DISSERTATION

ON

**A DATA MINING APPROACH FOR FILLING THE**

**METABOLIC PATHWAY HOLE**

*Thesis submitted in partial fulfillment of the requirements for the degree of*

**Bachelor of Technology**

**In**

**Computer Science & Engineering**

***By***

|  |  |
| --- | --- |
| **Namrata Kumari** | **Shreekant Agarwal** |
| Roll Number  10400112170 | Roll Number  10400111079 |
| Regisration Number  121040120002 of 2012 – 2013 | Regisration Number  111040110079 of 2011 - 2012 |

***Under the supervision of***

**Prof. Sanjay Chakraborty**

Assistant Professor

Computer Science & Engineering

Institute of Engineering & Management



Computer Science & Engineering

Institute of Engineering & Management, Kolkata

**CERTIFICATE**

It is certified that the work contained in the project report titled **A Data Mining Approach for Filling The Metabolic Pathway Hole** by **Namrata Kumari** and **Shreekant Agarwal** has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Prof. Sanjay Chakraborty Prof. (Dr.) Debika Bhattacharyya**

Mentor HOD, CSE

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Dr. A. K. Nayak**

Principal

**ACKNOWLEDGEMENT**

The sense of fulfillment that comes with the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragement had been driving force. The acknowledgement serves as a testimony for the extreme gratitude and thankfulness that bear for all the people who had been my guiding light for this humungous task.

I would like to convey my sincere and whole hearted appreciation to my project guide Prof. Sanjay Chakraborty, Department of Computer Science & Engineering, Institute Of Engineering & Management. West Bengal university of technology for his continuous guidance and constructive advice throughout the course of this thesis work. I am indeed indebted to my guide for his doubt clearing sessions that he had organized at frequent intervals for my better understanding of the concepts and unknown facts that helped me inculcate greater interest in the area of my thesis work.

Besides Prof. Sanjay Chakraborty, I would like to convey my regards to other professors of Department of Computer Science & Engineering , Institute Of Engineering & Management, West Bengal University Of Technology, as well who have contributed towards my thesis work by giving their invaluable time.

Finally, I would also like to take this opportunity to thank my friends, batch mates for their much coveted selfless support and most definitely my family for being truthfully a generous helping hand to explore this domain of technology that has enlightened and further expanded my horizon of knowledge. There are yet a lot of people whom I would like to express my thankfulness who have contributed towards my such an extensive learning process and will surely be the same in the forthcoming future. Lastly I would apologize if their names are not mentioned here just because of my bad memory or lack of space.

**ABSTRACT**

A key challenge in metabolic pathway hole problem is the reconstruction of the pathway data, reactions, enzymes and genes to use all this data to set the missing genes to the pathway which suffer from missing some genes in its reactions, we mean by the reconstruction here the relation between the enzymes and the genes in the pathway, However, most organism specific metabolic networks are left with a number of unknown enzymatic reactions, that is, many enzymes are missing in the known metabolic pathways, and these missing enzymes are defined as metabolic pathway holes , Although all reactions in some pathways are known, but also this pathways have a holes, the hole in this case means that we do not know the genes behind this reactions.

In our project we have tried to introduce a data mining approach to overcome from the metabolic pathway hole problem.

**TABLE OF CONTENTS**

**CHAPTER 1- INTRODUCTION 1-11**

1.1. Data Mining 1

1.2. Applications of Data Mining 1-2

1.3 Data Mining in Bio Informatics 2

1.4 Gene 2-3

1.5 Computational Biology 3-4

1.6 Biochemical Pathway 4-5

1.6.1 Gene Interaction 5-6

1.6.2 Signal Transduction Pathway 7

1.6.3 Protein Interaction Pathway 8

1.6.4 Metabolic Pathway 8

1.7 Computational Method for Metabolic Pathway prediction 8

1.7.1.Pathologic Program 8-9

1.7.2.In Silico Method 9

1.8 Sources of Metabolic Pathway 10

1.8.1 Integrated Relational Database 10

1.8.2 KEGG 10-11

1.8.3 Bio Cyc 11

**CHAPTER 2 -** **PREVIOUS WORK**

2. 1 Graph Theory 12

2.2 Artificial Neural Network 12

2.3 Existing Metabolic Network Analysis Tools 12-1

2.3.1 Wikipathways 12

2.3.2 CPath 12-13

2.3.3 Toppgene suite 13-14

2.3.4 GenMAPP 14-15

2.3.5 Pathway Painter 15

**CHAPTER 3- PROPOSED WORK**

3.1 Metabolic Pathway Hole 16

3.2 Soft Computing Method 16

3.2.1 Introduction to Fuzzy Logic 16

3.3 Fuzzy Logic to Fill Metabolic Pathway Hole 17-20

3.4 Algorithm 18

**CHAPTER 4- CONCLUSION**

4.1 Conclusion 22

4.2 Future Work 22

**REFERENCES**  23

**LIST OF TABLES**

Table 1. Value ranges of fuzzy set parameter 19

Table 2. Example of three genes of the possible genes 20

Table 3. Shotgun-score result summary 20

Table 4. Fuzzy model 1 21

Table 5. Fuzzy model 2 21

Table 6. Fuzzy model 3 21

**LIST OF FIGURES**

Figure 1. Gene 3

Figure 2. Biochemical Pathway 5

Figure 3. Metabolic Network 8

Figure 4. The block diagram for the proposed algorithm 17

**CHAPTER 1. INTRODUCTION**

1. **Introduction**

In bioinformatics field there is a lot of problem need to be solved. One of the most important problem is metabolic pathway problem, where metabolic pathway includes several problem like re-construction pathway.

