**Phylogeny Terminology**

- **Phylogeny** - the history of descent of a group of organisms from a common ancestor
  
  From Greek:
  
  - *phylon* = tribe, race
  - *genesis* = source

- **Taxonomy** - the science of classification of organisms. From Greek:
  
  - *taxis* = to arrange, classify

- **Clade** - A set of species which includes all of the species derived from a single common ancestor

**Phylogeny: Inference Tool**

- **Phylogeny** is the inference of evolutionary relationships.

- Traditionally, *phylogeny* relied on the comparison of morphological features between organisms.

- Today, *molecular sequence data* are also used for phylogenetic analyses.

**Importance of Phylogeny**

- How many genes are related to my favorite gene?

- Was the extinct quagga more like a zebra or a horse?

- Was Darwin correct when he stated that humans are the closest to chimps and gorillas?

- How related are whales and dolphins to cows?

- Where and when did HIV originate?

- What is the history of life on earth?
Phylogenetic Analysis

- A phylogenetic analysis of a family of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution.
- Two sequences that are very much alike will be located as neighboring outside branches (leaves) and will be joined by a common branch beneath them.

Aim of Phylogenetic Analysis

- The evolutionary relationships among the sequences are depicted by placing the sequences as outer branches on a tree.
- The branching relationships on the inner part of the tree then reflect the degree to which different sequences are related.
- The aim of phylogenetic analysis is to discover all of the branching relationships in the tree and the branch lengths.

Phylogenetic Trees

- Phylogenetic tree: diagram showing evolutionary paths of species/genes.
- Why do we construct phylogenetic trees?
  - To understand the path (lineage) of various species.
  - To understand how various functions evolved.
  - To perform multiple alignment.

Additional Uses of Phylogenetic Trees

- To study the evolutionary relationships of different species and to understand how species relate to one another.
- To predict the unknown gene’s function according to its phylogenetic relationship to other genes.

Globin Family Evolution

- Earth Forms
- First Micro-organism
- Dinosaurs

Time Line For Life

- Billions of years ago
- Humans 3.5 million
### Terminology

- **Leaves** represent objects (genes, species) being compared.
  - *Taxon* refers to the leaves when they represent species and broader classifications of organisms.
- **Internal nodes** are hypothetical ancestral units.
- In a **rooted tree**, the path from root to a node represents an evolutionary path.
- An **unrooted tree** specifies relationships among objects, but not evolutionary paths.

### Rooted and Unrooted Trees

- **All objects in a rooted tree** have a single common ancestor.
  - In general, rooted trees require more information to construct than unrooted ones.
- **Objects are leaves in an unrooted tree** and internal nodes are common ancestors.
  - In general, given any two leaves, we cannot tell if they have a common ancestor.

### Outgroup Properties

- **The outgroup** should be a sequence known to be less closely related to the rest of the sequences than they are to each other.
- It should ideally be as closely related as possible to the rest of the sequences while still satisfying the first condition.
- The root must be somewhere between the outgroup and the rest (either on the node or in a branch).
**Rooting a Tree**

The best place for having the root of the phylogenetic tree.

---

**Convergent and Parallel Evolution**

- **Convergent evolution**
  Independent evolution of similar traits due to similar selection pressure
  
  *Example:* wings in birds and bats

- **Parallel evolution**
  Independent evolution of common traits in organisms sharing distant relatives
  
  *Example:* patterns of butterfly wings.

---

**Gene Duplication & Speciation**

---

**Two Duplication Events**

---

**Building Phylogenetic Trees**

- **Sequence Selection:**
  - Identify a DNA or protein sequence.
  - Obtain related sequences by performing a database search.
- Perform multiple alignment.
- Build a phylogenetic tree.
- Check the robustness of the tree.
Distance and Character Based Trees

The construction of the tree is:
- **distance-based**: measures the distance between species/genes (eg. mutations, time, distance metric).
  - First calculate the overall distance between all pairs of sequences, then construct a tree based on the distances.
- **character-based**: morphological features (eg. number of legs), DNA/protein sequences.
  - Use the individual substitutions among sequences to determine the most likely ancestral relationships.

