Introduction to Phylogenetic Analysis

Irit Orr

WEIZMANN INSTITUTE OF SCIENCE

Israel National Node

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?

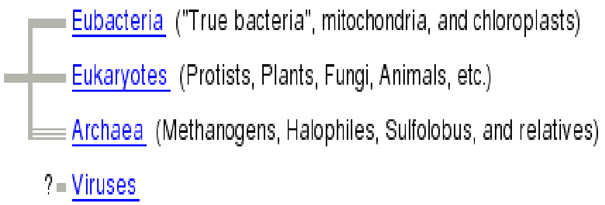
- Find evolutionary ties between organisms.

(Analyze changes occuring in different organisms during evolution).

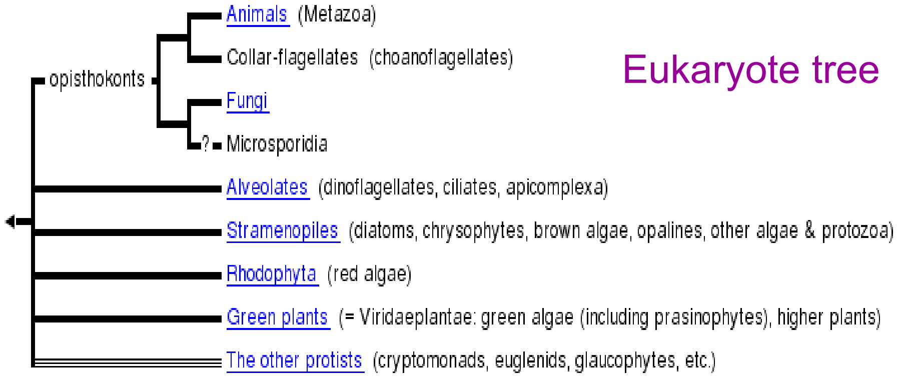
- Find (understand) relationships between an ancestral sequence and it descendants.

(Evolution of family of sequences)

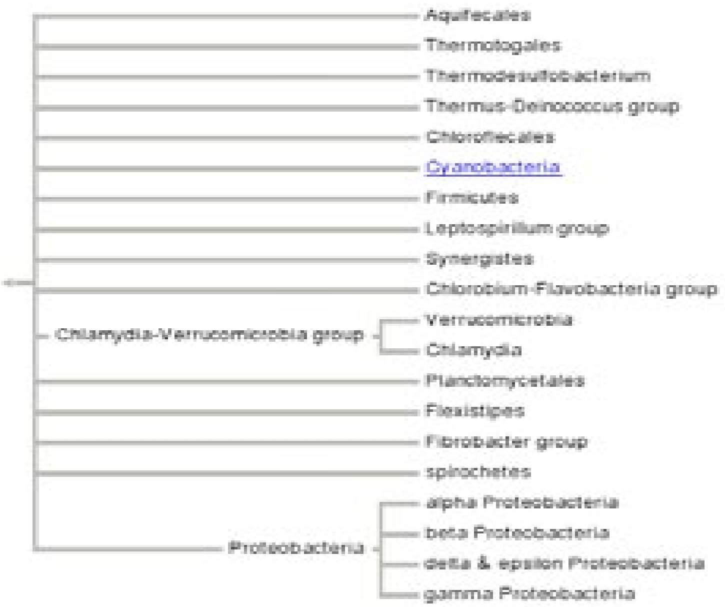
- Estimate time of divergence between a group of organisms that share a common ancestor.