Most organism specific metabolic networks are left were with a number of enzymatic reactions, that is, many enzymes are missing in known metabolic pathways, and these missing enzymes are defined as metabolic pathway holes, the hole in this case means here that we do not know the genes behind the reaction.

This thesis deals with pathway hole problem, to help in setting the correct gene in pathway which have a hole.

* 1. **Data Mining**

Generally, Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining refers to extracting or “mining” knowledge from large amounts of data. Data Mining (DM) is the science of finding new interesting patterns and relationship in huge amount of data. It is defined as “the process of discovering meaningful new correlations, patterns, and trends by digging into large amounts of data stored in warehouses”. Data mining is also sometimes called Knowledge Discovery in Databases (KDD). Data mining is not specific to any industry. It requires intelligent technologies and the willingness to explore the possibility of hidden knowledge that resides in the data.

* 1. **Applications of Data Mining.**

1. Banking: loan/credit card approval

predict good customers based on old customers.

1. Customer relationship management:

identify those who are likely to leave for a competitor.

1. Targeted marketing:

identify likely responders to promotions

1. Manufacturing and production: automatically adjust knobs when process parameter changes.
2. Medicine: disease outcome, effectiveness of treatments analyze patient disease history: find relationship between diseases.
3. Scientific data analysis: identify new galaxies by searching for sub clusters.
4. Gene Sequence analysis:Data Mining approach is also useful in Gene sequence Analysis.
   1. **Use of Data Minining In Bio Informatics.**

Medical applications are another fruitful area: data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications.

Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich, but lacks a comprehensive theory of life’s organization at the molecular level. The extensive databases of biological information create both challenges and opportunities for development of novel KDD methods. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience.

Applications of data mining to bioinformatics include gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

* 1. **Gene**

The word gene is derived from the Greek word *genesis* meaning "birth", or *genos* meaning "origin"**.** A gene is the molecular unit of heredity of a living organism. It is used extensively by the scientific community as a name given to some stretches of deoxyribonucleic acids (DNA) and ribonucleic acids (RNA) that code for a polypeptide or for an RNA chain that has a function in the organism. Living beings depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring. All organisms have genes corresponding to various biological traits, some of which are instantly visible, such as eye color or number of limbs, and some of which are not, such as blood type, increased risk for specific diseases, or the thousands of basic biochemical processes that comprise life.

Gene is made by 4 parts

**A-Adenine**

**G- Guanine**

**C- Cytosine**

**T- Thymine**

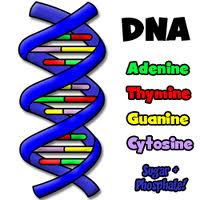
****

FIGURE. 1

* 1. **Computational Biology**

Computational Biology is the application of computer sciences and allied technologies to answer the questions of Biologists, about the mysteries of life. A mere application of computers to solve any problem of a biologist would not merit a separate discipline. It looks as if Computational Biology and Bioinformatics are mainly concerned with problems involving data emerging from within cells of living beings. It might be appropriate to say that Computational Biology and Bioinformatics deal with application of computers in solving problems of molecular biology, in this context.

What are these data emerging from a cell? Though not exhaustive, at the risk of over simplifying I will list 4 important data: DNA, RNA and Protein sequences and Micro array images. Surprisingly, first 3 of them are mere text data (strings, more formally) that can be opened with a text editor. The last one is a digital image.

* Analysing DNA sequence data to locate genes
* Analysing RNA sequence data to predict their structure
* Analysing protein sequence data to predict their location Inside cell
* Analysing gene expression images
* Identifying new Drug Molecules

Computational biology involves the development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological, behavioural, and social systems. Computational biology is different from biological computation, which is a subfield of computer science and computer engineering using bioengineering and biology to build computers, but is similar to bioinformatics, which is an interdisciplinary science using computers to store and process biological data.

Computational Biology, sometimes referred to as bioinformatics, is the science of using biological data to develop algorithms and relations among various biological systems. Prior to the advent of computational biology, biologists were unable to have access to large amounts of data. Researchers were able to develop analytical methods for interpreting biological information, but were unable to share them quickly among colleagues.

