DMPFinder - Finding differentiating pathways with gaps from two groups of metabolic networks*

Henry C.M. Leung
Department of Computer Science
The University of Hong Kong
Hong Kong
cmleung2@cs.hku.hk

S.Y. Leung
Department of Computer Science
The University of Hong Kong
Hong Kong
syleung@cs.hku.hk

Carlos L. Xiang
Department of Computer Science
The University of Hong Kong
Hong Kong
lxiang2@cs.hku.hk

S.M. Yiu
Department of Computer Science
The University of Hong Kong
Hong Kong
smyiu@cs.hku.hk

Francis Y.L. Chin
Department of Computer Science
The University of Hong Kong
Hong Kong
chin@cs.hku.hk

Abstract

Why some strains of a species exhibit a certain phenotype (e.g. drug resistant) but not the other strains of the same species is a critical question to answer. Studying the metabolism of the two groups of strains may discover the corresponding pathways that are conserved in the first group but not in the second group. However, only a few tools provide functions to compare two groups of metabolic networks which are usually limited to the reaction level, not the pathway level.

In this paper, we formulate the DMP (Differentiating Metabolic Pathway) problem for finding conserved pathways exist in first group, but not the second group. The problem also captures the mutation in pathways and derives a measure (*p*-value and e-score) for evaluating the confident of the pathways. We then developed an algorithm, DMPFinder, to solve the DMP problem. Experimental results show that DMPFinder is able to identify pathways that are critical for the first group to exhibit a certain phenotype which is absent in the other group. Some of these pathways cannot be identified by other tools which only consider reaction level or do not take into account possible mutations among species. The software is available at:
http://i.cs.hku.hk/~alse/hkubrg/projects/DMPFinder/

1. Introduction

Metabolism refers to the set of cellular processes. These processes are not isolated events, but interrelated and can be modeled by a metabolic network. A metabolic network captures the set of chemical reactions among substrates, compounds and enzymes that represent the metabolism within a cell. Conceptually, a metabolic network can be divided into functional pathways corresponding to different metabolic activities in the cell.

Some important metabolic activities that lead to a specific phenotype of a species, e.g. the drug resistance property of a pathogenic bacterium, cannot be identified easily from the metabolic network of the species. However, as more and more information about metabolic networks is now available in databases such as KEGG [1] and BioCyc [2], comparative analysis can be a promising direction. Previous studies have shown that by comparing metabolic networks from different species, it is possible for scientists to gain better understanding on the cellular machinery, the evolutionary events or even the pharmacology (drug design). For instance, Dandekar *et al.* did one of the earliest comparative analyses on glycolytic metabolic pathway, which reveals the plasticity of the pathway among different species [3]. Some other groups also tried to reconstruct the phylogenetic trees using metabolic networks and studied the impacts on a shift in the network during the evolution [4]. Thus, computational tools are needed for comparing two groups of metabolic networks.

Given the metabolic networks of two groups of species (or strains) with one known to have the phenotype while the other does not, we study the problem of identifying the conserved pathways (or sub-pathways) which exist in the first group but not the other. These sub-pathways may be critical for the phenotype being studied which can be further investigated by biologists. Solving this problem would also provide important insights to areas such as metabolic network engineering in synthetic biology and pharmacology. For instance, when biolgoists try to import new biological function, *e.g.* oxidation of methane activity found in methanotrophic bacteria, into a target engineering bacteria, including only the enzyme for known central reaction (monooxygenase in this case) is most likely not sufficient. Our method will help to identify all the sub-pathways that are unique in those methanotrophic bacteria when comparing to normal species. Therefore, those pathways which ensure the central conversion to take place can be found. Also in the pathogenic study, one may apply our method to find how

*This research is partially supported by HK GRF grant (HKU 7116/08E).

one species or strain gains or loses its pathogeny from its closely related species and the resulted pathways can always be good targets for the drug design purposes.

