Introduction

Phylogenetics

Tree Basics

Tree Topology
Counting Trees
Tree Rooting Methods
Inferring Trees from Distances

Tree Building Methods

Clustering Methods
Parsimony Methods

Quality Assessment by Resampling

Bootstrap

Functional Gene Annotation Systems

Gene Ontologies
Pathway Annotation Systems

References
Outline

Introduction

Phylogenetics

Tree Basics
  Tree Topology
  Counting Trees
  Tree Rooting Methods
  Inferring Trees from Distances

Tree Building Methods
  Clustering Methods
  Parsimony Methods

Quality Assessment by Resampling
  Bootstrap

Functional Gene Annotation Systems
  Gene Ontologies
  Pathway Annotation Systems

References
Outline

**Introduction**

**Phylogenetics**

**Tree Basics**
- Tree Topology
- Counting Trees
- Tree Rooting Methods
- Inferring Trees from Distances

**Tree Building Methods**
- Clustering Methods
- Parsimony Methods

**Quality Assessment by Resampling**
- Bootstrap

**Functional Gene Annotation Systems**
- Gene Ontologies
- Pathway Annotation Systems

**References**
Bioinformatics and Black Death in Europe

Understanding the causes of Black Death with next generation sequencing, sequence analysis and phylogenetics.

- Article in Nature
- Movie
- BBC News Article
Common Workflow of Phylogenetic Sequence Analyses

1. Select the Appropriate Sequences for a Phylogenetic Question
   **Important:** sequences should show significant similarity.

2. Create a Multiple Alignment for Chosen Sequences
   **Important:** unalignable sequence areas should be removed.

   - S1: FMFFSACKRIACEGAMELFLFLT 450
   - S2: FMFFSACKRIVGEALAMELFLFLT 450
   - S3: LAFCCCGAVCLGEPALRLEILVVLTLT 443
   - S4: SLFFCPCKCSCMWRLAELELQMLAAA 470
   - S5: YTPFCSOPNIGRFALNMKLAATI 457
   - Consensus: ..Pf.Gk.R.C.Ge.LA.mELf.Lt

3. Compute a Distance Matrix for Multiple Alignment

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.0</td>
<td>0.43</td>
<td>0.71</td>
<td>0.71</td>
<td>0.48</td>
</tr>
<tr>
<td>S2</td>
<td>0.0</td>
<td>0.57</td>
<td>0.57</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>0.0</td>
<td>0.29</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>0.0</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Calculate Phylogenetic Tree
   **Important:** choose a tree building method.

5. Tree Post Processing
   **Important:** tree rooting and bootstrapping.
Outline

Introduction
  Phylogenetics

Tree Basics
  Tree Topology
  Counting Trees
  Tree Rooting Methods
  Inferring Trees from Distances

Tree Building Methods
  Clustering Methods
  Parsimony Methods

Quality Assessment by Resampling
  Bootstrap

Functional Gene Annotation Systems
  Gene Ontologies
  Pathway Annotation Systems

References
**Tree Styles**

**Unrooted Tree**

**Rooted Trees**

Newick format: \(((A:0.5, B:0.5):0.1, (C:0.5, D:0.5):0.1);\)
Tree Styles: Circle Tree

Circle Tree
Significance of Branch Lengths in Cladograms, Phylograms and Ultrametric Trees

**Cladogram**
- Taxon A
- Taxon B
- Taxon C
- Taxon D

Branch lengths have no meaning.

**Phylogram**
- Taxon A
- Taxon B
- Taxon C
- Taxon D

Branch lengths are proportional to (genetic) change.

**Ultrametric Tree**
- Taxon A
- Taxon B
- Taxon C
- Taxon D

Branch lengths are proportional to time.