* + 1. **What kinds of problems do computational biologists work on?**

Much of computational biology is concerned with the analysis of molecular data, such as bio-sequences (DNA, RNA, or protein sequences), three-dimensional protein structures, gene expression data, or molecular biological networks (metabolic pathways, protein-protein interaction networks, or gene regulatory networks). A wide variety of problems can be addressed using these data, such as the identification of disease-causing genes, the reconstruction of the evolutionary histories of species, and the unlocking of the complex regulatory codes that turn genes on and off. Computational biology can also be concerned with non-molecular data, such as clinical or ecological data.

Problems investigated by computational biologists include topics as diverse as the genetics of disease susceptibility; comparing entire genomes to reveal the evolutionary history of life; predicting the structure, motions, and interactions of proteins; designing new therapeutic drugs; modelling the complex signalling mechanisms within cells; predicting how ecosystems will respond to climate change; and designing recovery plans for endangered species. The computational biologist must have skills in mathematics, statistics, machine learning, and the physical sciences as well as in biology. A key goal in training is to develop the ability to relate biological processes to computational models. Cornell faculty work primarily in six subareas of computational biology: 1. computational and statistical genomics, 2. population, comparative, and functional genomics, 3. bioinformatics, 4. proteomics, 5. ecology and evolutionary biology, and 6. statistical and computational methods for modelling biological systems.

* 1. **Biochemical Pathway**

A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off, or spur a cell to move. Some of the most common biological pathways are involved in metabolism, the regulation of gene expression and the transmission of signals. Pathways play key role in advanced studies of Genomics.

Biochemical pathway is totality of anabolic and catabolic reactions that occur in living cells. Enzymes catalyze the biochemical reaction. Examples of biochemical pathways are glycosis, beta-oxidation, amino-acid synthesis, krebs cycle etc. Just remember all chemical reactions taking place in living cells can be simply called biochemical reactions.

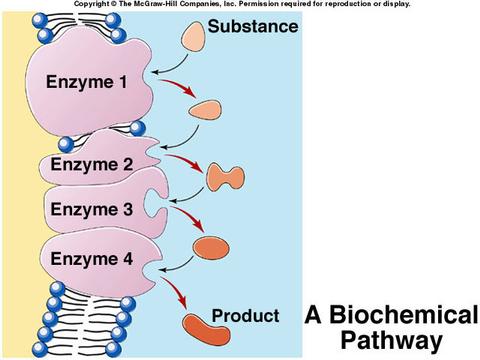


FIGURE. 2

* + 1. **Most common types of biological pathways**

1. **Gene regulation pathways**
2. **Signal transduction pathways**
3. **Protein interaction network**
4. **Metabolic pathways**
5. **Gene Regulatory Network**

A gene regulatory network or genetic regulatory network (GRN) is a collection of DNA segments in a cell which interact with each other indirectly (through their RNA and protein expression products) and with other substances in the cell, thereby governing the expression levels of mRNA and proteins. In general, each mRNA molecule goes on to make a specific protein (or set of proteins). In some cases this protein will be structural, and will accumulate at the cell membrane or within the cell to give it particular structural properties. In other cases the protein will be an enzyme, i.e., a micro-machine that catalyses a certain reaction, such as the breakdown of a food source or toxin. Some proteins though serve only to activate other genes, and these are the transcription factors that are the main players in regulatory networks or cascades. By binding to the promoter region at the start of other genes they turn them on, initiating the production of another protein, and so on. Some transcription factors are inhibitory.

In single-celled organisms, regulatory networks respond to the external environment, optimizing the cell at a given time for survival in the environment. Thus a yeast cell, finding itself in a sugar solution, will turn on genes to make enzymes that process the sugar to alcohol. This process, which we associate with wine-making, is how the yeast cell makes its living, gaining energy to multiply, which under normal circumstances would enhance its survival prospects.

In multi cellular animals the same principle has been put in the service of gene cascades that control body shape. Each time a cell divides, two cells result which, although they contain the same genome in full, can differ in which genes are turned on and making proteins. Sometimes a 'self-sustaining feedback loop' ensures that a cell maintains its identity and passes it on. Less understood is the mechanism of epigenetics by which chromatin modification may provide cellular memory by blocking or allowing transcription. A major feature of multi cellular animals is the use of morphogen gradients, which in effect provide a positioning system that tells a cell where in the body it is, and hence what sort of cell to become. A gene that is turned on in one cell may make a product that leaves the cell and diffuses through adjacent cells, entering them and turning on genes only when it is present above a certain threshold level. These cells are thus induced into a new fate, and may even generate other morphogens that signal back to the original cell. Over longer distances morphogens may use the active process of signal transduction. Such signalling controls embryogenesis, the building of a body plan from scratch through a series of sequential steps. They also control and maintain adult bodies through feedback processes, and the loss of such feedback because of a mutation can be responsible for the cell proliferation that is seen in cancer. In parallel with this process of building structure, the gene cascade turns on genes that make structural proteins that give each cell the physical properties it needs.

