Motion Planning In Metabolic Pathways Using Probabilistic Roadmap and A* Algorithms

Angela Makolo[1] and Obotu Ojobo[2]

[1] Department of Computer Science, University of Ibadan
Ibadan, Oyo State, (200284), Nigeria

[2] Department of Computer Science, University of Ibadan
Ibadan, Oyo State, (200284), Nigeria

Abstract

Motion planning and navigation strategies have found useful applications in many areas such as biological networks. This work applies motion planning algorithms to search for biochemically relevant pathways in metabolic pathways. The choice pathways are represented as graphs with its compounds as nodes (vertices), and the possible reactions between the compounds as the edges. The probabilistic roadmap (PRM) algorithm is then used to construct the roadmap (graph) using its local planner function while modelling a group of pool metabolites as obstacles. A* search algorithm queries the roadmap to get the most relevant (cost effective) path using the thermodynamic feasibility of the reactions as the weighting scheme. For ease of testing and evaluation, the system was implemented using python programming language. Choice pathways from KEGG database in KGML format (i.e. xml format for KEGG) were used to test the system, which revealed that the results were consistent with other pathway search tools with reasonable performance and can be adopted for pathfinding problems. Improvements in several areas can however better optimise the system, example include the aspect of weighting schemes utilized.

Keywords: *Metabolic pathway, Metabolite, Organism, Graph, Nodes or Vertices, Edges.*

1. Introduction

Motion planning is an area of robotics that deals with the best and most optimal paths of movement for a robot to achieve its intended goal through a series of tasks. It is also known as the navigation system and can be used for navigation optimization in networks and graphs. The system through which the navigation is to take place is modelled as a graph, G = (V, E) with nodes or vertices, V and edges, E. The ideas of [17] have helped summarise the motion planning concepts adopted in this work. The book records how successful sampling-based motion planning strategies have been in recent years for solving problems from robotics, manufacturing, and biological applications that involve thousands and even millions of geometric primitives. It is on the bases of successful application of sampling-based motion planning techniques to biological systems that this work aims to benefit from by applying same strategies in finding biochemically relevant paths in metabolic pathways. The concept is also referred to as metabolic pathfinding, route search etc. by various researchers and groups.

Metabolic pathways are linked series of steps in biochemical reactions that take place within cells of living organisms to convert molecules into forms that are needed and useful for the wellbeing of that organism. In other words, there are series of actions among molecules (metabolites) in a cell that leads to a certain product or a change in a cell. Metabolic pathways are actually a form of biological networks which can be studied computationally for detailed understanding and analysis of the pathway. Concepts of motion planning has found application in metabolic pathways as earlier suggested because they involve navigation problems due to the interaction between metabolites (compounds in the reaction). Hence this study applies motion planning algorithms to pathfinding problems in metabolic pathways (i.e. search for biochemically relevant paths). As would be seen in more details later, the composition of the metabolic pathways has made sampling based motion planning a suitable technique for best path search within pathways. This is so because; the pathways can be mapped onto graphs, G = (V, E) (they can actually be represented by networks and graphs) with the compound and reactions represented as nodes or vertices, V and edges, E respectively. And for efficiency of the search, a weighted graph would be more preferable in representing the pathways with some specific properties of the constituent compounds and reactions as weights as would be seen much later in the work.

There is the need to constantly optimize graph traversal strategies and algorithms for application in the vast areas of application for which they are feasible. Sampling-based techniques have found usefulness in biological networks but from research, their application in metabolic networks is limited hence the need to optimize them for such networks.

The aim of this work is to develop an optimal model for motion planning in metabolic pathways. The objectives to help achieve the aim are; (i) to determine the best paths for effective distribution of molecules in the pathway using PRM and A* algorithms. (ii) To design a model that reveals the significance of the constituents or components of the pathway and the consequence of their removal or manipulation on the pathway. (iii) To implement the model and evaluate it with existing ones.

Based on the constant need to improve on pathfinding strategies, the feat recorded by motion planning systems and its application in biological networks, utilizing sampling based motion planning using probabilistic roadmap (PRM) and A* search algorithms for optimised path finding in metabolic pathways is necessary. With A* revered for its performance and accuracy and probabilistic roadmap algorithm known to work well for high-dimensional configuration spaces such as metabolic pathways and other biological networks the probability that they will produce a definite robust pathfinding solution is very high.

1.1 Metabolic Pathways Data Sources

A lot of work has been done by researchers to help store biological data for subsequent easy access and analysis. The data generated from biological researches over time have brought about a very vast array of databases of heterogeneous nature. This also holds true for pathway databases (the focus of this work) which are subsets of the aforementioned. "As a consequence, all the data are stored in large and widely distributed databases, with heterogeneous data formats and access mechanisms. In most of the reviewed articles published on biological databases a critical analysis of the data is missing. However, to extract and use data from heterogeneous data sources it is essential for a user to have information about the origin of the data, updates, redundancies and reliability" [22]. A number of pathway databases exist among which are KEGG, EcoCyc, WIT, pathDB etc. each with its own method of data storage and representation. A closer look will be given to the KEGG and EcoCyc databases since this work mostly utilizes the data and tools they provide.