Eubacteria ("True bacteria", mitochondria, and chloroplasts)
Eukaryotes (Protists, Plants, Fungi, Animals, etc.)
Archaea (Methanogens, Halophiles, Sulfolobus, and relatives)
? Viruses

Basic tree Of Life

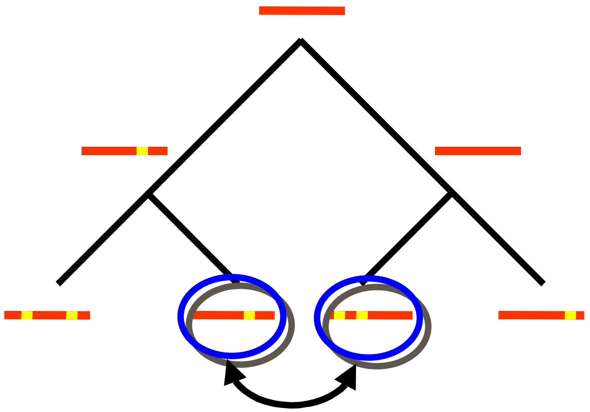

opisthokonts
Animals (Metazoa)
Collar-flagellates (choanoflagellates)
Fungi
? Microsporidia
Alveolates (dinoflagellates, ciliates, apicomplexa)
Stramenopiles (diatoms, chrysophytes, brown algae, opalines, other algae & protozoa)
Rhodophyta (red algae)
Green plants (= Viridaeplantae: green algae (including prasinophytes), higher plants)
The other protists (cryptomonads, euglenids, glaucophytes, etc.)

Eukaryote tree



Eubacteria tree

Similar sequences, common ancestor...



... common ancestor, similar function

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.

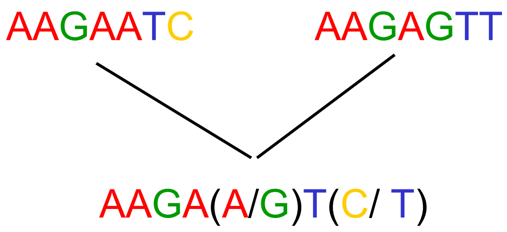
How we calculate the Degree of Divergence

- 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

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


AAGAATC

AAGAGTT

AAGA(A/G)T(C/ T)

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

Relationships of Phylogenetic Analysis and Sequences Analysis

- The progressive multiple alignment of a group of sequences, first aligns the most similar pair.
- Then it adds the more distant pairs.
- The alignment is influenced by the “most similar” pairs and arranged accordingly, but….it does not always correctly represent the evolutionary history of the occured changes.
- Not all phylogenetic methods work this way.

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

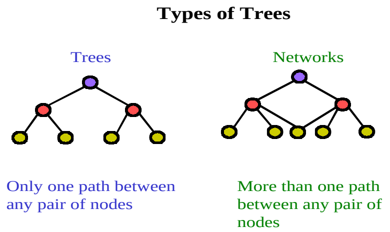
- Another approach to treat gaps is by using sequences similarity scores as the base for the phylogenetic analysis, instead of using the alignment itself, and trying to decide what happened at each position.
- The similarity scores based on scoring matrices (with gaps scores) are used by the DISTANCE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

What is a phylogenetic tree?

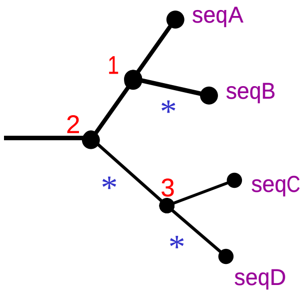
- An illustration of the evolutionary relationships among a group of organisms.
- Dendrogram is another name for a phylogenetic tree.
- A tree is composed of nodes and branches. One branch connects any two adjacent nodes. Nodes represent the taxonomic units. (sequences)

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.


Types of Trees
Trees
Networks
Only one path between any pair of nodes
More than one path between any pair of nodes