1. **Signal Transduction Pathways**

Signal transduction refers to any process by which a cell converts one kind of signal or stimulus into another. These processes most often involve ordered sequences of biochemical reactions inside the cell that are carried out by enzymes. In single-cell organisms, signal transduction pathways determine how the cell senses and responds to its environment. In multi-cellular organisms, a multitude of different signal transduction pathways are required for coordinating the behaviour of individual cells to support the function of the organism as a whole. As may be expected, the more complex the organism, the more complex the repertoire and interconnectivity of signal transduction pathways. Thus, sensing of both the external and internal environment at the cellular level relies on signal transduction.  
  
Many diseases such as diabetes, heart disease, inflammation and cancer arise from defects in signal transduction pathways, underscoring the critical importance of signal transduction in medicine as well as biology.

1. **Protein Interaction Network**

Protein interactions refer to intentional physical contacts established between two or more proteins as a result of biochemical events and/or electrostatic forces. In fact, proteins are vital macromolecules, at both cellular and systemic levels, but they rarely act alone. Diverse essential molecular processes within a cell are carried out by molecular machines that are built from a large number of protein components organized by their PPIs. Indeed, these interactions are at the core of the entire interactomics system of any living cell and so, unsurprisingly, aberrant PPIs are on the basis of multiple diseases, such as Creutzfeld-Jacob, Alzheimer’ disease, and cancer.

1. **Metabolic Network**

Metabolic pathways are series of chemical reactions occurring within a cell. A principal chemical is modified by a series of chemical reactions enzymes catalyze these reactions, and often require dietary minerals, vitamins, and other cofactors in order to function properly. Because of the many chemicals that may be involved, metabolic pathways can be quite elaborate. In addition, numerous distinct pathways co-exist within a cell. This collection of pathways is called the metabolic network.

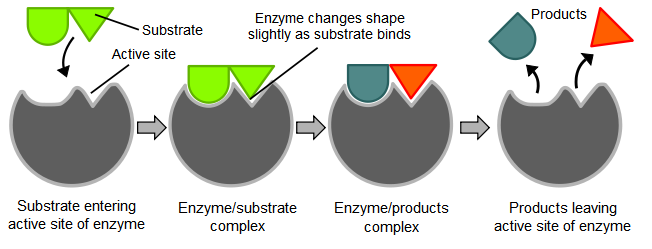
****

Figure 3. MetabolicNetwork

* 1. **Different computational approaches for Metabolic Pathway prediction**
     1. **Pathologic Program**

Pathologic program predicts the metabolic pathways of an organism from its annotated genome. The word ‘annotated gene refers to the process of attaching biological information to sequences. It takes place in two phases.

1. R**eactome inference phase**

The first step in pathway prediction is to infer the set of biochemical reactions that can be catalyzed by the enzymes encoded by the genome, that is to infer the rectome of the organism because reactions are the direct building block of the pathways. Once we know the reactome of the organism, we can more easily assess the evidence for what pathways are present.

1. **Pathway inference**

Pathway inference is based on the set of catalyzed reactions that were imported from metacyc during the phase described. All pathways inferred during the phase described have also come from the fixed reference database namely metacyc. All the inferred pathways are copied from metacyc to the PGDB generated for that organism. When we say a reaction is present in the generated PGDB, we mean the reaction has been inferred as present in the organism by the organism by the reactome inference phase or that reaction is marked as spontaneous metacyc.

* + 1. **IN SILICO METHOD OF METABOLIC PATHWAY**

In silico is an expression used to mean "performed on computer or via Computer Simulation”.  For example, in 2007 researchers developed an in silico model of tuberculosis to aid in drug discovery, with the prime benefit of being faster than real time simulated growth rates, allowing phenomena of interest to be observed in minutes rather than months.

The genome annotation, along with biochemical and strain-specific information, provides the information needed to reconstruct complete metabolic networks of these microbes. The reconstructed metabolic networks can be used to analyze, interpret, and predict the metabolic gene sequence.

**Basic Procedures**

First of all, a metabolic pathway goes under breakdown into their respective reactions and enzymes. The reconstruction collect all the relevant metabolic information of an organism and compiles it in a mathematical model.

In general, the process is as follows-

* Draft a reconstruction
* Refine the model
* Convert model into mathematical/ computational representation
* Evaluate and debug the model through experiments.
  1. **Sources of metabolic pathways**
     1. **Integrated relational database**

The explosion of structural data that has resulted from gene-sequencing and proteomic studies has emphasized the need to relate structures to functions. It is important that these structural data are integrated with the functional data on catalytic proteins that are available from a number of sources.

In addition tools have been developed to maintain the database and allow propagation of new and updated biochemical terminology, which has been integrated into the database. These tools are connected to the ChEBI database, which provides a definitive dictionary of compounds, to improve the quality of the IntEnz vocabulary. ChEBI stands for dictionary of Chemical Compounds of Biological Interest. The ChEBI database is also hosted at the European Bioinformatics Institute (EBI), but is still under development.

* + 1. **KEGG - Kyoto Encyclopedia of Genes and Genomes**

 KEGG is 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 a computer representation of the biological system, consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information) that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction and relation networks (systems information).

