**Introduction to Bioinformatics**

Prof. Dr. Nizamettin AYDIN  
naydin@yildiz.edu.tr

**Phylogenetics**

---

**What is Molecular Phylogenetics**

- **Phylogenetics** is the study of evolutionary relationships
- **Example:** relationship among species

---

**A Brief History of Molecular Phylogenetics**

- **1900s**
  - Immunochemical studies
    - Cross-reactions stronger for closely related organisms
  - Nuttall (1902) - apes are closest relatives to humans
- **1960s - 1970s**
  - Protein sequencing methods, electrophoresis, DNA hybridization and PCR contributed to a boom in molecular phylogeny
- **late 1970s to present**
  - Discoveries using molecular phylogeny
    - Endosymbiosis - Margulis, 1978
    - Divergence of phyla and kingdom - Woese, 1987
    - Many Tree of Life projects completed or underway

---

**Molecular data vs. Morphology/Physiology**

- **Molecular data**
  - Strictly heritable entities
  - Data is unambiguous
  - Regular & predictable evolution
  - Quantitative analyses
  - Relationship of distantly related organisms can be inferred
  - Abundant and easily generated with PCR and sequencing

- **Morphology/Physiology**
  - Can be influenced by environmental factors
  - Ambiguous modifiers: “reduced”, “slightly elongated”, “somewhat flattened”
  - Unpredictable evolution
  - Qualitative argumentation
  - Homology difficult to assess
  - Only close relationships can be confidently inferred
  - Problems when working with micro-organisms and where visible morphology is lacking

---

**Phylogenetic concepts: Interpreting a Phylogeny**

- Physical position in tree is not meaningful
- Swiveling can only be done at the nodes
- Only tree structure matters

---

**Phylogenetic concepts: Interpreting a Phylogeny**

- Physical position in tree is not meaningful
- Swiveling can only be done at the nodes
- Only tree structure matters
Tree Terminology

• Relationships are illustrated by a phylogenetic tree / dendrogram
  – Combination of Greek dendro/tree and gramma/drawing
  – A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.
  – Dendrograms are often used in computational biology to illustrate the clustering of genes or samples, sometimes on top of heatmaps.
• A cladogram is a type of phylogenetic tree that only shows tree topology
  – the shape indicating relatedness.
    • It shows that, say, humans are more closely related to chimpanzees than to gorillas, but not the time or genetic distance between the species.
    – Combination of Greek clados/branch and gramma/drawing

Tree Terminology

• The branching pattern is called the tree’s topology
• Trees can be represented in several forms:

Rectangular cladogram
Slanted cladogram

Tree Terminology

• Trees can be represented in several forms:
  Circular cladogram

Tree Terminology

• Same tree - seven different views:
  Rectangular Phylogram, Rectangular Cladogram, Slanted Cladogram, Circular Phylogram, Circular Cladogram, Radial Phylogram and Radial Cladogram

Tree Terminology

Operational taxonomic units (OTU) / Taxa

Internal nodes

Root

Branches

Polytomy

Terminal nodes

Sisters

Rooted trees

Unrooted trees

• Rooted trees:
  – Has a root that denotes common ancestry
• Unrooted trees:
  – Only specifies the degree of kinship among taxa but not the evolutionary path

Taxon, plural taxa (taxonomy): Any group or rank in a biological classification into which related organisms are classified.
Tree Terminology

- Scaled trees:
  - Branch lengths are proportional to the number of nucleotide/amino acid changes that occurred on that branch (usually a scale is included).

- Unscaled trees:
  - Branch lengths are not proportional to the number of nucleotide/amino acid changes (usually used to illustrate evolutionary relationships only).

Methods in Phylogenetic Reconstruction

- Distance methods
  - calculate pairwise distances between sequences, and group sequences that are most similar.
  - This approach has potential for computational simplicity and therefore speed

- Maximum Parsimony
  - assumes that shared characters in different entities result from common descent.
  - Groups are built on the basis of such shared characters, and the simplest explanation for the evolution of characters is taken to be the correct, or most parsimonious one.

