Introduction to Phylogenetic Analysis

Irit Orr

Mar 2004

Subjects of this lecture

1 Introducing some of the terminology of phylogenetics.

2 Introducing some of the most commonly used methods for phylogenetic analysis.

3 Explain how to construct phylogenetic trees.

Taxonomy - is the science of classification of organisms.

Phylogeny – is the evolution of a genetically related group of organisms.

Or: A study of relationships between collection of "things" (genes, proteins, organs..) that are derived from a common ancestor

Phylogenetics - WHY?

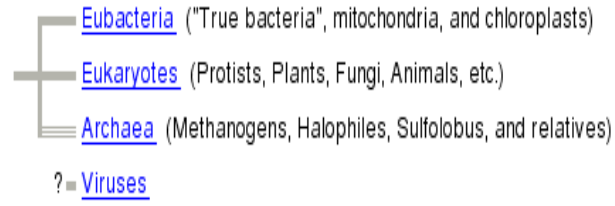
- Find evolutionary ties between organisms.

(Analyze changes that occured in different organisms during evolution).

- Find (understand) relationships between an ancestral sequence and it descendants.

(Evolution of family of sequences)

- Estimate time of divergence between a group of organisms that share a common ancestor.



Basic tree Of Life

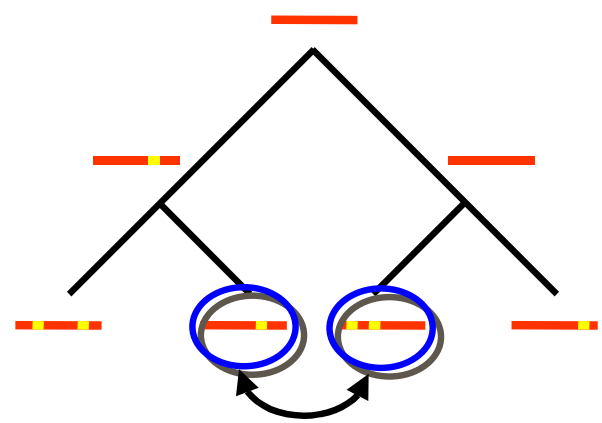
Eukaryote tree

From a common ancestor sequence, two DNA sequences are diverged.

Each of these two sequences start to accumulate nucleotide substitutions.

The number of these mutations are used in molecular evolution analysis.

Similar sequences, common ancestor...

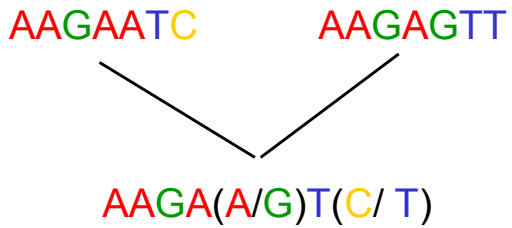


... common ancestor, similar function

- One of the most striking features of life is all living organisms share "highly conserved regions" in proteins, particularly in proteins that are involved in information processing (transcription and translation).
- All of course share the same genetic codes.
- This information lead us to accept the theory that all organisms known to us have evolved from a common ancestor.
- (Patrick Forterre coined this ancestor as LUCA (Last Universal Common Ancestor)

Relationships of Phylogenetic Analysis and Sequences Analysis

When 2 sequences found in 2 organisms are very similar, we assume that they have derived from one ancestor.



The sequences alignment reveal which positions are conserved from the ancestor sequence.

How we calculate the Degree of Divergence

- If two sequences of length N differ from each other at n sites, then their degree of divergence is:

  n/N or n/N*100%.

Relationships of Phylogenetic Analysis and Sequences Analysis

- Most phylogenetic methods assume that each position in a sequence can change independently from the other positions.
- Gaps in alignments represent mutations in sequences such as: insertion, deletion, genetic rearrangments.
- Gaps are treated in various ways by the phylogenetic methods. Most of them ignore gaps.