Rooted Phylogenetic Tree

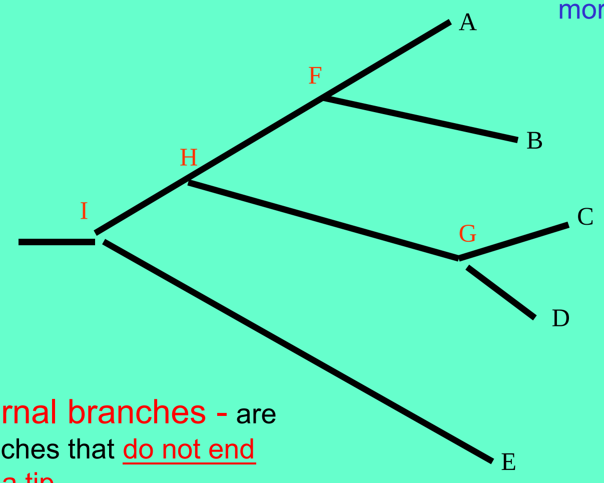


Leaves = Outer branches
Represent the taxa (sequences)

Nodes = 1 2 3
Represent the relationships
Among the taxa (sequences)
e.g Node 1 represent the ancestor
seq from which seqA and seqB derived.

Branches *
The length of the branch represent
the # of changes that occurred in the
seqs prior to the next level of separation.

External branches - are branches that end with a tip.
(FA,FB,GC,GD,IE)
more recent diversions



Internal branches - are branches that do not end with a tip.
(IH,HF,HG)
more ancient diversions

In a phylogenetic tree...

- ✓ Each NODE represents a speciation event in evolution. Beyond this point any sequence changes that occurred are specific for each branch (specie).
- ✓ The BRANCH connects 2 NODES of the tree. The length of each BRANCH between one NODE to the next, represents the # of changes that occurred until the next separation (speciation).

In a phylogenetic tree...

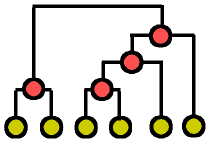
- ✓ NOTE: The amount of evolutionary time that passed from the separation of the 2 sequences is not known. The phylogenetic analysis can only estimate the # of changes that occurred from the time of separation.
- ✓ After the branching event, one taxon (sequence) can undergo more mutations then the other taxon.
- ✓ Topology of a tree is the branching pattern of a tree.

Tree structure

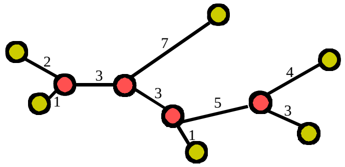
- ♠ Terminal nodes - represent the data (e.g sequences) under comparison (A,B,C,D,E), also known as OTUs, (Operational Taxonomic Units).
- ♠ Internal nodes - represent inferred ancestral units (usually without empirical data), also known as HTUs, (Hypothetical Taxonomic Units).

Slide taken from Dr. Itai Yanai

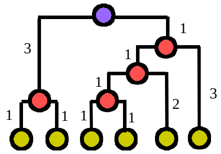
Different kinds of trees can be used to depict different aspects of evolutionary history



1. Cladogram:
   simply shows relative recency of common ancestry



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



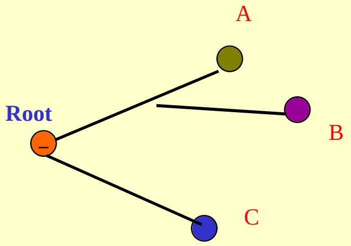
3. Ultrametric trees:
   (dendograms) special kind of additive tree in which the tips of the trees are all equidistant from the root

The Molecular Clock Hypothesis

- All the mutations occur in the same rate in all the tree branches.
- 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.

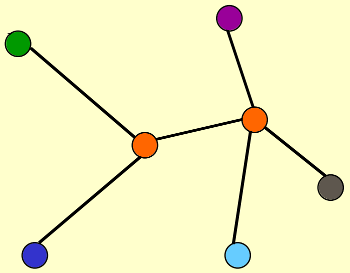
Rooted Tree = Cladogram

- A phylogenetic tree that all the "objects" on it share a known common ancestor (the root).
- There exists a particular root node.
- The paths from the root to the nodes correspond to evolutionary time.



