International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 5 Issue 8|| August 2016 || PP. 20-24

A new algorithm for Predicting Metabolic Pathways

Essam Al Daoud
(Computer Science/ Zarqa University, Jordan)

ABSTRACT*:* *The reconstruction of the metabolic network of an organism based on its genome sequence is a key challenge in systems biology. The aim of the work described here is to develop a new algorithm to predict pathway classes and individual pathways for a previously unknown query molecule. The main idea is to use a dense graph, where the compounds are represented as vertices and the enzymes are represented as edges, the weights are assigned to the edges according to the previous known pathways. The shortest path algorithm is applied for each missing enzyme in a pathway. A pathway is considered belong to an organism if the total cost between the initial and final compound is higher than a threshold. Validation experiments show that the suggested algorithm is capable to classify more than 90% of pathways correctly.*

Keywords: *Classification, compound, dense graph, enzyme, pathway.*

I. INTRODUCTION

Metabolic Pathway is a sequence of enzymatic reactions that begins with initial substrate, progresses through intermediates and ends with a final product. Catalysts reactions occur 1,000,000 times faster with enzymes. Approximately 2,700 different enzymes are found in human body. These enzymes can combine with coenzymes to form nearly one hundred thousand various chemicals that help us to see, hear, feel, move digest food, and think. Every organ, tissue and all the one hundred trillion cells in our body depend upon the reaction of enzymes and their energy factor. An enzyme brings reactants together by binding to them, without enzymes, collisions are random. Enzyme name ends with suffix "*-ase*", e.g., *glucose phosphorylase* is an enzyme that adds a phosphate to glucose. The enzymes are not part of reaction, not changed or affected by reaction and used over and over. Enzymes are responsible for digestion, absorption, transporting, metabolizing, and eliminating the waste from forty-five known nutrients: Carbohydrate lipids (fats), protein, water, 9 amino
The pathway prediction algorithm predicts pathways in a sequenced genome by recognizing in that genome previously known path-ways from the KEGG or/and MetaCyc database [1]. Pathways are recognized based on the enzymes present in the genome. It assumes that the genome has already been annotated using one of the many available genome annotation pipelines, such as [2]. Catabolic pathway (catabolism): breaking down of macromolecules, releases energy which may be used to produce ATP. Anabolic pathway (anabolism): building up of macromolecules, requires energy from ATP. Metabolism: the balance of catabolism and anabolism in the body. Conceptually, a metabolic network can be divided into functional pathways. Identifying different metabolic pathways of a species is an important topic in biological research. Any subtle shifts or malfunctions in metabolic pathway may result in diseases. For example, phenylketonuria (PKU) is a metabolic disorder caused by the lack of the enzyme, phenylalanine hydroxylase, which may cause mental retardation in a person. There may also be important metabolic activities that lead to the drug resistance property of pathogenic bacteria. This topic is particularly important for studying new species that have high impact, such as endophytic fungi that can produce fuel and pathogenic bacteria.

II. RELATED WORKS

Patho Logic is a famous tool can be used to predict the metabolic pathways in sequenced and annotated genomes. The reactome inference phase infers the reactions catalyzed by the organism from the set of enzymes present in the annotated genome [1]. The pathway inference phase infers the metabolic pathways present in the organism from the reactions catalyzed by the organism. Both phases draw on the MetaCyc database of metabolic reactions and pathways [2]. To quantitatively validate methods for pathway prediction, Dale et al. developed a large "gold standard" dataset of 5,610 pathway instances known to be present or absent in curated metabolic pathway databases for six organisms. they defined a collection of 123 pathway features, whose information content they evaluated with respect to the gold standard. Feature data were used as input to an extensive collection of machine learning (ML) methods, including naïve Bayes, decision trees, and logistic regression, together with feature selection and ensemble methods. they compared the ML methods to the previous PathoLogic algorithm for pathway prediction using the gold standard dataset. ML-based prediction methods can match the performance of the PathoLogic algorithm [3]. Oyelade et al. extract linear and non linear metabolic pathways from the malaria parasite. The weights are calculated using the metabolite degrees and relevant pathways are obtained using atom mapping information. Adopting the representation of the biochemical metabolic network, lead to accept metabolic network from other source apart from KEGG. This gives us opportunity to compare the metabolic pathways extracted from different metabolic networks [4]. HME3M is another metabolic pathway prediction tool, it first identifies frequently traversed network paths using a Markov mixture model. Then by employing a hierarchical mixture of experts, separate classifiers are built using information specific to each path and combined into an ensemble prediction for the response [5]. Cai and Chou constructed a positive and negative training datasets. The positive set consists of those pairs with each formed by one compound and one enzyme associated with the same reaction. The GO (gene ontology) and microarray data were used to represent the sample of an enzyme, and then the nearest neighbor algorithm is implemented to perform the prediction [6]. Thus, the sample of an enzyme-compound pair can be expressed as a vector with 1540+80+40=1660 dimensions; i. e.,

