CSCI1950-Z
Computational Methods for Biology
Lecture 6

Ben Raphael
February 9, 2009
http://cs.brown.edu/courses/csci1950-z/

Outline

• Neighbor joining
• Deriving distances
• Probabilistic Models of Phylogeny
  – Maximum likelihood
Neighbor Joining as a Pair Group Method

Iteratively combine leaves/groups minimizing selection criteria into larger groups.

\[ C \leftarrow \{ \{1\}, \ldots, \{n\} \} \]

While \(|C| > 2\) do

[Select pair of clusters.]

\[ s(C_x, C_y) = \min s(C_i, C_j). \]

\[ C_k \leftarrow C_x \cup C_y \]

[Replace \(C_x\) and \(C_y\) by \(C_k\).]

\[ C \leftarrow (C \setminus C_x \setminus C_y) \cup C_k. \]

NJ Selection Criterion

Let \(C = \{1, \ldots, n\}\) be current clusters/leaves.

Define: \(u_i = \sum_k D(i, k)\).

Intuitively, \(u_i\) measures separation of \(i\) from other leaves.

**Goal:** Minimize \(D(i, j)\) and maximize \(u_i + u_j\).

**Solution:** Find pair \((i, j)\) that minimizes:

\[ S_D(i, j) = (n - 2) D(i, j) - u_i - u_j \]

**Claim:** Given additive matrix \(D\).

\[ S_D(x, y) = \min S_D(i, j) \text{ if and only if } x \text{ and } y \text{ are neighbors in tree } T \text{ with } d_T = D. \]

**Proof**
Uniqueness of NJ Selection Criterion
(Bryant 2005)

\[ S_D(i, j) = (n - 2) \ D(i, j) - u_i - u_j \]
is the unique criterion satisfying:

1. Consistency: If \((x,y)\) are neighbors then, \(S_D(x,y)\) is minimal.
2. \(S_D(i,j)\) is linear and continuous wrt \(D\).
3. \(S_D\) is invariant under permutations of leaves.

Compact Additive Trees

Compact Additive Tree Problem

Given an \(n \times n\) distance matrix, determine if there is an additive tree for \(D\) with exactly \(n\) vertices?

Let \(G(D)\) be the complete graph with edge weights
\(w(\ (i,j) \ ) = D_{ij}.\)

**Theorem:** If there is a compact additive tree for \(D\), then \(T\) must be the unique minimum spanning tree of \(G(D)\).

Recall: A spanning tree is a tree containing all vertices. Minimum spanning tree has the least total weight.
**Distance Matrix**

Sequence a gene of length $m$ in $n$ species $\rightarrow n \times m$ alignment matrix.

<table>
<thead>
<tr>
<th></th>
<th>Mouse: ACAGTGACGCCACACACGT</th>
<th>Gorilla: CCTGCGACGTAAACACCGC</th>
<th>Chimpanzee: CCTGCCAGTTGCAAAACGC</th>
<th>Human: CCTGCCAGTTGCAACACGA</th>
</tr>
</thead>
</table>

Transform into...

```
0  7  11  10
7  0  4  6  n x n distance matrix
11 4 0 2
10 6 2 0
```

**Distances from Sequences**

Chimpanzee: CCTGCCAGTTGCAAAACGC

Ancestor: CCCGCGACTTAAACAAACGC

Human: CCTGCCAGTTGCAACACGA

Convergent  Multiple

Hamming distance = 3.

$D_s = \text{differences per site} = 3/20$

12 total mutations.
Observed Differences vs. Jukes-Cantor

Given a character matrix M, what is the “most likely” tree T that generated M?

Assume:
1. Characters evolve independently.
2. Constant rate of mutation on each branch.
3. State of a vertex depends only on parent and branch length: i.e. Pr[x | y, t] depends only on y and t. (Markov process)