The tree is constructed based on the gain or loss of traits.

Methods for Constructing Phylogenetic Trees

- Distance-Based Methods:
  - Unweighted Pair Group Method Using Arithmetic Averages (UPGMA)
  - Fitch Margoliash (FM)
  - Neighbor Joining (NJ)
- Character-Based Methods:
  - Maximum Parsimony (MP)
  - Maximum Likelihood (ML)

Other Methods for Constructing Trees

**Clustering Methods**
- Follow a set of steps (an algorithm) and arrive at a tree.
- Use distance data.
- Produce a single tree.
- Do not use objective functions to compare the current tree to other trees.

**Optimality Criterion**
- Use objective functions to compare different trees.
- First define an optimality criterion, i.e. minimum branch length, and then find the tree with the best value for the objective function.

Clustering Algorithms

- The strength of clustering algorithms is:
  - Their speed
  - Their robustness
  - Their ability to reconstruct trees for very large numbers (thousands) of sequences.
  - Most clustering methods reconstruct phylogenetic trees for a set of sequences on the basis of their pairwise evolutionary distances.

Classification of Tree Building Methods

- **Distance-Based**
  - UPGMA
  - Neighbor Joining
  - Fitch-Margoliash
- **Character-Based**
  - Maximum Parsimony
  - Maximum Likelihood

Distance-Based Method

- **Given**: an $n \times n$ matrix $M$, where $M(i,j)$ is the distance between objects $i$ and $j$
- **Build**: an edge-weighted tree such that the distances between leaves $i$ and $j$ correspond to $M(i,j)$
**UPGMA**

- **UPGMA** is a sequential clustering algorithm.
  - It works by clustering the sequences, at each stage amalgamating two operational taxonomic units (OTUs) and at the same time creating a new node in the tree.
  - The edge lengths are determined by the difference in the heights of the nodes at the top and bottom of an edge.

**The Molecular Clock**

- **UPGMA** assumes that:
  - the gene substitution rate is constant, in other words: divergence of sequences is assumed to occur at the same rate at all points in the tree.
  - Known as the **Molecular Clock**.
  - the distance is linear with evolutionary time.

**UPGMA Algorithm**

- The algorithm iteratively picks two clusters and merges them, thus creating a new node in the tree, denoted by T.
- The **average distance** between two clusters: C_i and C_j is determined by:
  \[ d_{ij} = \frac{1}{|C_i| \cdot |C_j|} \sum_{p \in C_i, q \in C_j} d_{pq} \]

**The UPGMA Algorithm**

- **Initialization**
  - Assign each sequence i to its own cluster C_i.
  - Define one leaf of T for each sequence; place at height zero.

- **Iteration** while more than two clusters, do
  - Determine the two clusters C_i, C_j for which \( d_{ij} \) is minimal.
  - Define a new cluster \( C_k = C_i \cup C_j \); compute \( d_{kl} \) for all l.
  - Define a node k with children i and j; place it at height \( d_{ij}/2 \).
  - Replace clusters C_i and C_j with C_k.

- **Termination**
  - Join last two clusters, C_i and C_j; place the root at height \( d_{ij}/2 \).

**UPGMA: Example (1st Iteration)**

Sequences A and D are the closest and are combined to create a new cluster V of height ½ in T.

**UPGMA: Example (2nd Iter.)**

The table of distances is updated to reflect the average distances from V to the other sequences.