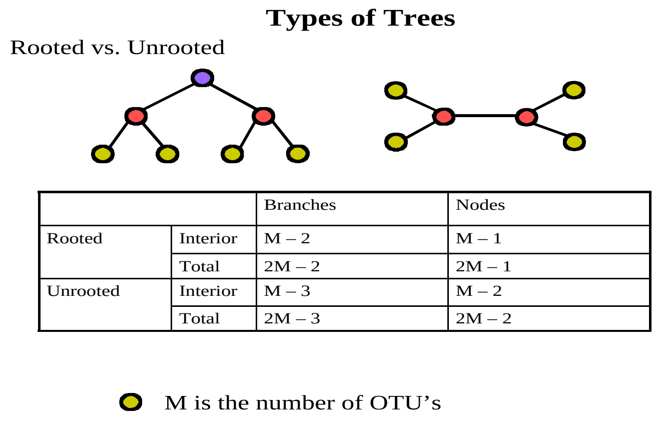
Unrooted Tree = Phenogram

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.

Slide taken from Dr. Itai Yanai



Types of Trees

Rooted vs. Unrooted

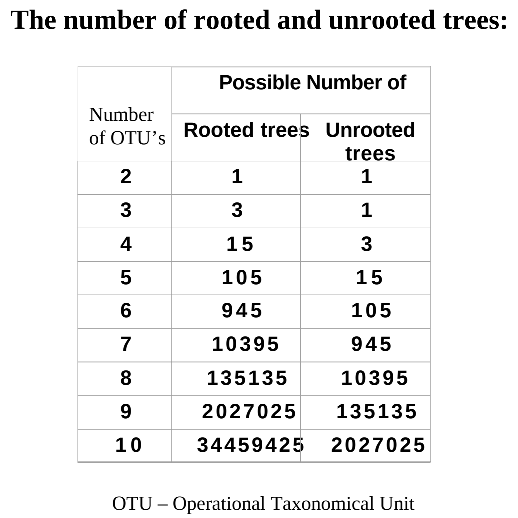
		Branches	Nodes
Rooted	Interior	M – 2	M – 1
	Total	2M – 2	2M – 1
Unrooted	Interior	M – 3	M – 2
	Total	2M – 3	2M – 2

● M is the number of OTU's

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.
- Therefore, the error of the unrooted tree topology is smaller than that of the 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 sequence to use, Protein or DNA?

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.

Selecting sequences for phylogenetic analysis

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

❖Proteins are much more conserved since they "need" to conserve their function.

So it is better to use sequences that mutate slowly (proteins) than DNA. However, if the genes are very small, or they mutate slowly, we can use them for building the trees.

Known Problems of Multiple Alignments

❷ Important sites could be misaligned by the software used for the sequence alignment. That will effect the significance of the site - and the tree.

- For example: ATG as start codon, or specific amino acids in functional domains.

❷ Gaps - Are treated differently by different alignment programs and should play no part in building trees.

Alignment of a coding region should be compared with the alignment of their protein sequences, to be sure about the placement of gaps.

```
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
```

Known Problems of Multiple Alignments

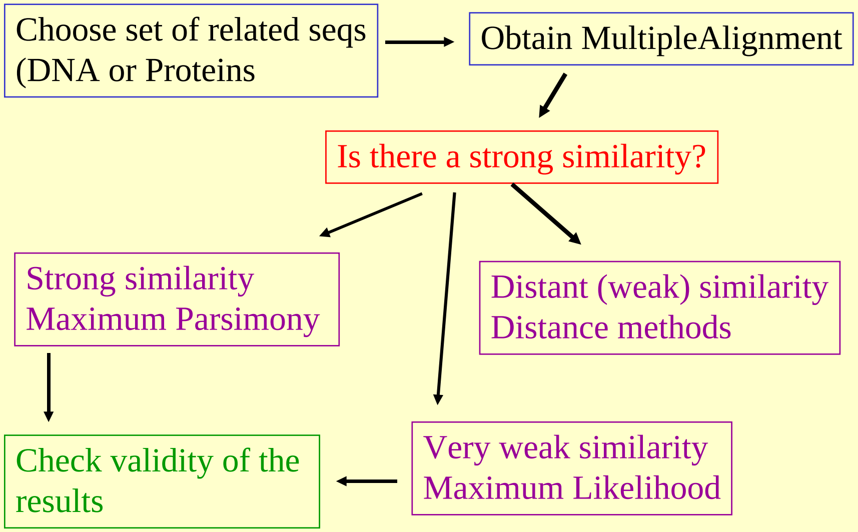
- ♥ Low complexity regions - effect the multiple alignment because they create random bias for various regions of the alignment.
- ♥ Low complexity regions should be removed from the alignment before building the tree.

