Phylogeny and Evolution

Gina Cannarozzi
ETH Zurich
Institute of Computational Science
Aristotle (384-322 BC) classified animals. He found that dolphins do not belong to the fish but to the mammals.

Carolus Linneus (1758) introduced binomial classification.

Darwin 1859 explained evolution as a process of random mutation and natural selection.

Zimmerman in the 1930s and Hennig in the 50’s began to define objective measures for reconstructing evolutionary history based on shared attributes of extant and fossil organisms. They worked on cladistics- the systematic classification of organisms based “shared derived properties”

1965 Zuckerkandl and Pauling were the first to use molecular sequences as indicators of phylogeny.
Introduction

Goal: reconstruct the evolutionary history of life

Phylogenetic Tree of Life

Carl Woese proposed the third domain or kingdom of life based on ribosomal RNA in 1990.
Motivation

Fig. 1.2: The case of the Florida dentist. Each branch represents the sequence from part of the envelope (env) gene of HIV-1. Viral sequences were obtained from the dentist and seven of his former patients (labelled A to G), all infected with the virus. Five of these patients (A, B, C, E and G), have sequences very closely related to those from the dentist (boxed), suggesting that he infected them. Two of his other former patients (D and F) had other risk factors for HIV infection and their viruses are separated from the dentist by sequences taken from local controls (LC) — HIV-infected individuals living within a 90-mile radius of the dentist’s surgery. Because HIV-1 is so variable, two different sequences are included for the dentist and patient A. Data taken from Ou et al. (1992).

Figure 3.1: An evolutionary tree for the word "snow", showing the reconstructed ancestral word *snigwh-
Proto-Indoeuropean >3000 B.C.

- snigu Church Russian
- sniș Church Romanian
- snaj Church Slovenian
- sne Church Slavonic

- Old Russian
- Old French
- Old Irish
- Old Norse
- Old High German
- Old English
- Gothic
- Greek

- Italian
- Spanish
- French

- Norwegian
- Danish
- Swedish
Topology

Unrooted Tree

Rooted Tree

- Root
- Internal node
- Leaf node

 topology - shape of tree, branching order between nodes

rotation about a branch does not change the topology
Tree representations

\[ ((A,B)(C,D)) = ((B,A)(C,D)) = ((C,D),(B,A)) \]

\[
\text{Tree(\text{Tree(Leaf(A,L1+L3,1),L3,Leaf(B,L2+L3,2)), 0, Tree(Leaf(D,L6+L4,4),L4,Leaf(C,L5+L4,3))))}
\]
Tree Components

• topology - branching pattern of a tree
• root- place on the tree from which everything evolves- common ancestor of everything at the leaves
• external nodes, leaves, taxonomic units
• internal nodes or hypothetical taxonomic units (HTU) represent speciation or gene duplication events
• branches or edges - can have a length
Rooting a tree

• Most phylogenetic methods produce unrooted trees. This is because they detect differences between sequences, but have no means to orient residue changes relatively to time.

• There are two ways to root an unrooted tree:
  • use an outgroup- include a group of sequences known to be outside the group of interest
  • assume a molecular clock- all lineages have evolved with the same rate from their common ancestor (usually not a good assumption)

Phylogenetic Tree of Life
Phylogenetic Trees:

graphical representation of the evolutionary history of a set of species

ancestor of mammals

ancestor of vertebrates
alignment implies an evolutionary relationship also represented by Phylogenetic Tree
aligns amino acids that diverged from the same residue in (hypothetical) most recent common ancestor
darwinian evolution is driven by random mutation and natural selection
our model allows for point mutations and insertions/deletions (indels)
mutations may be adaptive, neutral or deleterious
alignment shows accepted substitutions since divergence
proteins evolve under functional constraints - mutations that destroy function do not appear in database via organism death
"correct" alignment represents actual events- substitutions, indels
impossible to verify -> take alignment with the highest probability that the alignment is correct under our model
String Alignments

[Rice, Mosquito] triosephosphate isomerase

lengths=55,53 simil=117.9, PAM_dist=111, identity=36.4%
NGTTDQVDKIVKILNEGQIASTDVEVVEVVSPPYVFLPVVKSQRLPEIQVAAQNCW
||....!...!|!..|.|!|..|.!.:. .||||. | .!|.:.!|||...! ||||||!
NGDKASIADLCKVLTTGPLNAD__TEVVVGCPAPYTLARSQLPDSVCVAAQNCY

Similarity Score (Likelihood Based)
PAM distance (evolutionary distance)

For pairwise string alignments, the dynamic programming algorithm guarantees that the highest scoring alignment is found.