Day 2, Phylogenetics, Gene Annotations & Introduction to R

Tree Basics

Tree Topology
Rotating Branches in Phylogenetic Trees

Figure 5: Branch Rotations

- Rotations of internal nodes yield exactly the same tree.
There are five possibilities to root the unrooted tree on the top left.
Trees rooted on branches c and d are not shown.
Outline

Introduction
Phylogenetics

Tree Basics
Tree Topology
Counting Trees
Tree Rooting Methods
Inferring Trees from Distances

Tree Building Methods
Clustering Methods
Parsimony Methods

Quality Assessment by Resampling
Bootstrap

Functional Gene Annotation Systems
Gene Ontologies
Pathway Annotation Systems

References
The number of possible trees grows more than exponentially as the number of taxa $n$ increases:

\[ N \text{ unrooted trees} = \frac{(2n - 5)!}{2^{n-3}(n-3)!} = (2n - 5)!! \quad (1) \]

\[ N \text{ rooted trees} = \frac{(2n - 3)!}{2^{n-2}(n-2)!} = (2n - 3)!! \quad (2) \]

$n = \text{number of leaves (taxa)}$

$N = \text{number of possible trees}$

Example

<table>
<thead>
<tr>
<th>$n$</th>
<th>$N$ Unrooted Trees</th>
<th>$N$ Rooted Trees</th>
</tr>
</thead>
<tbody>
<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>7</td>
<td>954</td>
<td>10,395</td>
</tr>
<tr>
<td>10</td>
<td>2,027,025</td>
<td>34,459,425</td>
</tr>
</tbody>
</table>
Methods for Rooting Phylogenetic Trees

Outgroup Method
- Rooting by including one or more outgroup taxa/sequences.

Gene Duplication
- Paralogous gene duplication predating the common ancestor of a clade are used.

Midpoint Rooting
- Tree is rooted by midpoint between the two most distant branches.
Rooting by Outgroup

- Rooting is accomplished by including one or more outgroup (taxa/sequences) that differ from all ingroup members more than all the ingroup members among each other.

- The main assumption of this method is that outgroup taxa fall outside of the ingroup.
The root is placed between paralogous gene populations.
**Midpoint Rooting**

- Choose the midpoint between the two most distant branches.

- Midpoint rooting assumes that the rate of evolution is the same on the longest branches of the tree.
Outline

Introduction
  Phylogenetics

Tree Basics
  Tree Topology
  Counting Trees
  Tree Rooting Methods

Inferring Trees from Distances

Tree Building Methods
  Clustering Methods
  Parsimony Methods

Quality Assessment by Resampling
  Bootstrap

Functional Gene Annotation Systems
  Gene Ontologies
  Pathway Annotation Systems

References
Transforming Characters to Distances

DNA Alignment

G1  TTATTTAA
G2  AATTTAA
G3  AAAAAATA
G4  AAAAAAT

Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>0.43</td>
<td>0.71</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>G2</td>
<td>3</td>
<td>0.57</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td>5</td>
<td>4</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>G4</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Absolute distances in bottom triangle and uncorrected relative distances in top triangle.
Outline

Introduction
  Phylogenetics

Tree Basics
  Tree Topology
  Counting Trees
  Tree Rooting Methods
  Inferring Trees from Distances

Tree Building Methods
  Clustering Methods
  Parsimony Methods

Quality Assessment by Resampling
  Bootstrap

Functional Gene Annotation Systems
  Gene Ontologies
  Pathway Annotation Systems

References
Desirable Features of Tree Building Methods

- Consistency: will the method converge on the correct solution given enough data?
- Efficiency: how fast is the method?
- Robustness: will minor violations of the assumptions result in poor estimates of phylogeny?
Outline

Introduction
   Phylogenetics

Tree Basics
   Tree Topology
   Counting Trees
   Tree Rooting Methods
   Inferring Trees from Distances

Tree Building Methods
   Clustering Methods
   Parsimony Methods

Quality Assessment by Resampling
   Bootstrap

Functional Gene Annotation Systems
   Gene Ontologies
   Pathway Annotation Systems

References
Clustering Methods

- Algorithmic methods in which the algorithm itself defines the tree selection criterion.
- No optimality criteria applied.
- Advantage: tend to be very fast (efficient) computations that produce singular trees.
- Disadvantages:
  - Do not allow evaluation of competing hypotheses.
  - No objective function (e.g. likelihood, number of steps) is used to compare different trees to each other, even if numerous other trees could explain the data equally well.
- Examples that construct trees from distances:
  - UPGMA
  - Neighbor joining (NJ)
UPGMA stands for unweighted pair group method using arithmetic averages [Sokal & Michener 1958].

