**Phylogenetics**

Key Notes

1. Phlogenetics = similarities and differences among species can be used to infer evolutionary relationships (phylogenetics). This is because, if two species are very similar, they are likely to have shared a recent common ancestor. Many different types of character can be used in phylogenetic analysis, but nucleic acid and protein sequences are the most popular because they are common to all life forms (allowing both closely and distantly related taxa to be studied) and they are compared objectively. However, caution must be exercised when inferring phylogenetic from sequences because the rate of mutation may not be constant and sequences may be subject to differential selection.
2. Graphs and trees = a graph is a diagram showing relationships between two particular entities, for example evolutionary relationships between species. Evolutionary relationships are enerally represented by a special type of a graph called a tree, which has *n* nodes and *n-1* links.

Phylogenetics

Living organisms are classified into groups based on observed similarities and differences. A general principle of classification systems is that the more closely related species *a* is to be species *b,*the more likely they shared a recent common ancestor. In this way, similarities and differences between organisms can be used to infer **phylogenies (evolutionary relationships).** The branch of science that deals with resolving evolutionary relationships among organisms is **phylogenetics.**

**Phylogenetics** can be studied in three ways:

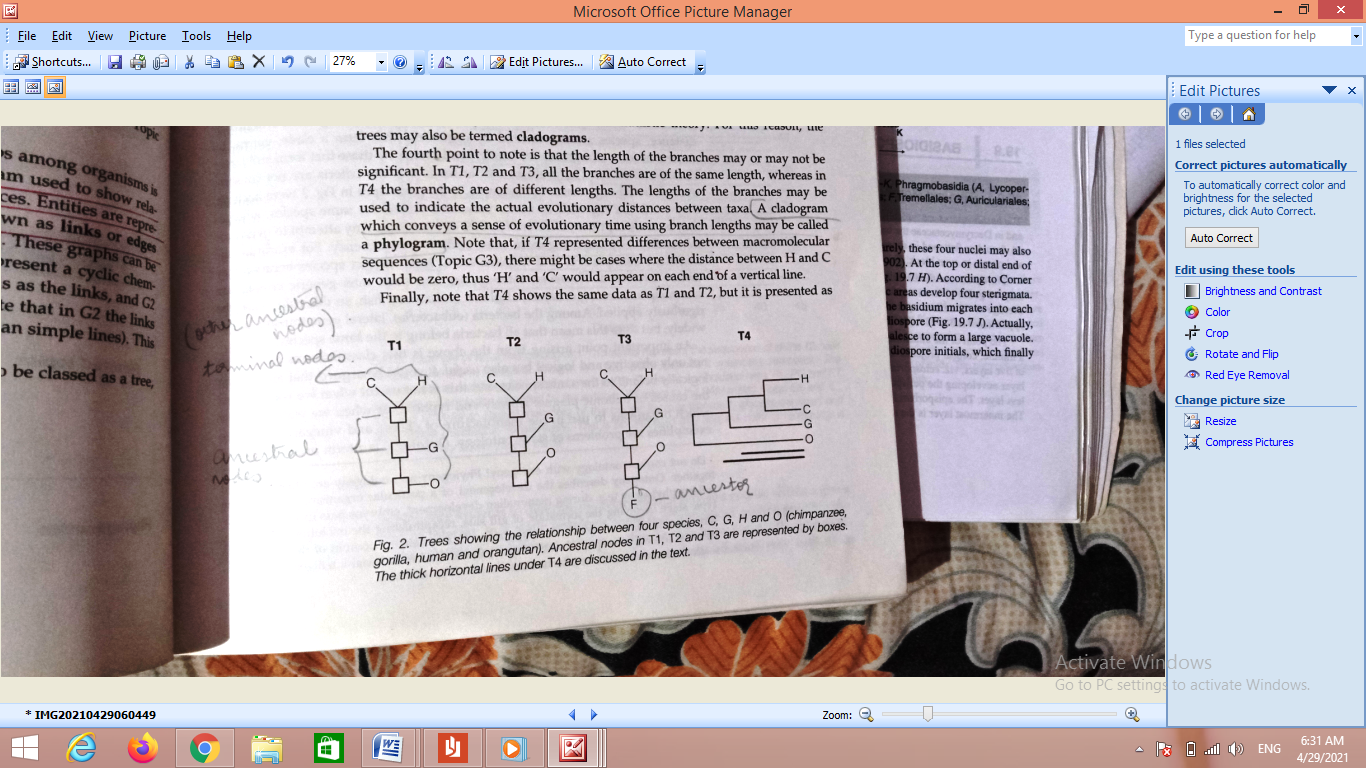
1. **Phenetics**
2. **Cladistics**
3. **Evolutionary systematic**
4. **Phenetics =** In **Phenetics,** species are grouped with others they resemble phenotypically and all characters are taken into account.
5. **Cladistics=** In **claditics,** species are grouped only with those that share *derived* characters, that is, characters that were not present in their distant ancestors.
6. **Evolutionary systematic =**  incorporates both phenetic and cladistic principles.

