RGB MAPS: a Proposed Database for Solving Metabolic Pathway Hole

Alaa Eldin Abdallah Yassin
Central Lab for Agricultural Expert Systems,
Ministry of Agriculture and Land Reclamation
Giza, Egypt
aboelmnzer@gmail.com

Laila Mohamed ElFangary
Faculty of Computers and Information - Helwan University
Cairo, Egypt
lailaelfangary@gmail.com

Ahmed Farouk Al-Sadek 1,2
1Central Lab for Agricultural Expert Systems, Giza, Egypt
2Faculty of computer science, October University for modern science and Art
afsadek@gmail.com

***Abstract*—Filling pathway hole is a point of research in the field of Bioinformatics especially in metabolic pathway where the analysis of metabolic pathways is an essential topic in understanding the relationship between genotype and phenotype [4].The pillar of the research cycle is the data collection which precedes the analysis phase to solve the pathway hole. The required data for this area is scattered among different data sources, which represent a problem for the researchers of this area. This paper provides a solution to this obstacle by collecting the required data from various data sources in one database RGB MAPS. Also we have developed a tool that could be used by other researchers to analyze the pathway holes.**

I. INTRODUCTION

In recent years, a large number of metabolic databases have been developed to cover the huge amount of genome sequencing, where a key challenge in systems biology is the reconstruction of an organism's metabolic network from its genome sequence[12].Once the sequences are obtained, functions must be assigned to these new sequences [1]. Metabolic databases like MetaCyc[13],KEGG[14], Reactome[15] and Model SEED[1]are not enough to help the researcher of metabolic pathway hole filling because he needs a different forms of data which are the E.C similarity between different organisms , and the genes similarity between the target organism and other organisms.

This paper is organized into the following sections; the first section will introduce pathway holes, its definition and problem, the second section will summarize the pathway hole filling approaches, the third section will survey the popular pathway databases where MetaCyc and KEGG are described in particular in some detail, the fourth section is the most important section in the paper which addresses to talk about our proposed database called RGB MAPS, then the evaluation section which evaluate the RGB MAPS data base, finally the conclusion of the paper.

II. PATHWAY HOLES

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 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 [2], 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 genes behind this reactions.
So we can shorten the metabolic pathway hole types to two types

- Unidentified enzymatic reactions in the pathway (figure1.A).
- Unknown genes behind the known reactions in the pathway (figure1.B).

The reason of these holes in the pathway is the huge amount of genome sequencing data, but on the other hand there are no laboratory experiments covering this size of data in all organisms, add to that the difficulty of conducting laboratory research on some organisms due to the length of its life cycle or their rarity or for other reasons, like Canes families, Macacafascicularis and Pan troglodytes. And do not forget to mention the expensive price of these laboratory experiments [1][3].

In other words we can say that "Pathway holes occur when a genome appears to lack the enzymes needed to catalyze reactions in a pathway"[2].

If a protein has not been assigned a specific function during the annotation process, any reaction catalyzed by that

protein will appear as a missing enzyme or pathway hole in a Pathway/Genome database. [2]

Finally we need to say that: there are two types of hole missing the gene whose product is predicted to catalyze the reaction and the EC number of the missing reaction.

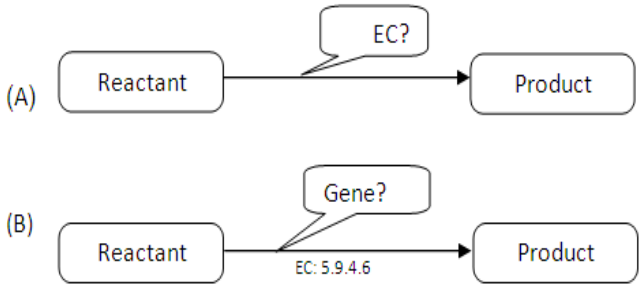


FIGURE 1. PATHWAY HOLE TYPES 1

III. APPROACHES FOR FILLING PATHWAY HOLES

Previous research uses similarity in pathway expression in related organisms to fill pathway holes because similar sequences usually have common descent, and therefore, similar structure and function.

With the emergence of metabolic pathways and their problems like holes, that accompanied the development of some algorithms to solve this problem taking advantage of the great development which computer science has reached, these algorithms depend on some approaches which most of them based on homology searches [3][6].

IV. PATHWAY DATABASES

With the huge amount of molecular data about different organisms has been accumulated and systematically stored in specific databases, in this section we will mention the popular databases. Popular databases are Kyoto Encyclopedia of Genes and Genomes (KEGG), BioCarta, MetaCyc, ExPASy, Reactome, Model SEED, BiGG family [1].

