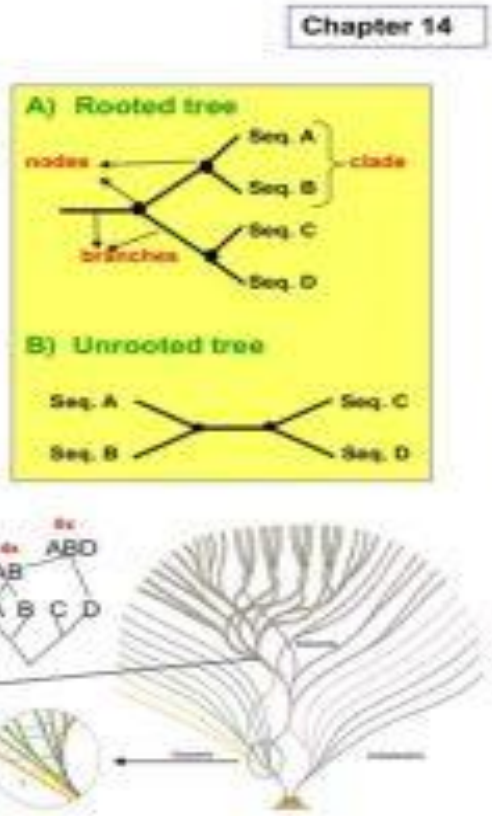
Phylogenetic Analysis -II

Phylogenetic Analysis

1. Introduction
2. Construction of Phylogenetic Trees
    1. Construction and editing of a MSA
    2. Selection of a substitution model
    3. Tree building
        1. Distance based methods
            1. UPGMA
            2. Neighbor-joining
        2. Character based methods
            1. Maximum Parsimony
            2. Maximum Likelihood
    4. Tree evaluation
3. Software
    1. MEGA
    2. PHYLIP
    3. PAUP



Summary of Part I

- Phylogenetics is the study of genetic relatedness of individuals of the same, or different, species. Through phylogenetics, evolutionary relationships can be inferred.
- A phylogenetic tree may be rooted or unrooted, depending on whether the ancestral root is known or unknown, respectively.
- A phylogenetic tree's root is the origin of evolution of the individuals studied. Branches between leaves show the evolutionary relationships between sequences, individuals, or species, and branch length represents evolutionary time.
- When constructing and analyzing phylogenetic trees, it is important to remember that the resulting tree is simply an estimate and is unlikely to represent the true evolutionary tree of life.

Building Phylogenetic Trees

Main methods:

- Distances matrix methods
  - Neighbour Joining, UPGMA
- Character based methods:
  - Parsimony methods
  - Maximum Likelihood method
- Validation method:
  - Bootstrapping
  - Jack Knife

Statistical Methods

- ✓ Bootstrapping Analysis –

Is a method for testing how good a dataset fits a evolutionary model.

This method can check the branch arrangement (topology) of a phylogenetic tree.

In Bootstrapping, the program re-samples columns in a multiple aligned group of sequences, and creates many new alignments, (with replacement the original dataset).

These new sets represent the population.

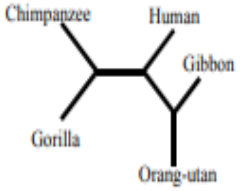
Statistical Methods

- The process is done at least 100 times.
- Phylogenetic trees are generated from all the sets.
- Part of the results will show the # of times a particular branch point occurred out of all the trees that were built.

The higher the # - the more valid the branching point.

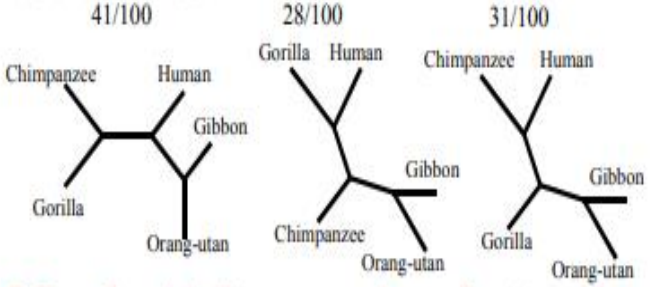
Taken from Dr. Itai Yanai

Given the following tree, estimate the confidence of the two internal branches

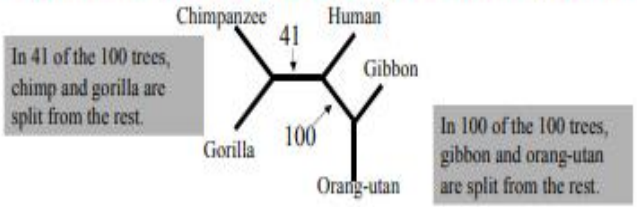


Estimating Confidence from the Resamplings

1. Of the 100 trees:



2. Upon the original tree we superimpose bootstrap values:



Statistical Methods

- Bootstrap values between 90-100 are considered statistically significant

Character Based Methods

All Character Based Methods assume that each character substitution is independent of its neighbors.