**Cladistic** is accepted as the best method available for phylogenetic analysis because it accepts and employs current evolutionary theory, that speciations occurs by bifurcation (cladogenesis)

Many different criteria can be used for phylogenetic analysis, including morphological characteristics, biochemical properties and, most recently, the analysis of macromolecular sequences (nucleic acid and protein sequences). Macromolecular sequences are particularly useful for comparison because they provide large unbiased data set, which extends across all known organisms, allowing the comparison of both closely related and distantly related taxa. Most importantly, however, the relatedness between sequences can be quantified objectively using sequence alignment algorithms. This is where bioinformatics plays an important role in phylogenetics.

**Phlyolgenetic Trees and clasograms**

**Phylogentic trees** (also called **dendograms)** are used to show evolutionary relationships. The nodes represent different organisms and links are used to show lines of descent. As an example, we consider the phylogenetic relationship between the entities C (chimpanzee), G (Gorilla), H (Human) and O (Organutan).



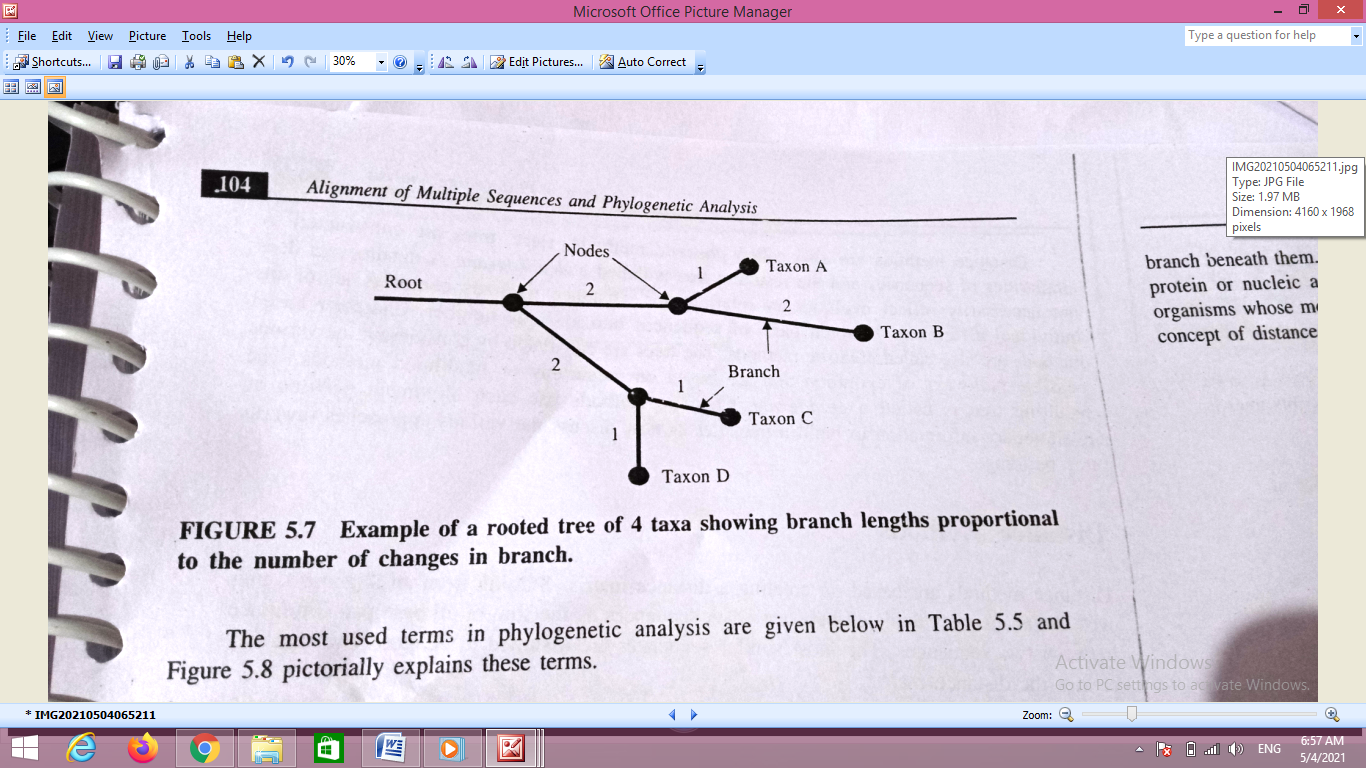
The first point to note about these trees is that there are two types of node. The **ancestral nodes (represented by boxes)** give rise to branches. These may link to other ancestral nodes, or they may link to **terminal nodes (shown as letters), which may also known as leaves or tips. Leaves represent known species and mark the end of the evolutionary pathway.**

