Alignment of metabolic pathways

MAYA PORTNOY
GUY SUDAY
What is Alignment?

- Arrangement in a straight line, or in correct or appropriate relative positions.
- It is useful for comparing objects and determining their similarity.
A real life example

• What did Trump mean?
• Option 1: Coverage
• Option 2: Conference
• Option 3: ?
What is Sequence Alignment?

- Arrangement of two or more sequences (DNA, RNA or protein) in parallel to each other.
- It is used for identifying regions of similarity that may be a consequence of functional, structural or evolutionary relationships between the sequences.
What is Sequence Alignment?

- In order to evaluate how good is an alignment in comparison to other alignments, we need to define a scoring method.
- The method will determine the score for a match, a mismatch and a gap.

**Example:**
- Match: +1
- Mismatch: -1
- Gap: -2

**Alignment score:**
0
What are Metabolic Pathways?

• A series of chemical reactions catalyzed by enzymes, in which the products of each reaction are the substrates of the next one.
Metabolic Pathways as graphs

• A metabolic pathway can be represented as a directed graph:
  • **Nodes**: enzymes
  • **Edges**: edge \((u,v)\) - the product of enzyme \(u\) is the substrate of enzyme \(v\)
Introduction

• Nowadays many organisms are being studied in order to understand their genome-scale metabolic networks.

• Examining similarities and differences between networks of different organisms contributes to that effort.

• A need arises for good tools capable of:
  • Searching for matches of a query pathway in a collection of known pathways.
  • Aligning two pathways to locate conserved pathway fragments (that did not undergo changes during evolution).
Similarity between pathways

- When computing the similarity between pathways, we take into account:
  1. The resemblance between any two corresponding nodes
  2. The likeness between the pathways’ network structure

- These criteria reflect:
  1. The similarity (rather than identity) between matched enzymes
  2. The topological similarity between the pathways in a biological context

- As in sequence alignment, the closeness between two pathways is reflected by a scoring method that measures the distance.
The model

The problem of finding structural similarity among pathways can be solved by finding isomorphism/homomorphism between graphs.

Unfortunately, this problem is NP-complete. All the more so, when searching for similar rather than identical matches.

Fortunately, taking into account the typical properties of metabolic pathway graphs can simplify the problem:

- Cycles are rare
- The graph is directed
- It is a DAG!
The model

• A **multi-source tree** is a DAG whose underlying graph is a tree, where some of the nodes can have several incoming and outcoming edges.

• Most metabolic pathways can be cast as multi-source trees.

• Hence, our algorithm will be based on this conclusion.

The algorithm’s purpose

• **The subtree isomorphism problem:** Given a pattern tree $P$ and a text tree $T$, find a subtree of $T$ that is isomorphic to $P$ (if exists).

• **The subtree homeomorphism problem:** A variant of the former, where 2-degree nodes can be deleted from the text tree $T$.

• The engine is based on the homeomorphism model because a sequence of enzymes can be functionally replaced by a single enzyme.

Definitions

- $\Delta$ is a node-to-node similarity score table.
- $\delta$ is a score for deleting a node from a tree (usually negative).
- $M[T_1, T_2]$ is a partial mapping from the nodes of $T_1$ to the nodes of $T_2$ that preserves the ancestor relations of the nodes.
Definitions

• The **LSH (Labeled Subtree Homeomorphism) score** denoted as $LSH(M[T_1,T_2])$ is:

$$LSH(M[T_1,T_2]) = \delta(|T_2| - |T_1|) + \sum_{\forall (u,v) \in M} \Delta[u,v].$$

• The **ALSH (Approximate Labeled Subtree Homeomorphism) problem**: given undirected labeled trees $P$ and $T$, a scoring table $\Delta$ and a deletion penalty $\delta$, find mapping $M[P,t]$ from $P$ to some subtree $t$ of $T$ such that $LSH(M[P,t])$ is maximal.