- Maximum Parsimony (minimum evolution) - in this method one tree will be given (built) with the fewest changes required to explain (tree) the differences observed in the data.

Character Based Methods

Q: How do you find the minimum # of changes needed to explain the data in a given tree?

A: The answer will be to construct a set of possible ways to get from one set to the other, and choose the "best". (for example: Maximum Parsimony)

```
CCGCCACGA
  P   P   R
CGGCCACGA
  R   P   R
```

Character Based Methods - Maximum Parsimony

- Not all sites are informative in parsimony.

- Informative site, is a site that has at least 2 characters, each appearing at least in 2 of the sequences of the dataset.

Maximum Parsimony

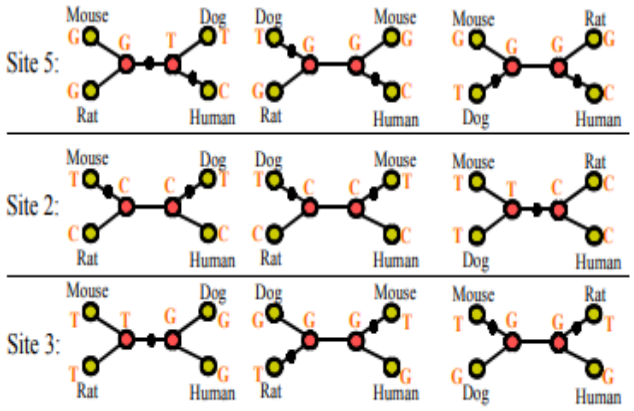
Start by classifying the sites:

```
                12345678901234567890
Mouse           CTTCGTTGGATCAGTTTGATA
Rat             CCTCGTTGGATCATTTTGATA
Dog             CTGCTTTGGATCAGTTTGAAC
Human           CCGCCTTGGATCAGTTTGAAC
-------------------------------------
Invariant     *   * ******** *****
Variant        ** *          *      **
-------------------------------------
Informative   **                    **
Non-inform.      *          *
```

Taken from
Dr. Itai Yanai

```
                12345678901234567890
Mouse           CTTCGTTGGATCAGTTTGATA
Rat             CCTCGTTGGATCATTTTGATA
Dog             CTGCTTTGGATCAGTTTGAAC
Human           CCGCCTTGGATCAGTTTGAAC
                 ** *
```

Taken from
Dr. Itai Yanai



Maximum Parsimony

```
                12345678901234567890
Mouse           CTTCGTTGGATCAGTTTGATA
Rat             CCTCGTTGGATCATTTTGATA
Dog             CTGCTTTGGATCAGTTTGAAC
Human           CCGCCTTGGATCAGTTTGAAC
Informative     **                  **
```

Taken from
Dr. Itai Yanai

3 0 1

Character Based Methods - Maximum Parsimony

- The Maximum Parsimony method is good for similar sequences, a sequences group with small amount of variations

Maximum Parsimony methods do not give the branch lengths only the branch order.

For larger set it is recommended to use the “branch and bound” method instead Of Maximum Parsimony.

Maximum Parsimony Methods are Available…

- For DNA in Programs:

paup, molphy,phylo_win

In the Phylip package:

DNAPars, DNAPenny, etc..

- For Protein in Programs:

paup, molphy,phylo_win

In the Phylip package:

PROTPars

Character Based Methods - Maximum Likelihood

- Basic idea of Maximum Likelihood method is building a tree based on mathemaical model.
- This method find a tree based on probability calculations that best accounts for the large amount of variations of the data (sequences) set.

- Maximum Likelihood method (like the Maximum Parsimony method) performs its analysis on each position of the multiple alignment.

  This is why this method is very heavy on CPU.

Character Based Methods - Maximum Likelihood

- Maximum Likelihood method – using a tree model for nucleotide substitutions, it will try to find the most likely tree (out of all the trees of the given dataset).
- The Maximum Likelihood methods are very slow and cpu consuming.
- Maximum Likelihood methods can be found in phylip, paup or puzzle.