How to Build a Phylogenetic Tree

Phylogenetics tree is a structure in which species are arranged on branches that link them according to their relationship and/or evolutionary descent. A typical rooted tree with scaled branches is illustrated in fig1 [1]

In this short article, a brief review of different *methods of tree building* is given and their validity, reliability and efficiency are compared. The calculation of *genetic distance* and methods of *trees rooting* are also discussed. For other issues such as tree evaluation and searching algorithms, see reference [2].

Genetic distance:

Genetic distance is the number of mutation/evolutionary events between species since their divergence. The simplest way to calculate it is to count the number of difference between two sequences. However, the method, which is often refer to as *hamming method*, does not reflect the evolutionary history of the sequence, because not all the mutation events are recorded in the current sequences. Figure 2 shows an example in which only 3 of 12 mutation events are detected between homologous sequences [3].

Jukes-Cantor Model, which was proposed in 1969, was the first attempt to cope with this problem. They gave a corrective formula $K = -\frac{3}{4}\ln(1-\frac{4}{3}p)$ (fig 3), where K is the genetic distance (the mean number of substitutions) and p is the percent difference. This relation reveals that the correction increase with the genetic distance and when the percent of difference saturates (75%), the true genetic distance is no longer extractable [4].

More recent models take into consideration of the different transition/transversion rate, different base frequencies and non-uniform substation rates between sites. All these models give similar results at low divergence. Nevertheless, at high percent of difference, choosing proper substitution model will be very crucial.

Tree-Building Methods

The most popular and frequently used methods of tree building can be classified into two major categories: phenetic methods based on *distances* and cladistic methods based on *characters*. The former measures the pair-wise distance/dissimilarity between two genes, the actual size of which depends on different definitions, and constructs the tree totally from the resultant distance matrix. The latter evaluate all possible trees and seek for the one that optimizes the evolution.

Distance-Based Methods:

The most popular distance-based methods are the unweighted pair group method with arithmetic mean (UPGMA), neighbor joining (NJ) and those that optimize the additivity of a distance tree (FM and ME) [2].

- UPGMA Method

This method follows a clustering procedure:

(1) Assume that initially each species is a cluster on its own.

(2) Join closest 2 clusters and recalculate distance of the joint pair by taking the average.

(3) Repeat this process until all species are connected in a single cluster.

Strictly speaking, this algorithm is phenetic, which does not aim to reflect evolutionary descent. It assigns equal weight on the distance and assumes a randomized molecular clock. WPGMA is a similar algorithm but assigns different weight on the distances.

UPGMS method is simple, fast and has been extensively used in literature. However, it behaves poorly at most cases where the above presumptions are not met.

- Neighbor Joining Method (NJ)

This algorithm does not make the assumption of molecular clock and adjust for the rate variation among branches. It begins with an unresolved star-like tree (fig 4(a)). Each pair is evaluated for being joined and the sum of all branches length is calculated of the resultant tree. The pair that yields the smallest sum is considered the closest neighbors and is thus joined .A new branch is inserted between them and the rest of the tree (fig4 (b)) and the branch length is recalculated. This process is repeated until only one terminal is present.

NJ method is comparatively rapid and generally gives better results than UPGMA method. But it produces only one tree and neglects other possible trees, which might be as good as NJ trees, if not significantly better. Moreover, since errors in distance estimates are exponentially larger for longer distances, under some condition, this method will yield a biased tree [5].

- Weighted Neighbor-Joining (Weighbor)

This is a new method proposed recently [5]. The Weighbor criterion consists of two terms; an additivity term (of external branches) and a positivity term (of internal branches), that quantifies the implications of joining the pair. Weighbor gives less weight to the longer distances in the distance matrix and the resulting trees are less sensitive to specific biases than NJ and relatively immune to the "long branches attraction/distraction" drawbacks observed with other methods.

- Fitch-Margoliash (FM) and Minimum Evolution (ME) Methods

Fitch and Margoliash proposed in 1967 a criteria (FM Method) for fitting trees to distance matrices [2]. This method seeks the least squared fit of all observed pair-wise distances to the expected distance of a tree. The ME method also seeks the tree with the minimum sum of branch lengths. But instead of using all the pair-wise distances as FM, it fixed the internal nodes by using the distance to external nodes and then optimizes the internal branch lengths

FM and ME methods perform best in the group of distance-based methods, but they work much more slowly than NJ, which generally yield a very close tree to these methods.

