The Effectiveness of Organism Selection in Filling Metabolic Pathway Hole Problem

Ahmed Farouk Al-Sadek [1,2], Alaa Eldin Abdallah Yassin[1]

1Central Lab for Agricultural Expert Systems, Giza, Egypt
2Faculty of computer science, October University for modern science and Art

afsadek@gmail.com, aboelmnzer@gmail.com

Abstract. *In recent years with the huge amount of biology data in bioinformatics field, especially with the Human Genome Project, urgent needs to analyze this data to exploit optimization. Biology data characterized from other data, it directly affects the human life dramatically and significantly. In bioinformatics field there are a lot of problems need to be solved. One of the most important problem is metabolic pathway hole problem, where solving this problem helps the biologist to set the correct gene in a pathway which have a hole where the path of this pathway is unknown in some parts of it, to use these result is several useful application as gene therapy. Until now there are no enough researches to solve missing gene problem. Previous researches used BLAST as the most popular similarity tool because similar sequences usually have common descent, and therefore, similar structure and function, but these researches select some organisms from the huge amount of available organisms. In this paper we will introduce our observations of the role of organism selection and how this selection affects on the results of filling pathway hole.*

Keywords

Metabolic pathway. Bioinformatics. Pathway hole. RGBMAPS database. BLAST.

1 Introduction

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 [1, 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 gene(s) that produce this enzyme.

With the up growth 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 [3,4]. If the sequences are similar, this means that they often derive from the same ancestral sequence, which means that, they probably have the same ancestor, share the same structure, and have a The importance to know this, that we can extrapolate data we know about a particular DNA or protein sequence to all similar DNA and protein sequences.[5].

Because previous researches used BLAST as the most popular similarity tool using some organisms in the similarity process as Ahmed ElSadek, Laila ElFangary and Alaa.Yassin team[1] used the seven organisms of RGBMAPS database [6]and done their own algorithm to solve pathway hole problem. In this paper we will focus on how the organism selection play an important role in the filling hole results.

2 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 [6, 7], 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).

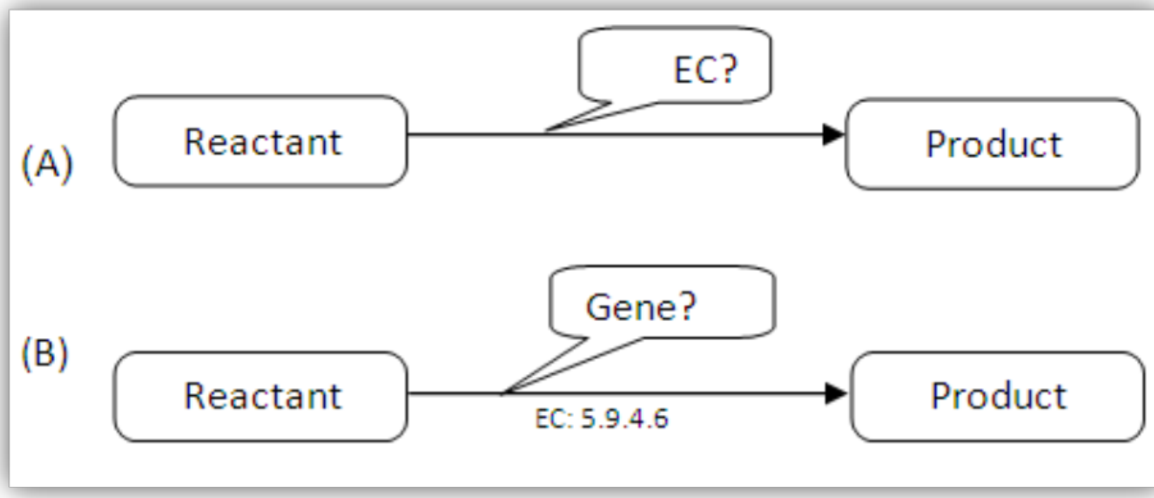

Fig.1 . Pathway hole types.

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 [8, 9].

3 BLAST and filling pathway hole

In the previous works to fill pathway hole, the researches depends on BLAST to do the similarity phase of their work after applying their own algorithms. The researchers select specific organisms to use it in the similarity process between these organisms and target organism.