✳ 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?



Choose set of related seqs (DNA or Proteins

Obtain MultipleAlignment

Is there a strong similarity?

Strong similarity Maximum Parsimony

Distant (weak) similarity Distance methods

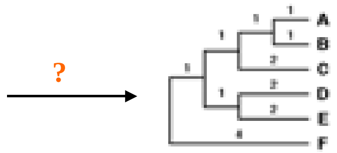
Very weak similarity Maximum Likelihood

Check validity of the results

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

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.

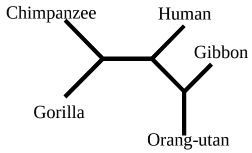
Statistical Methods

- 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.

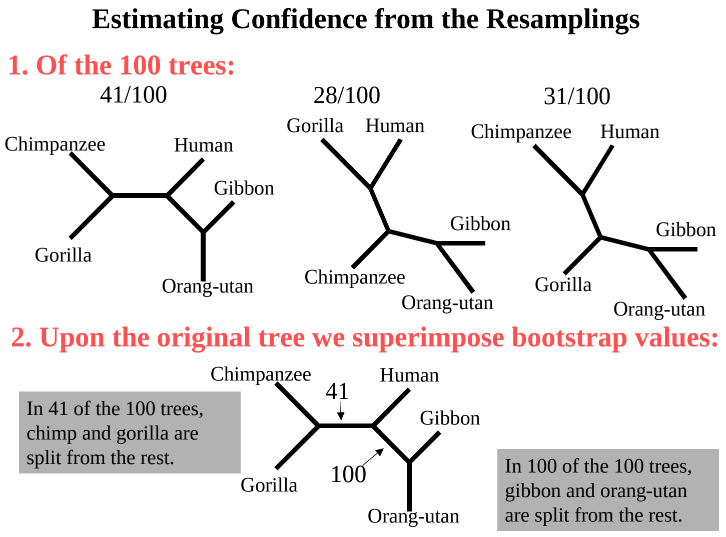
The higher the # - the more valid the branching point.

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

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

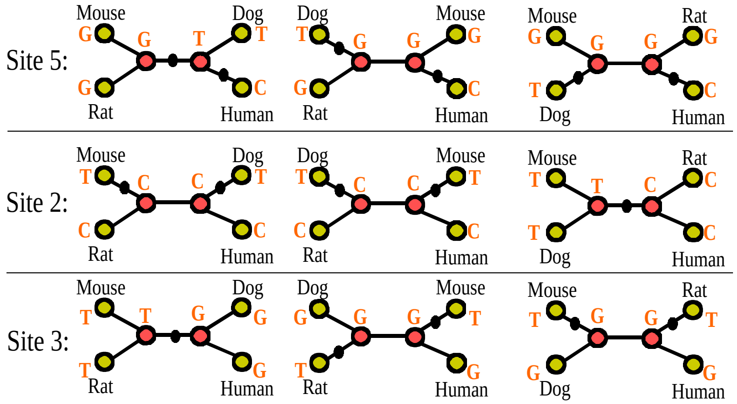
Start by classifying the sites:

```
                123456789012345678901
Mouse           CTTCGTTGGATCAGTTTGATA
Rat             CCTCGTTGGATCATTTTGATA
Dog             CTGCTTTGGATCAGTTTGAAC
Human           CCGCCTTGGATCAGTTTGAAC
-------------------------------------
Invariant       *  * ******** *****
Variant          ** *        *     **
-------------------------------------
Informative      **                **
Non-inform.         *        *
```