[15] gives a concise background of the KEGG database and describes KEGG which stands for Kyoto Encyclopedia of Genes and Genomes to be a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information. It is an integrated database resource consisting of eighteen databases broadly categorized into systems information, genomic information, chemical information and health information, which are distinguished by colour coding of web pages. Within these categories consists of sub-categories such as pathway maps, reactions, etc. from where this work extracts data for testing. The details of the database have been encoded for easy computer representation as will be much explained later under the methodology section 3.4: Data Extraction and Preparation.

EcoCyc is a scientific database for the bacterium Escherichia coli K-12 MG1655. It is a bioinformatics database that describes the genome and the biochemical machinery of Escherichia coli K-12 MG1655. The EcoCyc project performs literature-based curation of the entire genome, and of transcriptional regulation, transporters, and metabolic pathways. It is part of the larger BioCyc collection of Pathway/Genome Databases (PGDBs) [6] [14]. A good number of researches make use of E. coli reference pathways for testing and this work likewise adopts same to effectively check the method's consistency with others. This justifies the exploration of EcoCyc which is also affiliated with PathoLogic, a metabolic route search software for searching for best routes between two compounds (start and target) hence, a valuable evaluation tool. A little more on EcoCyc is given in subsequent sections of this work. Table 1 shows more pathway tools and databases.

Table 1: Pathway tools and types of model-organism databases (MODs). Source: [2].

Tool	**Task**	**Link Accessibility (URL)**
PathFinder	Dynamically represents and provides visualization of biochemical information	http://bibiserv.TechFak.Uni-Bielefeld.DE/pathfinder
PathMiner	Identifies plausible routes using known biochemical transformations	http://pathminer.uchsc.edu/
Pathway Hunter Tool (PHT)	Analyzes the shortest paths and calculates the average shortest paths	http://www.pht.uni-koeln.de
aMAZE	Web interface to the aMAZE relational database	http://www.amaze.ulb.ac.be
Pathway Prediction System (PPS)	Predicts microbial catabolism of organic compounds	http://umbbd.ahc.umn.edu/predict/
KEGG genes	Identifies the link between genomic information in the GENES database	http://www.genome.jp/kegg
MetaRoute	Explores genome-scale metabolic networks	http://www-bs.informatik.uni-tuebingen.de/Services/
KEGG pathway	Provides reference knowledge for pathway mapping	http://www.genome.jp/kegg/
Pathway Tools version 13.0	Allows the user to interrogate and explore relationships within the network	http://www.biocyc.org/download.shtml
Pathway projector	Provides intuitive browser pathway map with the addition of gene and enzyme nodes	http://www.g-language.org/PathwayProjector/

PathPred	Functions as knowledge-based prediction system	http://www.genome.jp/tools/pathpred/
Atom tracking system	Enables pathfinding algorithms to avoid unrealistic connections	http://www.kavrakilab.org/atommetanet
EcoCyc Database	Provides pathway/genome navigator software with the EcoCyc database	http://ecocyc.org/
MetaCyc Database	Provides a uniquely high-quality resource for metabolic pathways and enzymes	http://metacyc.org/

2. Related Works

2.1 Pathfinding Techniques

Pathfinding is a search for the most feasible path from a start to a goal configuration within a graph. The nature of the search is dependent on the nature of the system being modelled as a graph. Usually the path to be found lies within the specified start and target (goal) points on the mapped graph representing the system in question. Some pathfinding methods as reported by various researchers are hereby presented. Work done by [13] reveals that the PathoLogic component of the Pathway Tools software performs prediction of metabolic pathways in sequenced and annotated genomes using two phases: the reactome inference and pathway inference phases. An interaction with the online based PathoLogic software also reveals its capability of finding relevant paths within pathways. [2] worked on finding biochemically relevant metabolic pathways between two given compounds by tracking the movement of atomic groups through metabolic networks and use combined information of reaction thermodynamics and compound similarity to guide the search towards more feasible pathways and better performance. The method performed very with good compound and inclusion accuracies. However it infers pathways across all of the data in KEGG, but for some applications, researchers may be only interested in the metabolic network of a single organism or several related organisms. Combining atomic group tracking with weighted metabolite graph improves the quality of the found pathways. On the other hand, the use of combined information of reaction thermodynamics and compound similarity ensures meaningful pathways are found even without the option of tracking atomic groups. This helps to emphasize the importance of modelling the metabolic pathways as weighted graphs. [3] presented graph-based search algorithm based on atom mapping rules and path weighting schemes that returns relevant or textbook-like routes between a source and a product metabolite within seconds for genome-scale networks. Though the approach is limited by its failure to find routes passing pathways of the core metabolism (like glycolysis or TCA cycle) because frequently occurring compounds (like pyruvate or acetyl-coA) have to be traced, Its speed allows the algorithm to be used interactively through a web interface to visualize relevant routes and local networks for one or multiple organisms based on data from KEGG. The method proposed by [4] utilized atom mapping rule to search for lightest path in a degree-weighted network. The key component of which was a new method of computing optimal atom mapping rule. The method however, needed improvement in that; it should have been possible to integrate further relevant information such as thermodynamic efficiency for better search. [5] performed a study which increased pathfinding accuracy by computing shortest path on weighted metabolic graphs while filtering out the selection of highly connected compounds or nodes, which correspond to pool metabolites and co-factors (e.g. H2O, NADP and H^+). Paths inferred using this approach generally corresponds to biochemically valid pathways. The drawback of the procedure is that it ignores the various segmentation (i.e. the differences in the definition of pathways used in the different databases, such as EcoCyc, KEGG and WIT) of metabolism defined by different databases. It also limits users to the state of pathway annotation as at when the tool was built. Considering that, reaction stoichiometry has been neglected in many graph-based pathfinding approaches, this method proposed by [19] showed that reaction stoichiometry can be incorporated into path-finding approaches via mixed-integer linear programming. This resulted in improved prediction of topological and functional properties in metabolic networks. [20] examined the effectiveness of using compound node connectivity in a path finding approach and also presented a path finding approach based upon integer programming. The approach performed well, when a metabolic path was regarded as being from a source reaction to a target reaction and performed less well when a metabolic path was regarded as being from a source compound to a target compound.