Character-Based Methods

Distance-based methods are more rapid and less computationally intensive than character-based methods, but the actual characters are discarded once the distance matrix is derived. On the other hand, character-based methods make use of all known evolutionary information, i.e. the individual substitutions among the sequences, to determine the most likely ancestral relationships.

- Maximum parsimony (MP)

The criterion of MP method is that the simplest explanation of the data is preferred, because it requires the fewest conjectures. By this criterion, the MP tree is the one with fewest substitutions/evolutionary changes for all sequences to derive from a common ancestor.

For each site in the alignment, all possible trees are evaluated and are given a score based on the number of evolutionary changes needed to produce the observed sequence changes. The best tree is thus the one that minimized the overall number of mutation at *all* site.

MP works faster than ML and the weighted parsimony schemes can deal with most of the different models used by ML. However, this method yields little information about the branch lengths and suffers badly from long-branch attraction, that is the long branches have become artificially connected because of accumulation of inhomogous similarities, even if they are not at all phylogenetically related.

MP yields more than one tree with the same score.

- Maximum Likelihood (ML):

Like MP methods, ML method also uses each position in an alignment and evaluates all possible trees. It calculates the likelihood for each tree and seeks the one with the maximum likelihood.

For a given tree, at each site, the likelihood is determined by evaluating the probability that a certain evolutionary model has generated the observed data. The likelihood's for each site are then multiplied to provide likelihood for each tree.

ML method is the slowest and most computationally intensive method, though it seems to give the best result and the most informative tree.

Rooting trees:

Root is the common ancestor of the species under study. Most phylogenetic methods do not locate the root of a tree and the unrooted trees only reflect the relationship among species but not the evolutionary path. Fig5 (a) shows an unrooted tree of species A, B, C and D.

If one assumes that argument of molecular clock evolution is valid, which asserts that informational macromolecules evolve at rates that are constant through time and for different lineages, then the root is simply the mid-point of the longest span across the tree. For example, for the unrooted tree in fig 5(a), the root is just the mid-point of the longest span A-B (fig 5(b)).

Whether the molecular clock assumption is valid is still a controversial [6]. However, evidences from many research works have shown that such a model is too simplified and

sometimes behave in an erratic manner. A more commonly used method is to evaluate the rooting by an outgroup, i.e. a distantly related specie. For instance, mouse can be used an outgroup to root a tree of primates [3] (see fig5(c)). However, this method gives rise to a dilemma. If the outgroup species is very closely related to the ingroup, it may just be a mistakenly excluded member. (Of course, in our example, the mouse is obviously not a member of primate, but it is not always apparent in other cases). On the other hand, a very divergent outgroup may have accumulated so many inhomologous similarities that it seems to be artificially connected with the ingroup (see the long-branch attraction mentioned in MP method).

Reference:

[1] http://allserv.rug.ac.be/~avierstr/principles/phylogeny.html, *Phylogenetics* (1999)

[2] Andreas D. Baxevanis & B.F. Francis Ouellette, *Bioinformatics: a practical guide to the analysis of genes and proteins* (2001)

[3] P.Higgs & Manchester, *Introduction to Phylogenetics Methods (ITP series on-line seminars)* (2001)

[4] http://hiv-web.lanl.gov/, *How to make a phylogenetic tree* (1999)

[5] Bruno WJ, Socci ND, Halpern AL, *Mol Biol Evol,* **17(1),** 189-97 (2000)

[6] Wen-Hsiung Li & Dan Graur, *Fundamentals of Molecular Evolution* (1991)