Relationships of Phylogenetic Analysis and Sequences Analysis

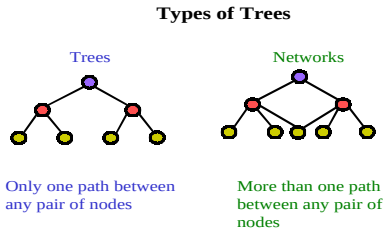
- Another approach to treat gaps is by using sequences similarity scores as the base for the phylogenetic analysis, instead of using the alignment itself, and instead of trying to decide what happened at each position of the sequence.
- Alignment similarity scores are based on scoring matrices (with gaps scores), and are used by the DISTANCES methods.

What is a phylogenetic tree?

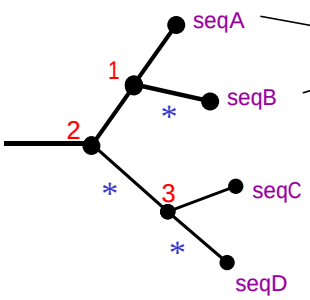
- An illustration of the evolutionary relationships among a group of organisms.
- Dendrogram is another name for a phylogenetic tree.
- A tree is composed of nodes and branches. One branch connects any two adjacent nodes. Nodes represent the taxonomic units. (sequences)

What is a phylogenetic tree?

- E.G: 2 very similar sequences will be neighbors on the outer branches and will be connected by a common internal branch.



Phylogenetic Tree



Tips (Leaves) = Represent the taxa (sequences)

Nodes = 1 2 3
Represent a speciation event.
Nodes (e.g 1) represent the ancestor seq from which seqA and seqB divrged.

Branches *
Connect Tips and Nodes on the tree. The length of the branch can represent the # of changes that occurred in the seqs prior to the next level of separation.

In a phylogenetic tree...

- ✓ Each NODE represents a speciation event in evolution. Beyond this point any sequence changes that occurred in each new specie, are specific for it (specie).
- ✓ The BRANCH connects NODES or TIPS on the tree. The length of each BRANCH between one NODE to the next, represents the # of changes that occurred until the next speciation (separation) event.
- ✓ After a branching event, one taxon (sequence) can undergo more mutations then the other taxon.

In a phylogenetic tree...

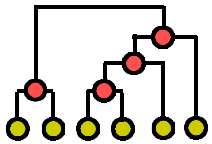
- ✓ NOTE: The amount of evolutionary time that passed from a diversion of 2 sequences from their common ancestor is not known.
- ✓ The phylogenetic analysis can only estimate the time, based on the # of changes that occurred from the time of separation.
- ✓ Topology of a tree is the branching pattern of a tree.

Tree structure

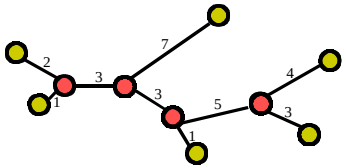
- ♠ Terminal nodes - represent the data (e.g sequences) under comparison (A,B,C,D,E), also known as OTUs, (Operational Taxonomic Units).
- ♠ Internal nodes - represent inferred ancestral units (usually without empirical data), also known as HTUs, (Hypothetical Taxonomic Units).

Slide taken from Dr. Itai Yanai

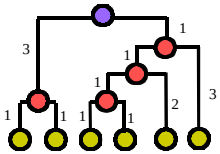
Different kinds of trees can be used to depict different aspects of evolutionary history



1. Cladogram:
   simply shows relative recency of common ancestry



2. Additive trees:
   a cladogram with branch lengths, also called phylograms and metric trees



3. Ultrametric trees:
   (dendograms) special kind of additive tree in which the tips of the trees are all equidistant from the root

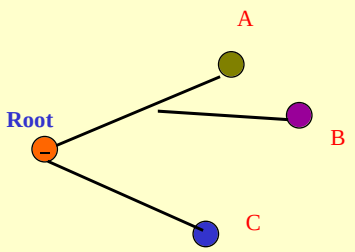
The Molecular Clock Hypothesis

- All the mutations occur in the same rate in all the taxa of a tree.
- The rate of the mutations is the same for all positions along the sequence.
- The Molecular Clock Hypothesis is most suitable for closely related species.