There are other very important factors to be considered in metabolic pathfinding. These factors help determine optimal paths within the pathway and are generally referred to as the parameters or heuristics that guide the search. No single parameter or heuristic is a perfect and universal choice for pathfinding because each has its own strength and / or weakness and situations where they best fit. This is buttressed by the following statement: "These parameters and heuristics capture information on the metabolic network structure, compound structures, reaction features, and organism-specificity of pathways. No one metabolic pathfinding algorithm or search parameter stands out as the best to use broadly for solving the pathfinding problem, as each method and parameter has its own strengths and shortcomings" [16]. They also went further to detail out

the parameters and heuristics which are broadly captured as Metabolic Network Structure, Structure of Compounds, Reactions and Organism. A brief highlight of each is given. **Metabolic Network Structure** consists of;

i. Graph connectivity: stemming from the graph-based representation of the pathway graph-based features and constraints will be used to model the pathways.
ii. Path length: pathways with smallest number of steps are often considered optimal by pathfinding algorithms as they tend to require less manipulation in a metabolic engineering context.

Structure of Compounds;

i. Atom tracking: atoms that are retained from start to target compounds can be tracked using some algorithms
ii. Chemical similarity: compounds with similar chemical structures may be connected by a common reaction. It could be chemical fingerprint or graph based comparison.

Reactions consist of;

i. Reaction rules: reactants and products that undergo same structural change may fall under same reaction rule
ii. Thermodynamics: paths are ranked based on thermodynamic feasibility or reactions.
iii. Stoichiometry: reaction stoichiometry may be used to limit the number of biologically irrelevant pathway in graph-based pathfinding.
iv. Enzyme efficiency and promiscuity: edges between compounds are weighted based on the frequency of enzyme catalysed reactions

Finally, many algorithms give users the opportunity to search for pathways based on choice of organisms.

2.1.1 Graph-Based Metabolic Pathfinding

There are lots of metabolic pathfinding techniques as can be seen from the preceding discussions each with its/their own benefit(s) and/or downside(s). Of interest to this work is the graph based method and as such a closer look at the graph based method will be done. “In graph-theoretic metabolic modelling, the main consideration is the connectivity of metabolism network. The basic abstraction level of metabolic networks can be represented as mathematical graphs, using nodes to represent metabolic components, and edges to represent their various types of inter-actions” [2]. They also added other graph concepts like the concept of weighted graphs where weights are assigned to the edges of the graphs representing the reactions and the shortest path is the set of edges with the least weight depending on the weighting scheme.

Meanwhile a subgraph is represented as a subset of nodes with a specific set of edges connecting them, and the total number of subgraphs while a bipartite graph contains two sets of nodes (substrates and products) that are linked by edges (enzymatic reaction). It was noted however that this model graph only facilitates drug discovery and ranking of choke points and load points, and both points are used to find enzymes (edges), which uniquely consume or produce a particular metabolite (nodes).

A hyper-graph is a generalization of an ordinary graph where an edge, called a hyper-edge, can connect more than two vertices. The vertices in the hyper-graph are the compounds, and the hyper-edges are the reactions connecting the compounds.

