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 biologists 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
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 $O(ab^{32}sg^{-3}ln^3)$ 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 products 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...
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.

\[
p \text{-value} = \frac{\min[k, f + |F| - k] + k}{(t + f) |T| + |F|}
\]

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 Algorithm

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 \( I \) (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 \( O(n^2) \) 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.
DMPFinder found 56 reactions (length=2 pathways) and 2 linear pathways (pathway with length > 2) with e-score > 0.34 when testing whether DMPFinder can find reactions in O(gl) time. Pathway P exists in network G if and only if \( A[i] \leq p \). It is remark that given a pathway \( P \) with e-score \( s \leq 1 \), we may append many different reactions occur frequently in both \( T \) and \( P \) to construct a longer pathway \( P' \) with e-score \( s' \leq 1 \). DMPFinder does not output these redundant pathways \( P' \) unless \( s' < s \).

## Table 1 KEGG Dataset

<table>
<thead>
<tr>
<th>Rank</th>
<th>Reaction</th>
<th>Occurrence</th>
<th>Score</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>D-Ribulose 1,5-bisphosphate -&gt; 3-Phospho-D-glycerate</td>
<td>9</td>
<td>2.06(\times)10^{-3}</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>Protochlorophyllide -&gt; Chlorophyllid</td>
<td>9</td>
<td>2.06(\times)10^{-3}</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>Phytolueene -&gt; zeta-Carotene</td>
<td>9</td>
<td>2.06(\times)10^{-3}</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td>zeta-Carotene -&gt; Neurosporene</td>
<td>9</td>
<td>2.06(\times)10^{-3}</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>D-Fructose 6-phosphate -&gt; D-Erythrose 4-phosphate</td>
<td>9(2)</td>
<td>2.06(\times)10^{-3}</td>
<td>0.05</td>
</tr>
<tr>
<td>6</td>
<td>Magnesium protoporphyrin monomethyl ester -&gt; C_{35}H_{34}MgN_{4}O_{5}</td>
<td>9(3)</td>
<td>2.06(\times)10^{-3}</td>
<td>0.05</td>
</tr>
<tr>
<td>7</td>
<td>C_{35}H_{34}MgN_{4}O_{5} -&gt; C_{35}H_{34}MgN_{4}O_{5}</td>
<td>9(3)</td>
<td>2.06(\times)10^{-3}</td>
<td>0.05</td>
</tr>
<tr>
<td>8</td>
<td>C_{35}H_{34}MgN_{4}O_{5} -&gt; Divinylprotochlorophyllide</td>
<td>9(3)</td>
<td>2.06(\times)10^{-3}</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 2 BioCyc Dataset

<table>
<thead>
<tr>
<th>Rank</th>
<th>Reaction</th>
<th>Occurrence</th>
<th>Score</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Protoheme IX -&gt; biliverdin-IX-alpha</td>
<td>9</td>
<td>2.06(\times)10^{-3}</td>
<td>0.03</td>
</tr>
<tr>
<td>10</td>
<td>chlorophyllide a -&gt; monovinyl protochlorophyllide a</td>
<td>9</td>
<td>2.06(\times)10^{-3}</td>
<td>0.03</td>
</tr>
<tr>
<td>11</td>
<td>all-trans-zeta-carotene -&gt; neurosporene</td>
<td>9</td>
<td>2.06(\times)10^{-3}</td>
<td>0.03</td>
</tr>
<tr>
<td>12</td>
<td>e' -&gt; A reduced ferredoxin</td>
<td>9</td>
<td>2.06(\times)10^{-3}</td>
<td>0.03</td>
</tr>
<tr>
<td>13</td>
<td>Magnesium protoporphyrin monomethyl ester -&gt; C_{35}H_{34}MgN_{4}O_{5}</td>
<td>8(2)</td>
<td>2.05(\times)10^{-4}</td>
<td>0.34</td>
</tr>
<tr>
<td>14</td>
<td>C_{35}H_{34}MgN_{4}O_{5} -&gt; C_{35}H_{34}MgN_{4}O_{5}</td>
<td>8(2)</td>
<td>2.05(\times)10^{-4}</td>
<td>0.34</td>
</tr>
<tr>
<td>15</td>
<td>C_{35}H_{34}MgN_{4}O_{5} -&gt; Divinylprotochlorophyllide</td>
<td>8(2)</td>
<td>2.05(\times)10^{-4}</td>
<td>0.34</td>
</tr>
</tbody>
</table>

### 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 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
reaction has been assigned the lysine biosynthesis function, previous study has shown that it is a trans-
knighted 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) cyanobacteria 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 cyanobacteria have these reactions while DMPFinder can find these reactions by constructing penalty blocks. For example, the 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 Calvin-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,
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