Rooted Tree = Cladogram

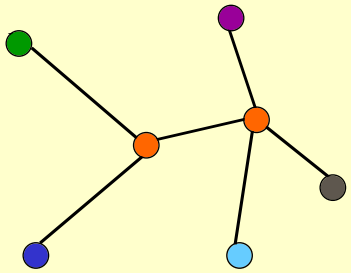
- A phylogenetic tree that all the "objects" on it share a known common ancestor (the root).
- There exists a particular root node.
- A root is a taxon (seq) that branched earlier of all the other taxa on the tree, but is related to them.

❖ The paths from the root to the nodes correspond to evolutionary time



Unrooted Tree = Phenogram

A phylogenetic tree where all the "objects" on it are related descendants - but there is not enough information to specify the common ancestor (root).



The path between nodes of the tree do not specify an evolutionary time.

Rooted versus Unrooted

❖ The number of tree topologies of rooted tree is much higher than that of the unrooted tree for the same number of OTUs (sequences).

❖ Therefore, the error of an unrooted tree topology is smaller than that of a rooted tree.

Slide taken from Dr. Itai Yanai

The number of rooted and unrooted trees:

Number of OTU's	Possible Number of	
	Rooted trees	Unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10395	945
8	135135	10395
9	2027025	135135
10	34459425	2027025

OTU – Operational Taxonomical Unit

- Orthologs - genes related by speciation events. Meaning same genes in different species.
- Paralogs - genes related by duplication events. Meaning duplicated genes in the same species.

Selecting sequences for phylogenetic analysis

What *type* of sequences are used for the phylogenetic analysis? Protein or DNA?

For DNA sequences, the rate of mutation is assumed to be the same in both coding and non-coding regions.

However, there is a difference in the substitution rate between coding and non-coding regions.

- Non-coding DNA regions have more substitution than coding regions.

Selecting sequences for phylogenetic analysis

- For Proteins, the rate of mutation is very low in the conserved region ,or "functional regions" as we relate to them.
- Most evolution algorithm can better analyze regions that mutate slowly, small number of changes in the multiple alignments.
- Regions that have "high number of changes" need special algorithm to deal with them successfully.

Known Problems of Multiple Alignments

- The base for a phylogenetic analysis is Multiple Alignments of the sequences.

SO…Good analysis is based on good alignments.

- Check the alignment to see that Important Sites are not misaligned by the software used for the sequence alignment.
- Misalignment can effect the significance of the site - and the tree.
  - For example: ATG as start codon, or specific amino acids in functional domains.

Known Problems of Multiple Alignments

- Gaps in multiple alignment are usually not scored, (or plainly ignored) by most programs.
- Gaps are not scored since there is no suitable model of evolution mechanism that produces them.

Known Problems of Multiple Alignments

```
T  Y  R  R  S  R        ACA  TAC  AGG  CGA
                         T    Y    R    R
T  Y  R  R  S  R        ACA  TAC  AGG  CGA
                         T    Y    R    R
T  Y  R  -  S  R        ACA  TAC  AGG  ---
                         T    Y    R    -
T  Y  R  -  S  R        ACA  TAC  ---  CGA
                         T    Y    -    R
T  Y  R  R  S  R        ACA  TAC  AGG  CGA
                         T    Y    R    R
```

Alignment of Proteins that contains gaps, should be compared with the alignment of their DNA coding regions.. The reason is to to be sure about the placement of gaps, since the degeneracy of the cod.