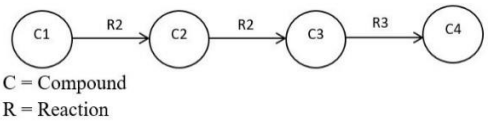

Fig. 1 Simple graph showing compounds and reactions

2.2 Probabilistic Roadmap and A* (A Star) Algorithms

Probabilistic roadmap (PRM) algorithm is a sampling based motion planning algorithm which achieves path planning by building or constructing a roadmap (which is the graph to be searched for optimal path), after which a search algorithm such as A* is used to query the generated graph or roadmap for the most optimal path with respect to system being modelled and the pattern of modelling. Having carried out a comprehensive study on sampling-based motion planning, [7] reached a consensus that sampling-based methods offer an efficient solution for what is otherwise a rather challenging dilemma of path planning and that the methods have been extended further away from basic robot planning into further difficult scenarios and diverse applications. They went further to say search algorithms such as Dijkstra and A* find an optimal solution in a connectivity graph. Having worked extensively on robot motion planning approaches, [21] were able to identify and classify the amount of the existing work for each motion planning approach and determine the percentage of the application of each approach. Of particular interest to this work is the review in probabilistic roadmap planner (PRM) under the sampling-based motion planning approach. According to them, the PRM demonstrated the tremendous potential of the sampling-based methods. PRM fully exploits the fact that it is cheap to check whether or not a single robot configuration is in Q_{free} (i.e. configurations not in obstacle in other words the free configuration space) and then creates a roadmap in Q_{free}. The probabilistic roadmap algorithm has many variants, each with its own merit due to research activity and the need to modify it to best suit the peculiarity of the system for which it is being modified [9].

2.3 Application of Sampling-Based Methods to Biological Networks

Though not much has been seen of the use of motion planning techniques in metabolic pathfinding, a couple of researches have been done on other biological networks using motion planning and in particular sampling-based motion planning. In the review of a few of such researches, the benefits of sampling-based methods were obvious. A few of such systems are presented. [18] presented a novel, sampling-based algorithm which adapts the probabilistic roadmap framework to compute transition paths between functionally-relevant protein states to achieve detailed characterization of the precise yet complex relationship between protein structure, dynamics and function. An activity that challenges both wet and dry laboratories as protein dynamics involves disparate temporal scales. The method leverages known structures to initialize its search and define a reduced conformation space for rapid sampling which helps to address the insufficient sampling issue suffered by sampling-based algorithms, then it embeds samples in a nearest-neighbour graph where transition paths can be efficiently computed via queries. The application and benefits of sampling based method in biological networks is again highlighted as [11] applied probabilistic graphical methods, based on Bayesian network and knowledgebase constraints using gene expression data to reconstruct soybean metabolic pathways. The results show that this method can predict new relationships between genes, improving on traditional reference pathway maps.

2.4 Limitations of Metabolic Pathfinding Techniques

Metabolic pathfinding has a lot of benefits in the biochemical domain however, it is not without its own drawback. [2] highlighted one major drawback of pathfinding approaches as the relative inflexible in terms of adding additional biologically meaningful constraints. The stoichiometric information for a metabolic path for instance must be extracted as a separate stage and noted that being able to add biologically-based constraints as an intrinsic part of the pathfinding process would significantly refine the search for biologically meaningful metabolic paths.

Another limitation of metabolic pathfinding in the graph-theoretical setting is that all metabolites are by default considered equal in importance thereby, allowing shortest paths to be computed through many currency and cofactor metabolites. However the biological relevance of such path is low.

3. Methodology

Using probabilistic roadmap (PRM) algorithm to find the most relevant or feasible path within a metabolic pathway summarily involves the following: configuration space sampling, roadmap building and roadmap querying. After the configuration sampling and roadmap building phases of the PRM the A* algorithm searches the derived roadmap for the most feasible path. The following sections give the details.

3.1 Configuration Space Sampling

Configuration space sampling involves collecting sample structured data from biological databases such as; Kyoto Encyclopedia of Genes and Genomes (KEGG) and plantcyc. These data sets form the configuration space, C for which the roadmap building and querying will be done. The sample data may be in the form of flat files or xml files. This work will adopt the xml file style provided by KEGG database known as KEGG Markup Language or KGML. Details of the data collection and preparation are given below.

The metabolites, for which relevant pathways are to be searched make up the free configuration space, C_{free} while the pool metabolites and intermediate compounds which do not affect the biochemical feasibility of the pathways, make up the configuration space in collision or obstacle, C_{obs}. These categories of metabolites are usually involved in a large number of reactions and form hub nodes when represented in the network.

3.2 Roadmap Building

In roadmap building, a graph G = (V, E) is constructed from the free configuration space, C_{free} by connecting each configuration $c \in C_{free}$ to several of its nearest neighbours. A brief explanation is given; V is the set of vertices or nodes (representing the metabolites) such that $u,v \in V(G)$, E is a set of edges (representing the reactions between metabolites i.e. $uv \in E(G)$). The probabilistic roadmap algorithm used for building the roadmap as given by [9] is presented in Algorithm 1 below;

Algorithm 1: ConstructRoadmap [9]

Let: $V \leftarrow \varnothing$; $E \leftarrow \varnothing$;
1: loop
2: $c \leftarrow$ a (useful) configuration in C_{free}
3: $V \leftarrow V \cup \{c\}$
4: $N_c \leftarrow$ a set of (useful) nodes chosen from V
5: for all $c' \in N_c$, in order of increasing distance from c do
6: if c' and c are not connected in G then
7: if the local planner finds a path between c' and c then
8: add the edge c' c to E

The input to the algorithm is a number of nodes and the output is a roadmap G = (V, E). As explained by [7] A roadmap is built in the learning phase thus;

1. A node, q_{rand}, is selected from the C-space using sample procedure.

2. q_{rand} is discarded, if it is in C_{obs}.
3. Otherwise, q_{rand} is added to the roadmap.
4. Find all nodes within a specific range to q_{rand}
5. Attempt to connect all neighbouring nodes using local planner to q_{rand}.
6. Check for collision and disconnect colliding paths
7. This process is repeated until a certain number of nodes have been sampled.