The second point to note is that **T1 and T2 are unrooted trees, whereas T3 is a rooted tree. T1 and T2 are identical** except that **T2 except that T2** is drawn in a conventional style with angled branches to look more like a real tree. **These are describes as unrooted trees because neither of them shows the position of the last common ancestor of all the species. In T3,** the position of this ancestor is indicated by the node F.

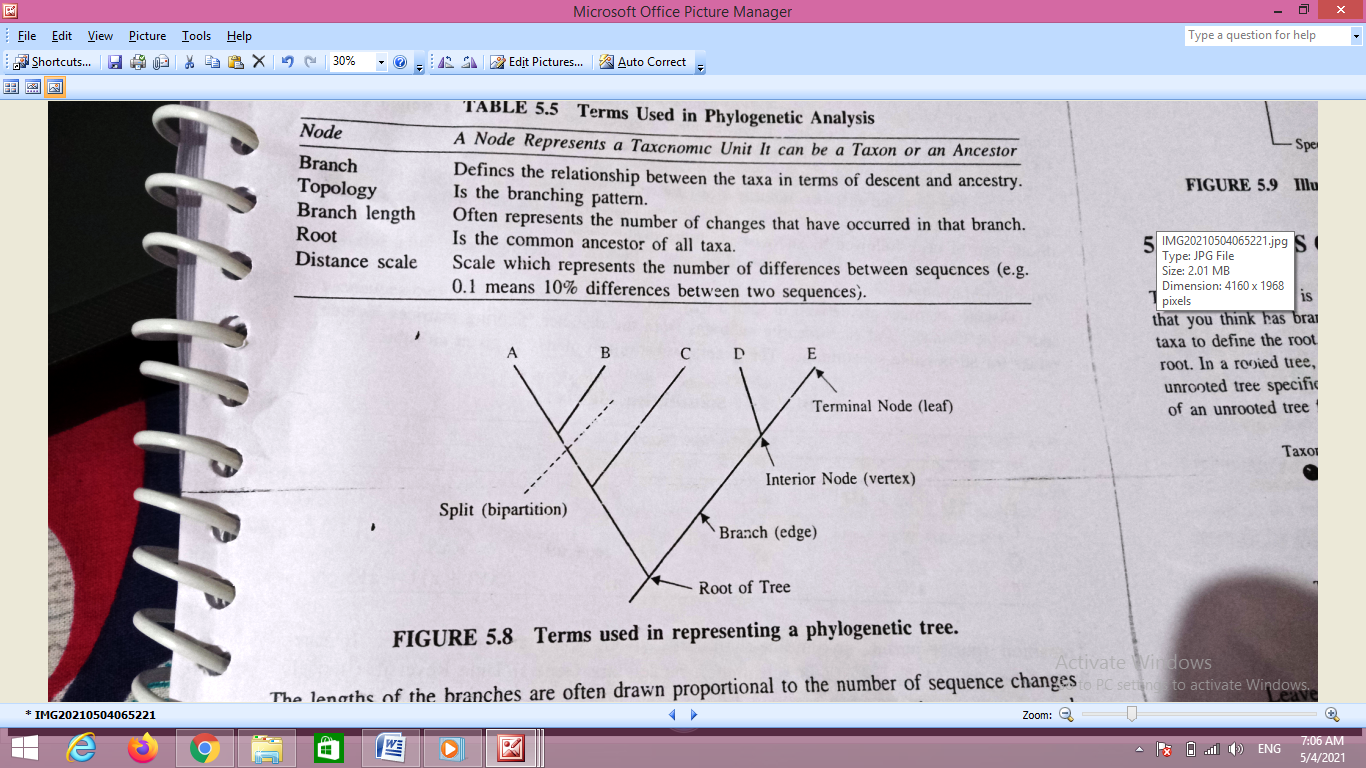
The third point to note is that each tree is binary, there , no ancestral nodes have more than two branches. Thus, the evolution of species is represented as a series of bifurcation, which fits in with cladistic theory. For this reason, the trees may also be termed **cladograms.**

