Review:

1. Character, states
2. Parsimony method
3. Ancestor construction
4. Perfect phylogeny
5. Maximum likelihood
Distance Based Methods

• Input of distance based methods is an $n \times n$ distance matrix $d(i,j)$.
• We want to compute a tree with $n$ leaves, with edge weights. $T(i,j)$ is the distance of two leaves on the tree.
• We want to minimize
• This is also NP-hard.
• So we use heuristics.

$$\sum_{i,j} |d(i,j) - T(i,j)|^2$$
\[ d(a, c) = \frac{d(A, c) + d(B, c)}{2} \]

### Table

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
• Unweighted Pair Group Method with Arithmetic mean.
• A heuristic method with no performance guarantee.
• At each time, it finds $i,j$ with the minimum distance.
• Merge the two taxa $i,j$ into a new one $u$. Update the distance matrix.
• For any $k$, let: $d(u,k) = (d(i,k)+d(j,k))/2$.
• Recursion.
Neighbor Joining


- Neighbor Joining uses a similar idea as UPGMA. But it uses a more sophisticated formula to determine the two neighbors to be joined.

\[ Q(i, j) = (r - 2)d(i, j) - \sum_{k=1}^{r} d(i, k) - \sum_{k=1}^{r} d(j, k) \]

- Find the minimum \( Q(i, j) \), merge i,j to a new node u.

- Update \( d(k, u) = \frac{d(k, i) + d(k, j) - d(i, j)}{2} \).

- Recursion on the remaining \( r-1 \) nodes.

\[
-Q(\hat{\mathbf{i}j}) = \left( \frac{2}{r} \sum_{k=1}^{r} (d(\hat{\mathbf{i}}, k) + d(\hat{\mathbf{j}}, k) - d(\hat{\mathbf{i}}, \hat{\mathbf{j}})) \right) + 2d(\hat{\mathbf{i}}, \hat{\mathbf{j}}).
\]
Neighbor Joining Idea

\[ d(k,u) = \frac{(d(k,i) + d(k,j) - d(i,j))}{2} \]
Methods for constructing phylogeny

• Exhaustive Search
  • Parsimony Method: Find a tree topology so that the total number of mutations on the edges is the smallest.
  • Perfect Phylogeny.
  • Maximum Likelihood.

• Distance Based Method
  • UPGMA
  • Neighbour Joining

• Quartet Method

• Whole Genome Phylogeny.
Quartet Methods

• For each group of four species, construct a tree of 4 (quartet), using your most favorite method, say maximum likelihood.

• Then find a tree that is most consistent with all the quartets.

• The problem is NP-hard (to find the tree with least error).

• There is a PTAS to do this (T. Jiang, P. Kearney, and M. Li. Orchestrating quartets: approximation and data (FOCS'98))
Challenges in Phylogeny of Chain Letter

• Parsimony or maximum likelihood: How do we know what is a “character” – a feature to look at?
  • In case of a chain letters?
  • In case of whole genome phylogeny?
• Distance based: What distance to use? edit distance doesn't work
• A “universal” solution -- information distance.
Information Distance

- Let x and y be two strings. P be the shortest program that takes x as input and computes y.
- The length of this shortest program defines $K(y|x)$. K is called the Kolmogorov Complexity.
- The information distance between x and y is defined as
  \[ d(x,y) = \max \{ K(x|y), K(y|x) \} / \max\{K(x),K(y)\}. \]
- Kolmogorov complexity is incomputable, but we can use compression in practice.
$k(x)$: shortest program length to output $x$.

$\pi = 3.14159265357\ldots$

$k(x(y))$: shortest program length to accept $y$ as input and output $x$.

$$d(x, y) = \frac{\max(k(x(y)), k(y|x))}{\max(1 + k(x), k(y))}$$
Reconstructing History of Chain Letters

For each pair of chain letters \((x, y)\) we computed \(d(x, y)\), hence a distance matrix.

A DNA compression program is used to compute the information distance.

Using Neighbor Joining to construct their evolutionary history based on the \(d(x, y)\) distance matrix.

The resulting tree is an almost perfect phylogeny: distinct features are all grouped together.
Phylogeny of 33 Chain Letters

- Adding the phrase "provided you send it back out."
- "And all things" title mutation
- "Love" title mutation
- "St. Jude & $470k" mutation
- "Kiss" title mutation
- "$50k lottery" mutation
- "St. Jude, ear"
- "Trust in the Lord"
- "And all things whatever..."
- With love all things are possible
- Kiss someone you love... and make magic.
- Netherlands; For no reason should this chain be broken.
Summary

• Exhaustive Search
  • Parsimony Method: Find a tree topology so that the total number of mutations on the edges is the smallest.
  • Perfect Phylogeny.
  • Maximum Likelihood.
• Distance Based Method
  • UPGMA
  • Neighbour Joining
• Quartet Method
• Whole Genome Phylogeny.
• Chain Letter and 2019 nCoV