3.3 Roadmap Querying

At this stage, A* (A star) search algorithm is employed to search through the constructed roadmap for the most relevant or feasible reaction path. This is done by supplying start and target (goal) compounds for a pathway as arguments to the search algorithm which will in turn query the roadmap for the most feasible pathways and returns same. The most feasible pathways here are those with the cumulative lowest cost which in this work is considered to be the thermodynamic feasibility of the reactions that make up a pathway. The A* search algorithm is hereby presented as given by [1] in an online material;

Algorithm 2: A* Algorithm [1]

1: make an openlist containing only the starting node
2: make an empty closed list
3: while (the destination node has not been reached):
4: consider the node with the lowest f score in the open list
5: if (this node is our destination node) :
6: we are finished
7: if not:
8: put the current node in the closed list and look at all of its neighbors
9: for (each neighbor of the current node):
10: if (neighbor has lower g value than current and is in the closed list):
11: replace the neighbor with the new, lower, g value
12: current node is now the neighbor's parent
13: else if (current g value is lower and this neighbor is in the open list):
14: replace the neighbor with the new, lower, g value
15: change the neighbor's parent to our current node
16: else if this neighbor is not in both lists:
17: add it to the open list and set its g

A* expands paths that are already less expensive by using this function: f(n) = g(n) + h(n).
Where;

1. f(n) = total estimated cost of path through node n.
2. g(n) = cost so far to reach node n.
3. h(n) = estimated cost from n to goal. This is the heuristic part of the cost function, so it is like a guess.

However for this work, f(n) = the thermodynamic feasibility of reactions,

The input to A* algorithm is a roadmap, a start goal and an end goal. The output is a low cost pathway.

3.4 Data Extraction and Preparation

Comprehensive transformations of metabolic pathways to various graph types have been done by researchers at different levels. Most of such data are accessible from various pathway tools and others are only referenced in journal articles. As a result, choice pathways are selectively collected for testing purposes from KEGG database in KGML format and transformed into a graph for querying purpose. KEGG database is an integrated database resource consisting of eighteen databases. These databases contain various data objects for computer representation of the biological systems. Thus, the database entry of each database is called the KEGG object, which is identified by the KEGG object identifier consisting of a database-dependent prefix and a five-digit number [15]. Among the databases is the KEGG Reaction database from where in particular, the reactions of the choice test pathways were collected. An example reaction in KEGG Reaction is R00658, an example compound say pyruvate is C00022 and so on. The choice pathways here will be consistent with those used by many researchers for test to ensure that the results of this method are properly evaluated. Having collected the KGML file, two tags name and title, corresponding to the attributes of the pathway tag are then created within the document, preferably just below the pathway tag. This is to enable the system easily capture the names and titles for the pathways being uploaded to aid further search. Figure 2 and 3 below show a sample pathway and KGML extracted from KEGG database respectively.

The A* search uses cumulative path weight of the compounds and reactions of a path to determine the most feasible path. The weighting scheme adopted here as earlier said is the thermodynamic feasibility of reactions in the graph as provided by [10]. Where available, the degree weighting scheme would also be efficient with this method. The degree of a node is number of in-coming and/or outgoing edges of the node in a graph. In other words, it is the level of connectivity of the node.