Known Problems of Multiple Alignments

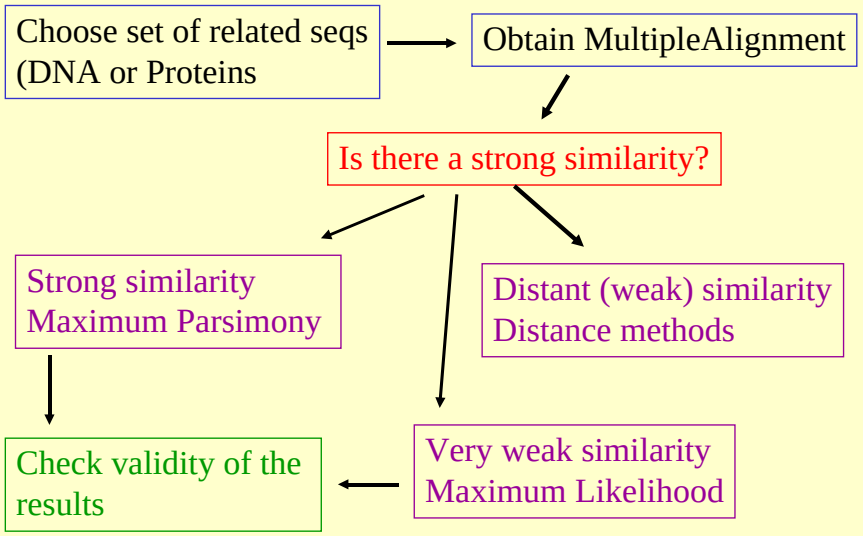
- Low complexity regions - effect the multiple alignment because they create random bias for various regions of the alignment.
- Low complexity regions should be removed from the alignment before building the tree.

- If you delete these regions you need to consider the affect of the deletions on the branch lengths of the whole tree.

Selecting sequences for phylogenetic analysis

- Sequences that are being compared belong together (orthologs).
- If no ancestral sequence is available you may use an "outgroup" as a reference to measure distances. In such a case, for an outgroup you need to choose a close relative to the group being compared.
  - For example: if the group is of mammalian sequences then the outgroup should be a sequence from birds and not plants.

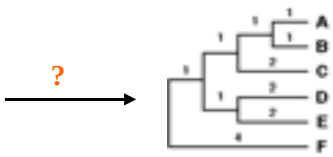
How to choose a phylogenetic method?



Taken from Dr.Itai Yanai

Given a multiple alignment, how do we construct the tree?

```
A - GCTTGTCCGTTACGAT
B – ACTTGTCTGTTACGAT
C – ACTTGTCCGAAACGAT
D - ACTTGACCGTTTCCTT
E – AGATGACCGTTTCGAT
F - ACTACACCCTTATGAG
```



Building Phylogenetic Trees

Main methods:

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

Character Based Methods

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

- In most of the programs based on Character Methods, the program will yield (build) one tree with the fewest changes required to explain (tree) the differences observed in the data.

Character Based Methods

For example the Maximum Parsimony (minimum evolution) program will find the tree where any changes can occur in any of the sequences (taxa) on the tree, that can turn them into any of the other sequences on the tree.

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 the sites are informative for the parsimony method.
- Informative site, is a site that has at least 2 characters, each appearing at least in 2 of the sequences of the dataset.
- With the parsimony method, only informative sites need to be analyzed.

Maximum Parsimony

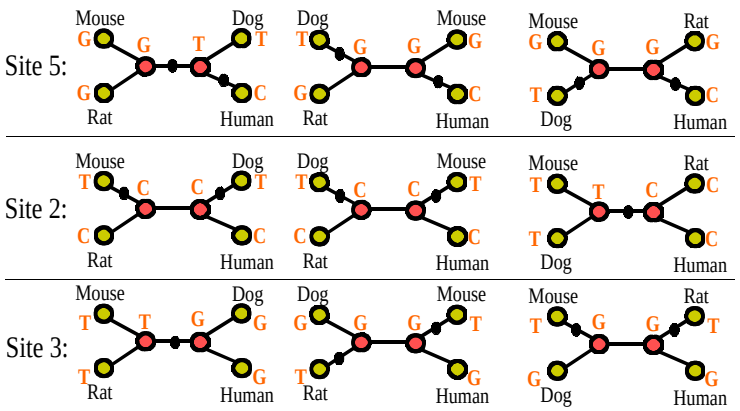
Start by classifying the sites:

```
                 12345678901234567890 1
Mouse            CTTCGTTGGATCAGTTTGATA
Rat              CCTCGTTGGATCATTTTGATA
Dog              CTGCTTTGGATCAGTTTGAAC
Human            CCGCCTTGGATCAGTTTGAAC
-------------------------------------
Invariant        *   * ******** *****
Variant           ** *         *     **
-------------------------------------
Informative      **                 **
Non-inform.         *         *
```

