A quick review

- Significance of similarity scores (P-values)
  - Empirical null score distribution
  - Extreme value distribution
- Multiple-testing correction (Bonferroni) and E-values

### Global alignment algorithm: *Needleman-Wunsch.*

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0</td>
<td>-4</td>
<td>-8</td>
<td>-12</td>
<td>-16</td>
</tr>
<tr>
<td>A</td>
<td>-4</td>
<td>-5</td>
<td>-9</td>
<td>-13</td>
<td>-12</td>
</tr>
<tr>
<td>T</td>
<td>-8</td>
<td>-4</td>
<td>5</td>
<td>1</td>
<td>-3</td>
</tr>
<tr>
<td>A</td>
<td>-12</td>
<td>-8</td>
<td>1</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>C</td>
<td>-16</td>
<td>-12</td>
<td>2</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

### Local alignment algorithm: *Smith-Waterman.*

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Phylogenetic Trees

Genome 373
Genomic Informatics
Elhanan Borenstein
I think

There between A & B various for of relation. C & B. The first gradation, B & D rather great. But there then genus would be formed. - fairly relation.
Defining what a “tree” means

**rooted tree (all real trees are rooted):**

- Branch points
- Root
- Ancestral sequence
- Branches

**unrooted tree:** (used when the root isn’t known):

- Leaves or tips (e.g., sequences)

Note:
- A tree has topology and distances

...sequence divergence is proportional to (horizontal) branch lengths

Illustration:
- Time radiates out from somewhere (probably near the center)
Are these topologically different trees?
Topologically, these are the SAME tree. In general, two trees are the same if they can be inter-converted by branch rotations.
Why is inferring phylogeny a hard problem?
The number of tree topologies grows extremely fast

In general, an unrooted tree with $N$ leaves has:
- $2N - 3$ total branches
- $N$ leaf branches
- $N - 3$ internal branches
- $N - 2$ internal nodes
- $3\times5\times7\times\ldots\times(2N-5) \sim O(N!)$ topologies
There are many rooted trees for each unrooted tree

For each unrooted tree, there are $2N - 3$ times as many rooted trees, where $N$ is the number of leaves ($\# \text{ branches} = 2N - 3$).

20 leaves - $564,480,989,588,730,591,336,960,000,000$ topologies
How can you compute a tree?

- Many methods available, we will talk about:
  - Distance trees
  - Parsimony trees

- Others include:
  - Maximum-likelihood trees
  - Bayesian trees