Parsimony

Genome 559: Introduction to Statistical and Computational Genomics

Elhanan Borenstein
Who am I?

- Faculty at Genome Sciences
- Computational systems biologist
- Training: CS, physics, hi-tech, biology
- Research interests: Complex biological networks | Evolutionary dynamics | Microbial communities and metagenomics

What will change?

- Not much!
- Informatics: From sequence to genes and to systems
- Programming:
  - More emphasis on design and coding practices
  - Tip of the day
  - Coding style
A quick review

- **Trees:**
  - Represent sequence relationships
  - A sequence tree has a topology and branch lengths (distances)
  - The number of tree topologies grows very fast!

- **Distance trees**
  - Compute pairwise corrected distances
  - Build tree by sequential clustering algorithm (UPGMA or Neighbor-Joining).
  - These algorithms don't consider all tree topologies, so they are very fast, even for large trees.
“Maximum Parsimony Algorithm”

A fundamentally different method:

Instead of reconstructing a tree, we will search for the best tree.
“Pluralitas non est ponenda sine necessitate”
(Maximum) Parsimony Principle

- "Pluralitas non est ponenda sine necessitate" (plurality should not be posited without necessity)
  
  William of Ockham

- Occam’s Razor: Of two equivalent theories or explanations, all other things being equal, the simpler one is to be preferred.

  William of Ockham (c. 1288 – c. 1348)

- "When you hear hoof beats, think horses, not zebras"

  Medical diagnosis

- The KISS principle: "Keep It Simple, Stupid!"

  Kelly Johnson, Engineer

- “Make everything as simple as possible, but not simpler”

  Albert Einstein
Parsimony principle for phylogenetic trees

Find the tree that requires the fewest evolutionary changes!
Consider 4 species

- human
- chimp
- gorilla
- orangutan
Consider 4 species

The same approach would work for any discrete property that can be associated with the various species:

- Gene content (presence/absence of each gene)
- Morphological features (e.g., “has wings”, purple or white flowers)
- Numerical features (e.g., number of bristles)
Consider 4 species

Sequence data:

<table>
<thead>
<tr>
<th></th>
<th>human</th>
<th>chimp</th>
<th>gorilla</th>
<th>orangutan</th>
</tr>
</thead>
<tbody>
<tr>
<td>positions in alignment (usually called &quot;sites&quot;)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>agtctc</td>
<td>agagtc</td>
<td>cggcag</td>
<td>cgggac</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Parsimony Algorithm**

1) Construct all possible trees

2) For each site in the alignment and for each tree count the minimal number of changes required

3) Add all sites up to obtain the total number of changes for each tree

4) Pick the tree with the lowest score
Consider 4 species

Sequence data:

<table>
<thead>
<tr>
<th>Species</th>
<th>123456</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>agtctc</td>
</tr>
<tr>
<td>chimp</td>
<td>agagtc</td>
</tr>
<tr>
<td>gorilla</td>
<td>cgccag</td>
</tr>
<tr>
<td>orangutan</td>
<td>cgggac</td>
</tr>
</tbody>
</table>

All possible unrooted trees:

- H closest to C
- or
- H closest to G

- H closest to O
- or

H closest to C

H closest to O
Consider 4 species

Sequence data: 123456

All unrooted trees:

H closest to C

For each site and for each tree count the minimal number of changes required:

H closest to G

or

H closest to O
Consider site 1

What is the minimal number of evolutionary changes that can account for the observed pattern?

(Note: This is the “small parsimony” problem)
Consider site 1

What is the minimal number of evolutionary changes that can account for the observed pattern?

(Note: This is the “small parsimony” problem)
Consider site 1

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>ag</td>
<td>t</td>
<td>c</td>
<td>t</td>
<td>c</td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>ag</td>
<td>a</td>
<td>g</td>
<td>t</td>
<td>c</td>
<td></td>
</tr>
<tr>
<td>gorilla</td>
<td>cg</td>
<td>g</td>
<td>c</td>
<td>a</td>
<td>g</td>
<td></td>
</tr>
<tr>
<td>orangutan</td>
<td>cg</td>
<td>g</td>
<td>g</td>
<td>a</td>
<td>c</td>
<td></td>
</tr>
</tbody>
</table>

Diagram of nucleotide sequences for human, chimp, gorilla, and orangutan.
Consider site 1
Consider site 2

- human: agtcttc
- chimp: agagttc
- gorilla: cggcag
- orangutan: cgggac

Uninformative (no changes)
Consider site 3
Put sites 1 and 3 together
Now put all of them together

Which tree is the most parsimonious?

parsimony score
The parsimony algorithm

1) Construct all possible trees

2) For each site in the alignment and for each tree count the minimal number of changes required

3) Add all sites up to obtain the total number of changes for each tree

4) Pick the tree with the lowest score
The parsimony algorithm

1) Construct all possible trees

2) For each site in the alignment and for each tree count the minimal number of changes required

3) Add all sites up to obtain the total number of changes for each tree

4) Pick the tree with the lowest score
The parsimony algorithm

1) Construct all possible trees

2) For each site in the alignment and for each tree count the minimal number of changes required

3) Add all sites up to obtain the total number of changes for each tree

4) Pick the tree with the lowest score
The parsimony algorithm

1) Construct all possible trees

2) For each site in the alignment and for each tree count the minimal number of changes required

3) Add all sites up to obtain the total number of changes for each tree

4) Pick the tree with the lowest score