Here we want to know: does the organism selection have an effects on the result, which used in filling hole problem or not? To answer this question we applied BLAST on 70 different enzymes which missed its genes. The similarity process occurred with the seven organisms, **R**attus norvegicus, **G**allus gallus, **B**os taurus, **M**us musculus, **A**rabidopsis thaliana, **P**ongo abelii and **S**accharomyces cerevisiae which included in RGBMAPS database [6] and derived its name from the first letter of each organism.

Table 1 presents the similarity process results with the selected organism and we captured the first hit of the BLAST result because, BLAST ranks the result according to the best E-Value. The similarity process repeated with the seven organisms in the 70 enzymes, so the similarity process repeated 490 times .column 2 of table 1 represents EC of the enzyme; column 3 represents the real pathway gene of this enzyme in the target organism which is human in our case, from column 4 to column 10 represents the genes that catalyze this enzyme but in the different organisms and the last column represents the candidate gene after shot-gun score voting.

Tab.1: Similarity process results in the seven organisms using BLAST

#	*EC*	*Pathway genes*	*Arab.*	*Bos.*	*Gallss*	*Mus*	*pongo*	*rate*	*Scr.*	*Candidate gene*
1	2.3.1.61	DLST	DLST	DLST		DLST		DLST	DLST	DLST
2	1.2.4.1	PDHA1	PDHB	PDHB		PDHA1	PDHB	PDHA1	PDHA1	PDHA1
3	1.8.1.4	DLD	DLD			DLD	DLD	DLD	DLD	DLD
4	2.3.1.12	DLAT	DLAT		DLAT	DLAT		DLAT	DLAT	DLAT
5	4.2.1.47	GMDS	GMDS			GMDS				GMDS
6	1.2.4.4	BCKDHA		BCKDH B		BCKDHA		BCKDH A		BCKDHA
7	2.3.1.168	DBT		DBT		DBT				DBT
8	2.4.1.174	CSGALN ACT1				CSGALN ACT1				CSGALNA CT1
9	2.4.1.175	CHSY**1**				CHSY**3**				CHSY**3**
10	2.4.1.226	CHSY**1**				CHSY**3**				CHSY**3**
11	2.7.7.15	PCYT1B				PCYT1A		PCYT1 A	PCYT1 A	PCYT1A
12	2.7.8.2	CEPT1		CHPT1	CEPT1	**PCYT1A**		CEPT1	**EPT1**	CEPT1
13	3.1.4.4	PLD1	PLD2	PLD2		PLD1	PLD3	PLD1	PLD2	PLD2
14	1.14.13.39	NOS1	**NOA1**	NOS3	NOS2	NOS2		NOS1		NOS2
15	6.3.4.5	ASS1	ASS1	ASS1	ASS1	ASS1		ASS1	ASS1	ASS1
16	4.3.2.1	ASL		ASL	ASL	ASL		ASL	ASL	ASL
17	2.5.1.21	FDFT1	FDFT1	FDFT1		FDFT1	FDFT1	FDFT1	FDFT1	FDFT1
18	1.14.99.7	SQLE	SQLE			SQLE		SQLE	SQLE	SQLE
19	1.1.1.1	ADH1B	ADH5	ADH5	**ADH1C**	**CACNA2 D2**	**ADH6**	**CACN A2D2**	ADH1B	**CACNA2D 2** ADH5
20	1.2.1.3	ALDH2	ALDH3A 1	ALDH2		ALDH2	ALDH 1B1	ALDH2	ALDH2	ALDH2
21	6.2.1.1	ACSS1		ACSS3		ACSS1	ACSS3		ACSS2	ACSS1
22	1.11.1.6	CAT	**ALDH3A 1**	CAT		CAT	CAT		CAT	CAT
23	5.3.3.2	IDI1	IDI1	IDI1		IDI1	IDI1	IDI1	IDI1	IDI1
24	2.5.1.1	FDPS	FDPS	GGPS1	FDPS	FDPS		FDPS	GGPS1	FDPS
25	2.5.1.10	FDPS	FDPS	GGPS1	FDPS	FDPS		FDPS	GGPS1	FDPS
26	4.1.1.15	GAD1	**SGPL1**	**GLUL**	**GLUL**	GAD1	GAD1	GAD2	**SGPL1**	GAD1 **SGPL1**
27	1.2.1.24	ALDH5A 1	ALDH5A 1			ALDH5A 1		ALDH5 A1		ALDH5A1
28	2.6.1.19	ABAT		ABAT		ABAT		ABAT	ABAT	ABAT
29	1.11.1.9	GPX1	**GPX4**	GPX1		GPX1		GPX1	**GPX4**	GPX1
30	1.8.1.7	GSR	GSR			GSR		GSR	GSR	GSR
31	1.11.1.12	GPX4	GPX4	GPX4		GPX4		GPX4		GPX4
32	1.4.4.2	GLDC	GLDC		GLDC	GLDC			GLDC	GLDC
33	2.1.2.10	AMT	AMT	AMT	AMT	AMT			AMT	AMT
34	1.8.1.4	DLD	DLD			DLD	DLD	DLD	DLD	DLD
35		BDH1		BDH1	BDH1	BDH1		BDH1		BDH1
36	2.8.3.5	OXCT**1**				OXCT**2**		OXCT**2**		OXCT**2**
37	2.3.1.9	ACAT**1**	ACAT**1**	ACAT**1**		**ACAT2**		**ACAT2**	ACAT**1**	ACAT**1**
38	6.4.1.3	PCCB		PCCB		PCCB		PCCA		PCCB
39	5.1.99.1	MCEE		**ACAT1**		MCEE				MCEE ACAT1
40	5.4.99.2	MUT		MUT		MUT	MUT			MUT
41	2.7.1.23	NADK	NADK			NADK			NADK	NADK
42	3.1.3.2	ACP**6**	**PAPL**	**ACP1**	**ACP1**	ACP6	**ACP2**	**ACP5**	MINPP 1	ACP1
43	1.6.1.2	NNT		NNT		NNT				NNT
44	1.1.1.49	G6PD	G6PD			G6PD		G6PD	G6PD	G6PD