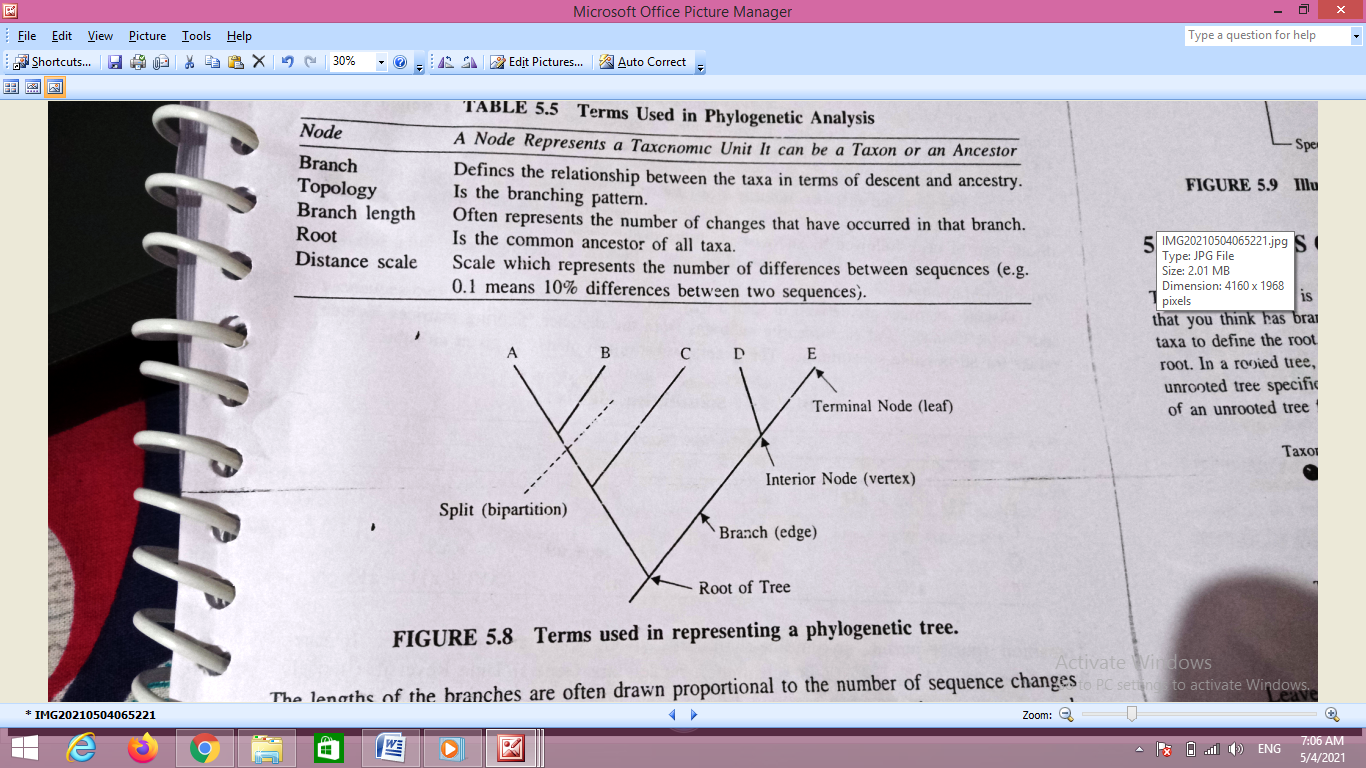
**The** fourth point to note that the length of the branches may or may not be significant. In T1, T2 and T3, all the branches are of the same length, whereas in T4 the branches are of different lengths. **The length of branches may be used o indicate the actual evolutionary distances between taxa. A cladogram which convey a sense of evolutionary time using branch lengths may be called a phylogram. Note that if T4 represented differences between macromolecular sequences, there might be cases where the distance between H and C would be zero, thus H and C would appear on each end of vertical line.**