$$EC=[g_1\, g_2\, ...\, g_{1540}\;\; i_1\;\; i_2\, ...\, i_{80}\;\; c_1\;\; c_2\, ...\, c_{40}]^T$$

CMP Finder is developed by Leung et al. first a weighted directed graph G is constructed where a vertex represents a common compound in the input graphs G_1, G_2,…, G_k and a directed edge (u, v) in G represents a building block producing compound v from compound u. The edge weight is 0 when the building block is an identical block and the weight is 1 when it is a penalty block. Hence, a path in G with total weight g represents a conserved path, i.e. an alignment of a path from each input graph with g penalty blocks, each of which has at most l penalties. Then CMP Finder will discover all conserved path in G from each initial substrate compounds to final product compounds using Floyd-War shall algorithm [7]. Mc Shan et al. present Path Miner, it predicts metabolic routes by reasoning over transformations using chemical and biological information. They build a biochemical state-space using data from known enzyme-catalyzed transformations in Ligand, including, 2917 unique transformations between 3890 different compounds. To predict metabolic pathways they explore this state-space by developing an informed search algorithm. For this purpose they develop a chemically motivated heuristic to guide the search. Since the algorithm does not depend on predefined pathways, it can efficiently identify plausible routes using known biochemical transformations [8]. The total cost for A* algorithm is given by:

$$F(0,m,L)= \sum_{i=1}^{i=m} \left|x^i - x^{i-1}\right| + \left|x^m - x^L\right|$$

Where x^0 is the initial state, x^L is the final state, x^m is an intermediate state.

III. THE NEW ALGORITHM

In this section we explain our new proposed algorithms, Pathway_Finder. The main idea in the suggested algorithm is to use a dense graph, where the compounds are represented by the vertices and the enzymes are represented by the edges. The initial weights for all edges are 1000, but the weights are divided by 10 for each enzyme in a pathway that is found in the training dataset. If the weight is one then the division operation is not implemented. A pathway is considered belong to an organism with high probability if all the enzymes for a given initial and final compound are in the organism and the weights are as low as possible. Whoever, if some enzymes are missing, then a shortest path algorithm between two compounds are applied. Algorithm 1 introduces the Pathway_Finder algorithm.

```
Algorithm 1. Pathway_Finder
Input: the Training dataset, the set of an organism enzymes, initial and final compounds
Output: The pathway and its rank (1-1000)
1-  Constructing the dense graph G:
a.  Initialize all the weights to 1000
b.  For each enzyme and for each pathway
If the weight >1 let weight=weight /10
2-  If all the enzymes for a given initial and final compounds pair in the set of an organism enzymes then
            $cost= \sum weight_i$   where $i=1..\#$ enzymes
3-  For each missing enzyme, find the shortest path in G between the two compounds
        $cost=Not\_Missing\ cost+ Missing\ cost$
        $Not\_missing\ cost = \sum weight_i$   where $i=1..\#$ not missing enzymes
        $Missing\ cost= \sum shortest\ path_i$   where $i=1..\#$missing enzymes
4-  Normalize the cost
        Rank=cost  /  #compounds
```