Only two of the listed databases are discussed in some detail: MetaCyc, because it is our choice to be the data source of our data set, that is visualized by our application, and KEGG to show an alternative pathway resource that strongly differs in the representation of the pathways, also because MetaCyc families and KEGG each contain huge amount of data [5][2].

A. *KEGG database*

- KEGG contain more than thousand organisms.
- KEGG database provide more than 9,000 biochemical reactions.
- KEGG database consists of nine components; four of them are major components that are basically self-contained, pathway, ligand, gene and brite**.**
- PATHWAY component in KEGG database hold the pathway maps for different organisms, and the LIGAND database holds chemical information like enzymes, compounds and reactions, the GENE component includes genes and proteins from various species.[5]

B. *Pathway in KEGG*

- KEGG represent metabolic pathway graphs with labeled arcs indicating the involved enzymes, KEGG pathway is a collection of manually drawn pathway maps [8] witch presented infigure2.
- KEGG database enable the user to reach the needed pathway by typing the name of the pathway or by selecting the diseases of these pathway, and the user can filter the result of the query by selecting the organism of the pathway.
- Also we can use KEGG to obtain the reactions that included in the pathway by typing the EC (ex: 1.6.5.1) of this reaction in the search tab.

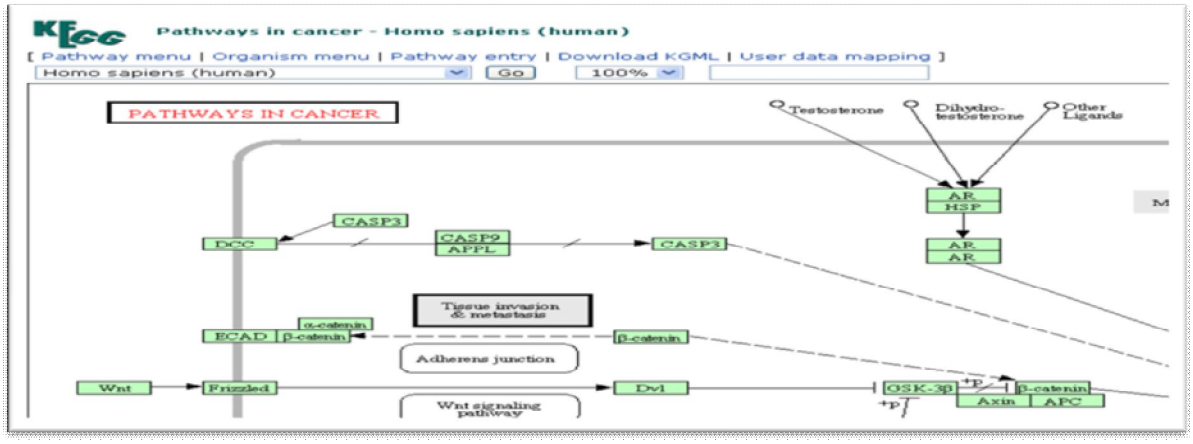


FIGURE 2. ATYPICAL KEGG PATHWAY DIAGRAM

The gene page in KEGG contains more information about the gene, like gene name, pathways that hold this gene, other databases that hold this gene, PDB structure, AA seq.and NT seq.all this information you can gain it by typing the gene name, or gene Id.

C. *MetaCyc database*

MetaCyc database used as a reference to some pathway tools software as pathologic component, which predicts the metabolic network of any organism[7][2], also MetaCyc used by Stanford Research Institute(SRI) because MetaCyc is multi organism database and also contains data that have been experimentally demonstrated in the scientific literature[2].

We can summarize MetaCyc family in several lines:

- MetaCyc contains more than thousand organisms.
- MetaCyc provides more than 9,000 biochemical reactions.
- The different DBs in MetaCyc share a common database schema.
- The same software is used to create, update, and query the DBs in the MetaCyc family.
- MetaCyc contains a rich array of data content for metabolic pathways, reactions, compounds, enzymes, and genes.
- MetaCyc is extensively linked to other biological databases containing protein and nucleic-acid sequence data bibliographic data and protein structures [9].
- Most of data have been manually curated and other curated PGDBs (pathologic database).
- Reactions of MetaCyc directly imported from ENZYME DB.
- MetaCyc ver. 14.6 contains 1,642 metabolic pathways.
- MetaCyc ver. 14.6 contains 8,983 metabolic reactions.
- MetaCyc ver. 14.6 contains 6,912 enzymes.
- MetaCyc ver. 14.6 contains 8,869 compounds [2].