V and E are the closest and are combined to create a new cluster W of height 1 in T.
UPGMA: Example (3rd Iter)

After updating the table of distances, B and F are the closest sequences and are combined to create a new cluster X of height 2 in T.

\[ d_{ij} \quad B \quad C \quad F \quad W \]

\[
\begin{array}{c|cccc}
  & B & C & F & W \\
\hline
B & & 8 & 4 & 6 \\
C & 8 & & 8 & \\
F & 6 & & & \\
W & & & & \\
\end{array}
\]

Understanding Bioinformatics by M. Zvelebil and J. Baum

UPGMA: Example (4th Iter)

Once more the table is updated. W and X are the closest sequences and are combined to create a new cluster Y of height 3 in T.

\[ d_{ij} \quad C \quad W \quad X \]

\[
\begin{array}{c|cccc}
  & C & W & X & \\
\hline
C & & & & 8 \\
W & & & & 8 \\
X & & & & 6 \\
\end{array}
\]

Understanding Bioinformatics by M. Zvelebil and J. Baum

UPGMA Tree: Second Example

The remaining 2 sequences, C and Y of distance 8 are combined to create a new cluster Z of height 4 in T.

<table>
<thead>
<tr>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>26</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>29</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

UPGMA assumes a uniform rate of mutation in the tree branches. At any given time, the two sequences should have the same number of changes separating them from the common ancestor.

Understanding Bioinformatics by M. Zvelebil and J. Baum

UPGMA’s Shortcoming

A tree where closest pair of leaves are not neighboring leaves.

\[ d(A,C) = 0.3 \]
\[ d(A,B) = 0.5 \]

So the neighboring pair A and B are further apart than the non-neighboring pair A and C.

Understanding Bioinformatics by M. Zvelebil and J. Baum

Fitch-Margoliash Method

- Fitch-Margoliash does not assume a constant mutation rate.
- With the Fitch-Margoliash Method, the sequences are combined in threes to define the branches of the predicted tree and to calculate the branch lengths of the tree.
- This method of averaging distances is most accurate for trees with short branches.
**Case With Three Sequences**

The distance \( c \) in the tree with three sequences \( A, B, \) and \( C \), can be computed as:

\[
c = \frac{d(A,C) + d(B,C) - d(A,B)}{2}
\]

The value of \( c \) can be obtained from:

\[
d(A,B) = a + b, \quad d(B,C) = b + c, \quad d(A,C) = a + c
\]

**Fitch-Margoliash Method**

- Reduce a tree with more than three species to \( A, B, \) and \( C \), where \( A \) and \( B \) are the closest species, and \( C \) is a composite average over all other species.
- Repeat the whole process for the next-closest species.
- Distances are computed as is done with UPGMA.

**Case With Three Sequences**

In a similar fashion, we can compute distances \( a \) and \( b \):

\[
a = \frac{d(A,B) + d(A,C) - d(B,C)}{2}
\]

\[
b = \frac{d(A,B) + d(B,C) - d(A,C)}{2}
\]
**Fitch-Margoliash (IV)**

(C) **STEP 3 (N = 3)**

\[
\begin{align*}
&d_{95} = 7 \\
&d_{07} = \frac{1}{2} \left( \frac{26}{3} + \frac{23}{3} - 7 \right) = \frac{14}{3} \\
&d_{07} = \frac{1}{2} \left( \frac{26}{3} + \frac{23}{3} - \frac{14}{3} \right) = \frac{10}{3} \\
&b_8 = \frac{1}{2} \left( \frac{26}{3} - \frac{23}{3} \right) = 4 \\
&b_9 = \frac{1}{2} \left( \frac{23}{3} - \frac{14}{3} \right) = 3
\end{align*}
\]

**Compute average over 3 paths**

**Introduction to Neighbor-Joining**

- **Neighbor-Joining** does not assume a constant rate of evolution.
  - The algorithm is based on the concept of minimum evolution; the true tree is the one for which the total branch length is minimum.
  - The resulting tree is not rooted and is additive.

**Neighbor-Joining Concept (I)**

- The neighbor joining algorithm starts with all sequences arising from the same node and gradually distinguishing pairs of nodes that are neighbors.

**Neighbor-Joining Concept (II)**

- Decompose the start-like tree by combining pairs of sequences. Here 1 and 2 are combined.
  - Each possible sequence pair is chosen and the sum of the branch lengths of the corresponding tree is calculated. The one with lowest sum is chosen.
**Neighbor-Joining Concept (III)**

- Once the choice is made calculate the branch lengths and update the distance tree using the Fitch-Margoliash computations.
- The process is then repeated until all sequences are exhausted.