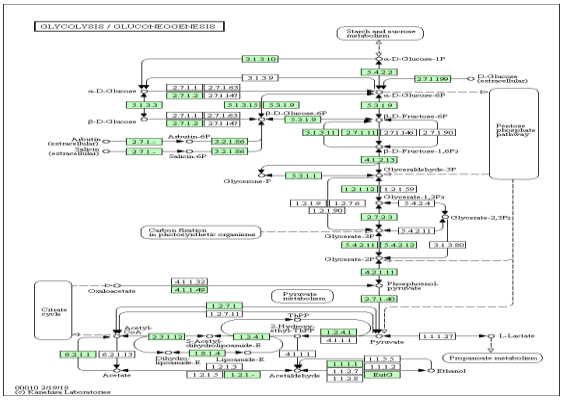

Fig. 2 Glycolysis / Gluconeogenesis pathway map of Escherichia coli K-12 MG1655. Source: http://www.genome.jp/kegg-bin/show_pathway?eco00010

```
    link="http://www.kegg.jp/dbget-bin/www_bget?K18978">
    <graphics name="K18978" fgcolor="#000000" bgcolor="#FFFFFF"
        type="rectangle" x="416" y="540" width="46" height="17"/>
</entry>
<entry id="144" name="ko:K00174 ko:K00175" type="ortholog" reaction="rn:R01196"
    link="http://www.kegg.jp/dbget-bin/www_bget?K00174+K00175">
    <graphics name="K00174..." fgcolor="#000000" bgcolor="#FFFFFF"
        type="rectangle" x="284" y="810" width="46" height="17"/>
</entry>
<entry id="146" name="ko:K00895 ko:K21071" type="ortholog" reaction="rn:R02073"
    link="http://www.kegg.jp/dbget-bin/www_bget?K00895+K21071">
    <graphics name="K00895..." fgcolor="#000000" bgcolor="#FFFFFF"
        type="rectangle" x="558" y="336" width="46" height="17"/>
</entry>
<entry id="147" name="ko:K22473 ko:K22474" type="ortholog" reaction="rn:R09479"
    link="http://www.kegg.jp/dbget-bin/www_bget?K22473+K22474">
    <graphics name="K22473..." fgcolor="#000000" bgcolor="#FFFFFF"
        type="rectangle" x="487" y="928" width="46" height="17"/>
</entry>
<relation entry1="70" entry2="73" type="ECrel">
    <subtype name="compound" value="86"/>
</relation>
<relation entry1="69" entry2="70" type="ECrel">
    <subtype name="compound" value="86"/>
</relation>
<relation entry1="69" entry2="73" type="ECrel">
    <subtype name="compound" value="86"/>
</relation>
<relation entry1="66" entry2="69" type="ECrel">
    <subtype name="compound" value="87"/>
</relation>
<relation entry1="66" entry2="81" type="ECrel">
    <subtype name="compound" value="89"/>
</relation>
<relation entry1="66" entry2="82" type="ECrel">
```

Fig. 3 Sample KGML file. Source: http://www.genome.jp/kegg-bin/show_pathway?eco00010

The entire method is summarized by Figure 4 presented below.

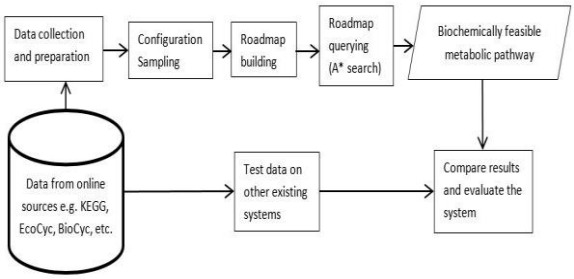

Fig. 4 Diagrammatic representation of the method

4. Implementation

Similar to many pathfinding applications, the method presented in this work has been implemented as a web based tool using python programming language and associated tools for scripting and MySQL database for data storage. It is implemented such that the KGML file is uploaded into a dedicated storage folder on the server side, from where the roadmap is built and uploaded into database for querying in order to get best path.

5. Results

5.1 Data Upload Phase

With the downloaded KGML file prepared as indicated in section 3.4, the file is then uploaded unto to server either through the python command line or through a web browser interface. The figures that follow show screenshots of the pre-upload and post-upload phases.