The big advantages of pathway in MetaCyc that MetaCyc inter-relates pathway information (including reactions and their substrates) with genes and their protein products.Figure.3 depicts the hyperlinks that are typically available within MetaCyc, allowing the user to navigate among pathways, genes, enzymes, etc.

MetaCyc has the flexibility in search thus if you don't know the pathway name, you can search by ontology, number of reactions, substrates present and other.

In MetaCyc you can search about reaction by reactant or product, or by genes that behind this reaction.

In MetaCyc you can search the gene by type its name or by typing the EC witch return all genes that catalyze that reaction.

D. *MetaCyc Compared to KEGG*

After presenting in the previous sections KEGG and MetaCyc database, the later is chosen in this research for the following: [9][10]

- MetaCyc version 13.5 (2009) contains 1,399 pathways, compared to the 155 pathways in KEGG version 50 (2009).
- MetaCyc assigns more than twice as many of its reactions (4,950 in version 13.5) to pathways as does KEGG (2,463 in version 50).
- MetaCyc pathways are closer to true biological pathways than are KEGG pathways.
- MetaCyc cites the primary literature sources from which pathway and enzyme data were obtained. KEGG contains no literature citations.

V. RGB MAPS DATABASE

As we mentioned in 'Pathway hole filling approaches' section, the similarity is the key to solving the pathway hole problem, so any researcher needs to do the three phases that are illustrated in figure 4 to collect the data sets required to the research operation.

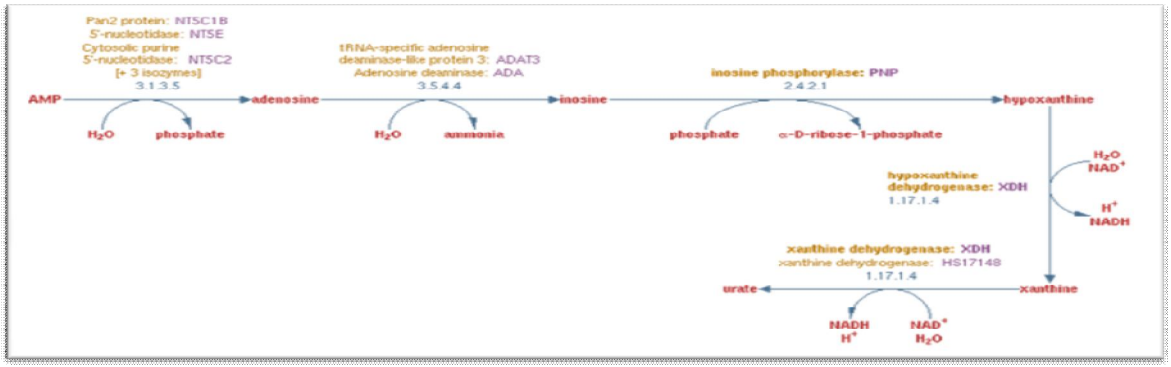


FIGURE 3. A TYPICAL METACYC PATHWAYS DIAGRAM, COMMENTARY AND OTHER DATA THAT IS INCLUDED IN THE PATHWAY PAGE ARE NOT SHOWN

