Phylogenetic Trees

Phylogenetic tree construction methods

- **A phylogenetic tree is characterised by its topology (form) and its length (sum of its branch lengths) ;**

- **Each node of a tree is an estimation of the ancestor of the elements included in this node;**

Phylogenetic tree construction methods

- **Parsimony**
- **Distance Methods**
- **Maximum likelihood Methods**

Key features of phylogenetic trees

The numbers of possible rooted (N_R) and unrooted (N_U) trees for n sequences are given by:

$$N_R = (2n-3)!/2^{n-2}(n-2)!$$

$$N_U = (2n-5)!/2^{n-3}(n-3)!$$

n	N_R	N_U
2	1	1
3	3	1
4	15	3
5	105	15
10	34459425	2027025

- **Note that only one of all possible trees can represent the true tree that represents phylogenetic relationships among the sequences.**

Phylogenetic tree construction methods

Methods directly based on sequences :

- **Maximum Parsimony**

explains the data, with as few evolutionary changes as possible.

- **Maximum likelihood**

probability of the genetic data given the tree.

Parsimony

The concept of parsimony is at the heart of all character-based methods of phylogenetic reconstruction.

The 2 fundamental ideas of biological parsimony are:

1- mutations are exceedingly rare events (?) ;

2- the more unlikely events a model invokes, the less likely

As a result, the relationship that requires the fewest number of mutations to explain the current state of the sequences being considered, is the relationship that is most likely to be correct.

Parsimony

Informative and Uninformative Sites:

Multiple sequence alignment, for a parsimony approach, contains positions that fall into two categories in terms of their information content : those that have information (are informative) and those that do not (are uninformative).

Example:

seq	1	2	3	4	5	6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

In general, for a position to be informative regardless of how many sequences are aligned, it has to have at least 2 different nucleotides, and each of these nucleotides has to be present at least twice.

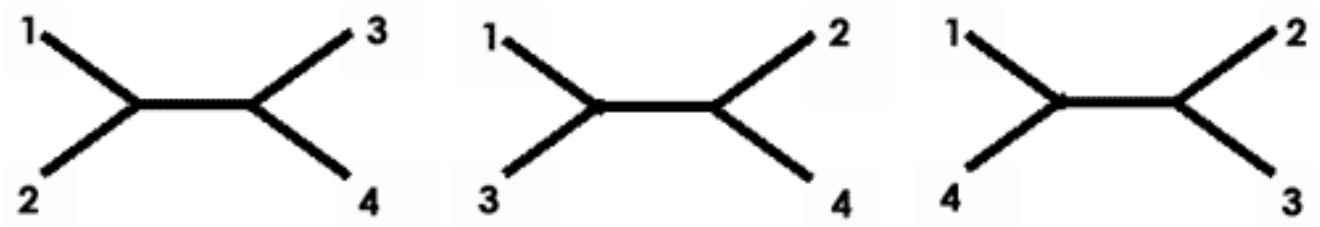
Position 1 is said invariant and therefore uninformative, because all trees invoke the same number of mutations (0);

Position 2 is uninformative because 1 mutation occurs in all three possible trees;