There exist a few tools that can compare two groups of metabolic networks. Most of them work on the reaction level, not on the pathway level. Clemente *et al.* [5] take into account the mutations that may occur in the species and defines a similarity measure to capture conserved reactions. However, they do not consider the mutation at pathway level. BioCyc [6] provides a more comprehensive set of comparative tools (Pathway Tools) for researchers to compare groups of metabolic networks. However, they do not emphasize on finding conserved reactions/pathways with mutations and only regard identical reactions to be conserved. Other related work (*e.g.* [7-9]) is mainly on identifying conserved metabolic pathways in multiple metabolic networks with or without given a query pathway, but not on comparing two groups of networks to identify sub-pathways that are conserved in one group, but not the other.

There are two recent works on comparing two groups of networks. Kastenmüller *et al.* [10] determine if any known pathway occurs commonly in one group, but not the other based on differences of the occurrences of the reactions inside the pathway. Although the method works well in known pathways, it cannot discover de novo pathways or critical sub-pathway in a known pathway. Instead of determining known pathways, Schmidt *et al.* [11] starts with a set of reactions that are more common in one group but rare in another group and expand these reactions into pathways without considering whether the expanded pathways are biased in one group. Thus, their approach is still based on reaction level and cannot discover pathways exist in one group but not the other. Also, as they do not consider mutations in pathway, some important pathways may be missed. There are other works (e.g. [12-18]) focus on the association between genotypes and phenotypes. However, they do not consider the networking effect among the genes.

Our contributions: In this paper, we formulate a computational problem, called Differentiating Metabolic Pathway (DMP) problem. Given two groups of metabolic networks, DMP problem is to identify sub-pathways (called differentiating pathways) that are conserved (not identical, but have the same initial substrates and resulting products and some intermediate compounds) in the first group of metabolic networks which do not exist in the second group of networks. We provide a solution, DMPFinder, to solve the problem and derive a measure (e-score) to evaluate how likely a pathway is biased to one group than the other by random in order to identify those significant pathways. We implemented our algorithm and evaluated the performance of our solution on nine cyanobacteria which can perform photosynthesis (the first group) and nine heterotrophic bacteria that cannot perform photosynthesis (the second group) based on two databases (KEGG and BioCYC). We successfully identified pathways which are related to photosynthesis. Some of these pathways cannot be found by only considering the reaction levels or without taking the pathway mutations into the model.

2. Methods

In this section, we define the *Differentiating Metabolic Pathway* (DMP) problem for finding metabolic pathways frequently exist in a group of species with a particular phenotype and rarely exist in another group of species without the phenotype. We first introduce the graph representation of metabolic pathway and metabolic network. Then we describe how to determine whether a pathway exists in a metabolic network by constructing building blocks. Last, we calculate a *p*-value and e-score to evaluate the significance of the relationship between a pathway and a phenotype. We have also developed an algorithm called DMPFinder for finding pathways with high relationship with a phenotype using $O(nb^{l/2}sgl+n^3)$ time where n is the maximum number of reactions and compounds in an input metabolic network, b is the largest branching factor in the network, s is the number of species in the two groups, g is maximum gap size and l is the length of the pathway under investigation.

2.1 Differentiating Metabolic Pathway Problem

A metabolic reaction converts a set of substrates into a set of products catalyzed by some enzyme(s). It can be represented by a directed graph where each compound and reaction is represented by a node and there is an edge from a compound node to a reaction node if the compound is a substrate of the reaction and there is an edge from a reaction node to a compound node if the compound is a product of the reaction. For those reversible reactions where substrates and produces can be converted to another by the same reaction, we treat it as two reactions and represent it by two reaction nodes.