The KEGG database has been in development by [Kanehisa Laboratories](http://www.kanehisa.jp/) since 1995, and is now a prominent reference knowledge base for integration and interpretation of large-scale molecular data sets generated by genome sequencing and other high-throughput experimental technologies.

* + 1. **BioCyc**

The BioCyc database collection  is a set of biological databases. Databases within BioCyc describe genome and pathway information for individual organisms. BioCyc is maintained by SRI International, in Menlo Park California.

Two databases within BioCyc are highly curated, meaning they have received extensive manual updating with information from the scientific literature. Those databases are EcoCyc and MetaCyc. The remaining BioCyc databases were generated computationally to predict what metabolic pathways are present in these organisms. In some cases, the databases were refined manually after their generation. As of October 2010, BioCyc contained databases for 1,004 genomes.

The BioCyc website contains a variety of software tools for searching, visualizing, comparing, and analyzing genome and pathway information. It includes a genome browser, and browsers for metabolic and regulatory networks. The website also includes tools for painting large-scale datasets onto metabolic and regulatory networks, and onto the genome.

BioCyc databases rely on a software system called Pathway Tools for their initial generation, subsequent updating, and for querying their content. The databases can also be installed locally.

**MetaCyc**

MetaCyc is a database that contain experimentally explained metabolic pathways.  In contrast to all other members of that collection, which are organism-specific DBs, MetaCyc is a *multiorganism* DB. MetaCyc contains more than 2151 pathways from more than 2515 different organisms, and is curated from the scientific experimental literature.

MetaCyc contains extensive data on individual enzymes, describing their subunit structure, cofactors, activators and inhibitors, substrate specificity, and, in some cases, kinetic constants. It also provides commentary and literature references.

The MetaCyc mission is to serve a broad community of researchers from genetics, molecular biology, microbiology, biochemistry, genomics, bioinformatics, metabolic engineering, and systems biology in support of the following tasks:

**Support computational metabolic network prediction**

One of MetaCyc’s primary applications is to serve as a reference database for computationally predicting the metabolic network of an organism from its annotated genome, such as by the PathoLogic algorithm.

**Provide an encyclopedic reference on pathways and enzymes**

MetaCyc is used as a readily accessible source of up-to-date, literature-curated information on metabolic pathways and enzymes by researchers for basic research and genome analysis, and by students and teachers for educational purposes.

**Support metabolic engineering**

Metabolic engineers use MetaCyc as an encyclopedia of metabolic pathways and enzymes that may be genetically engineered into an organism to alter its metabolism.

**CHAPTER 2. PREVIOUS WORK**

* 1. **Graph Theory and Analysis**

A metabolic pathway may be represented computationally using a combination of statistical methods and programming methods (Speed et al, 2002). Data-sets are obtained from a central repository database which consists of metabolic pathway information, from which graph-based representations are plotted.  These may also be used to represent the metabolic pathway and the altered enzymes could be identified through this method. Nodes may be represented are the compounds or the metabolites of the reaction or enzymes and edges may act as reaction links. Each node would contain a dataset which would help us segregate the list of similar entities. Statistical calculations such as, p-value calculation, matrix methods could be performed to rank the nodes.

|  |
| --- |
|  |
|  |

* 1. **Artificial Neural Network**

A neural network is an artificial representation of human brain. The main aim behind the development of neural network is to acquire human ability to adapt to changing circumstances and the environment. It is suitable for predicting both types of genes, that is, protein coding and RNA coding. Artificial neural network (ANN) is an interconnected group of artificial neurons. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurones) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons.

**2.3.** **Existing metabolic network analysis tool**

**2.3.1. Wikipathways**

WikiPathways was established to facilitate the contribution and maintenance of pathway information by the biology community. WikiPathways is an open, collaborative platform dedicated to the curation of biological pathways. WikiPathways thus presents a new model for pathway databases that enhances and complements ongoing efforts, such as KEGG, Reactome and Pathwa Commons. Building on the same MediaWiki software that powers Wikipedia, we added a custom graphical pathway editing tools and integrated databases covering major gene, protein, and small-molecule systems. The familiar web-based format of WikiPathways greatly reduces the barrier to participate in pathway curation. More importantly, the open, public approach of WikiPathways allows for broader participation by the entire community, ranging from students to senior experts in each field. This approach also shifts the bulk of peer review, editorial curation, and maintenance to the community.

**How Does it Work ?**

Each pathway at WikiPathways has a dedicated wiki page, displaying the current diagram, description, references, download options, version history, and component gene and protein lists. Any pathway can be edited from within its wiki page by activating an embedded pathway editor. After editing, an updated pathway image is displayed on the wiki page along with the version history and list of component genes and proteins. Users can easily monitor and undo changes, compare differences and search for overlapping pathways. Using the search feature, one can locate particular pathways by name, by the genes and proteins they contain, or by the text displayed in their descriptions and comments. One can also browse the collection of pathways with combinations of species names and ontology-based categories. The pathway content at WikiPathways is freely available for download in a variety of data and image formats, including GPML, which is a custom XML format compatible with pathway visualization and analysis tools such as Cytoscape, GenMAPP and PathVisio.