Clusters taxa (sequences) agglomeratively and creates at the same time a hierarchical tree.

The branch (edge) lengths and node positions are determined by the average distance between clusters.

There are variants of UPGMA that define the distance between clusters (linkage method) as the minimum or maximum of the distances between clusters, rather than the average.

The average linkage seems to have the best performance records.
UPGMA Algorithm

**Initialization**
- Assign each sequence to its own cluster.

**Iteration**
- Determine the clusters with the smallest distance (d). If there are several equidistant choices than pick one randomly.
- Join these closet clusters to a new (larger) cluster.
- Place a node for new cluster at the height corresponding to half of the distance between the joined clusters.
- Update distances by computing the distance of the new cluster to all other clusters.

**Termination**
- When only two clusters remain place the root at the height corresponding to half of their distance.
Illustration of UPGMA Algorithm

1. Iteration

2. Iteration

3-4. Iteration

Branch lengths have no meaning.
Branch lengths are proportional to (genetic) change.
Branch lengths are proportional to time.
[Million Years]

Tree Building Methods

Day 2, Phylogenetics, Gene Annotations & Introduction to R

Clustering Methods

Slide 29/56
Limitations of UPGMA Algorithm

- The edge lengths of an UPGMA tree correspond roughly to the times measured by a molecular clock with constant rate.
- The method assumes that the divergence of sequences occurs at all points in the tree with a constant rate and the distances are additive.
- If the molecular clock assumption applies to a given distance matrix, then UPGMA constructs the tree correctly.
- However, if this assumption does not apply to the underlying distance matrix, then UPGMA may construct the tree incorrectly.

Example Fig 6: correct tree on the left and incorrect UPGMA tree on the right:

Solution: test if the distance matrix is ultrametric, where in any triplet of distances one pair must be equal and the remaining one is the smallest.
Neighbor-Joining Method

- If the molecular clock property fails for a given data set, but additivity holds, then the neighbor-joining method can construct a correct tree.
- To overcome the problem that neighboring leaves can be more distant to each other than to non-neighboring leaves (see Fig 6), one can calculate the rate corrected distances $D_{ij}$ by subtracting from $d_{ij}$ the averaged distances to all other leaves:

$$D_{ij} = d_{ij} - (r_i + r_j), \text{ where } r_i = \frac{1}{|L| - 2} \sum_{k \in L} d_{ik} \quad (3)$$

- $|L|$ is the size of the leaf set $L$.
- Consequence: $i$ and $j$ are neighboring leaves if their $D_{ij}$ is minimal!
### Example: Rate Corrected Distance Matrix

#### Distance Matrix

<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>0.3</td>
<td>0.7</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-1.2</td>
<td>0.8</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-1.0</td>
<td>-1.0</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.2</td>
<td></td>
</tr>
</tbody>
</table>

- **Upper right triangle**: original distance values.
- **Lower left triangle**: rate corrected distances calculated by eq 3 as:
  
  $$r_A = (0.3 + 0.7 + 0.4)/2 = 0.7$$
  $$r_B = (0.3 + 0.8 + 0.5)/2 = 0.8$$
  $$r_C = (0.7 + 0.8 + 0.5)/2 = 1.0$$
  
  ... 
  
  $$D_{AB} = 0.3 - 0.7 - 0.8 = -1.2$$
  $$D_{AC} = 0.7 - 0.7 - 1.0 = -1.0$$
  
  ...
Main Differences Between UPGMA and Neighbor-Joining

**UPGMA**

1. Assumes additivity and ultrametricity.
2. Does not use rate corrected distances values for tree construction.
3. Results in a rooted tree with branch pairs reflecting the distance information.
4. Tree type: cladogram.
5. May fail to generate the correct tree from distance values that violate the ultrametricity rule.

**Neighbor-Joining**

1. Assumes additivity, but not ultrametricity.
2. Uses rate corrected distances values for tree construction.
3. Results in an unrooted tree with the branch lengths reflecting the distance information.
4. Tree type (after rooting): phylogram or cladogram.
5. Generates correct tree from distance values that may violate the ultrametricity rule.
Basic principle: find the tree that explains the observed sequences with the minimal number of substitutions.