Taken from Dr. Itai Yanai

```
                123456789012345678901
Mouse           CTTCGTTGGATCAGTTTGATA
Rat             CCTCGTTGGATCATTTTGATA
Dog             CTGCTTTGGATCAGTTTGAAC
Human           CCGCCTTGGATCAGTTTGAAC
                 ** *
```

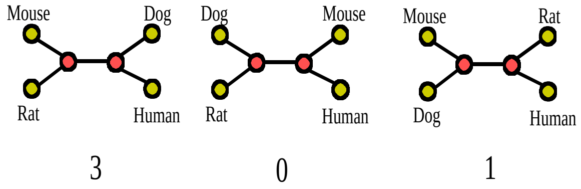
Taken from Dr. Itai Yanai



Maximum Parsimony

```
                123456789012345678901
Mouse           CTTCGTTGGATCAGTTTGATA
Rat             CCTCGTTGGATCATTTTGATA
Dog             CTGCTTTGGATCAGTTTGAAC
Human           CCGCCTTGGATCAGTTTGAAC
Informative      **                **
```

Taken from Dr. Itai Yanai



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.

Maximum Likelihood method

- Are available in the Programs:

paup or puzzle

In phylip package in programs:

DNAML and DNAMLK

Character Based Methods

- The Maximum Likelihood methods are very slow and cpu consuming (computer expensive).
- Maximum Likelihood methods can be found in phylip, paup or puzzle.

Distances Matrix Methods

- Distance methods assume a molecular clock, meaning that all mutations are neutral and therefore they happen at a random clocklike rate.
- 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 - the number of substitutions per site per time period.
- Evolutionary distance are calculated based on one of DNA evolutionary models.
- Neighbors – pairs of sequences that have the smallest number of substitutions between them.
- On a phylogenetic tree, neighbors are joined by a node (common ancestor).

Distances Matrix Methods

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

Distance method steps

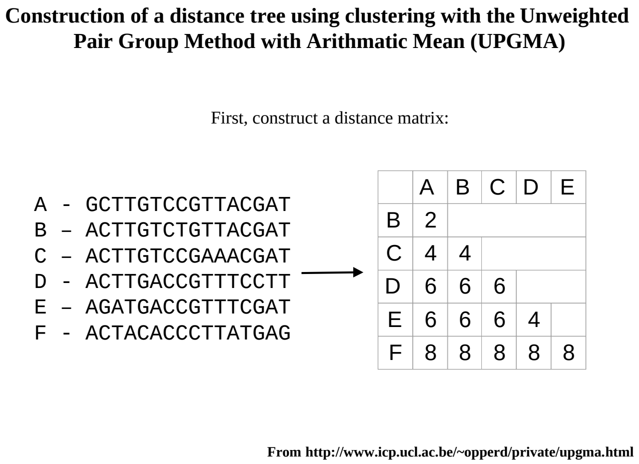
1 Multiple alignments - based on all against all pairwise comparisons.

2 Building distance matrix of all the compared sequences (all pair 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:

