metabolic pathway, prediction, alignment, PathAligner

Summary

Motivation: Comparing metabolic pathways are important to identify conservation and variations among different biology system. Alignment is a strong indicator of the biologically significant relationship.

Results: A definition of metabolic pathway is defined. An alignment algorithm and a computational system are developed to reveal the similarities between metabolic pathways.

Availability: http://bibiserv.techfak.uni-bielefeld.de/pathaligner

Introduction

Today a huge amount of molecular data about different organisms has been accumulated and systematically stored in specific databases (Collado-Vides, Hofestaedt, 2002). This rapid accumulation of biological data provides the possibility of studying metabolic pathways systematically. Analysis of metabolic pathways is an essential topic in understanding the relationship between genotype to phenotype (Dandekar et al., 1999).

Researches on genomic sequence alignment have been so far intensively conducted. Applications and tools, such as FASTA [http://www.ebi.ac.uk/fasta3] and BLAST [http://www.ncbi.nlm.nih.gov/BLAST] have been developed to further understand the biological homology and estimate evolutionary distance. Although the information provided by sequenced genomes can yield insights into their evolution and cellular metabolism, knowledge of the genome sequence alone is really only the start point of the real work.

Several approaches of metabolic pathway alignment are already made in the past years. Forst C.V. and Schulten K. (1999; 2001) extended the DNA sequence alignment methods to define distances between metabolic pathways by combining sequence information of involved genes. Dandekar et al. (1999) compared glycoslysis, Entner-Doudoroff pathway and pyruvate processing in 17 organisms based on the genomic and metabolic pathway data by aligning specific pathway related enzyme-encoding genes on the genomes. Tohsato Y. et al. (2000) proposed a multiple (local) alignment algorithm by utilizing information content that was extended to symbols having a hierarchical structure EC numbers. We are going to present a new pathway alignment strategy to analysis and fully characterize metabolic pathways in the cell.

Metabolic Pathway Definitions

A biochemical pathway is defined by Mavrovouniotis M.L. (1995) as an abstraction of a subset of intricate networks in the soup of interacting biomolecules. A prevailing definition is that a metabolic pathway is a special case of a metabolic network with distinct start and end points, initial and terminal vertices, respectively, and a unique path between them, i.e. a directed reaction graph with substrates as vertices and arcs denoting enzymatic reactions (Forst, Schulten, 1999). Typical metabolic pathways are given by the wall chart of Boehringer Mannheim [http://www.expasy.ch/tools/pathways/] and KEGG [http://www.genome.ad.jp/kegg/metabolism.html], which have been verified with a number of printed and on-line sources. Databases such as KEGG, WIT [http://
wit.mcs.anl.gov/WIT2/} represent metabolic pathway graphs with labeled arcs indicating the involved enzymes. Traditionally metabolic pathways have been defined in the context of their historical discovery, often named after key molecules (e.g. “glycolysis”, “urea cycle”, “pentose phosphate pathway” and “citric acid cycle” and so on). Schuster et al. (2000) provided a general definition of metabolic pathways based on the concept of elementary flux modes. The basic strategy to represent and compute pathways is the reactant-product binary relation. Properties of the pathway that rely upon the integration of two or more input molecules and unrelated output molecules and feedback effects are ignored.

Obviously, a metabolic pathway is a special part of complex network of reactants, products and enzymes with multiple interconnections representing reactions and regulation. One is called a pathway only if they are linear and non-branched. A pathway’s substrates are usually the products of another pathway, and there are junctions where pathways meet or cross. We consider that a metabolic pathway is a subset of reactions that describe the biochemical conversion of a given reactant to its desired end product.

Let \( M = \{m_1, \ldots, m_n\} \) be a set of metabolites in cells. Let \( e_i : M \rightarrow M \) be a function for enzymatic reactions taking place in the cells.

The fact that \( e_i \) is a function from a set of substrates \( S (S \subseteq M) \) into a set of products \( P (P \subseteq M) \). It can be written as follows:

\[
e_i : S \rightarrow P
\]

for all \( m, m, m, \in M \), the following property holds:

\[
e_i(m) = m, \text{ and } e_i(m) = m \Rightarrow e_i(e_i(m)) = m_i.
\]

Let \( e_i(m) = m, e_i(m) = m, \ldots, e_i(m) = m_i \), we define \( e_i e_i \ldots e_i(m) = e_i(e_i \ldots e_i(m)) = m_i \).