- Maximum Likelihood
  - compute the probability that a data set fits a tree derived from that data set, given a specified model of sequence evolution.

Comparison of Methods

Distance

- Uses only pairwise distances
- Minimizes distance between nearest neighbors
- Very fast
- Easily trapped in local optima
- Good for generating tentative tree, or choosing among multiple trees

Maximum parsimony

- Uses only shared derived characters
- Minimizes total distance
- Slow
- Assumptions fail when evolution is rapid
- Best option when tractable (<30 taxa, homoplasy rare)

Maximum likelihood

- Uses all data
- Maximizes tree likelihood given specific parameter values
- Very slow
- Highly dependent on assumed evolution model
- Good for very small data sets and for testing trees built using other methods

Methods in Phylogenetic Reconstruction

- Distance
  - Using a sequence alignment, pairwise distances are calculated
  - Creates a distance matrix
  - A phylogenetic tree is calculated with clustering algorithms, using the distance matrix.
  - Examples of clustering algorithms include the Unweighted Pair Group Method using Arithmetic averages (UPGMA) and Neighbor Joining clustering.

Methods in Phylogenetic Reconstruction

- Maximum Parsimony
  - All possible trees are determined for each position of the sequence alignment
  - Each tree is given a score based on the number of evolutionary step needed to produce said tree
  - The most parsimonious tree is the one that has the fewest evolutionary changes for all sequences to be derived from a common ancestor
  - Usually several equally parsimonious trees result from a single run.
Maximum parsimony: exhaustive stepwise addition

Step 1

Step 2

Step 3

Methods in Phylogenetic Reconstruction

• **Maximum Likelihood**
  - Creates all possible trees like Maximum Parsimony method but instead of retaining trees with shortest evolutionary steps, it employs a model of evolution whereby different rates of transition/transversion ratio can be used.
  - Each tree generated is calculated for the probability that it reflects each position of the sequence data.
  - Calculation is repeated for all nucleotide sites.
  - Finally, the tree with the best probability is shown as the maximum likelihood tree – usually only a single tree remains.
  - It is a more realistic tree estimation because it does not assume equal transition-transversion ratio for all branches.

How confident are we about the inferred phylogeny?

• **Bootstrapping**
  - Bootstrapping analysis is a kind of statistical analysis to test the reliability of certain branches in the evolutionary tree.
  - It involves resampling one’s own data, with replacement, to create a series of bootstrap samples of the same size as the original data.
  - In the case of nucleic acid (amino acid) sequences, the resampled data are the nucleotides (amino acids) of a sequence while the statistical significance of a specific cluster is given by the fraction of trees, based on the resampled data, containing that cluster.

The Bootstrap

• **Computational method** to estimate the confidence level of a certain phylogenetic tree.

Some Discoveries Made Using Molecular Phylogenetics

• **Universal Tree of Life**
  - Using rRNA sequences.
  - Able to study the relationships of uncultivated organisms, obtained from a hot spring in Yellowstone National Park.
Some Discoveries Made Using Molecular Phylogenetics

- **Endosymbiosis: Origin of the Mitochondrion and Chloroplast**

  <image>

  - **Mitochondria** and **chloroplasts** are derived from the α-purple bacteria and the cyanobacteria respectively, via separate endosymbiotic events.

Some Discoveries Made Using Molecular Phylogenetics

- **Relationships within species: HIV subtypes**

  <image>

Problems and Errors in Phylogenetic Reconstruction

- **Inherent strengths and weaknesses in different tree-making methodologies.**
- **More is better**
  - Errors in inferred phylogeny may be caused by small data sets and/or limited sampling.
- **Unsuitable sequences**
  - those undergoing rapid nucleotide changes or slow to zero changes overtime may skew phylogenetic estimations