Local alignment- find the highest scoring substring
Global alignment- find the highest score for aligning the complete strings
PAM distance

- Evolutionary distance (not time)
- definition: a 1 PAM transformation is an evolutionary step where 1% of the amino acids are expected to mutate
- M is a mutation matrix for which each element describes a probability of a mutation

\[ M_{ij} = \Pr(x_j \rightarrow x_i) \]

\[
M = \begin{pmatrix}
0.98 & 0.01 & \ldots & 0.01 \\
0 & 0.99 & \ldots & 0.002 \\
\vdots & \vdots & \ddots & \vdots \\
0.001 & 0 & \ldots & 0.97
\end{pmatrix}
\]

\[
\sum_{i=1}^{20} f_i (1 - M_{ii}) = 0.01
\]
Similarity score

Our score compares two events - the probability of alignment by reasons of common ancestry divided by the probability of alignment by random chance

\[
\frac{\text{CommonAncestry}}{\text{Chance}} = 10 \log_{10} \frac{f_A M_{AS}^2}{f_A f_S} = D_{AS}
\]

- -A- - sequence 1
- -X- - ancestor X.
- -S- - sequence 2

Match by Chance
\[
Pr\{A\} Pr\{S\} = f_A f_S
\]

Pr\{A and S from Ancestor X\}
\[
\sum_X f_X Pr\{X \rightarrow A\} Pr\{X \rightarrow S\}
\]

\[
= \sum_X f_X M_{AX} M_{SX}
\]

\[
= \sum_X f_S M_{AX} M_{XS}
\]

\[
= f_S M_{AS}^2
\]

\[
= f_A M_{SA}^2
\]
### Dayhoff Matrices

![Diagram showing the processes involved in Dayhoff Matrices]

#### 1 PAM

<table>
<thead>
<tr>
<th>Letter</th>
<th>C</th>
<th>S</th>
<th>T</th>
<th>P</th>
<th>A</th>
<th>G</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>Q</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>17.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>-18.5</td>
<td>12.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-21.6</td>
<td>12.7</td>
<td>12.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>-33.2</td>
<td>-18.6</td>
<td>-19.5</td>
<td>13.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-18.1</td>
<td>-14.3</td>
<td>-17.5</td>
<td>-18.8</td>
<td>11.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-25.2</td>
<td>-18.7</td>
<td>-25.3</td>
<td>-24.9</td>
<td>-18.2</td>
<td>11.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-24.1</td>
<td>-15.5</td>
<td>-17.5</td>
<td>-24.0</td>
<td>-22.3</td>
<td>-19.1</td>
<td>13.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-32.1</td>
<td>-18.7</td>
<td>-20.0</td>
<td>-22.7</td>
<td>-21.2</td>
<td>-20.5</td>
<td>-14.0</td>
<td>12.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-35.3</td>
<td>-19.4</td>
<td>-20.8</td>
<td>-21.6</td>
<td>-18.6</td>
<td>-23.7</td>
<td>-19.5</td>
<td>-12.8</td>
<td>12.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>-28.7</td>
<td>-18.4</td>
<td>-18.9</td>
<td>-19.7</td>
<td>-19.4</td>
<td>-22.8</td>
<td>-17.4</td>
<td>-18.7</td>
<td>-13.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 250 PAM