Metabolic reactions do not work alone. Several reactions can work together by first converting a set of substrates to some intermediate compounds and then converting them to other products by different reactions. The set of metabolic reactions corresponding to a particular function are called *metabolic pathway*. In general, a metabolic pathway can be represented by a connected component with the products of some reactions being the substrates of other reactions. In this paper, we focus on linear pathways where the substrates (products) of a reaction are the products (substrates) of at most one reaction in the pathway. However, the DMP problem and DMPFinder can be extended to model more complicated pathways easily. Similarly, a set of metabolic reactions occur in a species can be represented by connected

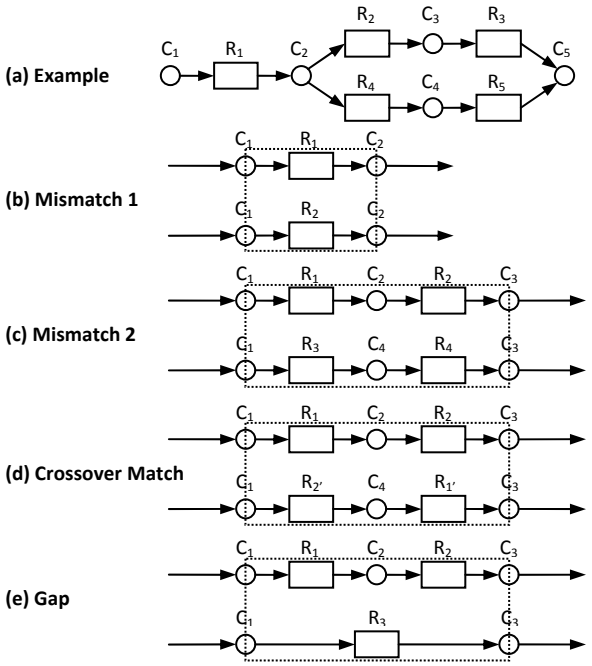

Figure 1 (a): An example of metabolic network. (b) – (e): All penalty blocks with gap size at most 4. For crossover match, Reaction 1 (Reaction 2) and Reaction 1' (Reaction 2') are reactions using similar enzymes, i.e., the penalty block represents two chains of reactions for producing compound C_3 from compound C_1 with different order of reactions. For gap and mismatch, the penalty block represents two different chains of reactions for producing compound C_3 from compound C_1.

components, called *metabolic network* of the species. Figure 1 (a) shows an example of a metabolic network.

Given a metabolic pathway *P* and a metabolic network *G*, we want to determine whether *P* exists in *G*. Because of evolution events, there are many cases that although pathway *P* does not exist in *G*, another pathway *P'* with the same set of substrates and products as *P* exists in *G*. In order to capture these evolution event, we applied the concept of metabolic network alignment using building block [7]. A building block is two aligned sub-paths such that the first nodes of both sub-paths refer to the same substrate compound *u* and the last nodes refer to the same substrate compound *v*. A building block is an identical building block if the lengths of the sub-paths are exactly two (i.e. represents exactly one reaction) and both reactions are catalyzed by the same enzyme. In order to capture the diversities such as gaps [19], mismatches and crossover mismatches [20-21], a penalty building block is defined as follows. A building block is a penalty building block if 1) both sub-paths are of length two and the reactions are catalyzed by different enzymes, or 2) the length of at least one sub-paths are larger than 2. The gap size of a building block is equal the length of the longer sub-path. Figure 1 (b) – (e) show the penalty blocks with gap sizes less than or equal to 4. Given a linear metabolic pathway *P* and a metabolic network *G*, *P* is considered existing in *G* if there is a path *P'* in *G* such that *P* and *P'* can be divided into sub-paths which can be aligned in order with at most *p* penalty blocks each have a maximum gap size of *g*. *p* and *g* are some predefined parameters.