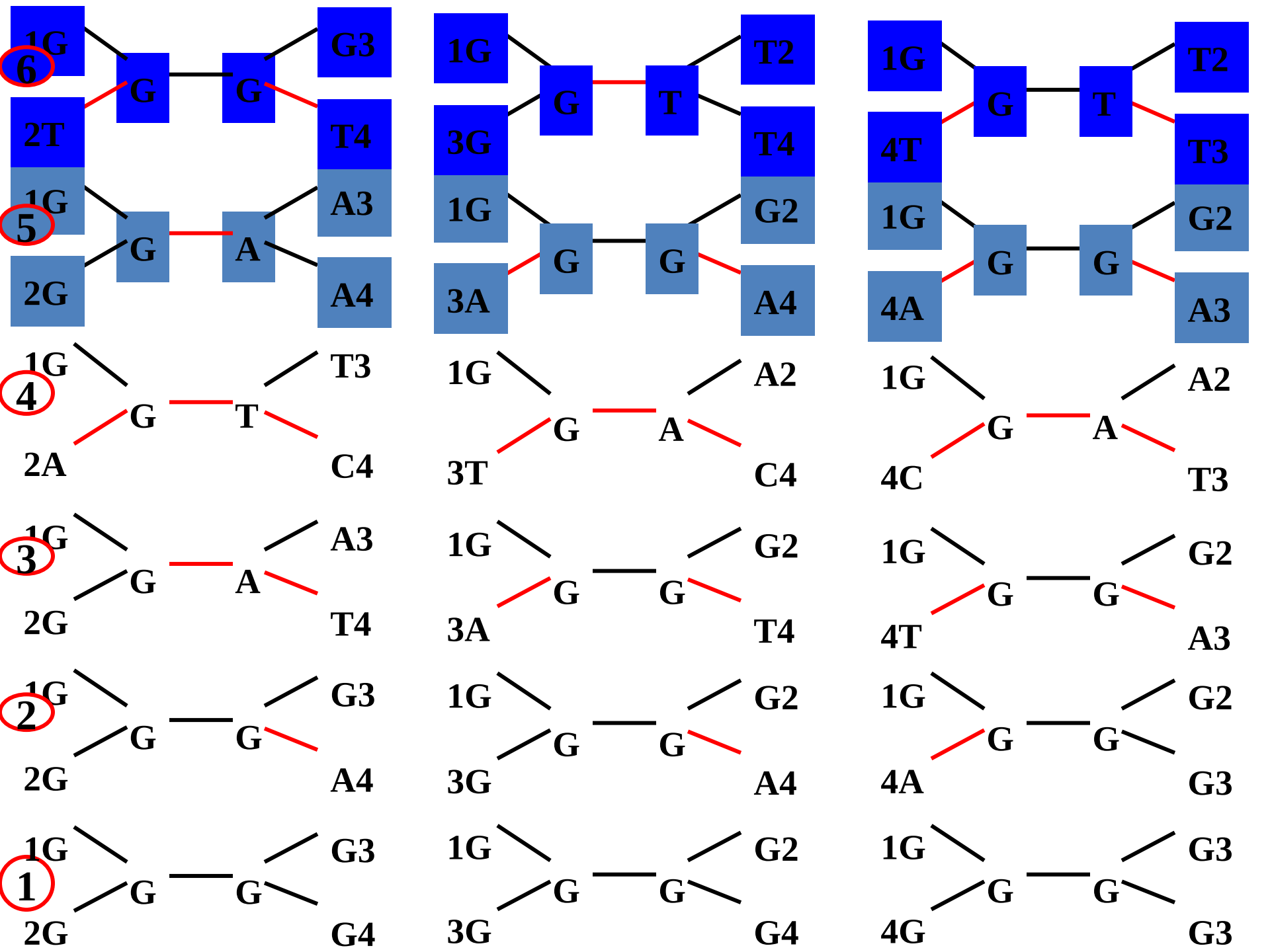
Position 3 idem, because 2 mutations occur; Position 4 requires 3 mutations in all possible trees.

Positions 5 and 6 are informative, because one of the trees invokes only one mutation and the other 2 alternative trees both require 2 mutations.

Parsimony



Only three different trees represent all possible relationships of four taxa

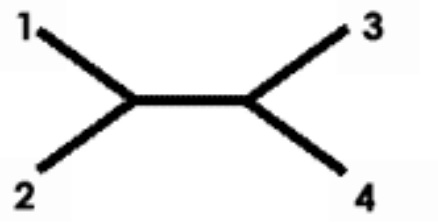

6
1G
2T
1G
2G
G
G
G
G3
T4
A3
A4
5
1G
3G
1G
3A
G
T
G
G
T2
T4
G2
A4
1G
4T
1G
4A
G
T
G
G
T2
T3
G2
A3
4
1G
2A
G
T
T3
C4
1G
3T
G
A
A2
C4
1G
4C
G
A
A2
T3
3
1G
2G
G
A
A3
T4
1G
3A
G
G
G2
T4
1G
4T
G
G
G2
A3
2
1G
2G
G
G
G3
A4
1G
3G
G
G
G2
A4
1G
4A
G
G
G2
G3
1
1G
2G
G
G
G3
G4
1G
3G
G
G
G2
G4
1G
4G
G
G
G3
G3

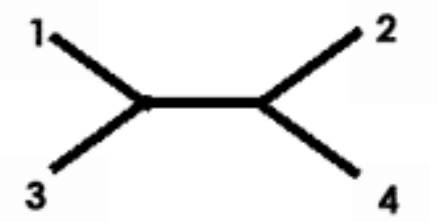
Parsimony

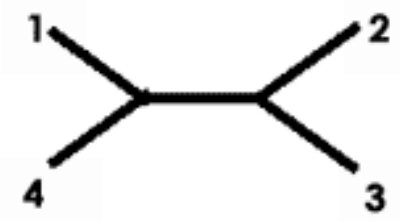
seq	1	2	3	4	5	6
1	A	G	T	G	C	A
2	C	G	T	G	C	G
3	T	A	T	C	C	A
4	G	A	T	C	C	G

Parsimony

seq	1	2	3	4	5	6
1	A	G	T	G	C	A
2	C	G	T	G	C	G
3	T	A	T	C	C	A
4	G	A	T	C	C	G







Maximum Parsimony (Fitch, 1977)

Parsimony criterion consists of determining the minimum number of changes (substitutions) required to transform a sequence to its nearest neighbor.

The maximum parsimony algorithm searches for the minimum number of genetic events (nucleotide substitutions or amino-acid changes) to infer the most parsimonious tree from a set of sequences.

The best tree is the one which needs the fewest changes.

Problems :

1. within practical computational limits, this often leads to the generation of tens or more "equally most parsimonious trees" which makes it difficult to justify the choice of a particular tree ;

2. long computation time is needed to construct a tree.