![Diagram](image)

**Picking Pairs with UPGMA and FM**

<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>0.3</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>B</td>
<td>0.3</td>
<td>0</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>C</td>
<td>0.5</td>
<td>0.6</td>
<td>0</td>
<td>0.9</td>
</tr>
<tr>
<td>D</td>
<td>0.6</td>
<td>0.5</td>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

**UPGMA and FM will pick A and C to pair**

**Why is the Wrong Pair Picked?**

- Neighbors A, B have one short edge (0.1) and one long edge (0.4)
- The problem is that the short edge (0.1) is closer to leaf C (0.3) than to the long edge (0.4)

![Diagram](image)

**Subtracted Averaged Distances**

- The solution is to subtract averaged distances to all other leaves.
- In essence, this will compensate for long edges.

\[ D_y = d_y - (r_i + r_j) \]

\[ r_j = \frac{1}{|L| - 2} \sum_{k \in L} d_{ik} \]

**Back to the Example**

<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>0.5</td>
<td>0.3</td>
<td>0.6</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td>0</td>
<td>0.6</td>
<td>0.9</td>
</tr>
<tr>
<td>C</td>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>D</td>
<td>0.6</td>
<td>0.9</td>
<td>0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

**Distances from New Node k**

- Given a new internal node k, the distance to another node m is given by:

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

![Diagram](image)
Distances in NJ Algorithm

To compute the distance from a leaf to a new internal node \( k \), we also use averaging:

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

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

where

\[
r_i = \frac{1}{|L| - 2} \sum_{m \in L} d_{im}
\]

Neighbor-Joining Algorithm

- \( T \) = set of leaf nodes
- while more than two subtrees in \( T \)
  - pick the pair \( i, j \) in \( T \) with minimal \( D_{ij} \)
  - define a new node \( k \) joining \( i \) and \( j \)
  - determine new distances as described
  - remove \( i \) and \( j \) from \( T \) and insert \( k \)
- join two remaining subtrees

Limitations of Distance-Based Phylogenetic Trees

The distance-based phylogenetic tree is derived from the pairwise distance of aligned sequences and not from the original sequence data. The distance information may not contain all the sequence information.

Observable Features

- Sometimes we do not have a distance metric between the species we are interested in.
- What we have instead, are observable features.
- We then use the observable features to build the tree. These trees are called Character-Based trees.

Character-Based Trees

- The building of the tree is based on morphological features and not on distances.
- Examples of morphological features:
  - has feathers
  - has a backbone
  - has a certain amino acid at a certain position in the sequence
  - whether or not a certain protein regulates another protein.

Maximum Parsimony Method

- The maximum parsimony method predicts the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences
  - This method is also known as the minimum evolution method.
- The maximum parsimony method is an exact method and is therefore used
  - for sequences that are quite similar, and
  - for small number of sequences.
**Maximum Parsimony**

- With **Maximum Parsimony** each character in the input sequences is examined separately.
- The following two assumptions are made:
  - Characters are mutually independent.
    - A change in one character has no effect on the distribution of the other characters.
  - After two species diverge in the tree, they continue to evolve separately.

**Finding Most Parsimonious Trees Exhaustive Search**

1. One tree with any three taxa.
2. Add fourth taxon (D) in each of three possible positions: 3 trees.
3. Add fifth taxon (E) in each of five possible positions on each of the three trees: 15 trees, and so on.

**Maximum Parsimony Algorithm**

- Construct all possible tree topology with the given sequences.
- Check each tree and count the minimum number of changes required to explain the data.
- Choose the tree with the smallest number of changes.

**Parsimony: An Example**

- Given four sequences:
  - Sequence 1: TGC
  - Sequence 2: TAC
  - Sequence 3: AGG
  - Sequence 4: AAG
- We want to find the tree with the smallest number of changes that explains the observed data.
- Draw all possible trees with 4 taxa.