* + 1. **CPath**

CPath is a robust, scalable, modular, professional- grade software platform for collecting, storing, and querying biological pathways. It can serve as the core data handling component in information systems for pathway visualization, analysis and modeling.

**How does it works:**

cPath an open source database and web application for collecting, storing, and querying biological pathway data. cPath makes it easy to aggregate custom pathway data sets available in standard exchange formats from multiple databases, present pathway data to biologists via a customizable web interface, and export pathway data via a web service to third-party software, such as Cytoscape, for visualization and analysis. cPath is software only, and does not include new pathway information. Key features include: a built-in identifier mapping service for linking identical interactors and linking to external resources; built-in support for PSI-MI and BioPAX standard pathway exchange formats; a web service interface for searching and retrieving pathway data sets; and thorough documentation. The cPath software is freely available under the LGPL open source licence for academic and commercial use.

* + 1. **Toppgene suite**

Toppgene suite is a one-step portal for gene list enrichment analysis and candidate gene prioritization based on functional annotation and protein interaction network. ToppGene Suite ([http://toppgene.cchmc.org](http://toppgene.cchmc.org/); this web site is free and open to all users and does not require a login to access) is a one-stop portal for (i) gene list functional enrichment, (ii) candidate gene prioritization using either functional annotations or network analysis and (iii) identification and prioritization of novel disease candidate genes in the interactome. Functional annotation-based disease candidate gene prioritization uses a fuzzy-based similarity measure to compute the similarity between any two genes based on semantic annotations. The similarity scores from individual features are combined into an overall score using statistical meta-analysis. A P-value of each annotation of a test gene is derived by random sampling of the whole genome. The protein–protein interaction network (PPIN)-based disease candidate gene prioritization uses social and Web networks analysis algorithms (extended versions of the PageRank and HITS algorithms, and the K-Step Markov method). We demonstrate the utility of ToppGene Suite using 20 recently reported GWAS-based gene–disease associations (including novel disease genes) representing five diseases. ToppGene ranked 19 of 20 (95%) candidate genes within the top 20%, while ToppNet ranked 12 of 16 (75%) candidate genes among the top 20%.

**Limitations:**

ToppGene or any functional annotation-based prioritization method has some limitations. First, when using a training set of genes, the assumption is that the disease genes we have yet to discover will be consistent with what is already known about a disease and/or its genetic basis, which may not always be the case. Second, the annotations and analyses, as well as the prioritization, can only be as accurate as the underlying online sources from which the annotations are retrieved. Similar to functional annotation-based methods, the performance of network-based prioritization methods (ToppNet) is also dependent on the quality of interaction data, which currently suffers from incompleteness and unreliability with missing interactions and false positives.

* + 1. **GenMAPP-GenMAPP**

Gene Map Annotator and Pathway Profiler is a free, open-source bioinformatics software tool designed to visualize and analyze genomic data in the context of pathways (metabolic, signalling), connecting gene-level datasets to biological processes and disease. First created in 2000, GenMAPP is developed by an open-source team based in an academic research laboratory. GenMAPP maintains databases of gene identifiers and collections of pathway maps in addition to visualization and analysis tools. Together with other public resources, GenMAPP aims to provide the research community with tools to gain insight into biology through the integration of data types ranging from genes to proteins to pathways to disease.

GenMAPP was designed and developed in the Conklin lab group at the J David Gladstone Institute, at the University of California, San Francisco. It assists in viewing and analyzing microarray, proteomics and other genome-scale data in terms of particular biological pathways. This facilitates fast, easy and systematic presentation and analysis of high-throughput expression data. In this work we have prepared a GenMAPP Gene Database (a species-specific library of gene information) for Saccharomyces pombe or fission yeast in collaboration with the GenMAPP team. In addition the pathway maps for S.pombe have been inferred from existing pathways for S. cerevisiae using homology information. The MAPPs can be used for MAPP finder analysis, which also integrates the annotations of the Gene Ontology project to create a global gene expression profile. GenMAPP provides tools to create, edit and annotate biological pathway maps

* + 1. **Pathway Painter**

Advanced Pathway Painter is a handy tool for bioscience students and researchers seeking displays of mapped genetic and protein pathways. It launches a compact, slightly crowded, but functional interface that's preloaded with a wide array of pathways, and updating the pathways through an Internet connection is a simple process. Researchers will appreciate its ease of use as they analyze gene and protein experiments that identify genetic variations related to disease. These are not only human pathways, but also pathways for a wide variety of plants and animals. Advanced Pathway Painter is free tool that helps you visualize pathways. The user has the possibility to display any kind of quantitative data from gene and protein experiments directly within the pathways (colors represent the value). The linking between the pathway items and the experiment data is done over the gene or protein names and their accession numbers. Furthermore the user has a quick overview on the gene/protein with the collected links in the Web-interface.