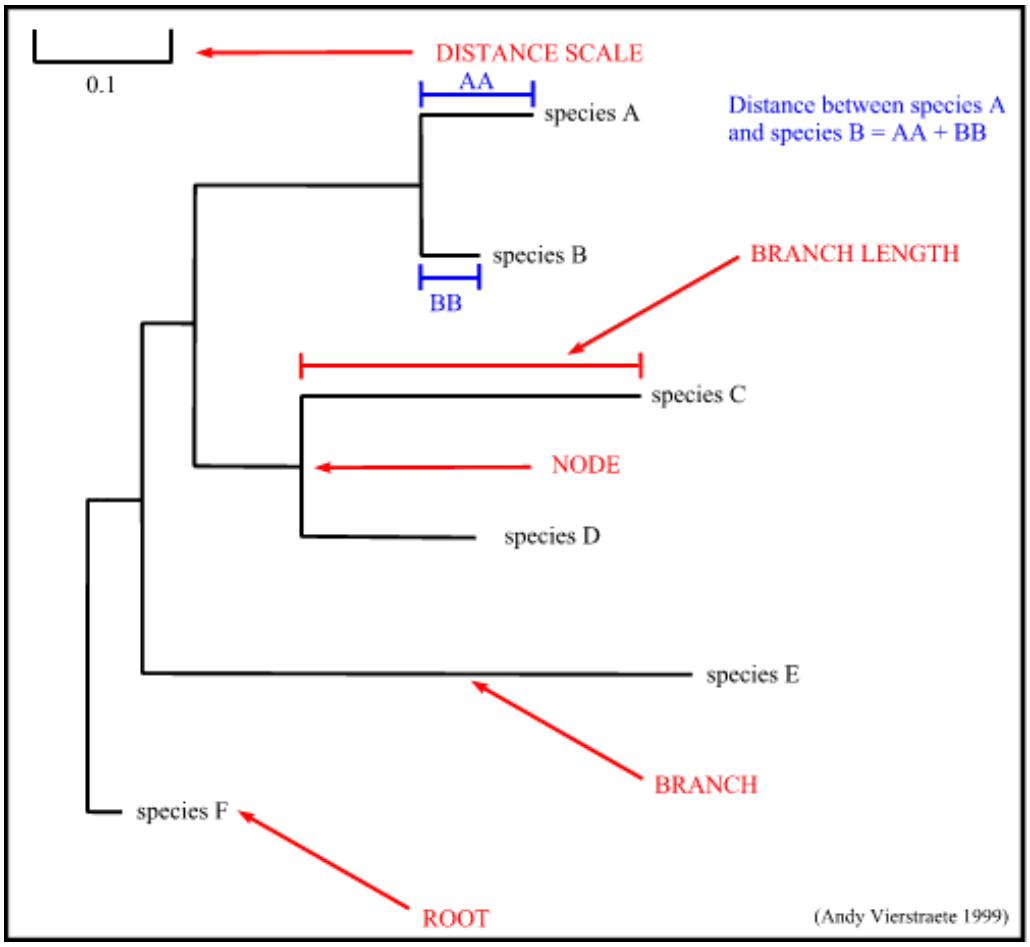

Fig 1. A typical rooted tree with scaled branches [1]. For definition of terminology, visit the website: http://allserv.rug.ac.be/~avierstr/principles/phylogeny.html)