#	*EC*	*Pathway genes*	*Arab.*	*Bos.*	*Gallss*	*Mus*	*pongo*	*rate*	*Scr.*	*Candidate gene*
45	3.1.1.31	PGLS	PGLS	PGLS		PGLS		PGLS	PGLS	PGLS
46	1.1.1.44	PGD				PGD		PGD	PGD	PGD
47	1.14.16.1	PAH		PAH		PAH		PAH		PAH
48	4.2.1.96	PCBD1		PCBD1	PCBD2	PCBD2	PCBD2	PCBD1		PCBD1 PCBD2
49	1.5.1.34	QDPR		QDPR		QDPR		QDPR		QDPR
50	2.7.1.32	CHKA				CHKA		**PCYT1 A**	CHKB	CHKA PCYT1A CHKB
51	2.7.7.15	PCYT1A				PCYT1A		PCYT1 A	PCYT1 A	PCYT1A
52	2.7.8.2	CHPT1		CHPT1	CEPT1	**PCYT1A**		CEPT1	**EPT1**	CEPT1
53	2.7.1.82	CHK**B**				**CHKA**		**CHKA**	**CHKA**	**CHKA**
54	2.7.7.14	PCYT2		PCYT2		PCYT2		PCYT2	PCYT2	PCYT2
55	2.7.8.1	**C**EPT1		**EPT1**	CEPT1	**EPT1**	**EPT1**	CEPT1	**EPT1**	**EPT1**
56	3.5.4.16	GCH1			GCH1	GCH1		GCH1	GCH1	GCH1
57	4.2.3.12	PTS				PTS	PTS	PTS		PTS
58	1.1.1.153	SPR		SPR		SPR		SPR		SPR
59	1.3.1.2	DPYD		DPYD		DPYD	DPYD	DPYD		DPYD
60	3.5.2.2	DPYS				DPYS		DPYS		DPYS
61	3.5.1.6	UPB1				UPB1	UPB1	UPB1		UPB1
62	1.2.1.18	ALDH6A 1		ALDH6 A1		ALDH6A 1		ALDH6 A1		ALDH6A1
63	2.6.1.1	GOT1	GOT2	GOT1	**MDH1**	GOT1	GOT1	GOT1	GOT1	GOT1
64	1.1.1.37	MDH2	MDH2	**GOT2**	MDH1	MDH2	MDH2	MDH2	MDH2	MDH2
65	1.2.1.8	ALDH7A 1	**ALDH2**	ALDH7 A1		ALDH7A 1		ALDH7 A1		ALDH7A1
66	1.1.99.1	CHDH				**ALDH7A 1**				**ALDH7A1**
67	2.3.1.38	FASN		FASN	FASN	FASN		FASN	**HSD17 B4**	FASN
68	2.3.1.41	OXSM	OXSM	FASN	FASN	FASN		FASN	**SYBU**	FASN
69	2.7.1.26	RFK				RFK			RFK	RFK
70	2.7.7.2	FLAD1				FLAD1	FLAD1		FLAD1	FLAD1