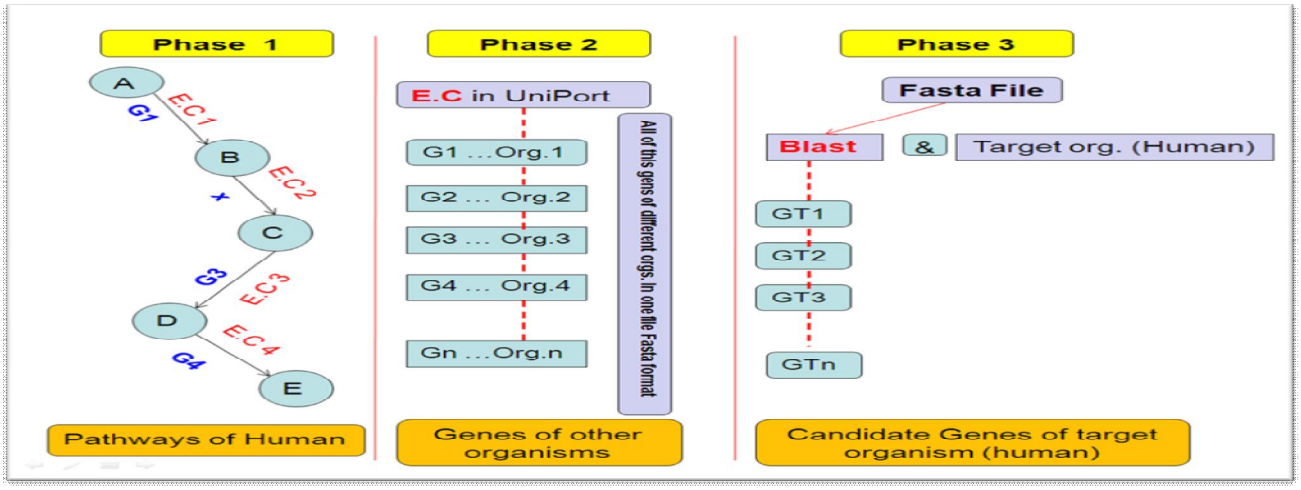


FIGURE 4. THE THREE PHASES OF RGB MAPS DATA COLLECTION

As we see in the figure 4, there is a hole in the second arc (EC2) because the gene of EC2 is not indentified, then we need to find the genes that catalyze this enzyme (EC2) but in related organisms, finally each gene from those related organisms will candidate some genes in target organism (human).

Our dataset that we built depends on three types of data: pathways, reactions and genes.

We collect the pathways and reactions from HumanCyc database (phase1) , The needed reactions are then accessed from the Uniprot database(phase 2),then in phase 3 we use the genes of that candidate organisms in Fasta format and candidate the human genes that are similar to it through BLAST(Basic Local Alignment Search Tool).[16]

In phase 2 we ask ourselves a question: what the other organisms that we will use to make the similarity operation between its genes and the target genome genes?

We note that taxonomy of human more than 100 organisms [11], so now we need strongly to minimize those organisms, our selecting criteria was two things, the first is the Rate and presence of this organism in the Enzyme FASTA file, and the second is the closeness of this organism to the human tree.

RGB MAPS is shortcut name for our database derived from the different seven organisms that we candidate in our database.

- R ……. **R**attus norvegicus
- G ……. **G**allus gallus
- B ……. **B**os taurus
- M …….**M**us musculus
- A …….**A**rabidopsis thaliana
- P ……. **P**ongo abelii
- S ……. **S**accharomyces cerevisiae

A. Collect the pathways

To fetch the different pathways that we need it to feed our dataset, we chose *HumanCyc*database http://humancyc.org/ which is one of the *MetaCyc* family database to be our pathway database, for the reasons that we mention in *MetaCyc* family section. We use Search/Filter by number of reactions to query our pathways, as presented in figure5.

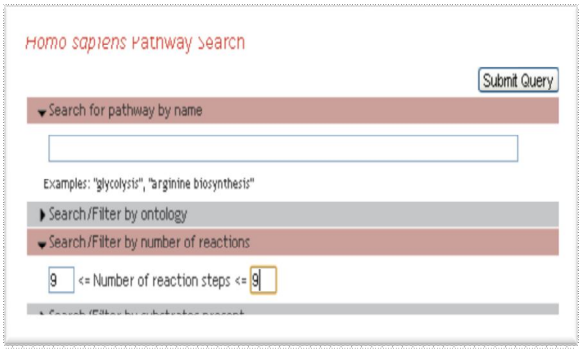


FIGURE 5. FILTER QUERY BY NUMBER OF REACTIONS

The results of this query represent a folder of our database that contains pathways in HumanCyc with a 9 reactions; we called this folder Human_9 the meaning of this name will be explained in: the RGB MAPS database structure section. We repeat this task for different number of reactions.

After that we now finished the first phase of the data collection. RGB MAPS database now contain 100 pathways and 338 reactions, as presented in table 1.

B. Gene collection

After we collect 100 human pathways and 338 reactions, now in this phase we need to collect the genes that catalyze these reactions but in different organisms to make these

genes as input genes in BLAST to hold the candidate genes in human that are similar to these genes.

In this phase we select uniprothttp://www.uniprot.org/to be our knowledgebase as that we mentioned above, because most of the enzymes in MetaCyc are linked to Uniprot [9].

No. of reactions in the pathway	No. of pathways	No. of reactions
9	1	8
8	1	8
7	1	7
6	7	38
5	18	81
4	15	57
3	27	82
2	30	64
	100	338

TABLE1. RGB MAPS PATHWAYS DATABASE

By typing the EC and the organism name in the Query box, (ex: 1.13.11.6 &Mus-musculus) of Uniprot database, the results are available to download in several formats like XML format, Excel Format and Fasta format.

