P. Glebe

metabolic paths

of

Multiple sequence alignment
Metabolic pathways

Enzyme Commission (EC) classification of enzymes

- 1. Oxidoreductase
- 2. Transferase
- 3. Hydrolase
- 4. Lyase
- 5. Isomerase

Hierarchy of total depth 4 having root * and first level given by Hashimoto, Matsuda, and Hashimoto.

Pathway analysis using enzyme hierarchy by Tohsato, Matsuda presentation of "A multiple alignment algorithm for metabolic pathways"
6. Ligase

For example, the enzyme with EC number ‘1.1.1.1’ is characterized as follows.

First level indicates enzyme type; so ‘1’ means oxidoreductase. Second level within oxidoreductase indicates the kind of donor which is oxidized; hence ‘1.1’ means an oxidoreductase acting on the carboxyl COOH group of donors.

Third level within oxidoreductase indicates the kind of acceptor; hence ‘1.1.1’ means that NAD+ or NADP+ is acceptor.

Fourth level is given by specific enzyme reaction of the form

\[ \text{Substrates} \rightarrow^{\text{enzyme}} \text{Products} \]
\[
\sum_{u} |s_u| = |S|
\]

Let \( |S| \) denote the number of occurrences of enzymes in pathway \( s \), i.e., if an enzyme catalyzes two separate reactions, it is counted twice. Let \( S = \{s_1, \ldots, s_n\} \) be a set of \( \ell \) linear metabolic pathways.

Kegg has 3705 enzymes with EC classifications (summer 2000).

In the subtree with node \( \gamma \), let \( |\gamma| \) be the number of enzymes of the subtree with node \( \gamma \), while \( |\gamma\rangle \) is the class of enzymes in reactions at leaves.

Enzyme class in \( [\gamma] \) is the class of enzymes in reactions at leaves.

Kegg database
Write instead of \((y)^I\) \(s\):

\[
(y)^s d \geq \log \left( \frac{|y|}{I} \right) \geq \log = (y)^s I
\]

And define the information content of enzyme class \([y]\) by

\[
\frac{\|S\|}{\sum_{e \in \text{occ}[y]} e} = (y)^s d
\]

If \(y\) is a node of the E C tree, let

If \(e\) is an enzyme in the reactions in metabolic pathway \(s\), then \(\text{occ}(e, s)\) is the number of occurrences.
matrices for amino acids.

similarity between enzymes by analogy to log odds similarity.

• The authors motivate their definition of information theoretical

\[(\eta)^S I < (\eta)^S I \quad \text{then} \quad [\eta] [\eta] \quad \text{if is subclass of [\eta]} \]

• Information is MONOTONIC DECREASING,

\[\log \frac{\log(\frac{1}{370})}{\log(\frac{1}{1.85})} = -1.185\]

• Example. I(\eta) = (1)$\log_{\eta}(1/370)$

EC tree:

where $\text{lea}(e_1, e_2)$ is the least common ancestor of $e_1, e_2$ in the

\[I(\text{lea}(e_1, e_2))\]

as

similarity between $e_1, e_2$ with respect to metabolic pathways.

• If $e_1, e_2$ are distinct enzymes with EC numbers, then define

\[S_{e_1, e_2} \]
where $q_\text{A} \times q_\text{B}$ is the background frequency of a resp. $q_a$
metabolic pathways.

- Probability that both enzymes match at random in set $S$ of
  number of enzymes in least common ancestor class), and
- Probability that two enzymes are similar (i.e., have small
  essentially the log odds ratio between

The authors claim that information theoretic similarity is

$$\log_2 \left( \frac{q^\nu b}{q^\nu d} \right)$$

is the log odds

model, and

is the evolutionary model versus the random

Hence
Consider two pathways having enzymes, given sequentially in order of pathway, given by 

\[ w_1, \ldots, w_n \]

and

\[ e_1, \ldots, e_n \]

Information theoretic similarity: alignment algorithm (Needleman-Wunsch), to maximize 

Given two linear metabolic pathways, apply global pairwise alignment.
otherwise replacing \((e, e')\) by \(\text{leq}(e, e')\) and
pathways by replacing \((-e', -e)\) and \((e, -e)\) by gap symbols \(\ast\), and

Define pattern obtained from an alignment of 2 metabolic

\[
((\text{leq}(e, e')), e') S = (e, e') I = (\text{leq}(e, e'), e') S
\]

Recall that \((*) I = -11.85\).

\[
-15 = (-e, e') S = (e, -e) S
\]

score as follows.

Given an alignment \(a_1, \ldots, a_p, q_1, \ldots, q_q\), compute alignment
Yieldings - 19.11

1.5 1.8 - 3.4 - 3.0 - 5.1
and total score (information content) given by

[2.4.7.4] [3.1.3.5] - [2.4.2.1] with pattern

[2.4.7.4.9] [3.1.3.5.6] - [2.4.2.1] [3.5.4.5] [3.1.3.5] [2.4.7.4.14]
Optimal alignment is given by

[2.4.7.4.9] [3.1.3.5] [2.4.2.1] [3.5.4.5] [3.1.3.5] [2.4.7.4.14]
Second pathway:

[2.4.2.3] [3.5.4.5] [3.1.3.5] [2.4.7.4.14] [3.5.4.5] [3.1.3.5]
Example. First pathway:
(d) I \cdot \sum_{u}^{d \in d} + (\gamma - u)\gamma m = (d)^{S} I

By using the average length of the pathways in $S$, $u$ is the number of patterns.

Given set $S$ of pathways, if $d$ is set of $\gamma$ patterns, the follows.

The information content of a set of patterns is NOT just the sum of the information of each pattern $d \in P$. Instead, do as higher classes have smaller information theoretic content, add content between enzymes differing by one level is on average $w$.

After applying pairwise alignment, enzymes classes in pattern originate from an average.

**Multiple alignment**
\[ d = n - u d \]

\[ \{d\} \cup \{\{d\}, \emptyset\} - d = \{d\} \]

\[ (\{d\}, \emptyset) V = d \]

\[ d \in \{d\} \quad \text{for each } d \in \{d\} \]

\[ d \in \{d\} \quad \text{for each } d \in \{d\} \]

\[ (d) I = \max \{S = 0 \mid S = d \mid |S| = n \} \]

\[ \text{Output. Set } P \text{ of patterns (in ideal situation, a unique output pattern).} \]

\[ \text{Input. Set } S \text{ of (linear) metabolic pathways.} \]

\textbf{Feng-Doolittle style MSA}
\[
\begin{align*}
\{ & \\
\gamma^{-u}d = d \\
\text{else} & \\
\{ & \\
\text{return} & \\
(1 - \gamma^{-u}d)I > (\gamma^{-u}d)I & \\
\{ & \\
\max & \\
(1d)I
\end{align*}
\]
extracts of sugar, DNA and amino acid metabolic pathways.

longest pathway. Authors report multiple alignment of linear

Time complexity is \( O \varepsilon \max \) where \( \varepsilon \max \) denotes length of