**Evolutionary relationships can be represented using Phylogenetic trees. A tree is a 2-D graph showing evolutionary relationship among organisms. The separate sources of sequences as taxa (singular taxon), defined as phylogenetically distinct units on the tree. The tree is composed of nodes (a point where branched bifurcate) representing the taxa and branches representing the relationships among the taxa. An example of a rooted tree of 4 taxa is shown in Figure 5.7**

****

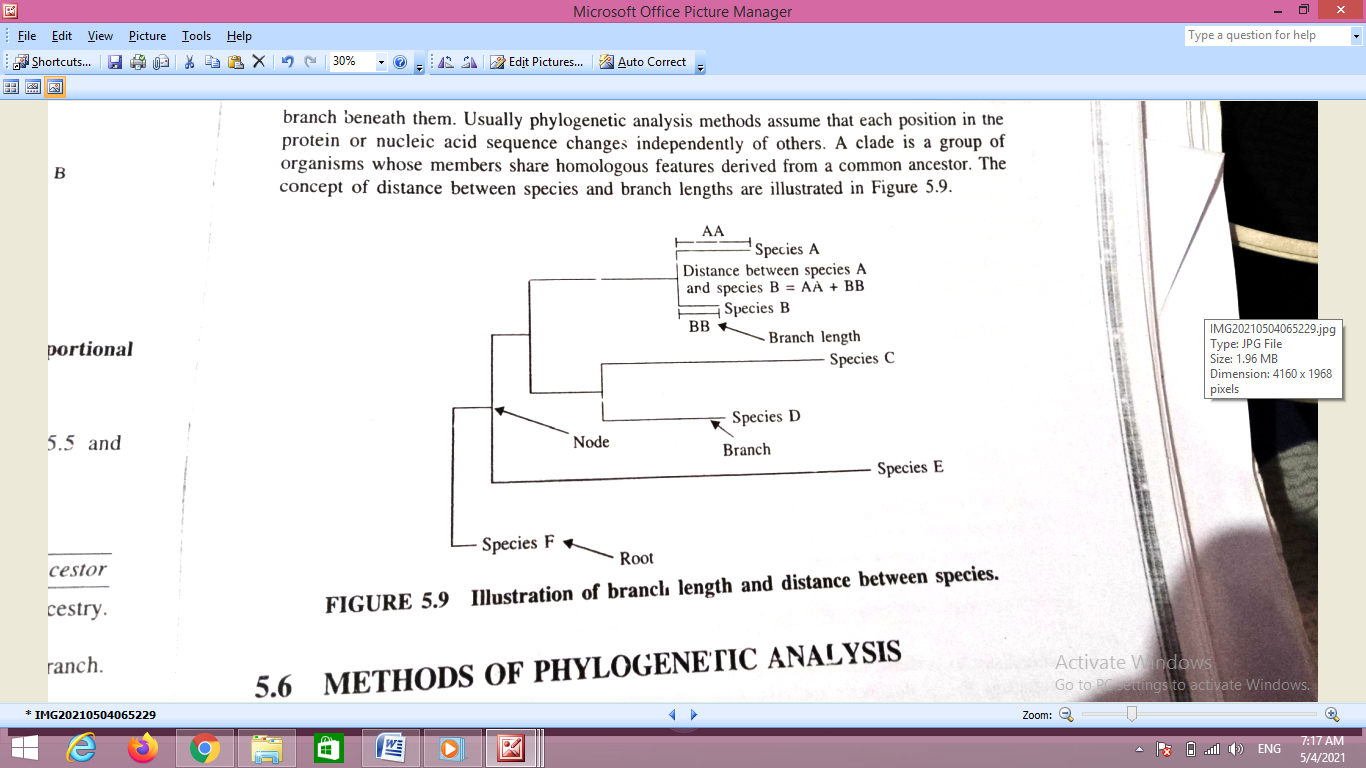
**The most used terms in phylogenetic analysis are given below in Table 5.5 and Figure 5.8 pictorially explains these terms.**