We select the Fasta format, because this is the standard format that the BALST deal with, Fasta format presented in figure 6(A).

In this phase we rename the downloaded file to "1.13.11.6_Mus-musculus.fasta" to be more readable, where 1.13.11.6 refer to the EC in the pathway, Mus-musculus refers to the organism of the genes that are included in the Fasta file.

Repeat the above steps with all reactions in the different organisms.

C. *Candidate Genes of target organism*

- In this phase we take the Fasta file that we download in phase two to be the input file in BLAST to make the similarity process with the target organism, human which the similarity result is our goal in this phase, this operation is presented in figure 6(B).
- In this phase we download the result in Excel format.
- Rename the downloaded file to be more readable ex: "*Afmid_Mus musculus.xls*", where *Afmid* refer to the gene name, *Mus-musculus* refer to the organism name and *xls* refer to the Excel file format **is** presented in figure 6(C).
- The Excel file contains all candidate genes of the similarity process between the different organisms and the target organism, human.
- Each candidate gene represents a row in the Excel file with a several data about this gene like E-Value, Score, Identity, Entry name and Pathway of this gene.
- All these gene information is a Treasure for researchers in the field of pathway hole filing, where they apply some of the computer science approaches, such as Machine Learning, Fuzzy or other to reach to the correct gene from this all candidate genes.

Fig. 6 (A)

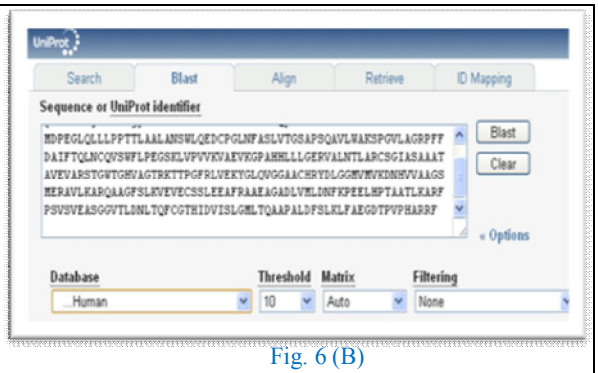
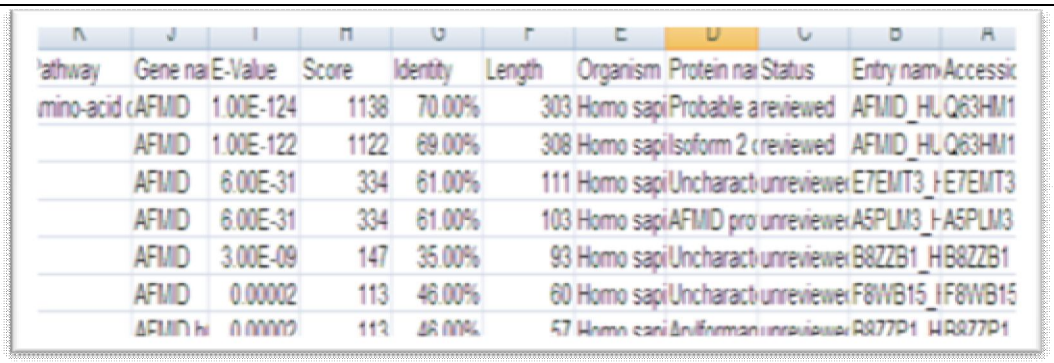


Fig. 6 (B)

Fig. 6 (C)



athway	Gene nai	E-Value	Score	Identity	Length	Organism	Protein nai	Status	Entry nam	Accessic
mino-acid (	AFMID	1.00E-124	1138	70.00%	303	Homo sapi	Probable a	reviewed	AFMID_HU	Q63HM1
	AFMID	1.00E-122	1122	69.00%	308	Homo sapi	Isoform 2 (	reviewed	AFMID_HU	Q63HM1
	AFMID	6.00E-31	334	61.00%	111	Homo sapi	Uncharact	unreviewe	E7EMT3_H	E7EMT3
	AFMID	6.00E-31	334	61.00%	103	Homo sapi	AFMID pro	unreviewe	A5PLM3_H	A5PLM3
	AFMID	3.00E-09	147	35.00%	93	Homo sapi	Uncharact	unreviewe	B8ZZB1_H	B8ZZB1
	AFMID	0.00002	113	46.00%	60	Homo sapi	Uncharact	unreviewe	F8WB15_H	F8WB15
	AFMID hu	0.00002	113	46.00%	57	Homo sapi	Arylformam	unreviewe	B8Z7P1_H	B8Z7P1

FIGURE 6. RGB MAPS DATA COLLECTION PROCESS: FASTA FILE FORMAT, BLAST OPERATION, CANDIDATE GENES OF HUMAN

- All these gene information is a Treasure for researchers in the field of pathway hole filing, where they apply some of the computer science approaches, such as Machine Learning, Fuzzy or other to reach to the correct gene from this all candidate genes.