A new proposed definition of the metabolic pathway is presented and discussed in the following paragraphs.

**Definition 1.** Given \( e_i : M \rightarrow M \), a metabolic pathway is defined as a subset of successive enzymatic reaction events \( P = e_i e_i \ldots e_i \).

Each enzymatic reaction \( e_i (1 \leq i \leq k) \) is catalyzed by a certain enzyme that is denoted as a unique EC number. The EC number is expressed with a 4-level hierarchical scheme that has been developing by the International Union of Biochemistry and Molecular Biology (IUBMB). The 4-digit EC number, \( d, d, d, d \) represents a sub-sub-sub-class indication of biochemical reaction. For instance, arginase is numbered by EC 3.5.3.1, which indicates that the enzyme is a hydrolase (EC 3.*.*.*), acts on the “carbon-nitrogen bonds, other than peptide bonds” (sub-class EC 3.5.*.*) in linear amidines (sub-sub-class EC 3.5.3.*). Thus we can adapt the EC number as a unique name for the responding enzyme catalyzed reaction.

**Metabolic Pathway Alignment**

**Theory Basics**

In order to score the similarity (percent identity) between two metabolic pathways, we define the similarity function.

**Definition 2.** Let \( E \) be a finite set of \( e \) functions, an edit operation is an ordered pair

\[
(\alpha, \beta) \in (E \cup \{e\}) \times (E \cup \{e\}) \setminus \{(e, e)\}.
\]

\( \alpha \) and \( \beta \) denote 4-digit EC strings of enzymatic reaction function, e.g. \( \alpha = e_{1,1,1,1} \) \( \beta = e_{3,4,5} \) \( e \) denotes the empty string for null function. However, if \( \alpha \neq e \) and \( \beta \neq e \), then the edit operation \( (\alpha, \beta) \) is identified with a pair of enzymatic reaction function.

An edit operation \( (\alpha, \beta) \) is written as \( \alpha \rightarrow \beta \) (we can simply written \( \alpha, \beta \) as EC numbers). There are
three kinds of edit operations:
$\alpha \rightarrow \varepsilon$ denotes the deletion of the enzymatic reaction function $\alpha$,
$\varepsilon \rightarrow \beta$ denotes the insertion of the enzymatic reaction function $\beta$,
$\alpha \rightarrow \beta$ denotes the replacement of the enzymatic reaction function $\alpha$ by the enzymatic reaction function $\beta$,
notice that $\varepsilon \rightarrow \varepsilon$ never happens.

**Definition 3.** Let $E_1 = e_1 ... e_m$ and $E_2 = e_1' ... e_n'$ be two metabolic pathways, an alignment of $E_1$ and $E_2$ is a pair sequence $(\alpha_1 \rightarrow \beta_1, ..., \alpha_h \rightarrow \beta_h)$ of edit operations such that $E_1' = \alpha_1, ..., \alpha_h$ and $E_2' = \beta_1, ..., \beta_h$.

**Example 1.** The alignment $A = (2.4.2.3 \rightarrow 2.4.2.4, 3.5.4.5 \rightarrow \varepsilon, 3.1.3.5 \rightarrow 3.1.3.5, \varepsilon \rightarrow 2.7.4.9)$ of the pathways $e_{2.4.2.3} e_{3.5.4.5} e_{3.1.3.5}$ and $e_{2.4.2.4} e_{3.1.3.5} e_{2.7.4.9}$ is written as follows, one over the other:

\[
\begin{pmatrix}
2.4.2.3 & 3.5.4.5 & 3.1.3.5 & \varepsilon \\
2.4.2.4 & \varepsilon & 3.1.3.5 & 2.7.4.9
\end{pmatrix}
\]