|  |
| --- |
|  |

**CHAPTER 3. PROPOSED WORK**

**3.1. Metabolic Pathway Hole**

Metabolic network is one of the important classes of biological networks, consisting of enzymatic reactions involving substrates and products. Recent developments in pathway databases enable us to analyze the known metabolic networks. However, most organisms specific metabolic networks are left with a number of unknown enzymatic reactions, that is many enzymes are missing in the known metabolic pathways, and these missing enzymes are defined as metabolic pathway holes, although all reactions in some pathways are known, but also this pathways have a holes, the hole in this case means here that, we do not know the gene(s) that produce this enzyme.

**3.2. Soft Computing Method**

Soft computing Technologies promise to become a powerful computational methodology for solving problems accurately and acceptably. Fuzzy logic is one of the soft computing components that could deal with uncertainty in real problem, due to the nature of continued data of our problem, because of the vagueness of boundaries between the concepts we preferred to use the fuzzy logic approach to overcome this problem.

**3.2.1. Introduction of Fuzzy Logic**

Fuzzy logic is a form of many valued logic; it deals with reasoning that is approximate rather than fixed and exact. Compared to traditional binary sets (where variables may take on true and false value), fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false.

Classical logic only permits propositions having a value of truth or falsity. The notion of whether 1+1=2 is an absolute, immutable, mathematical truth. However, there exist certain propositions with variable answers, such as asking various people to identify a color. The notion of truth doesn't fall by the wayside, but rather a means of representing and reasoning over partial knowledge is afforded, by aggregating all possible outcomes into a dimensional spectrum.

Both degrees of truth and probabilities range between 0 and 1 and hence may seem similar at first. For example, let a 100 ml glass contain 30 ml of water. Then we may consider two concepts: empty and full. The meaning of each of them can be represented by a certain fuzzy set. Then one might define the glass as being 0.7 empty and 0.3 full. Note that the concept of emptiness would be subjective and thus would depend on the observer or designer. Another designer might equally well design a set membership function where the glass would be considered full for all values down to 50 ml. It is essential to realize that fuzzy logic uses truth degrees as a mathematical model of the vagueness phenomenon while probability is a mathematical model of ignorance.

**3.3. Use Fuzzy logic to fill the Metabolic pathway Hole**

We designed a fuzzy based model to fill the metabolic pathway hole. From the Poole of soft computing we select fuzzy logic according to its ability to deal with vagueness and imprecise classes of the data. Due to the nature of continued data of our problem, because of the vagueness of boundaries between the concepts we preferred to use the fuzzy logic approach to overcome this problem. As we know, Fuzzy logic is an “approach to computing based on "degrees of truth" rather than the usual "true or false" (1 or 0) Boolean logic on which the modern computer is based.”, this does not mean that Fuzzy logic don’t 0 or 1, no Fuzzy already includes 0 and 1 as extreme cases of truth (or "fact") but also includes the various states of truth in between. Fuzzy logic system (FLS) consists of four main parts, Fuzzifier, Rules, Inference engine and Defuzzifier.

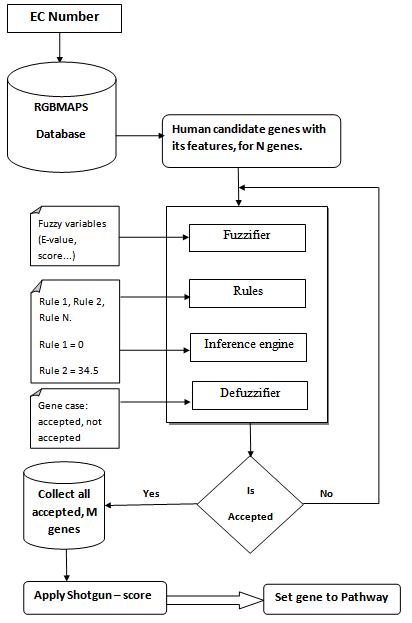


Figure 5. The block diagram for the proposed algorithm

**3.4. Overall Algorithm**

1. **Gene’s retrieval**

Retrieve from RGBMAPS database genes that catalyze the desired reaction in other organisms, then we retrieve all possible genes of the organism of interest using BLAST. RGBMAPS gives us directly all genes of a specific EC in the target organism human after passing three phases of collecting data, (i) collect the pathways of the interest organism with its reactions and ECs, (ii) retrieve all genes that act with this EC in the different organisms, (iii) using BLAST to retrieve all possible genes of the target organism which are similar to the genes of other organisms.

1. **Candidate genes to fill hole**

In this step of the algorithm we feed our proposed fuzzy model with all genes of the EC, to filter all these genes and candidate the genes only that can fill the pathway hole, we illustrated this step in the later section.

1. **Candidate evaluation**

In this step we applied shot-gun score to set the correct gene form these candidate genes that obtained from step2.