The alignment algorithm

- The alignment algorithm employs a bottom-up dynamic programming approach to solve the ALSH problem.
- First we will describe a basic algorithm for rooted, undirected trees.
- Then we will extend this algorithm to multi-source trees which are unrooted and directed.
The alignment algorithm: settings

- \( T^r = (V_T, E_T, r) \) – The text tree \( T \) which is rooted in \( r \)
- \( P^{r'} = (V_P, E_P, r') \) – The pattern tree \( P \) which is rooted in \( r' \)
- \( p_{u'}^{r'} \) – A subtree of \( P^{r'} \) which is rooted in \( u \)
- \( t_{v'}^r \) – A subtree of \( T^r \) which is rooted in \( v \)
- \( y_1, ..., y_{c(v)} \) - The children of \( v \in T^r \)
- \( x_1, ..., x_{c(u)} \) - The children of \( u \in P^{r'} \)
- \( c(u) \leq c(v) \) since no deletions are allowed from the pattern.

The alignment score

• For each node $v \in V_T$ and for each node $u \in V_P$, $\text{AlignmentScore}[u, v]$ is the maximal LSH score between any subtree $p^r_{u'}$ of $P^r$ and a matching subtree $t^r_v$ of $T^r$ (if exists).

• Otherwise, $\text{AlignmentScore}[u, v] = -\infty$.

Computing $\text{AlignmentScore}[u, v]$

- Recursively, in a post order traversal of $T^r$ (bottom-up).
- First, $\text{AlignmentScore}[u, v]$ is computed for all leaf nodes of $T^r$ and $P^{r'}$.
- Next, $\text{AlignmentScore}[u, v]$ is computed for each node pair $u \in V_P, v \in V_T$ based on previously computed scores for all their children.

Pseudo code and example

**Input**: A DP table with all values up to cell \((u, v)\) already set.
A Label-to-Label Scoring Table \(\Delta\).

**Output**: The score to be set to entry \((u, v)\) of the DP table.

<table>
<thead>
<tr>
<th>AS</th>
<th>(x_1)</th>
<th>(x_2)</th>
<th>(u)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y_1)</td>
<td>-2</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>(y_2)</td>
<td>-3</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>(y_3)</td>
<td>+1</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>(v)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(k\): the out-degree of node \(u\);
\(l\): the out-degree of node \(v\);

if \(k > l\) then
return \(-\infty\);

\[
\Delta[i,j]
\]

\[
\begin{array}{ccccccc}
A & B & C & D & E & F \\
-1 & -1 & -1 & -1 & -1 & -1 \\
+2 & -2 & -3 & -3 & -3 & -3 \\
-1 & +1 & -2 & +1 & -2 & -2 \\
-3 & -2 & -2 & -3 & +2 & +1 \\
-3 & -2 & -1 & +2 & +1 & +2 \\
\end{array}
\]

Pinter et al. "Alignment of metabolic pathways".
Pseudo code and example

else

\( G \): a bipartite graph with node bipartition \( X \) and \( Y \);
\( X \): the set of children \( \{x_1, \ldots, x_k\} \) of \( u \);
\( Y \): the set of children \( \{y_1, \ldots, y_\ell\} \) of \( v \);
node \( x_i \in X \) is connected to node \( y_j \in Y \) via an edge
whose weight \( w(x_i, y_j) \) is set to \( DP[x_i, y_j] \);

<table>
<thead>
<tr>
<th>AS</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( u )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_1 )</td>
<td>-2</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>( y_2 )</td>
<td>-3</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>( y_3 )</td>
<td>+1</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>( v )</td>
<td></td>
<td></td>
<td>?</td>
</tr>
</tbody>
</table>
Pseudo code and example

\[ \text{AS}(G): \text{the weighted assignment score of } G; \]
\[ \text{AS}(G) \leftarrow \max \sum_{(i,j) \in M} DP[x_i, y_j] \]
where \( M \) is a maximum matching;
end

\[ \text{AS}(G) = (+1) + (+1) = +2 \]