D. *RGB MAPS database structure*

- RGB MAPS contain 8 folders its name "Human_9", "Human_8" and … to "Human_2", Human… refer to the target organism of our database, _9 … refer to the No. of reactions in the pathway.
- In each folder we create folders with the different pathway names that have the No. of reactions, that follow the _ in the name of the main folder, (ex: folder with name: acetyl-D-glucosamine biosynthesis II).
- In each folder that holds the pathway name we create a folder with name: Organisms.
- This organisms folder contain a 7 folders, each folder hold a name of the different 7 organisms, as Mus-musculus.
- These folders hold all Fasta files that we download it with the Perl program.
- Finally RGB MAPS database contain 338 Fasta file represent the number of reactions.
- RGB MAPS database contain 2990 input genes.
- RGB MAPS database contain 189328 candidate genes.
- And so what in the RGB MAPS database?
- 100 txt file contain description to 100 pathways, this file download from HumanCyc site.
- 100 pictures that picked as a print screen to the 100 pathway.

VI. EVALUATION

In this section we will note shortly Evaluation to our RGB MAPS database.

Figure 7 presents the data of one pathway from the 100s that in the data set, the first column shows all reactions in this pathway, from column two to column eight shows the number of genes that behind this reaction in the sevens organisms, column nine illustrate the number of gens that will be the BLAST input Fasta gens and the last column illustrate in the last row of it the total number of candidate gens.

E.C	Mus	Rattus	Bos taurus	Gallus	Pongo	Arabidopsis	Saccharomyce	Total IN	Total Cand.
1.13.11.11	30	64	113					3	
3.5.1.9	33						32	2	
1.14.13.9	22	13					28	3	
3.7.1.3	20	14					27	3	
1.13.11.6	14	3	17				20	4	
2.4.2.19	5	21	11				22	4	
2.7.7.18	60	30	54		34			4	
6.3.5.1	17	9	13	8			10	5	
sum	201	154	208	8	34	0	139	28	744

FIGURE 7. RGB MAPS DATA SUMMARIZATION FOR EACH PATHWAY IN THE DATA BASE

No. of reactions in the pathway	No. of pathways	No. of reactions		No. of Input genes		No. of candidate genes
9	1	8	9	36	36	961
8	1	8	8	50	70	1817
7	1	7	7	53	80	1027
6	7	38	42	341	524	25189
5	18	81	90	767	1048	62864
4	15	57	60	374	624	18205
3	27	79	81	829	2058	57094
2	30	60	60	540	1106	22171
sum	**100**	**338**		**2990**		**189328**

TABLE 2: RGB MAPS DATABASE DESCRIPTION (ALL THE DATA BASE)

Tables 2 represent statistical for RGB MAPS at all, where the second column represents the total pathways in our data set and the third column represents the total number of reactions in the data set and the fourth column represents the total genes that act as input to BALST and the last column shows the total candidate genes to all pathways in the data set.

Our database produced from collecting (100) pathways for the human organism, and (338) reactions for these pathways. for each reaction there are (9) genes in average from the related 7 organisms, see figure 8 , finally the RGB MAPS database include for hole (60) human candidates genes in average that could be used to fill the hole of the pathway.

VII. CONCLUSION

No one can deny that all steps that done in phase two of collection data to make in manually way, that is very hard, waste effort and time, we have overcome the this problem by writing a small Perl program to make these steps easier.

We make BLAST with the amino acid sequence (AA Seq.) not by the nucleotide sequence (NT Seq.)Because the hole in pathway interested in the function (AA Seq.) not by the NT Seq.

Some Reactions acts with the same gene ,Ex : 1.14.15.4 and 1.14.15.5act with the same gene , 1.1.1.145 and 5.3.3.1act with the same gene,2.1.2.2 and 6.3.3.1 and 6.3.4.13act with the same gene, this note May be very useful in gene therapy.

APPENDIX A

To make our dataset easier to use we developed an application to facilitate the dataset use and to make it visual, here are some notes about our application.
The application consists of one form divided into four forms:

- From one show the reactions and genes of the selected pathway.
- Form two holds the pathway picture.
- Form three shows the genes in the seven organisms in a specific EC of the pathway.