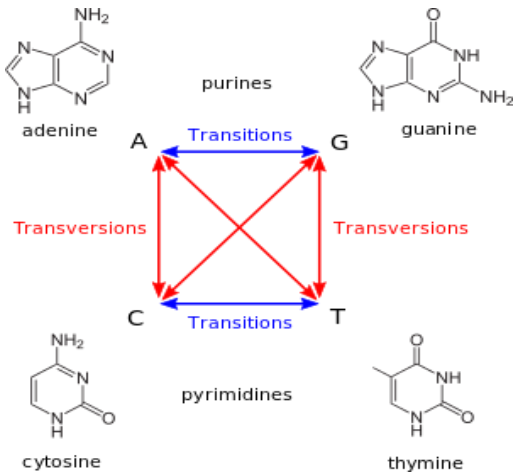
Maximum Parsimony (Fitch, 1977),...

The assumption, possibly erroneous, is that **evolution follows the shortest possible route** and that the correct phylogenetic tree is therefore the one that **requires the minimum number of nucleotide changes** to produce the observed differences between the sequences.

Maximum likelihood

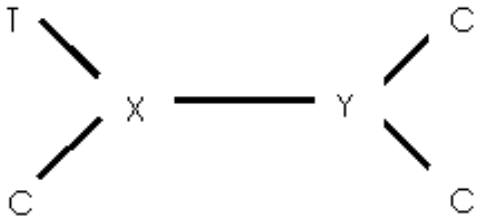
This approach is a purely statistically based method.

Probabilities are considered for every individual nucleotide substitution in a set of sequence alignment.



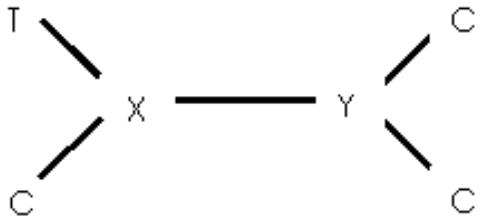
Still, objective criteria can be applied to calculating the probability for every site and for every possible tree that describes the relationships of the sequences in a multiple alignment.

Maximum likelihood



1. With ML, one must first estimate the probability of each kind of change in character state
    1. the probability of no change in a base, a transition, or a transversion
2. The likelihood L_n for the bases at each position *n* and for each tree is then calculated from these probabilities.
3. The logarithm of these values of *L* are then added to get the log likelihood (ln *L*) of each tree
4. The tree with the highest (least-negative) value of ln *L* is taken to be the most likely.

Maximum likelihood



The probability of no change in a base=0.7,

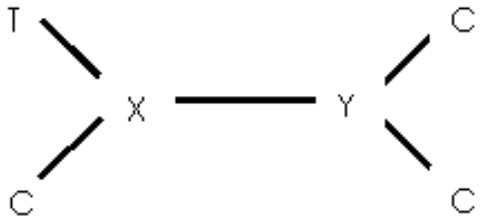
The probability of a transition=0.2,

The probability of a transversion =0.1

- *For taxon 1 the base at the first position is T, and for the other three taxa the base is C*
- *X and Y represent the bases at the first position for the two ancestral taxa*

1. *Both X and Y inherited C from their common ancestor*
2. *X became T and Y be-came C*
3. *It is also possible, but less likely, that X was A or G*

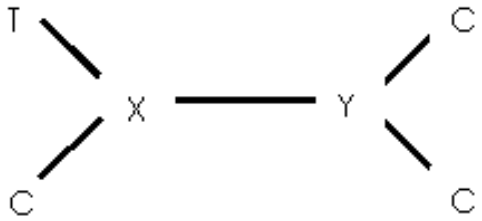
Maximum likelihood



If X and Y were both C

- *there were four branches with no change and one with a transition at that site.*
- *so the probability of each of the four bases being what they are is 0.7^4 x 0.2 = 0.04802.*
- *What if Y='C' and X='T'?*
- *What if Y='C' and X='A'?*
- *What if Y='C' and X='G'?*

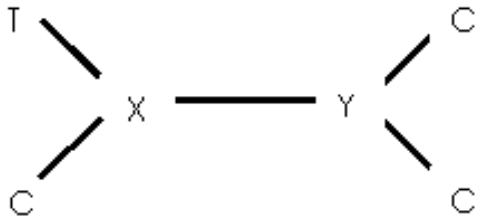
Maximum likelihood



If X and Y were both C

- *there were four branches with no change and one with a transition at that site.*
- *so the probability of each of the four bases being what they are is 0.7^4 x 0.2 = 0.04802.*
- *What if Y='C' and X='T' =>0.01372*
- *What if Y='C' and X='A' => 0.00049*
- *What if Y='C' and X='G' => 0.00049*

Maximum likelihood



	X = C	X = T	X = A	X = G
Y = C	0.04802	0.01372	0.00049	0.00049
Y = T	0.00112	0.00392	0.00004	0.00004
Y = A	0.00014	0.00014	0.00007	0.00002
Y = G	0.00014	0.00014	0.00002	0.00007

The sum of all 16 probabilities gives the likelihood $L_1 = 0.06858$ *and* $\ln L_1 = -2.680$