Given a metabolic pathway *P* and a set of metabolic networks *T* and *F* from a set of species with and without a particular phenotype. Assume *P* exists in *t* out of |*T*| networks in *T* and exists in *f* out of |*F*| networks in *F*, we can evaluate whether pathway *P* is related to the phenotype by comparing the values of *t*, *f*, |*T*| and |*F*|. *p*-value is defined as the probability that *P* exists in *t* or most networks in *T* under the null hypothesis that *P* is not related to the phenotype and exist in each network with equal probability.

$$p\text{-value} = \sum_{i=t}^{\min\{t+f,|T|\}} \binom{|T|}{i}\binom{|F|}{t+f-i} \Big/ \binom{|T|+|F|}{t+f}$$

When *p*-value is small, it means that pathway *P* exists in relatively more networks in set *T* than in set *F*. The null hypothesis is likely to be incorrect and *P* may relate to the phenotype. However, it is difficult to justify whether a *p*-value is small or not. Therefore, we further define the e-score(*l*) of a length-*l* pathway *P* which is the expected number of length-*l* pathways with *p*-value smaller than or equal to the *p*-value of *P*. *e*-score(*l*) equals the multiple of *p*-value and the number of length-*l* pathway. When the e-score is smaller than one, it is expected that no pathway with *p*-value smaller than or equal to the *p*-value of *P* by random. Pathway *P* is considered related to the phenotype.

Differentiating Metabolic Pathway (DMP) problem: Given two sets of metabolic networks *T* and *F*, the maximum penalty block *p* and the maximum gap size *g*, identify all linear pathways with e-score < 1.

2.2 DMPFinder Algortihm

We developed the algorithm DMPFinder for solving the DMP problem. It first constructs a metabolic network *U* by combining all metabolic reaction occurs in *T*. For each length-*l* linear pathway *P* in *U*, it checks whether *P* exists in each metabolic network based dynamic programming. Since a pathway obtains the smallest *p*-value when $t = |T|$ and $f = 0$ and the number of length-*l* linear pathways increase with *l*, we can calculate the upper bound of *l* (usually less than 12) such that the e-score of all pathways with length longer than *l* are larger than 1 which do not need to be enumerated. Based on the existences of *P*, DMPFinder calculates the e-score of *P* and outputs it if the e-score is smaller than 1.

For each metabolic network *G* in *T* or *F*, we calculate the pairwise shortest distance matrix *M*, *M*(*u*,*v*) is the shortest distance from compound *u* to compound *v* in *G*, using floyd algorithm [22] in $O(n^3)$ time. Given a linear length-*l* pathway *P*, let *P*[*i*] be the *i*-th node in *P*. *P*[*i*] is a compound node when *i* is odd, otherwise, it is a reaction

Rank	Reaction	Occurrence		Score		Annotation
		Cyan	Other	*p*-value	e-score	Pathways
	KEGG Dataset					
1	D-Ribulose 1,5-bisphosphate -> 3-Phospho-D-glycerate	9	0	2.06×10^{-5}	0.05	Carbon fixation
2	Protochlorophyllide -> Chlorophyllid	9	0	2.06×10^{-5}	0.05	Porphyrin & chlorophyll
3	Phytofluene -> zeta-Carotene	9	0	2.06×10^{-5}	0.05	Carotenoid biosynthesis
4	zeta-Carotene -> Neurosporene	9	0	2.06×10^{-5}	0.05	Carotenoid biosynthesis
5	D-Fructose 6-phosphate -> D-Erythrose 4-phosphate	9(2)	0	2.06×10^{-5}	0.05	Carbon fixation
6	Magnesium protoporphyrin monomethyl ester -> $C_{35}H_{34}MgN_4O_5$[†]	9(3)	0	2.06×10^{-5}	0.05	Porphyrin & chlorophyll
7	$C_{35}H_{34}MgN_4O_5$ -> $C_{35}H_{32}MgN_4O_5$[‡]	9(3)	0	2.06×10^{-5}	0.05	Porphyrin & chlorophyll
8	$C_{35}H_{32}MgN_4O_5$ -> Divinylprotochlorophyllide	9(3)	0	2.06×10^{-5}	0.05	Porphyrin & chlorophyll
	BioCyc Dataset					
9	Protoheme IX -> biliverdin-IX-alpha	9	0	2.06×10^{-5}	0.03	Porphyrin & chlorophyll
10	chlorophyllide a -> monovinyl protochlorophyllide a	9	0	2.06×10^{-5}	0.03	Porphyrin & chlorophyll
11	all-trans-zeta-carotene -> neurosporene	9	0	2.06×10^{-5}	0.03	Carotenoid biosynthesis
12	e^- -> A reduced ferredoxin	9	0	2.06×10^{-5}	0.03	Light reaction
13	Magnesium protoporphyrin monomethyl ester -> $C_{35}H_{34}MgN_4O_5$	8(2)	0	2.05×10^{-4}	0.34	Porphyrin & chlorophyll
14	$C_{35}H_{34}MgN_4O_5$ -> $C_{35}H_{32}MgN_4O_5$	8(2)	0	2.05×10^{-4}	0.34	Porphyrin & chlorophyll
15	$C_{35}H_{32}MgN_4O_5$ -> Divinylprotochlorophyllide	8(2)	0	2.05×10^{-4}	0.34	Porphyrin & chlorophyll