**Similarity Function**

**Definition 4.** A similarity function $\sigma$ assigns to each edit operation $(\alpha, \beta)$ a nonnegative real number. The similarity $\sigma(\alpha, \varepsilon)$ and $\sigma(\varepsilon, \beta)$ of the deletion operation $(\alpha, \varepsilon)$ and insertion operation $(\varepsilon, \beta)$ is 0. For all replacement operations $(\alpha, \beta)$ $\alpha \neq \varepsilon$, $\beta \neq \varepsilon$, say, $\alpha = d_1, d_2, d_3, d_4$ and $\beta = d_1', d_2', d_3', d_4'$, then the similarity function $\sigma(\alpha, \beta)$ is defined by:

\[
\sigma(\alpha, \beta) = \begin{cases} 
0, & \text{if } (d_1 \neq d_1'); \\
0.25, & \text{if } (d_1 = d_1' \text{ and } d_2 \neq d_2'); \\
0.5, & \text{if } (d_1 = d_1' \text{ and } d_2 = d_2' \text{ and } d_3 \neq d_3'); \\
0.75, & \text{if } (d_1 = d_1' \text{ and } d_2 = d_2' \text{ and } d_3 = d_3' \text{ and } d_4 \neq d_4'); \\
1, & \text{if } (d_1 = d_1' \text{ and } d_2 = d_2' \text{ and } d_3 = d_3' \text{ and } d_4 = d_4') \text{ i.e. } \alpha = \beta).
\end{cases}
\]

The definition does not exclude the possibility that $d_1, d_2, d_3, d_4$ can be respectively expressed as wide card symbols $*, **$ and $***$ which means no clear classification of the enzyme.

However single pair of EC string comparison just means to measure how different EC strings are. Often it is additionally of interest to analyze the total difference between two strings into $\sigma$ collection of individual elementary differences. The most important mode of such analyses is an alignment of the pathways. The function $s$ can be extended to alignments in a straightforward way: the similarity $\sigma(A)$ of an alignment $A = (\alpha_1 \rightarrow \beta_1, ..., \alpha_h \rightarrow \beta_h)$ is the sum of the similarities of the edit operations $A$ consists of:

\[
\sigma(A) = \sum_{i=1}^{h} \sigma(\alpha_i \rightarrow \beta_i).
\]

A alignment scoring scheme, $\text{Score}(E_1, E_2)$ of two metabolic pathways is the minimal mean similarity of their alignment

\[
\text{Score}(E_1, E_2) = \frac{1}{\max(m, n)} \sigma(A),
\]

where $m, n$ are the lengths of pathways.
Algorithms and Implementation

The pairwise alignment algorithm is as follows: 1. Initialize the set of unaligned EC number sequences, and the lengths; 2. Starting from both ends towards the middle, align one sequence to another and attempt to find all EC numbers with same 4-level hierarchical numbers. Score the similarities. Recall the alignment positions where EC number are identical and cut the sequences into more subsequences by removing the identical EC numbers; 3. Each pair of sub-sequences is initialized to begin a new round of 3-level hierarchical EC number matching till all pairs of sub-sequences are aligned. A similarity score is calculated afterwards; 4. Apply the same rule again, find the similarities of rest unaligned sub-sub-sequences based on 2-level hierarchical EC number matching and then sub-sub-sub-sequences on 1-level matching if any.

The algorithm has been implemented into the PathAligner system (http://bibiserv.techfak.unibielefeld.de/pathaligner) (Fig.).

Three web-based alignment interfaces are implemented: “E-E Alignment”, “M-E-M Alignment” and “Multiple Alignment”. “E-E Alignment” uses the basic algorithm to align two linear metabolic pathways (represented as EC number sequences). User can also align any such a metabolic pathway against our pool database to find a list of hits. “M-E-M Alignment” considers the differences of metabolites in two pathways, which are presented as ”Metabolite-EC number-Metabolite” patterns of sequence. It is possible to pick up two such pathways and align them to identify whether they are alternative pathways or partially are. “Multiple Alignment” allows the alignment of more than two metabolic pathways.

Fig. A screenshot of PathAligner.
Conclusion

Identification and analysis of metabolic networks is a complex task due to the complexity of the metabolic system. Abstract pathway defined as a linear reaction sequence is practical for our alignment algorithm. We have presented an algorithm to study the problem of metabolic pathway alignment. Our algorithm calculates the hierarchical similarities of EC numbers mapping from both ends of the sequences. The algorithm has been successfully implemented into the PathAligner system.

References


