 T: 2.6.1.1; 2.7.2.4; 1.2.1.11; 4.2.1.52; 1.3.1.26; 2.3.1.117; 2.6.1.17; 3.5.1.18; 5.1.1.7	EC 1.3.1.26	0.53 P10725 EC 5.1.1.1
3	P: Alanine biosynthesis I T: Superpathway of lysine threonine methionine and S-adenosyl-L-methionine biosynthesis P: 2.6.1.66; _____; _____; _____; _____; _____; _____; _____; 5.1.1.1 T: 2.6.1.1; 2.7.2.4; 1.2.1.11; 4.2.1.52; 1.3.1.26; 2.3.1.117; 2.6.1.17; 3.5.1.18; 5.1.1.7	EC 2.3.1.117	0.85 P10725 EC 5.1.1.1

Table 6: Pattern: Saccharomyces cerevisiae. Text: Bacillus subtilis. Fillings found with same EC number

N	P: Pattern Pathway T: Text Pathway Mapping of Pattern Enzymes with Text Enzymes	Missing Enzyme in Pattern	Similarity Score, Acc Num, EC Num
1	P: Lactate oxidation T: Sorbitol fermentation to lactate -formate- ethanol and acetate P: 1.2.4.1; _____; _____; _____; _____; 2.3.1.12 T: 1.2.1.12; 2.7.2.3; 5.4.2.1; 4.2.1.11; 2.7.1.40; 2.3.1.54	EC 2.7.1.40	0.68 P16387 EC 1.2.4.1
2	P: Lactate oxidation T: Sorbitol fermentation to lactate -formate- ethanol and acetate P: 1.1.1.27; _____; _____; 1.2.4.12 T: 1.1.1.140; 2.7.1.11; 4.1.2.13; 1.2.1.12	EC 4.1.2.13	0.85 P32473 EC 1.2.4.1

Table 7: Pattern: Saccharomyces cerevisiae. Text: Bacillus subtilis. Fillings found with Left/Right Neighbors.

N	P: Pattern Pathway T: Text Pathway Mapping of Pattern Enzymes with Text Enzymes	Missing Enzyme in Pattern	Similarity Score & Acc Num
1	P: Arginine degradation VIII T: Interconversion of arginine-ornithine and proline (Stickland_reaction) P: 3.5.3.1; _____; 2.6.1.13 T: 3.5.3.6; 2.1.3.3; 2.6.1.13	EC 2.1.3.3	0.58 P05150
2	P: Lactate oxidation T: Sorbitol fermentation to lactate -formate- ethanol and acetate P: 1.2.4.1; _____; _____; _____; _____; 2.3.1.12 T: 1.2.1.12; 2.7.2.3; 5.4.2.1; 4.2.1.11; 2.7.1.40; 2.3.1.54	EC 2.7.2.3	0.66 P00560
3	P: Lactate oxidation T: Sorbitol fermentation to lactate -formate- ethanol and acetate P: 1.2.4.1; _____; _____; _____; _____; 2.3.1.12 T: 1.2.1.12; 2.7.2.3; 5.4.2.1; 4.2.1.11; 2.7.1.40; 2.3.1.54	EC 5.4.2.1	0.8 Q12326
4	P: Lactate oxidation T: Sorbitol fermentation to lactate -formate- ethanol and acetate P: 1.2.4.1; _____; _____; _____; _____; .3.1.12 T: 1.2.1.12; 2.7.2.3; 5.4.2.1; 4.2.1.11; 2.7.1.40; 2.3.1.54	EC 4.2.1.11	0.69 P00924
5	P: Lactate oxidation T: Sorbitol fermentation to lactate -formate- ethanol and acetate P: 1.1.1.27; _____; _____; 1.2.4.1 T: 1.1.1.140; 2.7.1.11; 4.1.2.13; 1.2.1.12	EC 2.7.1.11	0.69 P16861

Table 8: Pattern: Halobacterium sp. Text: Saccharomycs.cerevisiae. Fillings found with Group Neighbors (same Number in first three positions, different in fourth position)