Compound names: [†]$C_{35}H_{34}MgN_4O_5$: 13(1)-Hydroxy-Mg-protoporphyrin IX 13-monomethyl ester, [‡]$C_{35}H_{32}MgN_4O_5$: 13(1)-Oxo-Mg-protoporphyrin IX 13-monomethyl ester.

Table 1 DMPFinder's output at a reaction level. This table shows part of the reactions found by DMPFinder. We count the occurrences of pathways in the two groups of species. In the parenthesis, a similar count based on the profile outputted from BioCyc Comparative Analysis Tool is given when it is different from DMPFinder. Because KEGG does not provide a comparative tool, we have implemented such tool using the same idea as BioCyc's.

node. We fill in a length-l array A where $A[i]$ represent the minimum number of penalty block when we align the length-i prefix of P with any linear pathway in G. Note that we need to fill in $A[i]$ when i is even only because each building block start and end with compound nodes. A[0] = 0 if and only if P[1] exists in G.

$$A[i] = \begin{cases} A[i-2] & P[i-1] \text{ exists in } G \\ \min_{2 \le k \le \min\{g,i\}} \{A[i]+1\} & M(P[i-k], P[i]) = 1 \\ \infty & \text{otherwise} \end{cases}$$

A[l] can be found in O(gl) time. pathway P exists in network G if and only if A[l] $\le p$. It is remark that given a pathway P with e-score $s \le 1$, we may append many different reactions occur frequently in both T and F to construct a longer pathway P' with e-score $s' \le 1$. DMPFinder does not output these redundant pathways P' unless $s' < s$.

3. Results and Discussion

In this section, we test whether DMPFinder can find pathways related to some phenotypes using KEGG databases (release 54.0) [1] and BioCyc databases (release 14.1) [2]. We selected nine cyanobacteria that can perform photosynthesis and nine heterotrophic bacteria that cannot perform photosynthesis. From each database, we extracted all spontaneous and enzymatic reactions of each bacterium to construct a metabolic network. Noted that for each metabolic reaction, we kept the primary compounds defined by the databases and removed those co-factors compounds, e.g. proton, water and NADP. DMPFinder was used to find those metabolic pathways related to photosynthesis from the nine metabolic networks of cyanobacteria T and another nine metabolic networks of heterotrophic bacteria F. The maximum number of penalty blocks allowed is 2 and the maximum gap size is 4.

3.1 Evaluation of DMPFinder

Photosynthesis is commonly known as a process that converts carbon dioxide into carbohydrate using the energy harvested from sunlight. During the photosynthesis, chlorophyll plays a crucial role in capturing the photons and transferring them to the reaction centers [23]. Carotenoids are integral constituents of the reaction centers, they have various functions such as light absorption, photooxydavtive stress protection etc [24]. The "light reactions" is followed by a series of reactions (known as Calvin-Benson-Bassham cycle) which make use of the NADPH and ATP (products of the "light reactions") to fix carbon dioxide into carbohydrate [25]. Therefore, we considered the porphyrin and chlorophyll pathway, carotenoid biosynthesis pathway and carbon fixation pathway as the pathways related to photosynthesis.