<table>
<thead>
<tr>
<th>Letter</th>
<th>C</th>
<th>S</th>
<th>T</th>
<th>P</th>
<th>A</th>
<th>G</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>Q</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>11.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>0.1</td>
<td>2.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-0.5</td>
<td>1.5</td>
<td>2.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>-3.1</td>
<td>0.4</td>
<td>0.1</td>
<td>7.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.5</td>
<td>1.1</td>
<td>0.6</td>
<td>0.3</td>
<td>2.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-2.0</td>
<td>0.4</td>
<td>-1.1</td>
<td>-1.6</td>
<td>0.5</td>
<td>6.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>-1.8</td>
<td>0.9</td>
<td>0.5</td>
<td>-0.9</td>
<td>-0.3</td>
<td>0.4</td>
<td>3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-3.2</td>
<td>0.5</td>
<td>-0.0</td>
<td>-0.7</td>
<td>-0.3</td>
<td>0.1</td>
<td>2.2</td>
<td>4.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-3.0</td>
<td>0.2</td>
<td>-0.1</td>
<td>-0.5</td>
<td>-0.0</td>
<td>-0.8</td>
<td>0.9</td>
<td>2.7</td>
<td>3.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>-2.4</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.2</td>
<td>-0.2</td>
<td>-1.0</td>
<td>0.7</td>
<td>0.9</td>
<td>1.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>-1.3</td>
<td>-0.2</td>
<td>-0.3</td>
<td>-1.1</td>
<td>-0.8</td>
<td>-1.4</td>
<td>1.2</td>
<td>0.4</td>
<td>0.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[www.biorecipes.com/Dayhoff/code.html]
Multiple Sequence alignments

<table>
<thead>
<tr>
<th></th>
<th>ATGCATGGGCAACATGACCAGGAGTTGCTTGTCGGTCCAAACACAGCGTT---GGCTCTCTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xenopus</td>
<td>ATGCATGGGCCAGCATGACCAGCAGGAGGTAGC---CAAAATAACACCAACATGCAAATG</td>
</tr>
<tr>
<td>Gallus</td>
<td>ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACCCAAAACAGCAACACGTGCAAATG</td>
</tr>
<tr>
<td>Bos</td>
<td>ATGCATCCGCCACCATGACCAGCAGGAGGTAGCTCTAAACAGCAACACGTGCAAATG</td>
</tr>
<tr>
<td>Homo</td>
<td>ATGCATCCGCCACCATGACCAGCAGGAGGTAGCACCTCAAAACAGCAACACGTGCAAATG</td>
</tr>
<tr>
<td>Mus</td>
<td>ATGCATCCGCCACCATGACCAGCAGGAGGTAGCTCTCAAAACAGCAACACGTGCAAATG</td>
</tr>
<tr>
<td>Rattus</td>
<td>ATGCATCCGCCACCATGACCAGCAGGAGGTAGCTCTCAAAACAGCAACACGTGCAAATG</td>
</tr>
</tbody>
</table>

- each column is descended from one position in the sequence of the common ancestor
- can not be built by algorithms which guarantee optimal score
- reasonable heuristic algorithms for constructing MSAs exist - clustal, MAlign, T-Coffee
Markovian Model of Evolution

• mutations occur with probability independent of previous substitutions
• substitutions occur independently at different positions in the polypeptide chain
• a single substitution matrix represents the probability of amino acid substitution at any position

Proteins do not have Markovian Behavior

distant residues come together in the 3D fold and influence each other
surface amino acids tolerate more variation than interior residues
biological function constrains accepted substitutions - active site conservation
back mutations are more probable  L -> I -> L
chemically similar substitutions are more probable

nature is too complex to model exactly
things that do not fit in our evolutionary model

- Lateral Gene Transfer
- Convergent evolution (flight evolved 5 different times)
- Reversals (snakes)
Phylogenetic Trees
How to build trees

- Starting point: molecular sequences (for this discussion)
- Goal: a phylogenetic tree describing the evolutionary relationships of the taxa
How many trees are there?

<table>
<thead>
<tr>
<th>Number of leaves</th>
<th>Number of unrooted trees</th>
<th>Number of rooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>945</td>
</tr>
<tr>
<td>10</td>
<td>2027025</td>
<td>34459425</td>
</tr>
<tr>
<td>20</td>
<td>2.216e+20</td>
<td>8.201e+21</td>
</tr>
<tr>
<td>50</td>
<td>2.838e+74</td>
<td>2.753e+76</td>
</tr>
<tr>
<td>(n)</td>
<td>((2n - 5)!!)</td>
<td>((2n - 3)!!)</td>
</tr>
</tbody>
</table>

Conclusion: We can not evaluate every tree topology when searching for the highest scoring tree.
Clustering Algorithms

For certain types of trees, clustering algorithms will work well

- Ultrametric Trees
- Additive Trees

Advantage: very fast
Disadvantage: most real trees do not satisfy these conditions.
Ultrametric Trees

\[ D_{AX} = D_{BX} = D_{CX} \]