N	P: Pattern Pathway T: Text Pathway Mapping of Pattern Enzymes with Text Enzymes	Missing Enzyme in Pattern	Similarity Score, Acc Num, EC Num	Percentile
1	P: Arginine_biosynthesis I T:De novo biosynthesis of purine nucleotidesII P: 2.1.3.3; _____; 6.3.4.5 T: 2.1.2.3; 3.5.4.10; 6.3.4.4	EC 3.5.4.10	0.63 Q9HPY4 EC 1.5.1.5 EC 3.5.4.9 (Bi functional protein)	0.61
2	P: Arginine_biosynthesis I T: Purine nucleotides(i)de novo(-i) biosynthesis II P: 6.3.5.5; _____; _____; _____; _____; 2.1.3.3 T: 6.3.5.3; 6.3.3.1; 4.1.1.21; 6.3.2.6; 4.3.2.2; 2.1.2.3	EC 4.1.1.21	0.57 Q5V1B4 EC 4.1.1.25	0.77

Table 9: Pattern: Halobacterium sp. Text: Escherichia coli. Fillings found with Group Neighbors (same Number in first three positions, different in fourth position)

N	P: Pattern Pathway T: Text Pathway Mapping of Pattern Enzymes with Text Enzymes	Missing Enzyme in Pattern	Similarity Score, Acc Num, EC Num	Percentile
1	P: aspartate_degradation I T: Superpathway of lysine-threonine-methionine- and S-adenosyl L-methionine biosynthesis P: 2.6.1.1; _____; _____; 1.1.1.37 T; 2.6.1.1; 2.7.2.4; 1.2.1.11; 1.1.1.3	EC 2.7.2.4	0.75 Q48295 EC 2.7.2.2	0.34

Table 10: Pattern: Halobacterium sp. Text: Bacillus subtilis. Fillings found with Group Neighbors (same number in first three positions, different in fourth position)

N	P: Pattern Pathway T: Text Pathway Mapping of Pattern Enzymes with Text Enzymes	Missing Enzyme in Pattern	Similarity Score, Acc Num, EC Num	Percentile
1	P: Arginine biosynthesis I T: Purine nucleotides (i) de_novo (-i) biosynthesis II P: 2.1.3.3; _____; 6.3.4.5 T: 2.1.2.3.; 3.5.4.10; 6.3.4.4	EC 3.5.4.10	1.0 Q9HPY4 EC 1.5.1.5 EC 3.5.4.9 (Bifunctional protein)	0.03
2	P: Arginine biosynthesis I T: Purine nucleotides (i) de_novo (-i) biosynthesis II P: 6.3.5.5; _____; _____; _____; _____; 2.1.3.3 T: 6.3.5.3; 6.3.3.1; 4.1.1.21; 6.3.2.6; 4.3.2.2; 2.1.2.3	EC 4.3.2.2	0.85 Q9HMQ3 EC 4.3.2.1	0.19
3	P: Arginine biosynthesis I T: Purine nucleotides (i) de_novo (-i) biosynthesis II P: 6.3.5.5; _____; _____; _____; _____; 2.1.3.3 T: 6.3.5.3; 6.3.3.1; 4.1.1.21; 6.3.2.6; 4.3.2.2; 2.1.2.3	EC 4.1.1.21	0.8 Q9HSA3 EC 4.1.1.25	0.28

Table 11: Pattern: Escherichia coli. Text: Saccharomyces cerevisiae. Fillings found with same EC number.

N	P: Pattern Pathway T: Text Pathway Mapping of Pattern Enzymes with Text Enzymes	Missing Enzyme in Pattern	Similarity Score, Acc Num, EC Num
1	P: Galactonate degradation T: Gluconeogenesis P: 4.2.1.6; _____; 2.7.1.58 T: 4.2.1.11; 5.4.2.1; 2.7.2.3	EC 5.4.2.1	0.75 Q8XDE9
2	P: Galactonate degradation T: Gluconeogenesis P: 2.7.1.58; _____; 4.1.2.21 T: 2.7.2.3; 1.2.1.12; 4.1.2.13	EC 1.2.1.12	0.80 P0A9B4

Table 12: Pattern: Escherichia coli. Text: Saccharomyces cerevisiae. Fillings found with group neighbors (same EC number in first three positions, different in the fourth position)

N	P: Pattern Pathway T: Text Pathway Mapping of Pattern Enzymes with Text Enzymes	Missing Enzyme in Pattern	Similarity Score, Acc Num, EC Num	Precentile
1	P: Galactitol degradation T: Mevalonate_pathway P: 2.7.1.144; _____; 44.1.2.40 T: 2.7.1.36; 2.7.4.2; 4.1.1.33	EC 2.7.4.2	1.00 Q8FF53 EC 2.7.4.6	0.03