3.2 KEGG Dataset

DMPFinder found 56 reactions (length-2 pathways) and 2 linear pathways (pathway with length > 2) with e-score < 1 for the KEGG dataset. Table 1 and Table 2 show the reactions and pathways found by DMPFinder with the lowest *e*-score. 82.1% of the reactions and 100% pathways found by DMPFinder are related to photosynthesis. The rest reactions represent some distinguish metabolic processes in cyanobacteria. For instance, the LL-diaminopimelate aminotransferase

Rank	Pathway	*p*-value	e-score	Annotation
	KEGG Dataset			
1	D-Fructose 1,6-bisphosphate -> D-Fructose 6-phosphate -> D-arabino-Hex-3-ulose 6-phosphate -> D-Ribulose 5-phosphate	2.06×10^{-5}	0.59	Carbon fixation, pentose and glucuronate interconversions & methane metabolism
1	D-Fructose 6-phosphate -> D-arabino-Hex-3-ulose 6-phosphate -> D-Ribulose 5-phosphate -> D-Ribulose 1,5-bisphosphate	2.06×10^{-5}	0.59	Carbon fixation, pentose and glucuronate interconversions & methane metabolism.

Table 2 DMPFinder's output at a pathway level. This table shows the top records have been identified by DMPFinder for a pathway longer than 2. The functional terms were added as a union of all the reactions' terms.

		Occurrence		Score		Annotation
Rank	Reaction	Cyan	Other	p-value	e-score	Pathways
	KEGG Dataset					
1	Sulfite -> Hydrogen sulfide	9	4(0)	1.47×10^{-2}	37.6	Sulfur metabolism
2	Nitrate -> Nitrite	8	5(0)	0.15	376.5	Nitrogen metabolism
3	L-Glutamine -> L-Glutamate	9	7(0)	0.24	602.4	Nitrogen metabolism
4	D-Glyceraldehyde 3-phosphate -> D-Xylulose 5-phosphate	9	9(0)	0	2560	Carbon fixation
5	D-Glyceraldehyde 3-phosphate -> D-Erythrose 4-phosphate	9	9(0)	0	2560	Carbon fixation

Table 3 Reactions found by BioCyc's Comparative Analysis Tool's that may be unrelated to photosynthesis.This table shows a list of reactions that are significant given the BioCyc's Comparative Analysis Tool's output, however, not included in the DMPFinder's output. We count the occurrences of each pathway judged by DMPFinder. In the parenthesis the corresponding count provided by the comparative tool is also given.

reaction has been assigned the lysine biosynthesis function, previous study has shown that it is a trans-kingdom enzyme found not only in plants but also in cyanobacteria [26].

Compared with the BioCyc Comparative Analysis Tool [6] which check whether a particular reaction exists in a particular species, there are 4 reactions (5-th to 8-th) related to photosynthesis found by DMPFinder cannot be found by the BioCyc Comparative Analysis Tool. It is because there are about 7 (out of 9) cyanobacterias do not contain these reactions but have other chains of reactions performing the same functions. Therefore, the BioCyc Comparative Analysis Tool considered only a few cyanbacteria have these reactions while DMPFinder can find these reactions by constructing penalty blocks. For example, BioCyc Comparative Analysis Tool considered that only Anabaena variabilis and Nostoc can convert D-Fructose 6-phosphate to D-Erythrose 4-phosphate which is a critical link in the Clavin-Benson cycle using fructose-6-phosphate phosphoketolase [EC:4.1.2.22] (5-th reaction). However, other cyanobacteria can perform the same reaction using transketolase [EC:2.2.1.1] [27]. For the 6-th, 7-th and 8-th reaction, they reveal some aerobic/anaerobic properties among different species of cyanobacteria because some of them are able to produce Divinyl-proto-chlorophyllide through an anaerobic pathway utilizing enzyme *BchE* [28-29]. Some linear pathways found by DMPFinder cannot be found by the BioCyc Comparative Analysis Tool because each reaction in the pathways has a high e-score and occurs in both cyanobacteria and heterotrophic bacteria. However, when comparing these reactions together, most cyanobacteria have the corresponding pathways which rarely occur in heterotrophic bacteria.