IV. DATASET

The number of metabolic pathways is very large, reflecting the fact that “life is extremely complicated”. The most important metabolic pathways for humans are glycolysis – glucose oxidation for obtaining ATP, citric acid cycle acetyl-CoA oxidation for obtaining GTP and valuable intermediates, oxidative phosphorylation, pentose phosphate pathways, urea cycle, and fatty acid. MetaCyc is a curated database of experimentally elucidated metabolic pathways from all domains of life. MetaCyc contains 2453 pathways from 2788 different organisms. MetaCyc contains pathways involved in both primary and secondary metabolism, as well as associated metabolites, reactions, enzymes, and genes . it contains two data fields to support pathway inference: the expected taxonomic range of each pathway, and a list of key reactions for pathways. The SRI BioCyc data-base collection contains Pathway/Genome Databases (PGDBs) for 1,004 genomes, and MicroCyc contains 535 genomes. Curated Path-way Tools-based PGDBs are available for Mus musculus [9], Saccharomyces cerevisiae, Arabi-dopsis thaliana, Drosophila melanogaster, Escherichia coli, and Homo sapiens. KEGG pathways is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for: Metabolism, cellular processes, organismal systems, human diseases and drug development. Fig. 1 shows a part from Glycolysis/Gluconeogenesis pathway. Table 1 and 2 summarize some important attributes for the enzyme Hexokinase and the compound D-Glucose-phosphate [10].

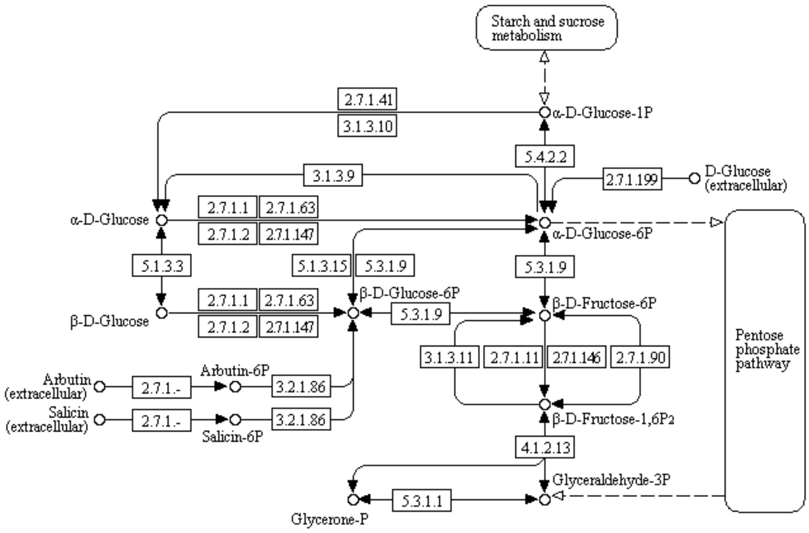

Figure1. Part from Glycolysis / Gluconeogenesis pathway

Table I. Some of important attributes of the Hexokinase enzyme

Name	Hexokinase
Comment	D-Glucose, D-mannose, D-fructose, sorbitol and D-glucosamine can act as acceptors; ITP and dATP can act as donors. The liver isoenzyme has sometimes been called glucokinase.
Pathway	Glycolysis / Gluconeogenesis, Fructose and mannose metabolism, Galactose, Amino sugar and nucleotide, Streptomycin, Butirosin and neomycin biosynthesis, Biosynthesis of antibiotics
Gene	HAS, PTR, PPS, GGO, PON, NLE, MCC, MCF, RRO, CJC

Table II. Some of important attributes of the D-Glucose-phosphate Compound

Name	D-Glucose-phosphate
Formula	C6H13O9P
Pathway	Glycolysis / Gluconeogenesis, Pentose and glucuronate interconversions, Galactose metabolism, Streptomycin biosynthesis, Biosynthesis of plant secondary metabolites, Polyketide sugar unit biosynthesis, Glucagon signaling pathway
Enzyme	2.4.1.1 2.4.1.7 2.4.1.20 2.4.1.30 2.4.1.31 2.4.1.49 2.4.1.97 2.4.1.139 2.4.1.231 2.4.1.321 2.4.1.329 2.4.1.333

In this study, pathways for five species are collected from KEGG and MetaCyc, the number of distinct enzymes and compounds for each species are summarized in Table 3.

Table III. Species and pathways from KEGG and MetaCyc datasets

Species	Pathways	Enzymes	Compounds
Homo Sapiens	110	1512	703
Escherichia Coli	200	2895	1233
Saccharomyces Cerevisiae	50	787	321
Mus Musculus	150	1826	828
Bos Taurus	100	1340	693

V. EXPERIMENTAL RESULTS

To test Pathway_Finder algorithm, two experiments are implemented, such that 80% and 90% from the dataset is used to construct the dense graph and the rest is used to test the accuracy, in addition, false pathways (false initial or final compounds) are used to calculate the sensitivity, specificity and Accuracy, pathways with low ranks are considered negative. Table 4 and 5 summarize the results

$$\text{Sensitivity (SN)= } \frac{TP}{TP + FN}$$

$$\text{Specificity(SP) = } \frac{FP}{TN + FP}$$

$$\text{Accuracy (ACC)= } \frac{TP + TN}{N}$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative.

Table IV. Sensitivity, Specificity and Accuracy using 80% from the dataset for training

Species	Training	Testing	SN	SP	ACC
Homo Sapiens	88	22	90.47	14.28	84.09
Escherichia Coli	160	40	92.10	11.90	90.00
Saccharomyces Cer.	40	10	100	9.09	95.00
Mus Musculus	120	30	89.65	12.90	88.33
Bos Taurus	80	20	94.73	9.52	92.50
Total/average	488	122	92.24	11.90	89.34

Table V. Sensitivity, Specificity and Accuracy using 90% from the dataset for training

Species	Training	Testing	SN	SP	ACC
Homo Sapiens	99	11	83.33	10.00	86.36
Escherichia Coli	180	20	95.00	5.00	95.00
Saccharomyces Cer.	45	5	83.33	0.00	90.00
Mus Musculus	135	15	87.50	7.14	90.00
Bos Taurus	90	10	100	9.00	95.00
Total/average	549	61	90.47	6.77	91.80

VI. CONCLUSION

Placing molecules in the context of known metabolic pathways might aid in understanding their biological function and will shed light on the presence of yet unidentified gene products that may be catalyzing relevant reactions. Thus, A number of databases containing biological pathway information are available. In this study, pathways for five species are collected from KEGG and MetaCyc, the new algorithm exhibits higher accuracy without sacrificing implementation time.

ACKNOWLEDGEMENTS

This research is funded by the Deanship of Scientific Research in Zarqa University /Jordan.

REFERENCES

[1] D. Armenta-Medina, L. Segovia, E. Perez-Rueda, Comparative genomics of nucleotide metabolism: a tour to the past of the three cellular domains of life, *BMC Genomics, 15,* 2014, 800-815.

[2] P. D. Karp, M. Latendresse, R. Caspi, The Pathway Tools Pathway Prediction Algorithm, *Standards in Genomic Sciences, 5,* 2011, 424-429

[3] J. M. Dale, L. Popescu, P. D Karp, Machine learning methods for metabolic pathway prediction, *BMC Bioinformatics, 11 (15),* 2010, 1-14

[4] J. Oyelade, E. Adebiyi, I. Ewejobi, B. Brors and R. Eils, Computational Identification of Signalling Pathways in Plasmodium falciparum, *Infection genetics and evolution journal of molecular epidemiology and evolutionary genetics in infectious diseases - Elsevier, Vol. 11(4)* , 2011, 755-764

[5] T. Hancock, H. Mamitsuka, A markov classification model for metabolic Pathways, *Algorithms for Molecular Biology* 2010, 5:10, pp 1-9

[6] Y. D. Cai, K. C. Chou, Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition, *Biochem Biophys Res Comm,* 305, 2003, 407-411.

[7] H. C. M. Leung, S. Y. Leung, F. Y. L. Chin, S. M. Yiu, C.L. Xiang, Predicting metabolic pathways from metabolic networks with limited biological knowledge, *IEEE International Conference on Bioinformatics and Biomedicine Workshops, Hong Kong, China*, 18 – 21 December 2010, 7-13.

[8] D.C. McShan, S. Rao, I. Shah, PathMiner: predicting metabolic pathways by heuristic search, *bioinformatics, 19 (13)*, 2003, 1692–1698.

[9] F. Ay, M. Kellis, T. Kahveci, SubMAP: aligning metabolic pathways with subnetwork mappings, *J Comput Biol*, *18*, 2011, 219–35.

[10] R. Caspi, T. Altman, R. Billington, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases, *Nucleic Acids Res 42,* 2014, 459–471