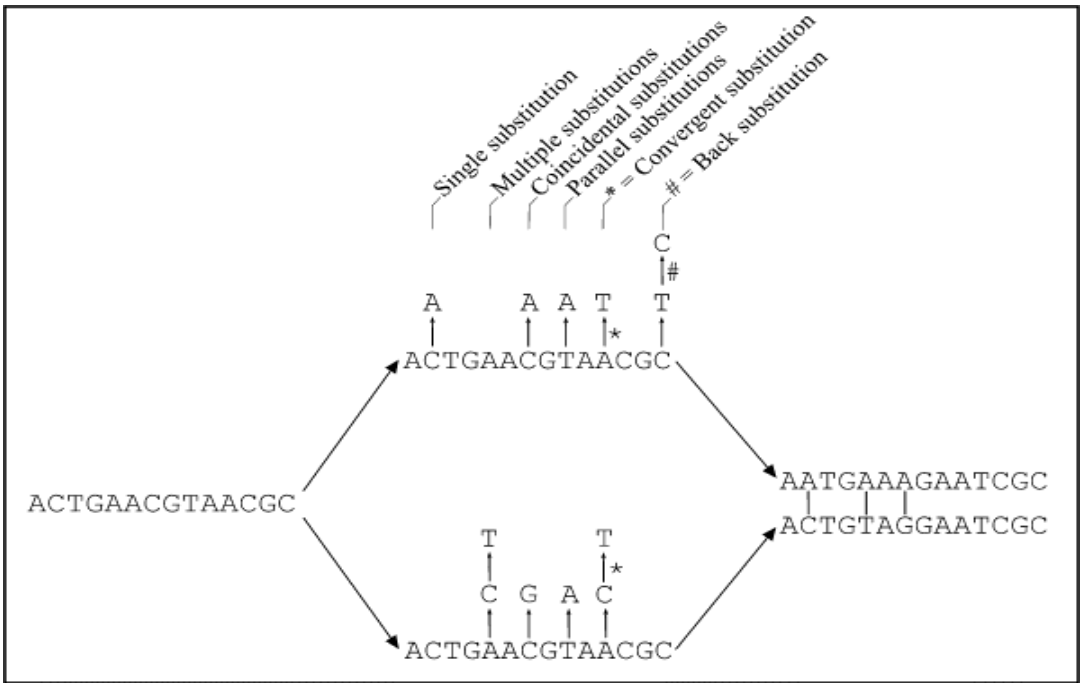

Fig 2 Two homologous DNA sequences (on the right) which descended form an ancestral sequence (on the left). Although 12 mutations have accumulated since their divergence from each other, only 3 of them are recorded by the current sequences [3].

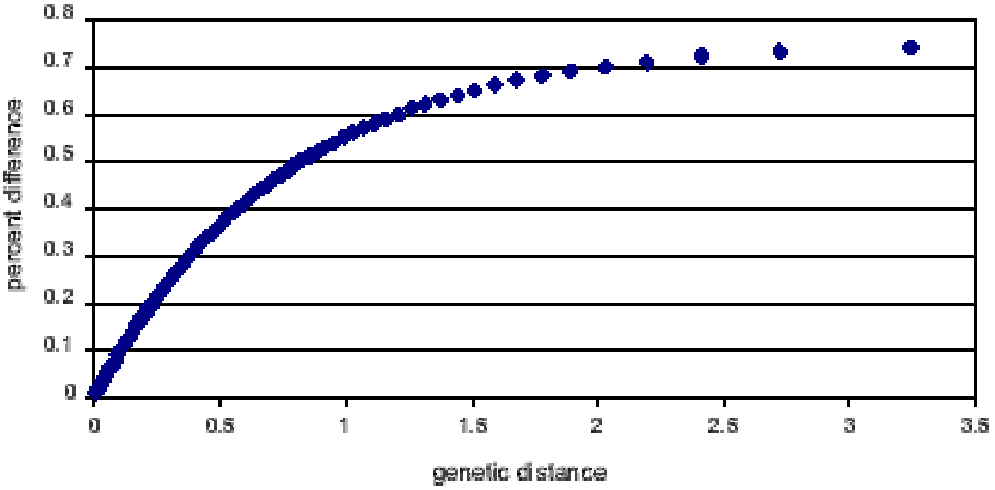

Fig 3. Jukes-Cantor Model of the relation between the percent difference p and the genetic distance K: $K = -\frac{3}{4}\ln(1-\frac{4}{3}p)$ [4]

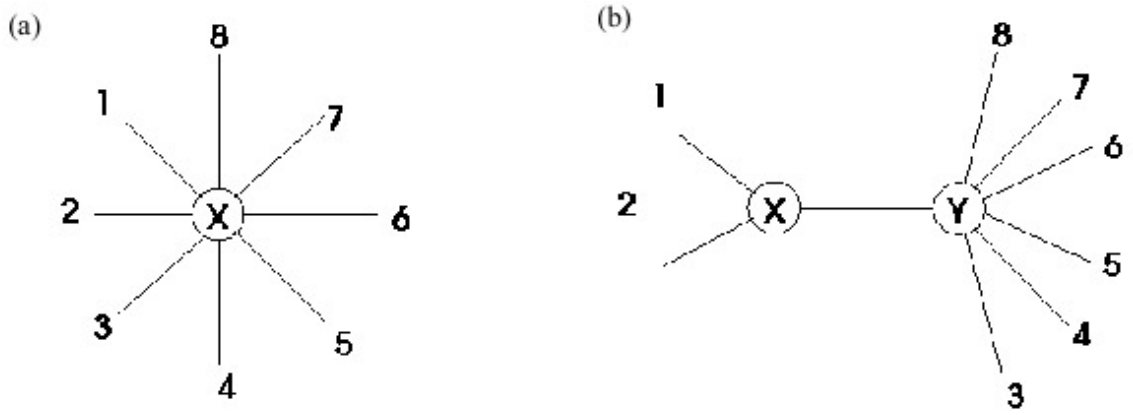

Fig 4 NJ method: (a) an unresolved star-like tree (b) the closest terminals(for definition, see text) 1 and 2 are joined and a branch is inserted between them and the remainder of the tree and the value of the branch is taken to be the mean of the two original values. This process is continued until only one terminal is left.