- Assume all evolution occurs at the same rate (molecular clock)
- Assume all distances are measured without error
- Assume all leaves are equidistant from the root
- UPGMA (unweighted pair group method with arithmetic averages) algorithm for tree building will usually work well for these trees (not mathematically guaranteed)
UPGMA

- Find i and j that have minimum entry D[i,j] in D
- Create new group (ij) which has n_{ij} = n_i + n_j members
- Connect i and j on the tree to a new node which corresponds to the group (ij). Give the two branches connecting i to (ij) and j to (ij) each length D_{ij}/2
- Compute distances of all nodes k to (ij) - as
  d[k,ij] = (n_i/(n_i+n_j))*d[k,i] + (n_j/(n_j+n_j))d[k,j]
- Repeat while number of matrix elements is > 1

```
join d and c

<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>12</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>24</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

join a and b

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c,d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>c,d</td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>a,b</th>
<th>c,d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a,b</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>c,d</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
```
Additive Trees

- assume that pairwise distances have no error
- assume that distances in matrix correspond exactly to branch lengths
- neighbor-joining algorithm is guaranteed to recover the true tree if the distance matrix is an exact reflection of the tree

\[ d(A, B) = L_1 + L_2 \]
\[ d(A, C) = L_1 + L_3 + L_4 \]
\[ d(B, C) = L_2 + L_3 + L_4 \]
**neighbor joining algorithm**

- does not assume clock-like evolution

- For each tip, compute \( u_i = \sum_{j:j \neq i} D_{ij} / (n - 2) \).

- Choose the \( i \) and \( j \) for which \( D_{ij} - u_i - u_j \) is the smallest

- Join items \( i \) and \( j \). Compute the branch length from \( i \) to the new node \((v_i)\) and from \( j \) to the new node \((v_j)\) as:

  \[
  v_i = 1/2D_{ij} + 1/2(u_i - u_j)
  \]

  \[
  v_j = 1/2D_{ij} + 1/2(u_j - u_i)
  \]

- Compute the distance between the new node \((ij)\) and each of the remaining tips as

  \[
  D_{ij,k} = (D_{ik} + D_{jk} - D_{ij})/2
  \]

- Delete tips \( i \) and \( j \) from the tables and replace them by the new node \((ij)\) which is now treated as a tip.

- if more than 2 nodes remain go back to step 1. Otherwise connect the 2 remaining nodes (say \( l \) and \( m \)) by a branch of length \( D_{lm} \).
Finding the Optimal Tree

- Construct an initial tree
  - Random tree
  - Heuristic for specific data types (Neighbor joining or UPGMA)
- Search for better scoring topologies using 4-, 5-, or 6-optim while evaluating the tree with a given scoring function (parsimony, distance, or likelihood)
- Continue to optimize under a scoring criterium until the score no longer improves
There are 3 different topologies with 4 subtrees.

- Divide the tree into 4 subtrees (A, B, C and D)
- Compute the quality for all possible topologies
- Select the best configuration
- Repeat for different subtrees until there is no improvement
5-optim and 6-optim

- 4-optim improves the topologies towards the leaves
- 5- and 6-optime improve towards the interior of the tree

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4-optim</td>
<td>4 subtrees</td>
<td>3 topologies</td>
</tr>
<tr>
<td>5-optim</td>
<td>5 subtrees</td>
<td>15 topologies</td>
</tr>
<tr>
<td>6-optim</td>
<td>6 subtrees</td>
<td>105 topologies</td>
</tr>
</tbody>
</table>
### Types of Tree Construction Methods

- **Character based - Parsimony**
- **Distance based - least squares**
- **Probability based - Maximum Likelihood or Bayesian**

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distance</strong></td>
<td>pairwise distance matrix</td>
<td>branch lengths topology</td>
</tr>
<tr>
<td><strong>Parsimony</strong></td>
<td>character tables (multiple sequence alignment)</td>
<td>topology</td>
</tr>
<tr>
<td><strong>Maximum Likelihood</strong></td>
<td>pairwise dist. matrix multiple sequence alignment</td>
<td>branch lengths topology</td>
</tr>
</tbody>
</table>
**Distance trees**

- Input: Distance matrix $D$ describing the measured distance between all taxa of interest