Results: by comparing column 3 and column 10 we can calculate the total accuracy of BLAST, where the first one represents the real gene and the other one represents the candidate one after applying shotgun score on the seven organisms. So the total accuracy of BLAST on 70 enzymes = the number of correct genes / the number of real genes = 61/70 = 87%.

4 Observations

In the previous section we showed the total results of BLAST after applying shot-gun score , now we will answer the question we ask before , does the organism selection have an effects on the result, which used in filling hole problem or not? To answer this question we summarize table 1 in table 2, showing the results of BLAST in each organism.

As shown Table and figure 2 answer the question clearly , where the second column in table 2 represents the number of correct genes which BLAST give in this organism , the third one give us the number of error results , the fourth column represents the number of cases which this organism cant candidate genes at all to this enzyme , the five column sum the two previous columns and the last one with the accuracy label represent the accuracy of this organism with the 70 enzyme by divide the number of the correct genes in the second column on 70.

Tab. 2: BLAST results in each organism

organism	*No. of correct genes*	*No. of error genes*	*Not found genes*	*Total Error*	*accuracy*
Arab.	27/70	6/70	37/70	43/70	38.5%
Bos.	39/70	5/70	26/70	31/70	56%
Gallss.	17/70	4/70	49/70	53/70	24%
Mus.	63/70	7/70	0/70	7/70	90%
Pongo.	18/70	3/70	49/70	52/70	26%
Rate.	49/70	5/70	16/70	21/70	70%
Scr.	36/70	8/70	26/70	34/70	51%

