Maximum parsimony

Kai Müller

Phylogenetic trees

- Cladogram
  - Branch lengths without meaning
  - Slanted/rectangular
  - Network/unrooted

- Parenthetical notation (Newick)

Free rotation around all nodes

A phylogram = the same phylogram
Polytomies - multifurcations

- Polytomies
  - hard
  - soft

Order within polytomies without meaning

How many trees are there?
- for n taxa, there are \((2n-5)! = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-5)\)

Too many!
- for n taxa, there are \((2n-3)! = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-3)\)

Types of characters
- Morphological
- Molecular
  - DNA sequences
  - Amino acid sequences
  - Presence of genes in Genome
  - Genomic rearrangements
  - Presence of introns
  - Single substitutions (SNP, RFLP, AFLP etc.)
Different types of sequence homology

• Imagine we are unaware of the gene duplication and sample only alpha-chick, alpha-frog, beta-mouse:

erroneously inferred by a mixture of paralogous & orthologous sequences
true organismal relationships

Alignment formats

<table>
<thead>
<tr>
<th>format</th>
<th>mfl programs</th>
<th>&gt; to format</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>BiEdit (v.0.3.2), clustal (v.1.83), g隐形evm, phylop, mrbayes, treediv</td>
<td>GenBank, NEXUS, mrbayes, phylop, TreeDiv</td>
</tr>
<tr>
<td>FASTA</td>
<td>clustal (v.1.83), g隐形evm, phylop, mrbayes, treediv</td>
<td>GenBank, NEXUS, mrbayes, phylop, TreeDiv</td>
</tr>
<tr>
<td>FASTA, GOMSBF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FASTA, CLUSTAL, PHYLIP, mrbayes, treediv</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ABI, CLUSTAL, FASTA, NEXUS, PHYLIP, MRBAYES, mrbayes, phylop</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GOBSMF, HENNING, NEXUS, PHYLIP, mrbayes, treediv</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ABI, FASTA, NEXUS, SCF, GZ, SCF</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

sequence alignment = primary homology assessment
tree building = secondary homology assessment

• Recognition of congruence among the primary homologies as a result of a treebuilding analysis of the data—the shared derived character states (synapomorphies) on the phylogenetic tree represent homologies
• an assessment of ‘congruence’ that explains the sameness as resulting from common ancestry

• conjecture, before data analysis, that similarity among certain characters and character states may be evidence for evolutionary groupings of the taxa
• an assessment of ‘essential sameness’ without reference to an ancestor
Overview: treebuilding methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Data types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Likelihood (ML)</td>
<td>discrete characters vs. distances</td>
</tr>
<tr>
<td>Neighbor Joining (NJ)</td>
<td>discrete characters vs. distances</td>
</tr>
<tr>
<td>Phylogeny Inference (PI)</td>
<td>discrete characters vs. distances</td>
</tr>
<tr>
<td>Distance-Matrix Methods (DMM)</td>
<td>distances</td>
</tr>
</tbody>
</table>

### Maximum parsimony

- parsimony principle
  - most parsimonious explanation should be preferred to others
  - Ockham’s razor
  - parsimony analysis ≈ cladistics
  - the oldest phyl. reconstr. method
  - still widely used because
    - principle/optimality criterion easy to grasp
    - quick
    - without convincing alternatives for morphological data

### Data types: discrete characters vs. distances

#### Alignment:

<table>
<thead>
<tr>
<th>Taxa</th>
<th>1 2 3 4 5 6 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tax1</td>
<td>G A G G G A A</td>
</tr>
<tr>
<td>Tax2</td>
<td>A A A A A A A</td>
</tr>
<tr>
<td>Tax3</td>
<td>A A A A A A A</td>
</tr>
<tr>
<td>Tax4</td>
<td>G G G G G G G</td>
</tr>
</tbody>
</table>

#### Distanzen:

<table>
<thead>
<tr>
<th>Taxa</th>
<th>1 2 3 4 5 6 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tax1</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>Tax2</td>
<td>2 3 4 5 6 7 1</td>
</tr>
<tr>
<td>Tax3</td>
<td>3 4 5 6 7 1 2</td>
</tr>
<tr>
<td>Tax4</td>
<td>4 5 6 7 1 2 3</td>
</tr>
</tbody>
</table>

### Parsimony-informative characters

- >= 2 states, >= 2 of which occur in >=2 taxa
Character conflict

- Character state distributions of different characters do not all support the same topology
- Example for the following tree:

```
    A
   /   \
  tax2 -- tax1
    \
     tax4
      \
       tax3
```

Character conflict: example

- Suggested topology No 1: 10 steps

```
    A
   /   \
  tax2-- tax1
    \
     tax4
      \
       tax3
```

Character conflict: example

- Suggested topology No 2: 9 steps

```
    A
   /   \
  tax2-- tax1
    \
     tax4
      \
       tax3
```