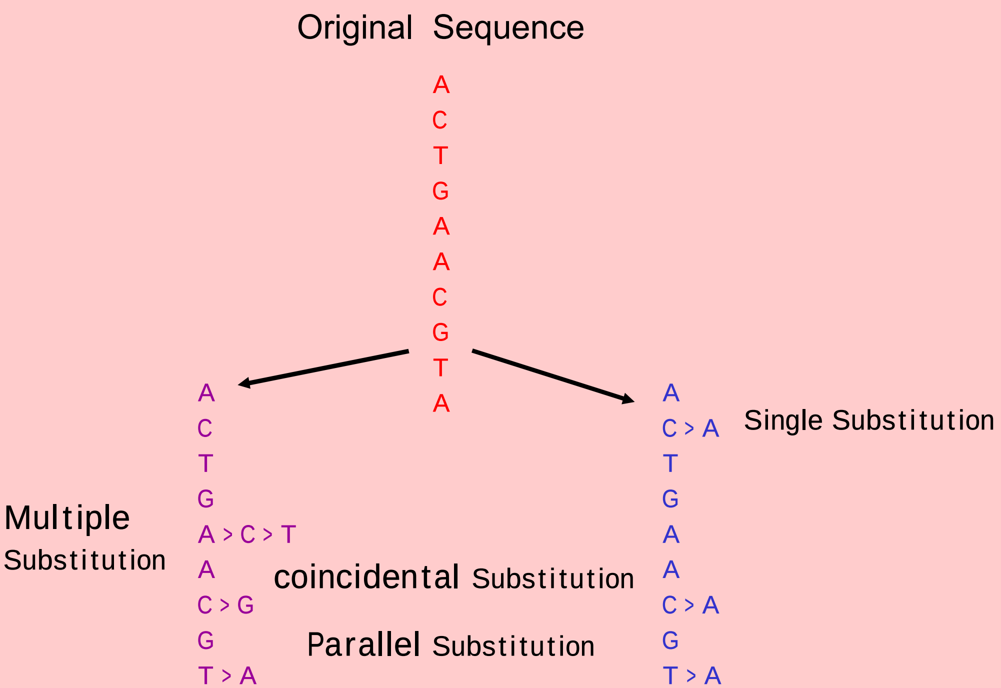
  DNADist, PROTDist, Fitch, Kitch, Neighbor

Mutations as data source for evolutionary analysis

- Mutation - an error in DNA replication or DNA repair.
- Only mutations that occur in germline cells play a rule in evolution. However, in some organisms there is no distinction between germline or somatic mutation.
- Only mutations that were fixed in the population are called substitutions.

Correction of Distances between DNA sequences

- In order to detect changes in DNA sequences we compare them to each other.
- 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.



Original Sequence

A C T G A A C G T A

Multiple Substitution

A C T G A > C > T A C > G G T > A

coincidental Substitution

Parallel Substitution

Single Substitution

A C > A T G A A C > A G T > A

Mutations - Substitutions

Q: What do we measure by sequence alignment?

A: Substitutions in the aligned sequences.

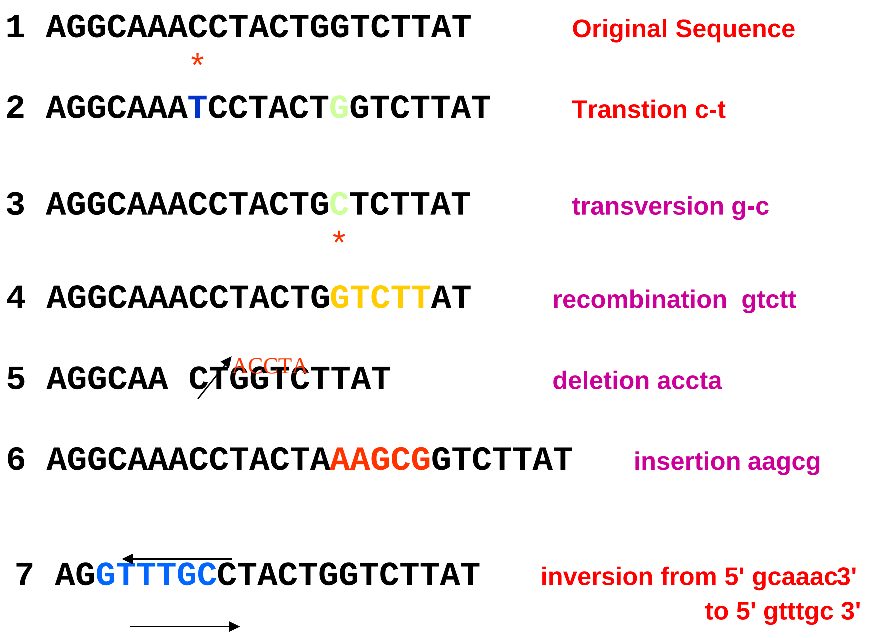
- The rate of substitution in regions that evolve under no constraints are assumed to be equal to the mutation rate.