1. **Gene’s retrival**

As we present above, in this step we need to obtain all possible genes to a specific EC in different organisms to be the input data to the fuzzy system, we select RGBMAPS database to be our source of data.

1. **Candidate genes to fill hole**

The purpose of this step is to filter these entire possible genes which are 86 genes in our example EC: 2.3.1.61, throw our fuzzy system to produce the candidate genes for this EC. To do that we need to determine the fuzzy input sets and the rules that the system will use .In the following internal sections we will illustrate that in some details.

* **Fuzzy input sets**

We chose three variables to be our fuzzy sets, Evalue, score and identity. We designed our fuzzy variables by setting the variables ranges and its mathematical shape.

* **Fuzzy variables ranges**

As presented in table 1, we show the different ranges of fuzzy variables that we assign to this system, as shown in table we assign three sets to each variable, high, medium and low.



Table.1: value ranges of fuzzy set parameters

The three ranges high, medium and low are arranged according to the value ranges of each fuzzy variable, where E-value the value of it is between 1 and 10, so we suggest that the high from 0 and 1 where medium between 0.8 and 2 and low is from 1.7 and upper, but score has a big value ranges because its value ranges is huge than Evalue and identity.

* **Fuzzy rule system**

In our work we create three models of rules and applied it on the data to elect the more efficient one, the election processes are discussed in “evaluation” section and the election model is presented in the given rule:

Rule 1: If score is high and identity is high then accepted.

Rule 2: if E-value is low and score is low and identity is low then not accepted.

Rule 3: if E-value is high and score is medium and identity is high then accepted.

Rule 4: if E-value is high and score is low and identity is low then not accepted.

Rule 5: if E-value is high and score is medium and identity is medium then may be accepted.

* **Applying fuzzy rules**

In this step we apply the elected fuzzy rule model on the data that obtained from the first step of the algorithm, which is the all possible genes of the EC which was in our example 86 genes of EC: 2.3.1.61. After applying the rules on 86 genes, the fuzzy system classify the genes into three classes accepted genes, may be genes and not accepted genes and this according to the ranges of the fuzzy variables and the different values of each genes. As presented in table 2, present sample to three different cases form the 86.



**Table 2**: Example of three genes of the possible genes.

After applying the rules on all possible genes, we have 42 accepted genes, 14 not accepted, 20 may be gens, 10 NAN, we select only the accepted genes in all organisms, so we have 42 candidate genes, all this genes can be the correct gene that can fill the hole, so now we need another level of filtration to obtain only the correct gene.

* **Candidate Evaluation**

As we present above we need to evaluate the candidate genes produced by the fuzzy model, to elect one gene only to set the pathway hole, we applied Shotgun score on these genes, and ranking from the biggest score to the lowest, where shotgun-score is “the number of query sequences whose fuzzy system output included the candidate sequences”.

In our example we have 42 hits (only accepted genes); we need to decide which one of them is the correct gene to fill the hole.



**Table 3**: Shotgun-score result summary

1. **Fuzzy rule evaluation**

We summarize the three models in the following tables 4, 5 and 6, where H mean high ,M mean medium and L mean low , which represent fuzzy variables ranges as presented before.



**Table 4**: Fuzzy model 1



**Table 5**: Fuzzy model 2

****

Table 6: Fuzzy model 3

**b)Filing hole evaluation**

In this section we evaluate the shotgun-score level which is the last step in our algorithm to set the correct gene to fill the pathway hole, so this evaluation reflect the evaluation of our proposed system at all, as presented in table 7 we applied our proposed system on 70 sample to set the correct gene the total percent is 85.71%.

**CHAPTER 4. CONCLUSION & FUTURE WORK**

**4.1. Conclusion**

* The proposed system has the ability to solve pathway hole problem using the proposed database RGBMAPS and the proposed fuzzy system.
* In data collection phase of our database, to do this task in manually way, that is very hard, waste effort and time; we have overcome the problem by using the proposed Fuzzy Logic method.
* Using fuzzy system is very favorable in pathway problems, because it’s gaining strength through its seemed closer to the way our brains work, which make the researchers closer to the data.

**4.2.Future Work**

Even though we have put in our best and worked hard, there are certain areas which are yet unexplored and demand more work to be done on it.In future woks we will use machine learning to build the fuzzy system rule and also we will re-evaluate our proposed system after changing the fuzzy variables ranges, (E-value, score and identity).

**References**

* 1. http://en.wikipedia.org/wiki/Metabolic\_network\_modelling
  2. http://gmod.org/wiki/Pathway\_Tools
  3. Lakshminarasimhan N, Tagore S (2009) Damages in Metabolic Pathways: Computational Approaches. J Comput Sci Syst Biol 2: 208-215. doi:10.4172/jcsb.100003
  4. IJCSI International Journal of Computer Science Issues.
  5. KEGG: Kyoto Encyclopedia *of genes and genomes*, Minoru Karehisa & Susumu Goto, Institute of Chemical Research, Kyoto University, Uji, Japan.