Taken from Dr. Itai Yanai

```
                 12345678901234567890 1
Mouse            CTTCGTTGGATCAGTTTGATA
Rat              CCTCGTTGGATCATTTTGATA
Dog              CTGCTTTGGATCAGTTTGAAC
Human            CCGCCTTGGATCAGTTTGAAC
                  ** *
```

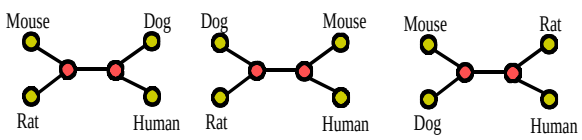
Taken from Dr. Itai Yanai



Maximum Parsimony

```
                 12345678901234567890 1
Mouse            CTTCGTTGGATCAGTTTGATA
Rat              CCTCGTTGGATCATTTTGATA
Dog              CTGCTTTGGATCAGTTTGAAC
Human            CCGCCTTGGATCAGTTTGAAC
Informative       **                 **
```

Taken from Dr. Itai Yanai



Character Based Methods - Maximum Parsimony

- The possible optimal tree is built by adding the number of changes at each informative site for each tree.
- The tree that requires the least number of changes is chosen.

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 mathematical 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 –

Starts with a set of sequences, and finds estimates for the variability in each position.

It checks the rates of transitions and transversions (using both models Kimura and Jukes & Cantor).

At the end, after all positions in the sequences alignment were checked, the likelihood of the whole tree is reported.

Maximum Likelihood method

- The Maximum Likelihood methods are very slow and cpu consuming.
- Maximum Likelihood methods can be found in phylip, paup or puzzle.
- In phylip package in programs:

  DNAML and DNAMLK

Distances Methods

- Distance - the number of substitutions per site per time period.
- Evolutionary distance are calculated based on one of DNA evolutionary models.
- Neighbors – pair of sequences, in a sequence set, that have the smallest number of changes (substitutions) between them.
- On a phylogenetic tree, neighbors are joined by a branch to the same node (common ancestor).

Distances Matrix Methods

- Some of the Distances methods assume a molecular clock, such as the Neighbor-Joining UPGMA. Most of the other Distances programs do not.
- This assumption is not true for several reasons:
  - Different environmental conditions affect mutation rates.
  - This assumption ignores selection issues which are different with different time periods.

Distances Matrix Methods

- Distance methods try to place the correct positions of all the neighbors, and find the correct branches lengths. However, they vary in the way they construct the tree.
- *Distance based clustering methods*:
  - **Neighbor-Joining (unrooted tree)**
  - **UPGMA (rooted tree)**

Common steps to build a TREE used by Distances method

1 Multiple alignments - based on all against all pairwise comparisons.

2 Building distance matrix of all the compared sequences (all pairs of OTUs).

3 Disregard of the actual sequences.

4 Constructing a guide tree by clustering the distances. Iteratively build the relations (branches and internal nodes) between all OTUs.

Distance method steps

Construction of a distance tree using clustering with the Unweighted Pair Group Method with Arithmatic Mean (UPGMA)

First, construct a distance matrix:

```
A - GCTTGTCCGTTACGAT
B – ACTTGTCTGTTACGAT
C – ACTTGTCCGAAACGAT
D - ACTTGACCGTTTCCTT
E – AGATGACCGTTTCGAT
F - ACTACACCCTTATGAG
```

→

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

From http://www.icp.ucl.ac.be/~opperd/private/upgma.html

Distances Matrix Methods

- Distances matrix methods can be found in the following Programs:

  Clustalw, Phylo_win, Paup

  In the GCG software package:

  Paupsearch, distances

  In the Phylip package:

  DNADist, PROTDist, Fitch, Kitch, Neighbor

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.