<table>
<thead>
<tr>
<th>AS</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( u )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_1 )</td>
<td>-2</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>( y_2 )</td>
<td>-3</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>( y_3 )</td>
<td>+1</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>( v )</td>
<td></td>
<td></td>
<td>-?</td>
</tr>
</tbody>
</table>

\[ \Delta[i,j] \]

\[
\begin{array}{ccccccc}
A & B & C & D & E & F \\
\hline
-1 & -1 & -1 & -1 & -1 & -1 \\
a & +2 & -2 & -3 & -2 & -3 & -3 \\
b & -1 & +1 & -2 & +1 & -2 & -2 \\
e & -3 & -2 & -2 & -3 & +2 & +1 \\
f & -3 & -2 & -1 & +2 & +1 & +2 \\
\end{array}
\]

Pseudo code and example

\[ \text{BestChild}(u, v) \text{: the child of node } v \text{ whose ALSH score with } u \text{ is the highest;} \]

\[ \text{BestChild}(u, v) \leftarrow \max_{j = 1}^{\ell} \Delta[u, y_j]; \]

\[ \text{BestChild}(u, v) = -7 \]

<table>
<thead>
<tr>
<th>AS</th>
<th>(x_1)</th>
<th>(x_2)</th>
<th>(u)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y_1)</td>
<td>-2</td>
<td>+1</td>
<td>(-\infty)</td>
</tr>
<tr>
<td>(y_2)</td>
<td>-3</td>
<td>-2</td>
<td>(-7)</td>
</tr>
<tr>
<td>(y_3)</td>
<td>+1</td>
<td>+1</td>
<td>(-\infty)</td>
</tr>
<tr>
<td>(v)</td>
<td></td>
<td></td>
<td>?</td>
</tr>
</tbody>
</table>

Pseudo code and example

\[ \delta: \text{the deletion penalty from } \Delta; \]
\[ \text{return } \max \{ \Delta[u,v] + AS(G), \text{ BestChild}(u,v) + \delta \}; \]

\[ \text{return } \max\{(+2) + (+2), (-7) + (-1)\} = (+4) \]
Extended algorithm

• Now we will extend the rooted algorithm to multi-source trees which are unrooted and directed:
  • Select an arbitrary node $r \in V_T$.
  • For each $u \in V_P$ compute the rooted ALSH between $P^u$ and $T^r$.
  • Filter out subtree alignments that mapped together edges of conflicting directions.

MetaPathwayHunter

- **MetaPathwayHunter** is a tool that implements this algorithm. It reports all the matches found for a given query in a given collection, ranked by similarity and statistical significance.

- MetaPathwayHunter was employed to study the similarities and differences in the metabolic networks of *E.coli* and *S.cerevisiae* as representatives of the prokaryotic and eukaryotic kingdoms.
Choosing \( \Delta \) table

- Each enzyme was associated with its EC classification – a numbering system that categorizes the catalyzed chemical reaction.

- For an enzyme class \( h \), \( C(h) \) denotes the number of enzymes whose classes are included under \( h \).

- The information content of \( h \) is defined as \( I(h) = -\log_2 C(h) \)

- \( \Delta[i,j] = I(h_{i,j}) \) where \( h_{i,j} \) is the lowest common upper class of enzymes \( e_i, e_j \).

- The table was built this way in order to reflect the enzymes’ functionality rather than their sequence.
Results: Inter-species

• MetaPathwayHunter found at least one statistically significant alignment for 63% of the *E.coli* pathways and 66% of the *S.cerevisiae* pathways.

• Out of 80 analogous pathways in both species, 62 were found to be statistically significant.

• These results imply that despite the evolutionary distance between the species, a considerable fraction of their metabolic networks is conserved.
Results: Inter-species

Results: Inter-species

Results: Intra-species

MetaPathway queries

• MetaPathwayHunter has another mode - **MetaPathway query** - in which the pattern consists of several enzymes and a suggested structure of their interactions.

• The output in this mode will be the entire pathway scheme.

• The gap penalty in this mode should be lower, in order to increase the chances of finding more matches.

• This mode can be useful in two scenarios:
  • When a user wishes to discover whether a group of enzymes are metabolically connected.
  • When a user has limited knowledge of a certain pathway.
MetaPathway queries

References


• https://en.wikipedia.org/wiki/Sequence_alignment

• https://www.slideshare.net/rvosa/perl-for-phyloinformatics
THANKS FOR LISTENING
ANY QUESTIONS?