We have also spotted that DMPFinder do not output five reactions that considered as occurring in all cyanobacteria but none heterotrophic bacteria according to BioCyc Comparative Analysis Tool (Table 3). It is because many heterotrophic bacteria can achieve the same reaction by different enzymes. Three out of these five reactions are not related to photosynthesis and the rest two reactions participate in multiple pathways in addition to the carbon fixation pathway.

We have also combined all reactions and linear pathways with e-score < 1 to form 20 connected components (see Supplementary Figure 2(a)[*]). 15 out of these 20 components represent sub-pathways of the photosynthesis pathways. For the rest of them, they are mainly single reactions that cannot be concatenated to form a longer pathway.

3.3 Biocyc Dataset

DMPFinder found 16 reactions with *e*-score < 1 for the Biocyc dataset. Table 1 shows the reactions with the lowest e-score. 87.5% of the reactions found by DMPFinder are related to photosynthesis. Similar as in the KEGG dataset, there are three reactions (13-th to 15-th, same as the 6-th to 8-th) related to photosynthesis found by DMPFinder that cannot be found by the BioCyc Comparative Analysis Tool. Since there are more missing reaction information in Biocyc databases than in KEGG database, DMPFinder cannot find any linear pathway with e-score < 1. After combining these reactions with e-score < 1, we get 7 connected components (see Figure 2(b)). 5 out of there 7 components represent sub-pathways of the photosynthesis pathways.

4. Conclusions

In this paper, we provide a computational tool,

[*]Supplementary Data is available at: http://i.cs.hku.hk/~alse/hkubrg/projects/DMPFinder/

DMPFinder, to identify pathways exist in one group of metabolic networks but rarely occur in another group of metabolic networks. It is believed that these pathways may be critical for certain phenotypes that only exist in the first group of species. The proposed algorithm has taken into account there may be errors in the networks, mutations in the species in the same group, and consider the pathway level instead of reaction level when locating these sub-pathways.

5. References

[1] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Research* 2010, **38**.

[2] Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahrén D, Tsoka S, Darzentas N, Kunin V, López-Bigas N: **Expansion of the Biocyc collection of pathway/genome databases to 160 genomes.** *Nucleic Acids Research* 2005, **33:**6083-6089.

[3] Dandekar T, Schuster S, Snel B, Huynen M, Bork P: **Pathway alignment: Application to the comparative analysis of glycolytic enzymes.** *Biochemical Journal* 1999, **343:**115-124.

[4] Forst CV, Schulten K: **Phylogenetic analysis of metabolic pathways.** *Journal of Molecular Evolution* 2001, **52:**471-489.

[5] Clemente JC, Satou K, Valiente G: **Finding conserved and non-conserved reactions using a metabolic pathway alignment algorithm.** *Genome informatics International Conference on Genome Informatics* 2006, **17:**46-56.

[6] Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, et al: **Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.** *Brief Bioinform* 2010, **11:**40-79.

[7] Li Y, Ridder Dd, Groot MJLd, Reinders MJT: **Metabolic Pathway Alignment (M-Pal) reveals diversity and alternatives in conserved networks.** In *APBC*. 2008

[8] Pinter RY, Rokhlenko O, Yeger-Lotem E, Ziv-Ukelson M: **Alignment of metabolic pathways.** *Bioinformatics* 2005, **21:**3401-3408.