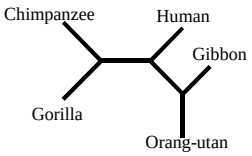
Statistical Methods

- 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

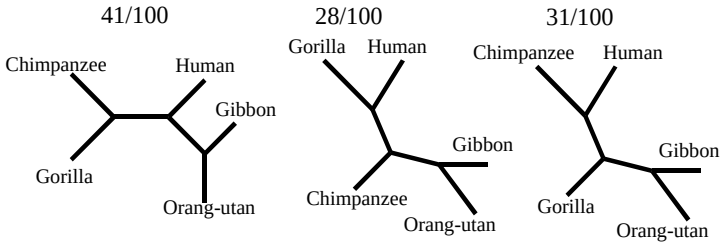
Given the following tree, estimate the confidence of the two internal branches



Taken from Dr. Itai Yanai

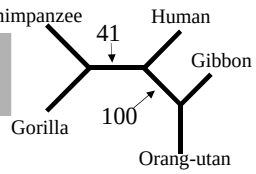
Estimating Confidence from the Resamplings

1. Of the 100 trees:



2. Upon the original tree we superimpose bootstrap values:

In 41 of the 100 trees, chimp and gorilla are split from the rest.



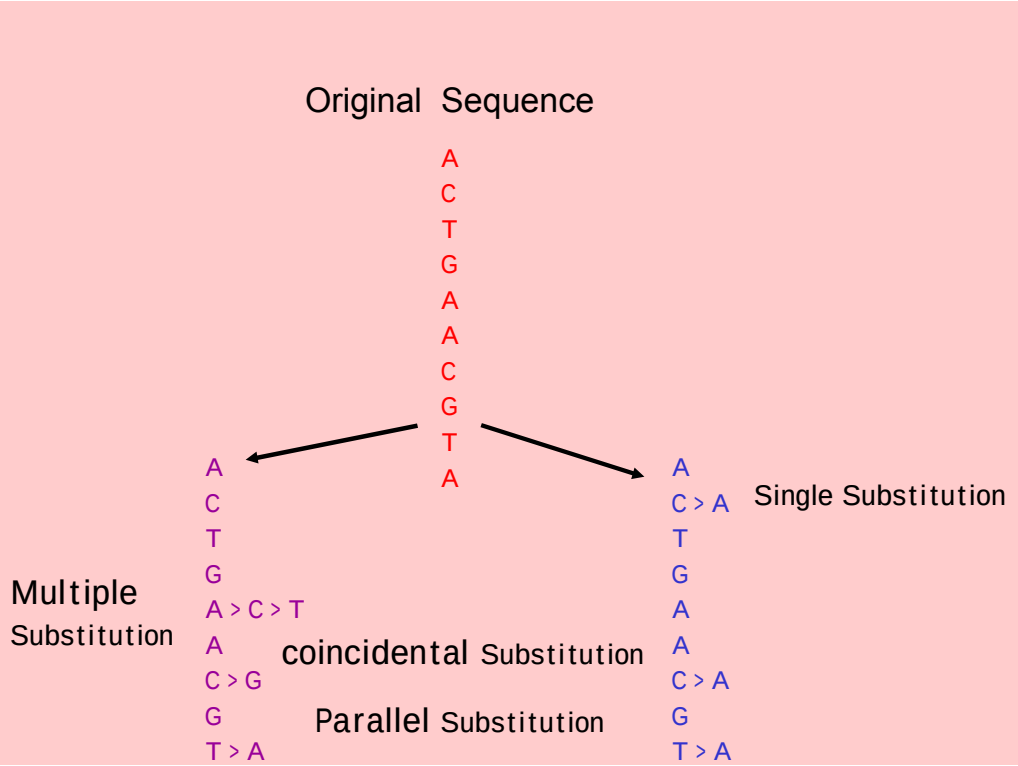
In 100 of the 100 trees, gibbon and orang-utan are split from the rest.

Statistical Methods

- Bootstrap values between 90-100 are considered statistically significant

Mutations in DNA as a source for evolutionary analysis

- Only mutations that were fixed in the population are called substitutions.
- We assume that each observed change in similar sequences, represent a "single mutation event".
- The greater the number of changes, the more possible types of mutations.