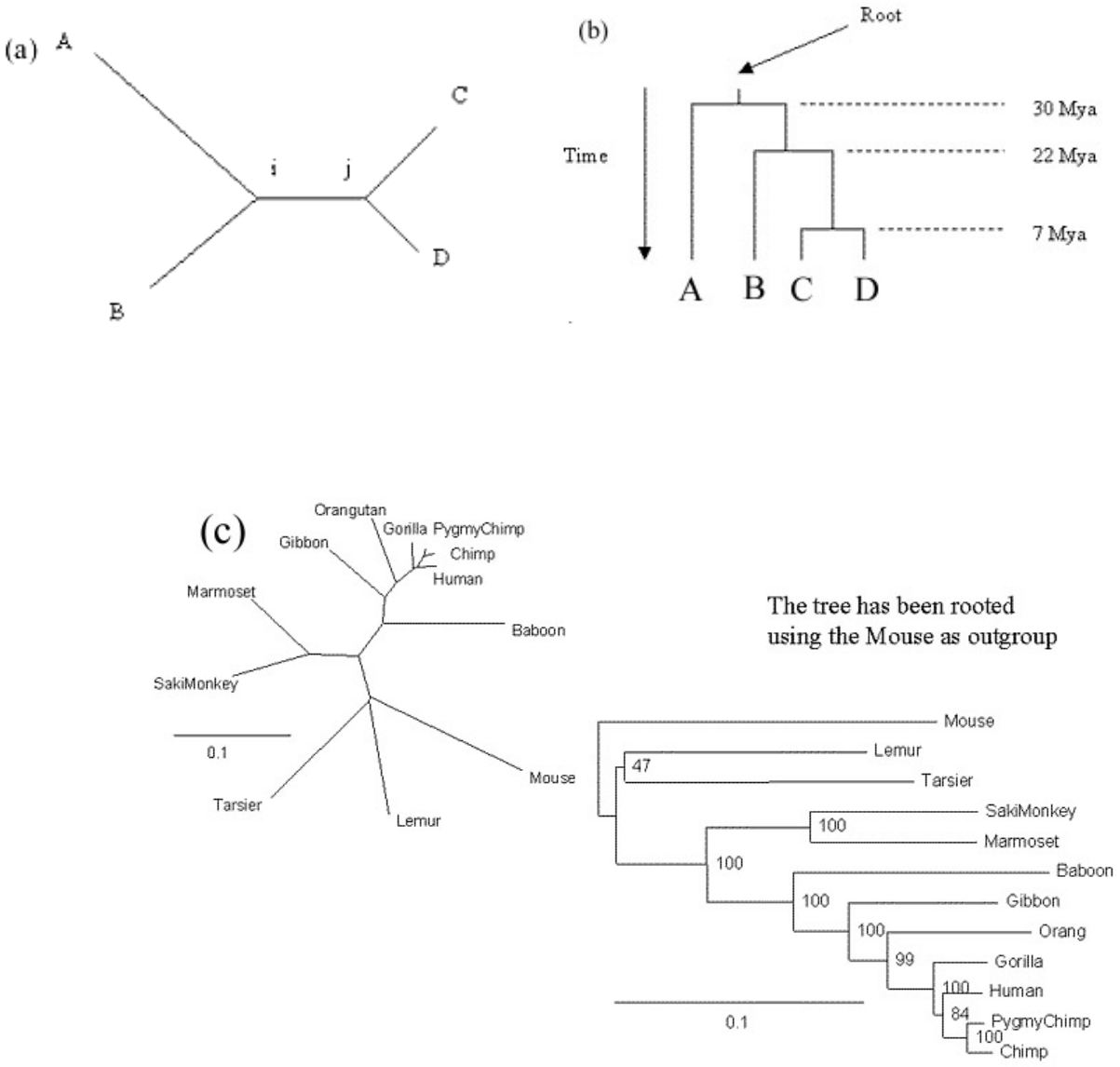

Fig5 (a) an unrooted tree (b) is rooted under the assumption of molecular clock
(c) Rooting a tree by using mouse as outgroup [3].