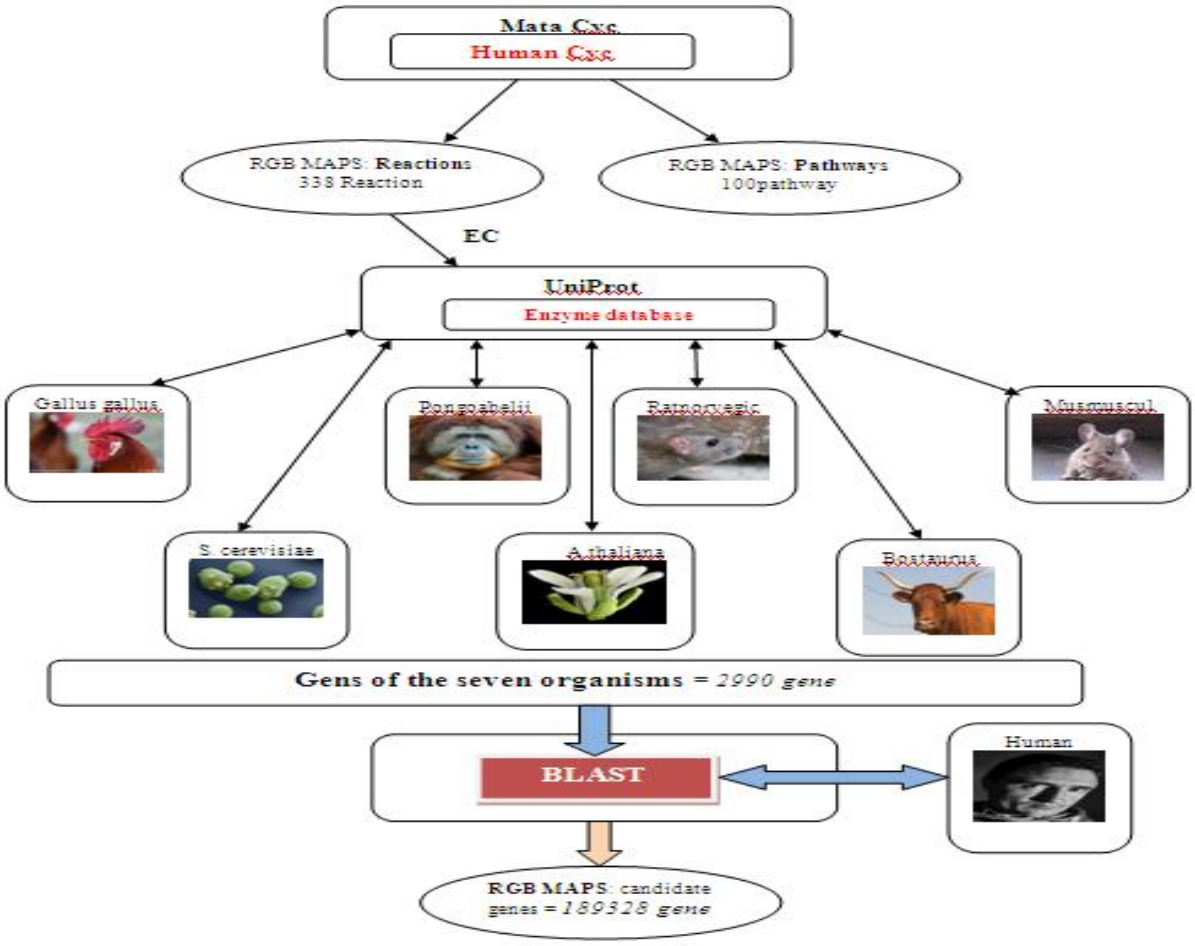
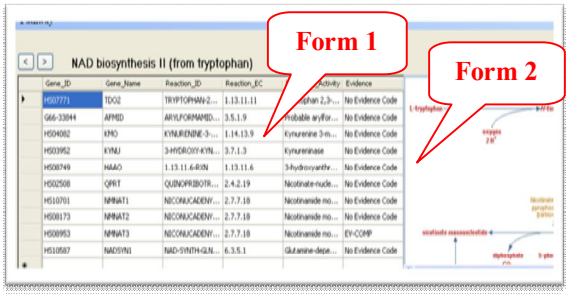


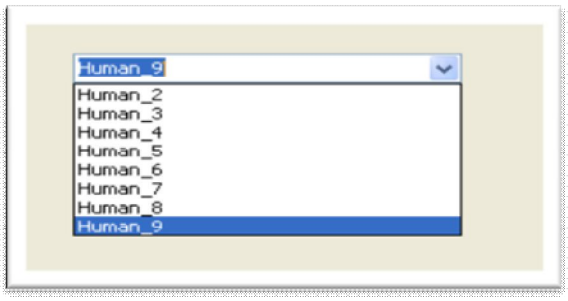
FIGURE 8.RGB MAPS EVALUATION

- Form four shows the candidate genes that BLAST produced it as a result to the similarity phase between the input gene in each organism of the related organisms and with the target organism (human).