Mutations - Substitutions

- Point mutation - mutation in a single nucleotide.
- Segmental mutation - mutation in several adjacent nucleotides.
- Substitution mutation - replacement of one nucleotide with another.
- Recombination - exchange of a sequence with another.

Mutations

- Deletion - removal of one or more nucleotides from the DNA.
- Insertion - addition of one or more nucleotides to the DNA.
- Inversion - rotation by 180 of a double-stranded DNA segment comprising 2 or more base-pairs.



1 AGGCAAACCTACTGGTCTTAT Original Sequence

*

2 AGGCAAATCCTACTGGTCTTAT Transtion c-t

3 AGGCAAACCTACTGCTCTTAT transversion g-c

*

4 AGGCAAACCTACTGGTCTTAT recombination gtctt

5 AGGCAA CTGGTCTTAT deletion accta

6 AGGCAAACCTACTAAAGCGGTCTTAT insertion aagcg

7 AGGTTTGCCTACTGGTCTTAT inversion from 5' gcaaac3' to 5' gtttgc 3'

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.*

Let's assume that..

A mutation of a sense codon to another sense codon, occur in equal frequency for all the codons.

If so we can now compute the expected proportion of different types of substitution mutations from the genetic code.

Where do the mutations occur?

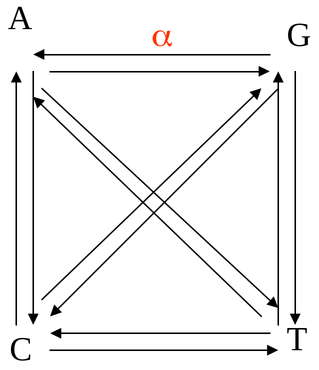
- Synonymous substitutions mostly occur at the *3rd* position of the codon.
- Nonsynonymous substitutions mostly occur at the *2nd* position of the codon.
- Any substitution in the *2nd* position is nonsynonymous

Correction of Distances between DNA sequences

- There are several evolutionary models used to correct 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.

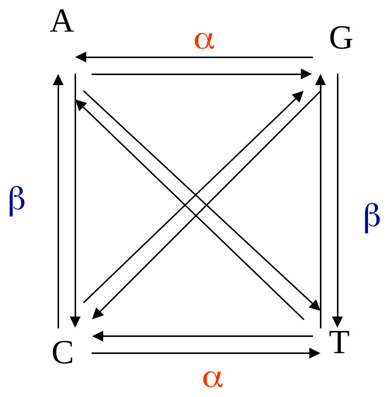


α Is the rate of substitutions In each of the 3 directions For one base.

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.

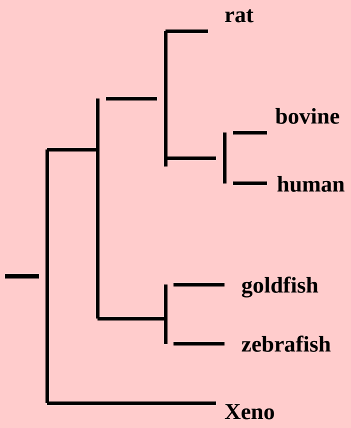
- These evolutionary models improve the distance calculations between the sequences.
- These evolutionary models have less effect in phylogenetic predictions of closely related sequences.
- These evolutionary models have better effect with distant related sequences.

DNA Evolution Models

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

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

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



How to read the tree?

Start at the base and follow The progression of the branch points (nodes)

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 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 The amount of data used for the tree construction is not always "informative". For example, if we compare proteins that are 100% similar in a site, we do not have information to infer to phylogenetic relationships between these proteins.

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.