Instead of building a tree, like with distance methods, parsimony assigns a cost to a given tree.

This requires searching through all possible trees or a subset of trees that contains the best or close to best tree topology.

Parsimony algorithms consist of two major steps:

1. Computation of the cost for a given tree.
2. Search for the tree with minimum cost.
Example: Computing the Cost of a Given Tree

- Given a multiple alignment and a tree topology, one can count the number of substitution needed for each tree.
- The following figure shows two possible trees for the alignment on the left. Trees differ in the order the sequences are assigned to the leaves.
- Hypothetical sequences have been assigned to the ancestral nodes that minimize the number of substitution needed in the entire tree.

![Alignment](alignment.png)

### Alignment

- AAG
- AAA
- GGA
- AAG

### Two Possible Parsimony Trees

- The tree on the left is more parsimonious than the one on the right because it requires only 3 instead of 4 changes.
- As shown here, parsimony treats each site independently and then sums up the substitutions needed for all sites.
Several tree search methods can be considered:

1. **Exhaustive Searches**: All trees are evaluated only possible for trees with less than 20 taxa.
2. **Heuristic Searches**: Not guaranteed to find the best tree (e.g. random branch changes and re-scoring of tree).
3. **Branch and bound algorithm**: This does not evaluate all trees, but guarantees to find the best tree.
4. **Many other approaches**.

**Branch and bound algorithm**

- It exploits the idea that the cost (number of substitutions) of a subtree can only increase by adding an extra edge.
- It systematically builds trees with increasing numbers of leaves and abandons avenues of tree building whenever an incomplete tree exceeds the smallest cost of a complete tree.
Outline

Introduction
Phylogenetics

Tree Basics
Tree Topology
Counting Trees
Tree Rooting Methods
Inferring Trees from Distances

Tree Building Methods
Clustering Methods
Parsimony Methods

Quality Assessment by Resampling
Bootstrap

Functional Gene Annotation Systems
Gene Ontologies
Pathway Annotation Systems

References
Nonparametric resampling methods are often used to estimate the variance associated with a statistic when the underlying sampling distribution for a statistic is either unknown or difficult to derive analytically.

Resampling methods include the bootstrap and the jackknife, both of which operate by repeatedly resampling data from the original data set to estimate the variance of the sampling distribution.

Although both methods have been used for evaluating the reliability of branches, the bootstrap method is more commonly applied.
Data points are randomly resampled from the original data set, with replacement, until new data sets with the original number of observations are obtained.

Statistic of interest (e.g. a tree) is computed for each replicated data set.

Agreement among the resulting trees is summarized with a majority-rule consensus tree (agreement $> 50\%$).

A bootstrap proportion (BP) is the frequency of occurrence of a clade (for all replicated data sets) and is a measure of support for a group.
Example: Bootstrapping

<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>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>R</td>
<td>R</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>B</td>
<td>R</td>
<td>R</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>C</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>D</td>
<td>Y</td>
<td>Y</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>Outgp</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
</tbody>
</table>

Original matrix (alignment)

<table>
<thead>
<tr>
<th>Taxa</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>B</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>C</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>D</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>Outgp</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
</tbody>
</table>

Resampled matrix (alignment)

Majority rule tree with bootstrap proportions (BP) at branch nodes.
Phylogenetics Software (Selection!)

- **PAUP**: complex phylogenetic tool collection (partially commercial)
- **PHYLIP**: complex phylogenetic tool collection (free)
- **MrBayes**: popular package for Bayesian inference of phylogenies.
- **PAML**: many utilities for dealing with different evolutionary rates.
- Many more: see for instance this Link Collection.
Outline

Introduction

Phylogenetics

Tree Basics

Tree Topology
Counting Trees
Tree Rooting Methods
Inferring Trees from Distances

Tree Building Methods

Clustering Methods
Parsimony Methods

Quality Assessment by Resampling

Bootstrap

Functional Gene Annotation Systems

Gene Ontologies
Pathway Annotation Systems

References
Overview of Genome Annotation Process

Sequence Assembly

Genomic Library → Sequencing → Assembly by Alignment → Contig Sequences → Additional Data → Finishing Steps → Assembled Genome

Genome Annotation

Assembled Genome → Gene Finding → Functional Gene Annotation
Outline

Introduction
  Phylogenetics

Tree Basics
  Tree Topology
  Counting Trees
  Tree Rooting Methods
  Inferring Trees from Distances

Tree Building Methods
  Clustering Methods
  Parsimony Methods

Quality Assessment by Resampling
  Bootstrap

Functional Gene Annotation Systems
  Gene Ontologies
  Pathway Annotation Systems

References
What Are Ontologies?