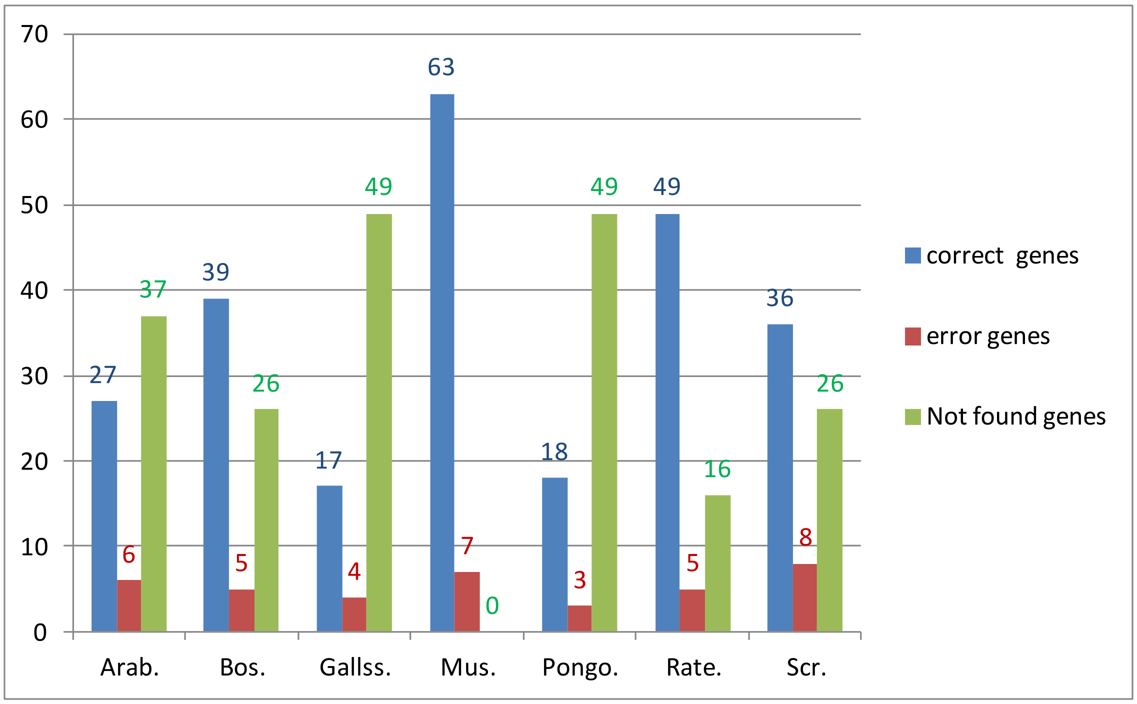

Fig. 2: chart of the seven organisms using BLAST.

The big notation which appears on the results above, that the number of not found genes affect on the final accuracy. Table 3 represents the percent of completeness data of each organism.

Tab. 3: The percent of completeness data of each organism

organism	*Found /total*	*percent*
Arabidopsis thaliana	33/70	47%
Bos taurus	44/70	63%
Gallus gallus	21/70	30%
Mus musculus	70/70	100%
Pongo abelii	21/70	30%
Rattus norvegicus	54/70	77%
Saccharomyces	44/70	63%

5 Organisms Ranking

From table 2 and 3, we observe that ranking of the organisms by the accuracy are equally likely to the ranking by the completeness of its data as presented in table 4.

Tab. 4: Organisms ranking summarization

#	*Ranking by the accuracy*	*%*	*Ranking by completeness of data*	*%*
1	Mus musculus	90%	Mus musculus	100%
2	Rattus norvegicus	70%	Rattus norvegicus	77%
3	Bos Taurus	56%	Bos Taurus	63%
4	Saccharomyces	51%	Saccharomyces	63%
5	Arabidopsis thaliana	38.5%	Arabidopsis thaliana	47%
6	Pongo abelii	26%	Pongo abelii	30%
7	Gallus gallus	24%	Gallus gallus	30%

From table 4 we observe that, the organism selection affect directly on solving filling pathway hole problem, where the organisms which in the same taxonomy with human give a good results as Mus musculus, Rattus norvegicus and Bos Taurus, but we must keep in mind the data size these organisms, because we observed that some organisms are fare from human in taxonomy like Saccharomyces and Arabidopsis thaliana but give better results than other organisms which are close to human like Pongo abelii , the reason is the data size.

6 Conclusion

We advice the researchers who need to try to solve pathway hole problem to select the organisms which are very close in taxonomy to the target organism and also have a suitable data size as Mus musculus and Rattus norvegicus, and also they may have a good results with the organisms which have a big data size regardless the taxonomy factor.

References

[1] Ahmed ElSadek, Laila ElFangary and Alaa.Yassin : “*Fuzzy-based Approach for Filling the Metabolic Pathway Hole*”, International Journal of Computer Science Issues , IJCSI, Volume 9, Issue 6, November,2012'.

[2] Marc S, Bakkr “*Metabolic Pathway Visualization Using Gene-Expression Data*”, Master’s Thesis 2007,Institute for computer Graphics and Vision Graz University of Technology ,Graz.

[3] Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD,” *Computational prediction of human metabolic pathways from the complete human genome*”, Genome Biology 2004, 6:R2.

[4] Yamanishi Y, Attori M, Kotera M,Goto S, Kanehisa M:”***E-zyme:predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs***”, Bioinformatics,2009.

[5] Jean-Michel Claverie,and Cedric Notredame, “***Bioinformatics for Dummies***” – 2nd Edition,2007. Published by Wiley Publishing, Inc.

[6] Alaa.yassin, Laila.Elfangary and Ahmed ElSadek : “***RGB MAPS: a Proposed Database for solving Metabolic Pathway Hole***”, The 8th International Conference on Informatics and Systems (INFOS2012) ,Cairo – Egypt, 14 – 16 May, 2012,pp 63.

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

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

[9] 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.