****

****

**The lengths of the branches are often drawn proportional to the number of sequences changes in the branch and hence can represent the divergence.**

**Two sequences that are very much alike will be located as neighbouring outside branches and would be joined to a common branch beneath them. Usually phylogenetic analysis methods assume that each position in the protein or nucleic acid sequence changes independently of others. A clade is a group of organisms whose members share homologous features derived from a common ancestor. The concept of distance between species and branch lengths are illustrated in figure 5.9**

****

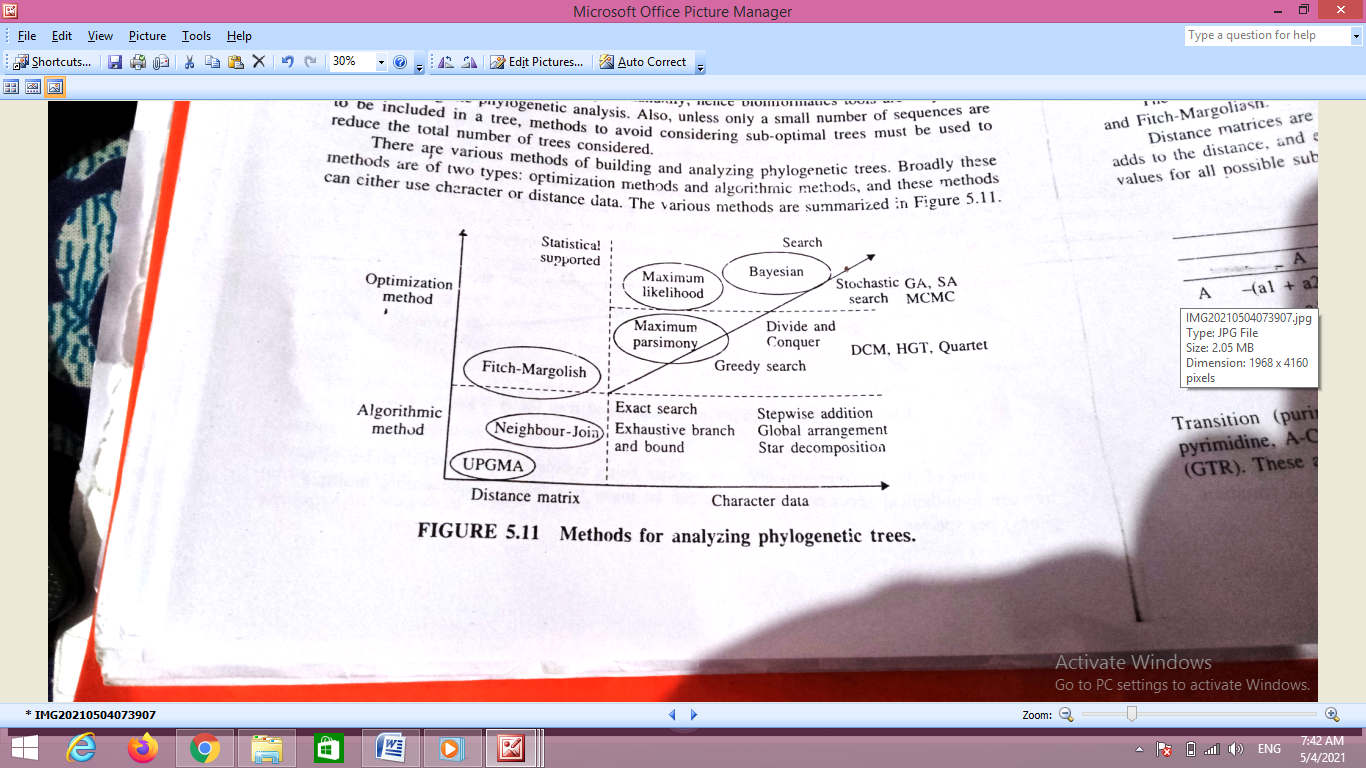
**Methods for analyzing phylogenetic trees.**

**There are various methods of building and analyzing phylogenetic trees.**

**Broadly these methods are of two types:**

1. **Optimization methods**
2. **Algorithmic methods**

**And these methods can be used character or distance data. The various methods are summarized in Figure 5.11**

****

Distance methods are also called ***phenetic methods.*** Here, trees are constructed by similarities of sequences and the resulting tree is called a ***Dendrogram.*** A Dendogram does not necessarily reflect evolutionary relationships. Character based methods are also called **cladistic methods.** The resulting tree is called **cladogram.** Cladistic methods use each alignment positions as evolutionary information to build a tree.

