

A Graph Analysis Method to Detect Metabolic Sub-networks Based on Phylogenetic Profile

Shoko Miyake Yoichi Takenaka Hideo Matsuda

Department of Bioinformatic Engineering,

Graduate School of Information Science and Technology, Osaka University

1-3 Machikaneyama, Toyonaka, Osaka 560-8531 Japan

E-mail: {s-miyake,takenaka,matsuda}@ist.osaka-u.ac.jp

Abstract

To elucidate fundamental constituting principle of functional modules or building blocks of metabolic networks, computational methods to analyze the network structure of metabolism are getting much attention. We propose a graph search method to extract highly conserved sub-networks of metabolic networks based on phylogenetic profile. We formulated reaction-conservation score for the measure of the phylogenetic conservation of reactions. We also formulated compound-conservation score to eliminate biologically-meaningless compounds and reduce the size of the networks. By applying our approach to the metabolic networks of 19 representative organisms selected from bacteria, archaea, and eukaryotes in the KEGG database, we detected some highly conserved sub-networks among the organisms. Comparing them to the metabolic maps in KEGG, we found they were mainly included in energy metabolism, sugar metabolism, and amino acid metabolism.

1 Introduction

Metabolic networks are structurally best-characterized biological networks, which are expected to uncover the fundamental design principles of the underlying functional organization in all cells. We propose a graph search method to explore sub-networks conserved among the wide range of organisms. So far, Ravasz *et al.*[1] and other researchers developed methods to extract sub-networks from metabolic networks only based on the topology of the networks, e.g. their degrees of nodes. Application of phylogenetic profiles to the task helps us to obtain more functionally linked sub-networks, because proteins with identical

patterns of occurrence across the organisms tend to function together in a pathway or structural complex. Our method consists of three steps: (1) represent a metabolic network as a bipartite graph, which contains two types of nodes representing compounds and enzymatic reactions, respectively, (2) assign a reaction-conservation score to each enzymatic reaction node and a compound-conservation score to each compound node according to the phylogenetic profile of the enzymatic reactions, and (3) explore sub-networks which are globally conserved among organisms based on the basic depth-first graph search algorithm with pre-defined thresholds for the conservation scores. Using the algorithm, we looked for all meaningful subgraphs throughout metabolic networks of various organisms.

2 Methods

2.1 Graph representation

We represent a metabolic network as an undirected bipartite graph $G = (C, R, L)$ where C is the set of nodes representing compounds, R is a set of nodes representing enzymatic reactions, and L is a set of undirected links-pairs of one compound node in C and one reaction node in R . For example, if $c_1, \dots, c_n \in C$ is involved in a reaction $r \in R$, then $(c_1, r), \dots, (c_n, r) \in L$. Here, the degree k_c of a compound $c \in C$ is the number of links in L that are connected to c .

2.2 Selecting species

We selected 19 representative organisms from all three domains of life in KEGG[2] database. We picked up the species with their genomes completed from eucaryota, bacteria, and archaea evenly.

2.3 Reaction-conservation score

We constructed a phylogenetic profile [3] for each enzyme by using genome data of the selected 19 organisms. Phylogenetic profile is $n \times m$ matrix M , and here, each column represents an enzyme and each row represents a genome. If ortholog or protein i exists in genome j , $M(i, j)$ becomes to 1.

We calculated reaction-conservation score of each enzymatic reaction based on the phylogenetic profile. The decision of whether a reaction is held by a organism is according to the existence of the corresponding enzymes in the genome of the organism. Each score for an enzymatic reaction r is formulated as following:

$$S_R(r) = \text{no. of organisms holding a reaction } r. \quad (1)$$

2.4 Compound-conservation score

We formulated compound-conservation score $S_C(c)$ for a compound c as follows :

$$S_C(c) = \frac{\sum_{r \in R(c)} S_R(r)}{k_c}, \quad (2)$$

where $R(c) = \{r \in R \mid (c, r) \in L\}$.

This is an average value of the reaction-conservation score of reactions that involve a target compound. If most reactions hold high score, the compound conservation score is also high, regardless of the degree.

2.5 Graph searching method

In graph searching step, we use a threshold for reaction conservation-scores T_R , and counterpart for compound-conservation scores T_C to reduce the network. The algorithm is based on depth-first graph search algorithm and consists of following steps:

1. Start with the unvisited compound c that have max value of S_C and iterate following two steps.
2. Visit unvisited reaction r where $S_R(r)$ is max score of adjacent reactions of c and $S_R(r) > T_R$.
3. Visit unvisited compound c where $S_C(c)$ is max score of adjacent compounds of r and $S_C(c) > T_C$.

3 Results

We applied our method to the pathway data in KEGG. The number of reactions was 5649, and the number of compounds was 4629. As a result, we obtained 51 sub-networks which have at least 2 reactions

when we set the thresholds $T_R = 10$ and $T_C = 10$. In Table1, we show a list of most conserved metabolic maps in KEGG. Here, "Reaction" is the number of reactions in the metabolic map, and "Cover" is the ratio of the reactions in the obtained sub-networks per map. We also obtained purine and pyrimidine metabolism with highest number of reactions in the results, 32 and 22, respectively. In general, metabolic maps of energy metabolism, sugar degradation, cofactor biosynthesis, and processing of amino acids and nucleotides are known as conserved ones, which correspond to many of our resulting sub-networks.

Table 1. The most conserved metabolic pathways in KEGG.

Metabolic Map Name	Reaction	Cover
Valine, leucine and isoleucine biosynthesis	27	0.63
Fatty acid biosynthesis (path 2)	33	0.55
One carbon pool by folate	31	0.45
Folate biosynthesis	45	0.42
Fatty acid biosynthesis (path 1)	40	0.4
Phenylalanine, tyrosine and tryptophan biosynthesis	35	0.4

4 Conclusions

We developed a method to extract highly conserved metabolic sub-networks based on a simple graph search algorithm using phylogenetic profiles. By applying our approach to the networks of selected 19 representative organisms, we detected some sub-networks conserved among them. This type of information is expected to help us to understand organizational and evolutionary principles of metabolic networks.

References

- [1] E. Ravasz, A.L. Somera, Z.N. Oltvai, and A.L. Barabasi, "Hierarchical Organization of Modularity in Metabolic Networks", *Science*, Vol.297, 2002, pp.1551-1555.
- [2] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, "The KEGG databases at GenomeNet", *Nucleic Acids Research*, Vol.30, No.1, 2002, pp.42-46.
- [3] M. Pellegrini, E. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates, "Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles", *Proc. Natl. Acad. Sci. USA*, Vol.96, 1999, pp.4285-4288.