X. SUMMARY

Bioinformatics is a data-intensive field of research and development. Data mining is a way to find the knowledge that we have lost in this big pool of information and the wisdom that we can find in this knowledge. Data mining is a method of reducing this complex bioinformatics data into more meaningful and useful patterns and relationships in data. Pathway holes identified by the network-mapping tool used for comparing biological pathways can be fixed by filling in appropriate enzymes or alternative pathways using this data mining framework. Pathway holes are caused if a gene to encode an enzyme, that is needed to catalyze a reaction in a pathway, is not identified in an organism's genome. Specific function of a protein may not be identified during annotation. Reactions catalyzed by those proteins, may also result in pathway holes [3][4]. With further research, some of those proteins have obtained specific annotations; therefore pathway descriptions should be updated in the Pathway/Genome databases. The pathway holes need to be validated, to see if they are real pathway holes or errors in pathway description. The framework designed for filling pathway holes will leverage the network-mapping tool to validate the pathway holes based on reaction properties of enzymes and

protein sequence homology and are therefore capable of updating pathway/genome databases. The fillings suggested by this framework can be used as candidate enzymes to conduct wet lab experiments to find their actual functions.

XI. BIBLIOGRAPHY

[1] Rediscovering Biology, Online Text Book, Unit 2, Proteins and Proteomics. http://www.learner.org/channel/courses/biology/textbook/proteo/proteo_12.html (accessed August, 2007).

[2] Cheng Q., Harrison R., Zelikovsky A., Homomorphisms of Multisource Trees into Networks with Applications to Metabolic Pathways, IEEE 7th International Symposium on BioInformatics and BioEngineering (BIBE 07), Boston, MA, 2007.

[3] Green M.L. and Karp P.D.*, Using genome-context data to identify specific types of functional associations in pathway/genome databases Bioinformatics Research Group, SRI International, Menlo Park, CA 94025, USA, July 2007, 23(13):i205-11, PMID: 17646298.

[4] Green M.L. and Karp P.D., A bayesian method for identifying missing enzymes in predicted metabolic pathway databases, BMC Bioinformatics, Sep. 2004.

[5] Kharchenko P., Chen L., Freund Y., Vitkup D., Church G.M., Identifying metabolic enzymes with multiple types of association evidence, BMC Bioinformatics, March 2006.

[6] Kharchenko P., Vitkup D., Church G.M. Filling gaps in a metabolic network using expression information. Bioinformatics. August 2004, Suppl 1:i178-85.

[7] Abascal F., Valencia A. Automatic annotation of protein function based on family identification. Proteins. November 2003, 53(3):683-92.

[8] BioCyc, Database collection. http://www.biocyc.org (accessed August 2007).

[9] Keseler I.M., Collado V.J., Gama C.S., Ingraham J., Paley. S., Paulsen I.T., Peralta-Gil M., and Karp P.D. Ecocyc: a comprehensive database resource for Escherichia coli, NucleicAcids Research, 2006, 33(1):D334-337.

[10] Krieger C. J., Zhang. P., Mueller L. A., Wang A., Paley S., Arnaud M., Pick J., Rhee S., and Karp P.D. Metacyc: A multiorganism database of metabolic pathways and enzymes, Nucleic Acids Research, 2006, 32(1):D438-42.

[11] Maillet I., Berndt P. Malo C., Rodriguez S., Brunisholz R.A., Pragai Z., Arnold S., Langen H., Wyss M. From the genome sequence to the proteome and back: evaluation of E. coli genome annotation with a 2-D gel-based proteomics approach. Proteomics. 2007 Apr;7(7):1097-106 DSM.

[12] http://www.chem.admu.edu.ph/~nina/rosby/ecnumber.htm (accessed August 2007).

[13] Trupti Joshi and Dong Xu, and Christopher S., Quantitative assessment of relationship between sequence similarity and function similarity, Bond Life Sciences Center, University of Missouri, Columbia, BMC Genomics. 2007; 8: 222. doi: 10.1186/1471-2164-8-222.

[14] Jose B Pereira-Leal, Emmanuel D Levy, and Sarah A Teichmann, The origins and evolution of functional modules: lessons from protein complexes, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK Philos Trans R Soc Lond B Biol Sci. 2006 March 29; 361(1467): 507–517

[15] Gattiker A., Gasteiger E. and Bairoch A.;ScanProsite: a reference implementation of a PROSITE scanning tool; Applied Bioinformatics 1:107-108(2002)

[16] The UniProt Consortium; **The Universal Protein Resource (UniProt);** Nucleic Acids Res. 35:D193-D197(2007).

[17] http://www.ncbi.nlm.nih.gov/blast/html/BLASThomehelp.html#NTBLAST; (accessed August 2007 to download Blastp);