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$D(A,B)$</td>
<td>$D(A,C)$</td>
<td>$D(A,D)$</td>
</tr>
<tr>
<td>B</td>
<td>$D(B,C)$</td>
<td>$D(B,D)$</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>$D(C,D)$</td>
</tr>
</tbody>
</table>

D's come from pairwise sequence alignments

$d(A,B) = L_1 + L_2$

$d(A,C) = L_1 + L_3 + L_4 + L_5$

$d(A,D) = L_1 + L_3 + L_4 + L_6$

$d(B,C) = L_2 + L_3 + L_4 + L_5$

$d(B,D) = L_2 + L_3 + L_4 + L_6$

$d(C,D) = L_5 + L_6$

The Ls are fit
What to minimize

\[ Q = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} (D_{i,j} - d_{i,j})^2 \]

where \( Q \) is what we are trying to minimize, \( n \) is the number of leaves, \( w_{i,j} \) is a weighting factor, 1 over the Pam variance, \((1/\sigma^2)\), \( D \) is the matrix of experimentally determined distances from the pairwise alignments (for example), \( d \) is a matrix of distances calculated from the fit tree.
Distance Methods

- consider pairwise distances as estimates of the branch length separating two species
- each distance infers the best unrooted tree for that pair of species
- in effect, we have many estimated 2-species trees and we try to find the best n-species tree implied by them
- individual distances are not exactly the path lengths in the full n-species tree between any two species
- we search for the full tree that does the best job of approximating these individual two-species trees
- search for the branch lengths and topologies that minimize difference between approximated branch lengths and experimental branch lengths
- for a given topology, it is possible to solve for the branch lengths that minimize Q using standard least squares methods
## Character Based Methods

### What is a character?
- finite number of states
- discrete

<table>
<thead>
<tr>
<th></th>
<th>backbone</th>
<th>skull opening</th>
<th>hip socket</th>
<th>grasping</th>
<th>warm-blooded</th>
</tr>
</thead>
<tbody>
<tr>
<td>alligator</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T. rex</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sparrow</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chimp</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Perfect Phylogeny

each character fits on one branch of a phylogenetic tree

changes in character happen only once

species with the same character are all under the same subtree

<table>
<thead>
<tr>
<th></th>
<th>backbone</th>
<th>skull opening</th>
<th>hip socket</th>
<th>grasping</th>
<th>warm-blooded</th>
</tr>
</thead>
<tbody>
<tr>
<td>alligator</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T. rex</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sparrow</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chimp</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>cat</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Parsimony

For molecular sequence data, each column of the MSA will be considered a character.

<table>
<thead>
<tr>
<th>Animal</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xenopus</td>
<td>ATGCATGGGCCAACATGACCAGGAGTTGGTGTCGGTCCAAACAGCGTT---GGCTCTCTA</td>
</tr>
<tr>
<td>Gallus</td>
<td>ATGCATGGGCCAGCATGACCAGCAGGAGGTAGC---CAAATAACACCAACATGCAAATG</td>
</tr>
<tr>
<td>Bos</td>
<td>ATGCATCCGCCACCATGACCAGCAGGAGGTGACACCACACGCACACGGTGCAATG</td>
</tr>
<tr>
<td>Homo</td>
<td>ATGCATCCGCCACCATGACCAGCAGGAGGTAGCCTCAAAACAGCAACCGTGCAATG</td>
</tr>
<tr>
<td>Mus</td>
<td>ATGCATCCGCACCCATGACCAGCAGGAGGTAGCCTCAAAACAGCAACCGTGCAATG</td>
</tr>
<tr>
<td>Rattus</td>
<td>ATGCATCCGCACCCATGACCAGCAGGAGGTAGCCTCAAAACAGCAACCGTGCAATG</td>
</tr>
</tbody>
</table>

*****  ****  *********  *  ***  *   * *** * *             *
Parsimony

The parsimony score is the number of changes of state on the evolutionary tree. The most parsimonious tree is that which minimizes the amount of evolutionary change.

The topology is given, parsimony is a method for finding the tree with the least amount of state changes.

The highest scoring tree minimizes the number of changes.

