Algorithmic issues in (co)phylogenetic analysis

blerina sinaimeri
evolution
Phylogenetic tree

Phylogenetic Tree
Phylogenetic tree
Phylogenetic Trees

- rooted / unrooted
- binary / k-ary
- labeled from a set / labeled from a multiset
- unweighted / weighted (branch lengths)
- unordered / ordered
Phylogenetic Trees

• Maximum Parsimony
• Maximum Likelihood method
### Maximum Parsimony

<table>
<thead>
<tr>
<th>Sequence 1</th>
<th>TGC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 2</td>
<td>TAC</td>
</tr>
<tr>
<td>Sequence 3</td>
<td>AGG</td>
</tr>
<tr>
<td>Sequence 4</td>
<td>AAG</td>
</tr>
</tbody>
</table>

Find the “**best**” tree... but what does “best” mean?

```
<table>
<thead>
<tr>
<th>S1</th>
<th>TGC</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2</td>
<td>TAC</td>
</tr>
<tr>
<td>S3</td>
<td>AGG</td>
</tr>
<tr>
<td>S4</td>
<td>AAG</td>
</tr>
</tbody>
</table>
```
Maximum Parsimony

Sequence 1: TGC
Sequence 2: TAC
Sequence 3: AGG
Sequence 4: AAG

Find the “best” tree...but what does “best” mean?

In Maximum Parsimony: Minimize the number of mutations across the edges
Maximum Parsimony

Sequence 1     TGC
Sequence 2     TAC
Sequence 3     AGG
Sequence 4     AAG

Find the “best” tree...but what does “best” mean?

In Maximum Parsimony: Minimize the number of mutations across the edges.
Maximum Parsimony

Sequence 1  TGC
Sequence 2  TAC
Sequence 3  AGG
Sequence 4  AAG

Find the “best” tree...but what does “best” mean?

In Maximum Parsimony: Minimize the number of mutations across the edges
Maximum Parsimony

The problem

- Input: \( n \) DNA sequences
- Goal: Find the tree that minimizes the number of mutations along the edges.

Possible unrooted trees: \( \frac{(2n-5)!}{2^{n-3}(n-3)!} \)
Maximum Parsimony

The problem

• Input: n DNA sequences
• Goal: Find the tree that minimizes the number of mutations along the edges.

Finding one optimal tree is NP-hard!
Maximum Likelihood

• Given certain rules about how sequences change over time, the best tree should reflect the most likely sequence of evolutionary events.

• maximize the probability that a given tree could have produced the observed data (i.e., the likelihood)

Differences with the parsimonious method

• Use of an explicit evolutionary model

• Allows variable substitution rates for each branch
Comparing Trees

How similar two phylogenetic trees are?

• Robinson-Foulds
• Triplet distance
• Maximum agreement subtree
• Edit distances (SPR, TBR, NNI)
• …
The SPR distance ($d_{\text{SPR}}$) is the minimal number of moves that transforms one tree into the other.

SPR (Subtree Prune and Regraft)

NP-hard
3-approximation algorithm
Maximum Agreement Forest

Technical detail

\[ \rho \]
Maximum Agreement Forest

A Forest of $T$ is a disjoint collection of phylogenetic subtrees whose union of leaf sets is $X \cup \rho$. 

A Forest

\[
\begin{align*}
A & \quad B \\
C & \quad D \\
E & \quad F
\end{align*}
\]
Maximum Agreement Forest

$T_1$

$T_2$

AF

MAF

7 components

3 components
Maximum Agreement Forest

$T_1$ and $T_2$ are two trees.

$AF$ and $MAF$ are components of the trees.

$AF$ has 7 components.

$MAF$ has 3 components.
Maximum Agreement Forest

\[ m(T_1, T_2) = \text{size of maximum agreement forest} \]

**Theorem. (BS04)**

Let \( T_1 \) and \( T_2 \) be two binary phylogenetic X-trees. Then

\[ d_{\text{SPR}}(T_1, T_2) = m(T_1, T_2) - 1 \]
Exercise 1

• Explain the role of $\rho$ in the relation between SPR and MAF. Does the theorem hold without introducing $\rho$?
Maximum Agreement Forest

3-approximation algorithm for MAF

T₁

T₂

T₁

T₂

T₁

T₂

T₁

T₂
coevolution
Symbiosis

**Mutualism**

Human Microbiota

**Parasitism**

Gopher

Lice
Interspecific interaction

Plant diversity

Mimicry

Parasitism

Human Microbiota

Mutualism