Distance based method are based on creating a distance matrix. Starting from an alignment, pairwise distances are calculated between DNA sequences as the sum of all base pair differences between two sequences. The most similar sequences are assumed to be closely related. This creates the distance matrix.

**Distance matrices are derived in such a way that each mismatch between two sequences adds to the distance, and each identity subtracts from the distance. Scoring matrices include values for all possible substitutions.**

**The various approaches towards tree building are as follows-**

1. **Distance methods**

**1a. UPGMA**

**1b. Neighbour joining (NJ)**

1. **Character based methods**

**2a. Maximum parsimony (MP)**

**2b. Maximum likelihood (ML)**

1. **Distance method**

**1a. UPGMA**

* **The unweighted pair-group method with arithmetic mean (UPGMA) is a popular distance analysis method.**
* **UPGMA method is a popular because of its simplicity, but it makes trees with two important assumed properties that are particularly important when trees are made from macromolecular sequence data.**
* **First, it is assumed that evolution occurs at the same rate on all tree branches (this is known as the assumption of a molecular clock)**
* **Second, it is assumed that distance in the trees are additives. The additive assumption is that the distance between any two leaves is the sum of distances on edges connecting them. As a consequence of these assumptions, the method can create incorrect trees. For example, two sequences might be very similar not because they have a direct common ancestor, but simply because they are evolving very slowly by comparison with other sequences being analyzed. UPGMA would produce a tree in which they had a direct common ancestor.**

**The neighbour- joining method**

**The neighbour-joining method is a distance clustering method. It is begin by choosing the two most closely related sequences, and then adding the next most distant sequence as a third branch to the tree. Consider an example tree with three sequences A, B and C and the distance between nodes x, y and z given below in Table 5.8**

**TABLE 5.8 Example of three sequences**

B C

A 24 28

B 27

The representation of these sequences is given in Figure 5.16

Figure 5.16 Example for Neighbour joining method

y

z

x

A B C

Simultaneous linear equations as given below can be used to calculate the branch lengths:

Thus with 3 equations and 3 unknowns you can calculate that x = 10, y = 14 and z = 18. Branches are added until all sequences are included in the tree.

A to B: x + y = 24

A to C: x + z = 28

B to C:y + z = 32.

**Character-based Methods**

**There are two major types of cladistics methods- based on Parsimony and based on Maximum likelihood.**

**In Parsimony methods, for each position 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 sequences changes. The most Parsimonious tree is the one with fewest evolutionary changes for all sequences to derive from a common ancestor. This is more time consuming method than all distance methods.**

**The Maximum Likelihood method also uses each position in an alignment. It evaluates all possible trees, and calculates the likelihood for each tree using an explicit model of evolution. The likelihoods for each aligned position are then multiplied to provide the composite likelihood for each tree. The tree with maximum likelihood is the most probable tree. This is the slowest method of all but may give the best result and the most information about tree.**

**These two methods are discussed below-**

1. **Maximum Parsimony (MP)- In distance matrix methods, all possible sequence alignments are carried out to determine the most closely related sequences, and phylogenetic trees are constructed on the basis of these distance measurements.**

**Maximum Parsimony methods (MP) can be used, in which trees are constructed on the basis of the minimum number of mutations required to convert one sequence into another. In proteins, this is achieved by multiple sequence alignment followed by the identification and analysis of corresponding positions in each sequence. For each aligned residue, the minimum number of base substitutions required to convert one amino acid into another is calculated. The final tree is generated by grouping those sequences that can be interconverted with the smallest number of overall changes. This is attractive intellectually, but, like the maximum likelihood method below, can be expensive in computer time.**

1. **Maximum likelihood methods- Maximum likelihood methods also involve multiple sequence alignment and the analysis of changes at each position of the sequence. However, the difference between maximum likelihood and maximum parsimony is that the former incorporates an expected model of sequence changes, which weights the probability of any residue being converted into any other. This model can be set by the experimenter. For each possible tree, the likelihood of different sequence changes at each position is calculated, and these values are multiplied to provide an overall likelihood for each tree.**