Occam's Razor- William of Occam (1300-1349):
Entities should not be multiplied more than necessary- the fewer assumptions an explanation of a phenomenon depends on, the better it is.
Parsimony Algorithm

Use labels at leaves to reconstruct the possible labels at internal nodes

- Compare the labels at each of the two children of each node.
- If there is an intersection of the two sets of labels, the parent node is labeled with the result of the intersection and there is no penalty.
- If the intersection is empty, then the node is labeled with the union of the two sets of labels and the penalty increases by +1.
- Continue from the leaves to the root until all nodes have been labeled.
Parsimony

Number of Changes: 3

Characters

A B

C D

E F

Number of Changes: 3
Optimizing under parsimony

- For a given topology and alignment position, determine what ancestral residues require the least amount of changes.
- Compute this for each alignment column (character). Add the number of changes for each position together to obtain the parsimony score (length of the tree).
- Compute this score for many tree topologies and keep the one(s) with the lowest score.
Assigning ancestral states

• start at the root, if the set contains more than one character, pick one at random
• Move from the root towards the leaves. If an intersection exists between the chosen state of the parent and the child, choose it. If not, choose another character at random
• Many trees may exist with the same parsimony score

Characters

\[ \begin{align*} 
\text{A} & \quad \text{R} \\
\text{B} & \quad \text{T} \\
\text{C} & \quad \text{T} \\
\text{D} & \quad \text{T} \\
\text{E} & \quad \text{G} \\
\text{F} & \quad \text{G} \\
\end{align*} \]

\[ \begin{align*} 
\text{G,T} & +1 \\
\text{T} & \quad \text{T} \\
\end{align*} \]
Parsimony problems

Inconsistency

true tree  \rightarrow \text{parsimony tree}

Backflips

there is no information on branch length, only change or no change
Maximum Likelihood

• Maximum Likelihood: general parameter estimation procedure
• Parameters are estimated from the data D such that the likelihood $L$ of the data given the parameters is maximized
• parameters - tree topology and branch lengths
• input data - aligned molecular sequences
• goal: find the topology and branch lengths that maximize the likelihood of the data
• Use Dayhoff matrices to obtain the likelihood of a transition for a given period of time (PAM distance).
Maximum Likelihood

\[
L(T)_i = \sum_{X_i} Pr(X_i) \times \sum_{Y_i} Pr_{L3}(X_i \rightarrow Y_i) Pr_{L1}(Y_i \rightarrow A_i) Pr_{L2}(Y_i \rightarrow B_i) \\
\times \sum_{Z_i} Pr_{L4}(X_i \rightarrow Z_i) Pr_{L5}(Z_i \rightarrow C_i) Pr_{L6}(Z_i \rightarrow D_i)
\]
Selecting data to Reconstruct Species Trees

- Sequences must be derived from a common ancestor (Homologous)
- Orthologs - sequences related by a speciation event
- Paralogs - sequences related by a gene-duplication event
Tree of Life

Archaea

Bacteria

Eukaryots

lives at pH 2
Thermoplasma acidophilum

Methanobacterium thermocautrophilicum

Pyrococcus furiosus

likes it the most at 100 °C
Thermoplasma acidophilum

Archaea

Bacteria

endures extreme radiation
Deinococcus radiodurans

Mycobacterium bovis

Escherichia coli

Yersinia pestis

Shewanella oneidensis

Buchnera aphidicola

Ralstonia solanacearum

Serratia pomeroyi

Agrobacterium tumefaciens

Bradyrhizobium japonicum

infests plants
N₂ fixation

important for bread and beer
the things flying around in the kitchen

causes malaria

Lactobacillus lactis

Plasmodium falciparum

Dryza salvia

Drosophila melanogaster

Homo sapiens

Mus musculus

Canis familiaris

Gallus gallus

Xenopus tropicalis

Candida albicans

Saccharomyces cerevisiae

Baker's yeast

you!

raises bread

helps in cheese production

Rickettsia typhi

H. pylori

Borrelia burgdorferi

Mycoplasma genitalium

Mycoplasma fermentans

Neisseria gonorrhoeae

Mycobacterium tuberculosis

Lysobacter expressus

Mesorhizobium loti

is a bacterium that fixes nitrogen

is an example of a photosynthetic bacterium

is an example of a photosynthetic eukaryote

is an example of a protist

is an example of a fungus

is an example of an animal

is an example of a plant

All genomes, Wed Aug 3 11:11:38 2005, 197 species, 132024 orthologs, 11 used here, distance tree, ToComple=36, Fit=0.4633; modified for educational purposes (drm)