- Problem: there are no consistent naming conventions for gene and protein functions. This limits the information content of the provided annotations, because they are often not ‘machine readable’.

- Solution: controlled vocabulary system developed and maintained by Gene Ontology (GO) Consortium.

- Ontologies are a machine interpretable representation of the different components and features of a system (e.g. organism).

- The Gene Ontology (GO) annotation systems consists of two main components:
  - A dictionary of controlled functional terms.
  - A graph (tree) system represents the relationships among the terms.

- Result: a unified functional annotation language that connects different gene annotation databases.
Database Unification by Gene Ontologies

Gene Ontology widely adopted

FlyBase

SGD

SMD

tair

MGI

Go

UniProt

NCBI

InterPro

AstraZeneca

GoBase

AgBase

GRAMENE

dictyBase

Day 2, Phylogenetics, Gene Annotations & Introduction to R

Functional Gene Annotation Systems

Gene Ontologies
Gene Ontologies (GO) are an annotation system that describes functions of genes and their products.

GO consists of three ontologies:

1. **Molecular Function (MF)**: molecular activity/task of genes
   Examples: carbohydrate binding and ATPase activity

2. **Biological Process (BP)**: broader biological function
   Examples: involved in mitosis or purine metabolism, accomplished by ordered assemblies of molecular functions

3. **Cellular Component (CC)**: subcellular location or component of complex
   Examples: nucleus, mitochondria or part of RNA polymerase II holoenzyme
The ontologies are structured as three directed acyclic graphs (DAGs):

1. DAG for molecular function (MF) terms
2. DAG for biological process (BP) terms
3. DAG for cellular component (CC) terms

DAGs are similar to hierarchical trees, but a child term can have multiple parent terms. For example, the biological process term hexose biosynthesis has two parents, hexose metabolism and monosaccharide biosynthesis.
DAG Example
Evidence Codes

The supporting data type for GO annotations can be tracked by their evidence codes. The current set of evidence codes are:

- **IDA**: Inferred from direct assay
- **IPI**: Inferred from physical interaction
- **IMP**: Inferred from mutant phenotype
- **IGI**: Inferred from genetic interaction
- **IEP**: Inferred from expression pattern
- **IEA**: Inferred from electronic annotation
- **ISS**: Inferred from sequence or structural similarity
- **TAS**: Traceable author statement
- **NAS**: Non-traceable author statement
- **IC**: Inferred by curator
- **RCA**: Reviewed Computational Analysis
- **ND**: no data available
Annotating Genes with GO

Gene Product

Reference

P05147

GO:0047519

IDA

PMID: 2976880

GO Term

Evidence

AmiGO

GO:0047519

IDA

PMID: 2976880

Functional Gene Annotation Systems

Gene Ontologies
Outline

Introduction
   Phylogenetics

Tree Basics
   Tree Topology
   Counting Trees
   Tree Rooting Methods
   Inferring Trees from Distances

Tree Building Methods
   Clustering Methods
   Parsimony Methods

Quality Assessment by Resampling
   Bootstrap

Functional Gene Annotation Systems
   Gene Ontologies
   Pathway Annotation Systems

References
Pathway Annotation Systems

- Pathway annotation systems assign gene products to pathway maps.
  - KEGG (Kyoto Encyclopedia of Genes and Genomes): pathway database for all organisms.
  - Reactome: general pathway database.
  - EcoCyc: pathway database for bacteria.
  - AraCyc: pathway database for Arabidopsis.
  - Many more...

- Advantages over GO: simpler data structure, less duplications.
- Disadvantages: low genome coverage (≈30%), contains only molecular information content.
References and Books


URL http://www.hubmed.org/display.cgi?uids=2531898