Problems and Errors in Phylogenetic Reconstruction

- **Mutations:**
  - Duplications, inversions, insertions, deletions etc. can give inaccurate signals
- **Genomic hotspots:**
  - small regions of rapid evolution are not easily detected
- **Homoplasy:**
  - nucleotide changes that are similar but occurred independently in separate lineages are mistakenly assumed as inherited changes
- **Sample contamination / mislabeling:**
  - always a possibility when working with large data sets

Maximum Parsimony - example

- **Maximum parsimony methods predict the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences.**
  - First, a multiple sequence alignment must first be obtained.
- For each aligned position, phylogenetic trees that require the smallest number of evolutionary changes to produce the observed sequence changes are identified.
Maximum Parsimony - example

The left tree is preferred over the right tree.

\[
\begin{align*}
\text{Total #substitutions} & = 3 \\
\text{Total #substitutions} & = 4
\end{align*}
\]

• Assuming we have 4 sequences
  – There are 3 possible trees:

\[
\begin{align*}
\text{Total #substitutions} & = 3 \\
\text{Total #substitutions} & = 4
\end{align*}
\]

• The optimal tree is obtained by adding the number of changes at each informative site for each tree, and picking the tree requiring the least total number of changes.

• For a large number of sequences the number of trees to examine becomes so large that it might not be possible to examine all possible trees.

**Maximum Parsimony - example**

Consider the following sequences

- \(S_1\): C ACC C C C C T T
- \(S_2\): A ACC C C A T
- \(S_3\): C A C T G C T
- \(S_4\): A A C T G C T

\((S_1, S_2), (S_3, S_4)\) 2 0 0 1 1 0 1 1 6

\((S_1, S_3), (S_2, S_4)\) 1 0 0 2 2 0 1 1 7

**Distance Methods -example**

For phylogenetic analysis, the distance score counted as

- either the number of mismatched positions in the alignment
- the number of sequence positions that must be changed to generate the other sequence is used.

The Fitch and Margoliash method uses a distance table.

- The sequences are combined in threes to define the branches of the predicted tree and to calculate the branch lengths of the tree.

**Phylogeny reconstruction for 3 sequences**

- There is a single tree topology

\[
\begin{align*}
a + b & = D_{AB} \\
b + c & = D_{BC} \\
a + c & = D_{AC}
\end{align*}
\]

\[
\begin{align*}
A & \cdots a & \cdots \text{arc} \\
B & \cdots b & \cdots \text{brc} \\
C & \cdots c & \cdots \cdots
\end{align*}
\]

- Input:
  - \(D_{AB}, D_{BC}\) and \(D_{AC}\) (pairwise distances)

- Output:
  \[
  \begin{align*}
a & = (D_{AB} + D_{AC} - D_{BC}) / 2 \\
b & = (D_{AB} + D_{BC} - D_{AC}) / 2 \\
c & = (D_{AC} + D_{BC} - D_{AB}) / 2
\end{align*}
\]
Distance Methods - example

• Distance matrix of 3 sequences and unrooted tree

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>--</td>
<td>22</td>
<td>39</td>
</tr>
<tr>
<td>B</td>
<td>--</td>
<td>--</td>
<td>41</td>
</tr>
<tr>
<td>C</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

- distance from A to B = a + b = 22 \quad (1)
- distance from A to C = a + c = 39 \quad (2)
- distance from B to C = b + c = 41 \quad (3)

• subtracting (3) from (2) yields:
\[ b + c - (b + c) = b - a = 41 - 39 = 2 \quad (4) \]

Distance Methods - example

• adding (1) and (4) yields
\[
a + b + b - a = 2b = 22 + 2 = 24
\]
\[
b = 24 / 2 = 12
\]
so
\[
a + b = a + 12 = 22;
a = 22 - 12 = 10
\]
finally
\[
a + c = 10 + c = 39;
c = 39 - 10 = 29
\]

Distance Methods - example

• Consider the alignment:

<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>-</td>
<td>3</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>-</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>4</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>-</td>
</tr>
</tbody>
</table>

The distances between these sequences can be shown as a 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>-</td>
<td>3</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

• Using this information, an unrooted tree showing the relationship between these sequences can be drawn: