Evolutionary Trees

1) Parsimony

2) Distance-based

3) Maximum likelihood

Distance-based methods

Given $n \times n$ matrix of all pair-wise distances between the $n$ sequences

\[ d_{ij} \]

Hierarchical Agglomerative clustering

1) Distance are over a "metric space"

2) Additive

\[ d_{ij} = d_{ik} + d_{kj} \]
UPGMA: Unweighted Pair Group Method using Arithmetic Average

1) pick the \( \min \{d_{ij}\} \)

2) \( c_k = c_i + c_j \)

3) remove \( c_i \) & \( c_j \) from consideration
   update \( d_{km} \) from new cluster to everyone else

\[
d(c_i; c_j) = \text{avg pairwise distance (UPGMA)}
= \frac{1}{|c_i|} \frac{1}{|c_j|} \sum \sum d_{ij}
\]
\[ d(C_k, C_m) = \frac{|c_i| \cdot d_{im} + |c_j| \cdot d_{jm}}{|c_i| + |c_j|} \]

\[
\begin{array}{cccc}
2 & 3 & 4 \\
1 & 0.3 & 0.5 & 0.6 \\
2 & & 0.6 & 0.5 \\
3 & & & 0.9 \\
\end{array}
\]

\[
\begin{array}{cccc}
k & 12 & & \\
1 & & 0.55 & 0.55 \\
2 & & & 0.9 \\
\end{array}
\]

\[
d(12, 3) = \frac{d(1, 3) + d(2, 3)}{2}
\]

\[
d(123, 4) = \frac{2d(12, 4) + d(3, 4)}{3}
\]

\[
= 1.1 + 0.9 = \frac{2}{3}
\]
Other distance notions

1) UPMA: mean pairwise distance
   \[ d(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{s_i} d_{ij} \]

2) Single link (min): \[ d(C_i, C_j) = \min_{q,a,b,c} \{ d_{ab} \} \]

3) MAX: complete link
   \[ d(C_i, C_j) = \max_{a,b} \{ d_{ab} \} \]
distance between means

\[ d(C_i, C_j) = \| \mu_i - \mu_j \|_2 \]

Euclidean Distance

additive

not ultrametric

UPGMA (only works for ultrametric distance)

1) distance between means

not applicable to the distance matrix
**Neighbor Joining**

```
```

**"avg" distance**

\[ \gamma_1 = 1.5/2 = 0.7 \]
\[ \gamma_2 = 0.7 \]
\[ \gamma_3 = 3.0/2 = 1.0 \]
\[ \gamma_4 = 1.0 \]

\[ \gamma_i = \frac{1}{n-2} \sum_{j \neq i} d_{ij} \]

\[ d_{ik} = \frac{1}{2} (d_{ij} + \gamma_i - \gamma_j) \]

\( k = (1, 3) \)

\[ d_{13k} = \frac{1}{2} (d_{13} + \gamma_1 - \gamma_3) \]

\[ = \frac{1}{2} (0.5 + 0.7 - 1.0) \]

\[ = \frac{1}{2} (0.2) = 0.1 \]

\[ d_{jk} = d_{ij} - d_{ik} \]

\[ = 0.4 \]
\[ d_{km} = \frac{1}{2} (d_{im} + d_{jm} - d_{ij}) \]

\[
\begin{array}{c|c|c}
 & 2 & 4 \\
\hline
1 & 0.2 & 0.5 \\
\hline
2 & 0.5 & \\
\end{array}
\]

\[ k = (1, \frac{1}{3}) \]

\[ = \frac{1}{2} (0.3 + 0.6 - 0.5) = \frac{1}{2} (0.4) = 0.2 \]

\[ d_{k4} = \frac{1}{2} (1.5 - 0.5) = 0.5 \]
Maximum Likelihood Approach

Data: $s_1, s_2, \ldots, s_n$ => find the best $T$ (evolutionary tree)

$$P(T \mid \text{data}) = \frac{P(\text{data} \mid T) \cdot P(T)}{P(\text{data})}$$

Posterior Probability of $T$

1) Search over trees $T$

2) maximize the likelihood given $T$

$T$ is given means only its topology

Parameter space $(x, y, t_i)$

$$P(\ x, y \mid T) = P(x \to A \mid t_1) \cdot P(x \to C \mid t_2) \cdot P(y \to A \mid t_3) \cdot P(y \to a | t_4)$$
\[ P(\text{dak}|T, t_i) = \sum_{x} \sum_{y} P(x, y | T) \]

\[ \theta = \sum t_i \]

Gene level changes → Insert, Delete, Substitutions

Genome level events

Translocation

Inversion

Duplicated event

Complemented
\[ d(H, M) \text{: Inversion distance} = \text{least number of inversions to convert } M \text{ into } H \]

Refrence: H: 1 2 3 4 5 6 7 8 9 10

M: 4 5 6 1 7 [2 9 8] 10 3

Reversals/Inversions are allowed.