[9] Yang Q, Sze SH: **Path matching and graph matching in biological networks.** *Journal of Computational Biology* 2007, **14:**56-67.

[10] Kastenmüller G, Schenk ME, Gasteiger J, Mewes HW: **Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes.** *Genome Biology* 2009, **10**.

[11] Schmidt MC, Samatova NF: **An algorithm for the discovery of phenotype related metabolic pathways.** In.; 2009: 60-65.

[12] Jim K, Parmar K, Singh M, Tavazoie S: **A Cross-Genomic Approach for Systematic Mapping of Phenotypic Traits to Genes.** *Genome Research* 2004, **14:**109-115.

[13] Levesque M, Shasha D, Kim W, Surette MG, Benfey PN: **Trait-to-Gene: A Computational Method for Predicting the Function of Uncharacterized Genes.** *Current Biology* 2003, **13:**129-133.

[14] Makarova KS, Wolf YI, Koonin EV: **Potential genomic determinants of hyperthermophily.** *Trends in Genetics* 2003, **19:**172-176.

[15] Martín MJ, Herrero J, Mateos A, Dopazo J: **Comparing bacterial genomes through conservation profiles.** *Genome Research* 2003, **13:**991-998.

[16] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles.** *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96:**4285-4288.

[17] Slonim N, Elemento O, Tavazoie S: **Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks.** *Molecular Systems Biology* 2006, **2**.

[18] Tamura M, D'Haeseleer P: **Microbial genotype-phenotype mapping by class association rule mining.** *Bioinformatics* 2008, **24:**1523-1529.

[19] Roy S: **Multifunctional enzymes and evolution of biosynthetic pathways: Retro- evolution by jumps.** *Proteins: Structure, Function and Genetics* 1999, **37:**303-309.

[20] Jensen RA: **Enzyme recruitment in evolution of new function.** *Annual Review of Microbiology* 1976, **30:**409-425.

[21] Schmidt S, Sunyaev S, Bork P, Dandekar T: **Metabolites: A helping hand for pathway evolution?** *Trends in Biochemical Sciences* 2003, **28:**336-341.

[22] Floyd RW: **Algorithm 97: Shortest path.** *Commun ACM* 1962, **5:**345.

[23] Krause GH, Weis E: **Chlorophyll fluorescence and photosynthesis: The basics.** *Annual Review of Plant Physiology and Plant Molecular Biology* 1991, **42:**313-349.

[24] Yasushi K: **New trends in photobiology. Structures and functions of carotenoids in photosynthetic systems.** *Journal of Photochemistry and Photobiology, B: Biology* 1991, **9:**265-280.

[25] Grotjohann I, Fromme P: **Structure of cyanobacterial Photosystem I.** *Photosynthesis Research* 2005, **85:**51-72.

[26] McCoy AJ, Adams NE, Hudson AO, Gilvarg C, Leustek T, Maurelli AT: **l,l-diaminopimelate aminotransferase, a trans-kingdom enzyme shared by Chlamydia and plants for synthesis of diaminopimelate/lysine.** *Proceedings of the National Academy of Sciences* 2006, **103:**17909-17914.

[27] **Carbon fixation in photosynthetic organisms - Reference pathway (Reaction)** [http://www.kegg.com/kegg-bin/show_pathway?rn00710+R01067+R00761]

[28] Walker CJ, Mansfield KE, Smith KM, Castelfranco PA: **Incorporation of atmospheric oxygen into the carbonyl functionality of the protochlorophyllide isocyclic ring.** *Biochemical Journal* 1989, **257:**599-602.

[29] Porra RJ, Schäfer W, Katheder I, Scheer H: **The derivation of the oxygen atoms of the 131-oxo and 3-acetyl groups of bacteriochlorophyll a from water in Rhodobacter sphaeroides cells adapting from respiratory to photosynthetic conditions: evidence for an anaerobic pathway for the formation of isocyclic ring E.** *FEBS Letters* 1995, **371:**21-24.