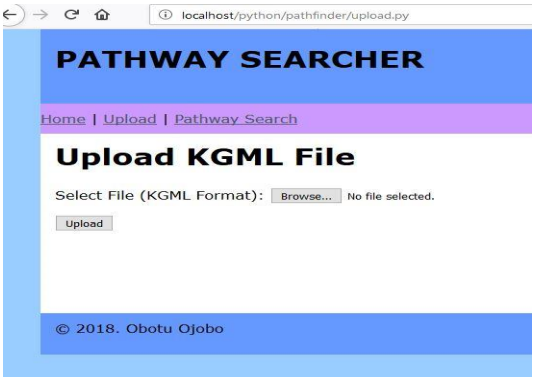

Fig. 5 KGML file pre-upload interface

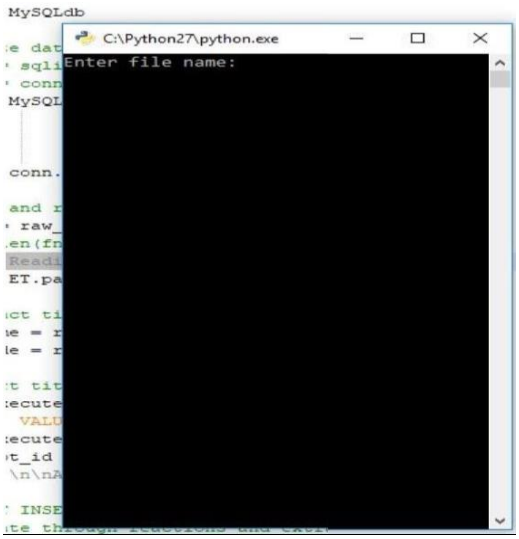

Fig. 6 KGML file pre-upload command line

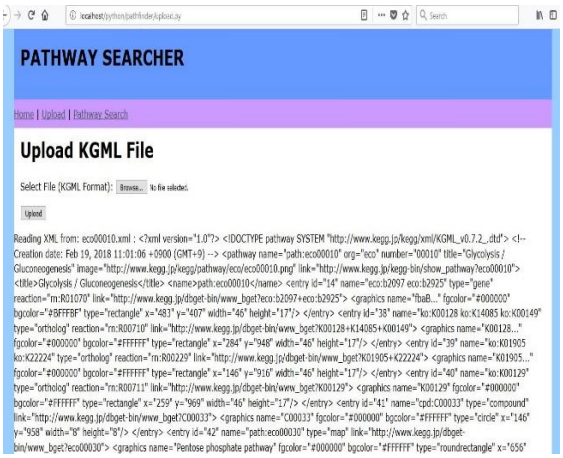

Fig. 7 KGML file post-upload interface

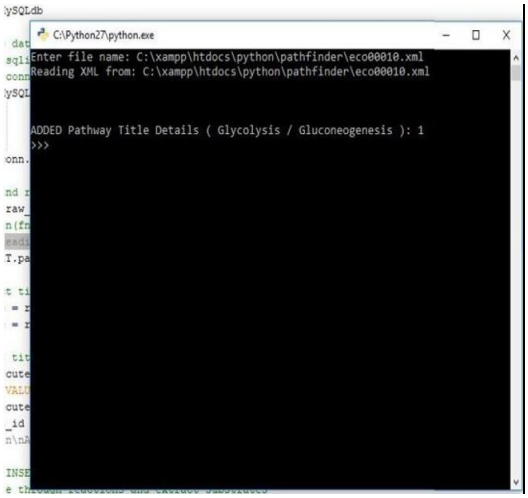

Fig. 8 KGML file post-upload command line

5.2 Pathway Search phase

On successful upload of the KGML file, a search for best route can then be initiated by providing a set of start and target nodes. The system is designed such that any pathway data upload at a prior date can be searched for by allowing the user to select from a list of organism as well as pathway before proceeding with the search. With all search criteria adequately met after proceeding with the search, the best route found is delivered in the search result page. The figures that follow illustrate more.

Fig. 10 Pathway search interface

Fig.11 Pathway search result interface

6. Evaluation

In order to evaluate the result obtained using this method, samples from E. Coli reference pathway were used. The details of the pathway samples were gotten from the PathoLogic: a pathway tools software of BioCyc and EcoCyc. The derived pathways were cross-checked against those from EcoCyc's PathoLogic for consistency and further evaluation as indicated below.

The result gotten using this method was consistent with pathway search result suggested by other researchers and tools and was evaluated based on the following evaluation metrics to determine its actual performance rate. The metrics are as presented in various literatures among which are [20]. The metrics given is as follows;

True positives (TP): The total number of reactions and compounds found in the computed path that are also in the metabolic path. The source and target nodes, whether reaction or compound, are not considered.

False positives (FP): The total number of reactions and compounds found in the computed path that are not in the metabolic path.

False negatives (FN): The total number of reactions and compounds found in the metabolic path that are not in the computed path.

Sensitivity (Sn):= TP/(TP + FN), is the fraction of the reactions and compounds in the metabolic path (excluding source and target) that are in the computed path.

Positive Predictive Value (PPV):= TP/(TP + FP), is the fraction of the reactions and compounds in the computed path (excluding source and target) that are in the metabolic path.

Accuracy (Ac):= (Sn + PPV)/2, is the average of the previous two values.

Using some reference pathways of E. Coli identified from KEGG database against the evaluation metrics presented above, the following values were obtained for the derived paths

Table 2: Result summary

E. Coli Reference Pathways	True positives (TP)	False positives (FP)	False negatives (FN)	Sensitivity (Sn)	Positive predictive value (PPV)	Accuracy (Ac)
Glycolysis/Gluconeogenesis	5	2	8	0.385	0.714	0.549
Citrate (TCA) Cycle	6	2	8	0.429	0.750	0.590
Pentose phosphate pathway	6	1	7	0.462	0.857	0.660

From the evaluation presented, it can be seen that the values derived may be more or less depending on the pathway being considered. However, looking at the metrics in totality, the overall performance can be said to be reasonable. Hence utilizing probabilistic roadmap with A* search yields a consistent pathfinding solution which can still be improved upon for better results.

7. Summary

This work was aimed at utilizing robotics graph traversal techniques for optimizing metabolic pathways route search. This contributes to maximizing the potentials and strengths of motion planning (robotics) graph traversal techniques in best paths finding problems especially in metabolic networks for which it is scarcely utilized. Meanwhile research has shown the successes that motion planning approaches have delivered in domains where they were applied, biological networks inclusive. Looking at other metabolic pathfinding works done; various strategies have been applied including graph techniques and algorithms all in a bid to find a more optimal metabolic pathfinding solution. All those have spurred the need to experiment on utilizing motion planning techniques for same purpose to contribute to the quest for an optimum metabolic pathfinding solution.

7.1 Conclusion

From the result, it can be deduced that the precision and accuracy with which robotics method (motion planning using sampling based algorithms and probabilistic roadmap techniques in particular) solves navigation issues in complex configuration space, it can be used in modelling and solving complex biochemical network problem such as metabolic pathways route search and thereby, benefit from the advantages of the method.

The method however is not without its own limitations. Further research may reveal areas which if worked upon, the efficiency of the method will be improved. In the meantime the limitation of the speed with which probabilistic roadmap (PRM) algorithm builds the roadmap (noticed especially with large pathway data represented in this case by large xml files) exists. This characteristic limitation of PRM however, does not always affect pathfinding speed because once the roadmap is built for a given pathway; it is stored in a database for subsequent querying or path search. Performing search on a previously uploaded pathway is fast irrespective of the size of the pathway because the roadmap does not need to be rebuilt for every search. It only needs to be rebuilt if there is an update to the pathway in question hence, a possible solution may be to device means for frequent periodic uploads to happen.

7.2 Recommendations

Applying motion planning techniques to pathfinding problems in metabolic pathways can be a great feat in bioinformatics especially with further research done to improve upon it. In particular one area of note needing review is the holistic way to synergistically utilize multiple weighting schemes in the search process. It is believed that if that is done, the approach could yield better, faster and more optimal results in general.

References

[1] Abiy et al. (n.d.). A* Search, https://brilliant.org/wiki/a-star-search downloaded on 17 March, 2018.
[2] Algfoor, Z. A., et al. (2017). Identification of metabolic pathways using pathfinding approaches: a systematic review. Brief Funct Genomics. Vol. 16 no. 2 pp 87-98. doi: 10.1093/bfgp/elw002.
[3] Blum, T. & Kohlbacher, O. (2008). MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. Bioinformatics, Vol. 24, No. 18, pp 2108–2109. http://doi.org/10.1093/bioinformatics/btn360
[4] Blum, T., & Kohlbacher, O. (2008). Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. Journal of computational biology, Vol. 15, No. 6, pp. 565–576. doi: 10.1089/cmb.2008.0044
[5] Croes, D., et al. (2005). Metabolic PathFinding: inferring relevant pathways in biochemical networks. Nucleic Acids Research, Vol. 33(Web Server issue), pp W326–W330. http://doi.org/10.1093/nar/gki437
[6] EcoCyc E. coli Database. https://ecocyc.org/ downloaded on 7th April, 2018.
[7] Elbanhawi, M. and Simic, M. (2014). Sampling-based robot motion planning: a review, IEEE Access, vol. 2, pp 56-77.
[8] Faust, K., et al. (2010). Pathway discovery in metabolic networks by subgraph extraction. Bioinformatics, Vol 26, No 9, pp 1211–1218. http://doi.org/10.1093/bioinformatics/btq105
[9] Geraerts R. and Overmars M. H. (2004). A comparative study of probabilistic roadmap planners, Algorithmic Foundations of Robotics V, pp 43-58. Berlin/Heidelberg: Springer.

[10] Ghosh, S. et al. (2013). Weighting schemes in metabolic graphs for identifying biochemical routes, Systems and Synthetic Biology, Vol. 8, Issue 1, pp 47–57.
[11] Hou, J., et al. (2015).Exploring soybean metabolic pathways based on probabilistic graphical model and knowledge-based methods. EURASIP Journal on Bioinformatics and Systems Biology. Springer International Publishing. DOI 10.1186/s13637-015-0026-5
[12] Huang, Y. et al. (2017). A method for finding metabolic pathways using atomic group tracking. PLoS ONE 12(1):e0168725. doi:10.1371/journal.pone.0168725
[13] Karp P. D. et al. (2011). The Pathway Tools Pathway Prediction Algorithm, Standards in Genomic Sciences, vol. 5, pp 424-429
[14] Karp P, et al. (2014). The EcoCyc Database, EcoSal Plus 2014; doi:10.1128/ecosalplus.ESP-0009-2013
[15] KEGG Overview, http://www.genome.jp/kegg/kegg1a.html downloaded on 7th April, 2018.
[16] Kim, S.M., et al. (2017). A review of parameters and heuristics for guiding metabolic pathfinding. Journal of Cheminformatics. Springer International Publishing, Vol. 9 No. 51. https://doi.org/10.1186/s13321-017-0239-6
[17] LaValle, S. M. (2006). Planning algorithms. Cambridge University Press. Chapter 5
[18] Maximova, T. et al. (2015). Computing Transition Paths in Multiple-Basin Proteins with a Probabilistic Roadmap Algorithm Guided by Structure Data, Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference, pp 35-42.
[19] Pey, J., et al. (2011). Path finding methods accounting for stoichiometry in metabolic networks. Genome Biology, https://doi.org/10.1186/gb-2011-12-5-r49
[20] Planes, F.J. & Beasley, J.E. (2009). Path finding approaches and metabolic pathways, Discrete Applied Mathematics, Vol 157, Issue 10, pp 2244-2256
[21] Tang, S. H., et al. (2012). A review on robot motion planning approaches. Pertanika Journal of Science and Technology, Vol. 20, Issue 1, pp 15-29
[22] Wittig, U. & Beuckelaer, A. D. (2001). Analysis and comparison of metabolic pathway databases. Briefings in Bioinformatics, Vol. 2, Issue 2, pp 126–142. https://doi.org/10.1093/bib/2.2.126

First Author Angela Makolo is with Computer Science Department of the University of Ibadan and the University of Ibadan Bioinformatics Research Group (ui.bioinformatics.edu.ng). She has a PhD in Computer Science with Bioinformatics Option from the University of Ibadan.

Second Author Obotu Ojobo holds his B.Sc. degree at Benue State University and his M.Sc. degree at University of Ibadan. The degrees were all in Computer Science.