The shortest tree

- Length of tree = #steps = #character state transitions
- Out of the two trees tried, 9 is the best score
- Trying all other out of the 15 possible trees shows: there are more that require 9, but none with fewer steps
Consensus trees

- strict
- loose (semi-strict)
- majority rule
- Adams
- supertrees

Types of parsimony

- different types of parsimony
  - Wagner parsimony
    - binary or ordered multistate characters
    - i.e., states are on a scale and changes occur via intermediate steps only
    - free reversible, $0 \rightarrow 1 = 1 \rightarrow 0$
  - Fitch parsimony
    - binary or unordered multistate characters
    - not on a scale, no intermediate steps required: $\text{cost}(0 \rightarrow 2) = \text{cost}(0 \rightarrow 1)$
    - free reversible
  - Dollo parsimony
    - polarity: ancestral/primitive vs. derived states defined a priori
    - derived states may occur only once on a tree, but can undergo $>1$ reversals
  - Generalized parsimony
    - cost (step) matrices define state transition costs

Character types

<table>
<thead>
<tr>
<th>Character types</th>
<th>Weighted parsimony</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 1 1 1 1 E</td>
<td>differential weight given to each character</td>
</tr>
<tr>
<td>B 1 1 1 1 C</td>
<td>relative importance of upweighted characters grows</td>
</tr>
<tr>
<td>C 1 1 1 1 D</td>
<td>other tree length results: $L = \sum_i l_i w_i$</td>
</tr>
<tr>
<td>D 1 1 1 1 E</td>
<td>different trees may be preferred now</td>
</tr>
<tr>
<td>E 1 1 1 1 D</td>
<td>successive weighting</td>
</tr>
<tr>
<td></td>
<td>problematic</td>
</tr>
<tr>
<td></td>
<td>– risk of circularity</td>
</tr>
</tbody>
</table>

Character types

- Fitch-Parsimonie
  - (ungeordnete Merkmale)
  - (ungeordnete Merkmale - symmetrische Matrix)
- Wagner-Parsimonie
  - (geordnete Merkmale)
  - (geordnete Merkmale - asymmetrische Matrix)
- Dollo-Parsimonie
  - (geordnete Merkmale - asymmetrische Matrix)
Calculating tree length

- initialize \( L = 0 \); root arbitrarily (root taxon);
- postorder traversal: assign each node \( i \) a state set \( S_i \) and adjust \( L \):
  - the set for terminal nodes contains the state of this leaf
  - cycle through all internal nodes \( k \) for which \( S_k \) has not yet been defined, unlike for its child nodes \( m \) and \( n \); if
    \[ S_m \cap S_n \neq \{ \} \rightarrow S_k = S_m \cap S_n; L_m = L_m + 1 \]
    - if \( k \) is parent node to root node, stop; if state at root node not in current \( S_k \),
      \[ L = L + 1 \]

Example:

\[
\begin{align*}
\{T\} \cap \{G\} = \{ \} & \rightarrow S_1 = \{T, G\} \\
\{T, G\} \cap \{T\} \neq \{ \} & \rightarrow S_2 = \{T\} \\
\{A\} \cap \{A\} = \{ \} & \rightarrow S_3 = \{A\} \\
\{T\} \cap \{A\} = \{ \} & \rightarrow S_4 = \{T, A\}
\end{align*}
\]

Tree searching: exhaustive search

- branch addition algorithm
Branch and bound

- \( L_{\text{min}} = L(\text{random tree}) \)
- "search tree" as in branch addition
  - at each level, if \( L < L_{\text{min}} \) go back one level to try another path
  - if at last level, \( L_{\text{min}} = L \) and go back to first level unless all paths have been tried already

Heuristic searches

- stepwise addition
  - as branch addition, but on each level only the path that follows the shortest tree at this level is searched

Star decomposition

Branch swapping

- NNI: nearest neighbour interchanges
- SPR: subtree pruning and regrafting
- TBR: tree bisection and reconnection

Tree inference with many terminals

- general problem of getting trapped in local optima
- searches under parsimony: parsimony ratchet
- searches under likelihood: estimation of
  - substitution model parameters
  - branch lengths
  - topology

Parsimony ratchet

1) generate start tree
2) TBR on this and the original matrix
3) perturb characters by randomly upweighting 5-25%. TBR on best tree found under 2). Go to 2) [200+ times]
4) once more TBR on current best tree & original matrix
5) get best trees from those collected in steps 2) and 4]
Bootstrapping

- estimates properties of an estimator (such as its variance) by constructing a number of resamples of the observed dataset (and of equal size to the observed dataset), each of which is obtained by random sampling with replacement from the original dataset.

Bootstrapping

- variants
  - FWR (Frequencies within replicates)
  - SC (strict consensus)

Bremer support / decay

- Bremer support (decay analysis) is the number of extra steps needed to "collapse" a branch.
- searches under reverse constraints: keep trees only that do NOT contain a given node
- Takes longer than bootstrapping: parsimony ratchet beneficial (~20 iterations)