DNA Substitution Mutations

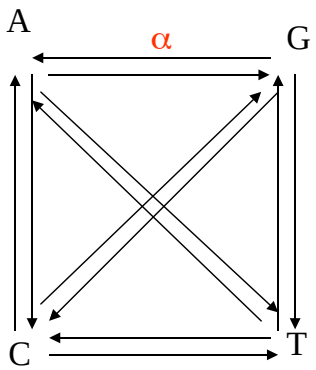
- Transition - a change between purines (A,G) or between pyrimidines (T,C).
- Transversion - a change between purines (A,G) to pyrimidines (T,C).
- *Substitution mutations usually arise from mispairing of bases during replication.*

Correction for likelihood of mutations in DNA sequences

- There are several evolutionary models used for correction for the likelihood of multiple mutations and reversions in DNA sequences.
- These evolutionary models use a normalized distance measurement that is the average degree of change per length of aligned sequences.

Jukes & Cantor one-parameter model

This model assumes that substitutions between the 4 bases occur with equal frequency. Meaning no bias in the direction of the change.

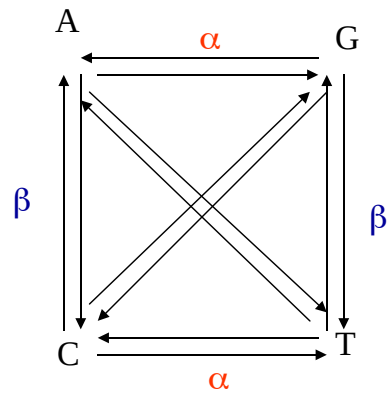


α Is the rate of substitutions In each of the 3 directions For one base.

α (Is the one parameter).

Kimura two-parameter model

This model assumes that transitions (A - G or T - C) occur more often than transversions (purine -pyrimidine).



α Is the rate of transitional Substitutions.

β Is the rate of transversional substitutions.

- These evolutionary models improve the distance calculations between the sequences.
- These evolutionary models have less effect in phylogenetic predictions of closely related sequences.
- These evolutionary models have better effect with distant related sequences.

DNA Evolution Models

① A basic process in DNA sequence evolution is the substitution of one nucleotide with another.

① This process is slow and can not be observed directly.

① The study of DNA changes is used to estimate the rate of evolution, and the evolution history of organisms.

How to draw Trees? (Building trees software)

* Unrooted trees should be plotted using the DRAWGRAM program (phylip), or similar.

* Rooted trees should be plotted using the DRAWTREE program (phylip), or similar.

* On a PC/Mac use the TreeView program

A Tip...

- For DNA sequences use the Kimura's model in the building trees programs.

- For PROTEINS the differences lie with the scoring (substitution) matrices used. For more distant sequences you should use BLOSUM with lower # (i.e., for distant proteins use blosum45 and for similar proteins use blosum60).

Known problems of Phylogenetic Analysis

- Order of the input data (sequences) -
  The order of the input sequences effects the tree construction. You can "correct" this effect in some of the programs (like phylip), using the Jumble option. (J in phylip set to 10).

- The number of possible trees is huge for large datasets. Often it is not possible to construct all trees, but can guarantee only "a good" tree not the "best tree".

Known problems of Phylogenetic Analysis

- The definition of "best tree" is ambiguous.
  It might mean the most likely tree, or a tree with the fewest changes, or a tree best fit to a known model, etc..

  The trees that result from various methods differ from each other. Never the less, in order to compare trees, one need to assume some evolutionary model so that the trees may be tested.

Known problems of Phylogenetic Analysis

- Δ Pay attention to the data used for the tree construction, use “informative” data, without large gaps.
- Δ Population effects are often to be considered, especially if we have a lot of variety (large # of alleles for one protein).

How many trees to build?

- ! For each dataset it is recommended to build more than one tree. Build a tree using a distance method and if possible also use a character-based method, like maximum parsimony.
- ! The core of the tree should be similar in both methods, otherwise you may suspect that your tree is incorrect.