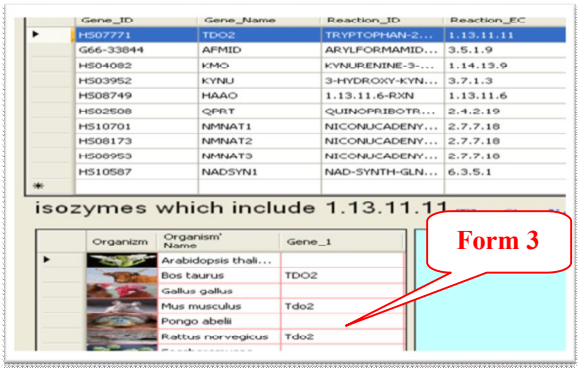


The steps to use RGB MAPS application:

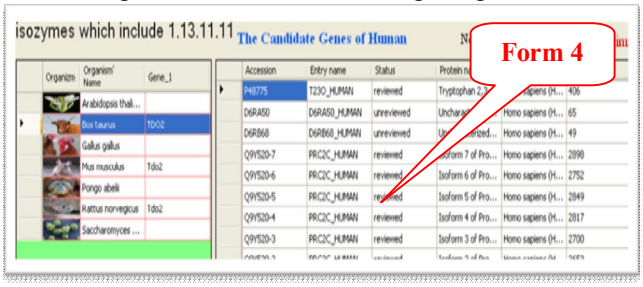
- Select the pathway by number of reactions.



- Click on the row in form 1 to explain the genes that catalyze that EC in the seven organisms.



- Click on the row in form 3 to show in form 4 the candidate genes that BLAST produced as a result of the similarity process between the related organisms and human as a target organism.



REFERENCES

[1] Karp PD, Caspi R:A survey of metabolic databases emphasizing the MetaCyc family,Arch Toxicol2011.

[2] Green ML, Karp PD:A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases, BMC Bioinformatics 2004, 5:76.

[3] Romero P, Wagg J, Green ML, Kaiser D,KrummenackerM, KarpPD: Computational prediction of human metabolic pathways from the complete human genome,Genome Biology 2004, 6:R2.

[4] Hofestaedt R, Chen M: Metabolic pathway perdition/alignment, BGRS 2004.

[5] Marc S, Bakk r:Metabolic Pathway Visualization Using Gene-Expression Data, Master's Thesis 2007.

[6] YamanishiY, AttoriM, KoteraM,Goto S,KanehisaM:E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs, Bioinformatics2009.

[7] Karp PD, Paley SM,Krummenacker M, LatendresseM, Dale JM,Lee TJ, Kaipa P, GilhamF,Spaulding A, Popescu L, Altman T, Paulsen L, Keseler IM,CaspiR:PathwayTools version 13.0: integrated software for pathway/genome informatics and systems biology, Bioinformatics 2009.

[8] http://www.genome.jp/kegg/pathway.html

[9] http://metacyc.org/

[10] Caspi R, AltmanT, Dale JM, Dreher K, Fulcher CA, GilhamF, KaipaP, S. arthikeyan A, Kothari A, Krummenacker M, Latendresse MMueller LA, Paley SM, Popescu L, Pujar AShearer AG, Zhang P,Karp PD: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, Nucleic Acids Research, 2010, Vol. 38, Database issue2009.

[11] http://www.uniprot.org/

[12] Dale JM, Popescu L, Karp PD:Machine learning methods for metabolic pathway Prediction, BMC Bioinformatics 2010, 11:15.

[13] Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, LatendresseM,Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC , Zhang P and Karp PD : The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases ,SRI 2007.

[14] KanehisaM,Goto S, Sato Y,FurumichiM and Tanabe M : KEGG for integration and interpretation of large-scale molecular data sets,2011.

[15] G. Joshi-Tope, M.Gillespie, I.Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono,B. Jassal, G.R. Gopinath, G.R. Wu, L. Matthews, S. Lewis, E. Birney and L. Stein :Reactome: a knowledgebase of biological pathways,2004.

[16] http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=ProgSelectionGuide.