**Parsimony Example**

- Position 1:
  - Only one change is introduced if seq1 and seq2 are grouped; and 2 changes if seq1 and seq3 or seq1 and seq4 are grouped.
- Position 2:
  - Only one change is introduced if seq1 and seq2 are grouped; and 2 changes if seq1 and seq2 or seq1 and seq4 are grouped.
- Position 3:
  - Only one change is introduced if seq1 and seq3 are grouped; and 2 changes if seq1 and seq3 or seq1 and seq4 are grouped.

The shortest tree is ((1,2),(3,4)).

**Informative Sites**

- A site that provides information for distinguishing between different topologies is said to be an **informative site**.
- Only **informative sites** need to be analyzed.
- A site is phylogenetically informative only when there are at least two different kinds of characters, each represented at least two times.
Informative Sites: An Example

<table>
<thead>
<tr>
<th>Taxa</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>G</td>
</tr>
</tbody>
</table>

Only sites at columns 5, 7, and 9 are informative sites.

Different Weights for Mutations

- Weights given to mutations do not have to be the same (as is the case in Matrix (A)).
- Matrix (B) favors transitions over transversions by a factor of R.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(B)</td>
<td>1</td>
<td>2</td>
<td>R</td>
<td></td>
</tr>
</tbody>
</table>

Steps of Tree Reconstruction

1. Choose a set of related sequences
2. Obtain a multiple sequence alignment
3. Is there strong sequence similarity?
   - Yes: Maximum Parsimony methods
   - No:
     1. Is there a clearly recognizable sequence similarity?
        - Yes: Distance methods
        - No:
          1. Analyze how well the data support prediction

Bootstrapping

Bootstrapping is a statistical technique that uses computer intensive random resampling of data to determine sampling error or confidence intervals for some estimated parameter.

Bootstrap: An Example (I)

Site: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Species
- Pmc2 (chimp) CTTAGAAAATTCCTAGATA
- Pmc3 (human) TTCTAGAAACTCTCATAATT
- Pmc5 (rhesus) TTTTAGAAATTTCTAGATA
- Pmc6 (bonobo) TTTTAGAAATTTCTAGATA
- Pmc7 (gibbon) TTCTAGAAACTCTCATAATT
- Pmc8 (orangutan) CTTAGAAAATTCCTAGATA
- Pmc9 (gibbon) CTTCTAGAAACTCTCATAATT
- Pmc10 (rhesus) TTTTAGAAATTTCTAGATA
- Pmc11 (chimp) TTTTAGAAATTTCTAGATA
- Pmc12 (orangutan) CTTCTAGAAACTCTCATAATT

Part of the data matrix of aligned nucleotide sequences for the malaria parasite Plasmodium. Only the first 20 columns of the 11 x 221 matrix are shown.

Bootstrap: An Example (II)

Randomly select, with replacement, 221 columns from the original matrix. Tree-building algorithm is applied to give a bootstrap tree. Repeat the process 200 times.

The numbers at the branches are confidence values based on Felsenstein’s bootstrap method.
A condensed tree showing well-supported features can be derived by applying the Bootstrap procedure.

Simply remove the internal branches that are supported by less than 60% of the bootstrap trees.

**Software Tools**

- **PHYLIP**
  - Phylogeny Inference Package.
  - Free.
  - Developed by Dr. Joe Felsenstein from the Department of Genome Sciences at the University of Washington.
  - Source code is written in ANSI C.
  - Executables are available for different platforms:
    - Windows, UNIX and Macintosh.

- **PAUP***
  - Phylogenetic Analysis Using Parsimony.
  - http://paup.csit.fsu.edu/
  - Developed by Dr. David Swofford of the Laboratory of Molecular Systematics, National Museum of Natural History.
  - The most sophisticated parsimony program.

**Phylogeny Programs**

392 phylogeny packages and 54 free web servers


**Workflow of Distance-Based Packages**

1. Obtain Homologous Sequences
2. Perform Sequence Alignment
3. Create Distance Matrix
4. Create Phylogenetic Tree

![Workflow Diagram](http://evolution.genetics.washington.edu/phylip/software.html)