Introduction, Advantages, Phylogenetic Trees, Tree topologies, Methods for phylogenetic analysis- Distance Matrix methods, Character based methods.

HMM (Hidden Markov Model): Introduction to HMM, Forward algorithm, Viterbi algorithm, applications in Bioinformatics

Introduction to phylogenetics

Phylogenetics is the scientific study of phylogeny. Phylogeny pertains to the evolutionary history of a taxonomic group of organisms. Thus, phylogenetics is mainly concerned with the

- relationships of an organism to other organisms according to evolutionary similarities and differences.
- identification and classification of organisms.
- provides information to, which is a branch of science concerned also in finding, describing, classifying, and naming organisms, including the studying of the relationships between taxa and the principles underlying such a classification.

In phylogenetics, DNA sequencing methods are used to analyze the observable heritable traits. It also makes use of phylogenetic trees

Application of phylogenetic analysis

Classification: Phylogenetics based on sequence data provides us with more accurate descriptions of patterns of relatedness than was available before the advent of molecular sequencing. Phylogenetics now informs the Linnaean classification of new species.

Forensics: Phylogenetics is used to assess DNA evidence presented in court cases to inform situations, e.g. where someone has committed a crime, when food is contaminated, or where the father of a child is unknown.

Identifying the origin of pathogens: Molecular sequencing technologies and phylogenetic approaches can be used to learn more about a new pathogen outbreak. This includes finding out about which species the pathogen is related to and subsequently the likely source of transmission. This can lead to new recommendations for public health policy.

Conservation: Phylogenetics can help to inform conservation policy when conservation biologists have to make tough decisions about which species they try to prevent from becoming extinct.

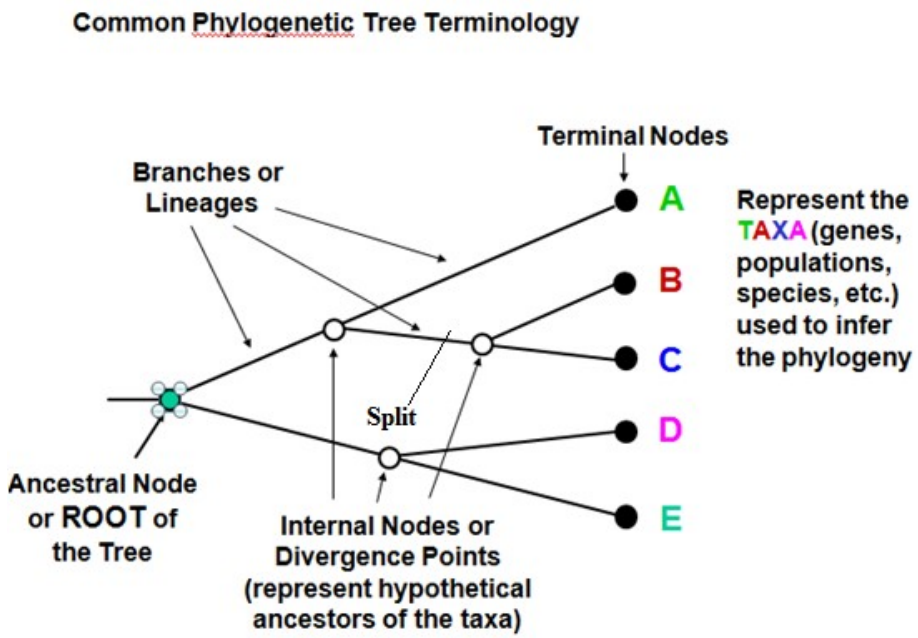
Bioinformatics and computing: Many of the algorithms developed for phylogenetics have been used to develop software in other fields.

Advantages

Phylogenetics analysis enriches our understanding of how genes, genomes, species (and molecular sequences more generally) evolve. Through phylogenetics, we learn not only how the sequences came to be the way they are today, but also general principles that enable us to predict how they will change in the future.

Phylogenetic trees

Pylogenetic tree is a 2D graph showing evolutionary histories and relationships among groups of organisms.We refer the separate source of sequences as taxa(singular taxon).Taxas are the distinct units on the tree. The phylogenetic tree has been used to understand biodiversity, genetics, evolutions, and ecology of organisms.



The tree composed of following

- **node :** a node represents a taxonomic unit. This can be a taxon (an existing species) or an ancestor (unknown species : represents the ancestor of 2 or more species).
- **branch :** defines the relationship between the taxa in terms of descent and ancestry.
- **topology :** is the branching pattern.

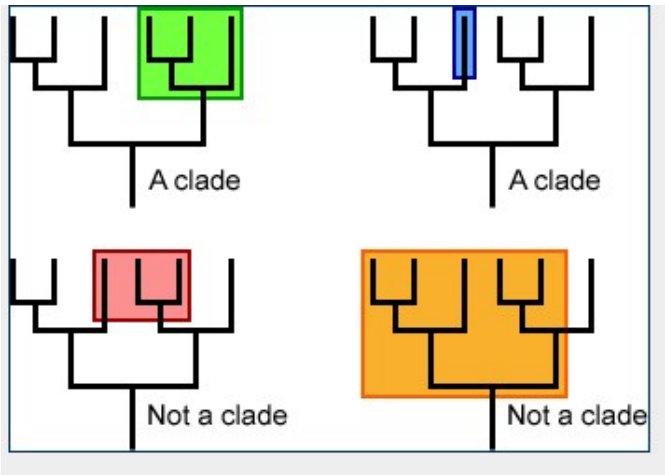
- **branch length :** often represents the number of changes that have occurred in that branch.
- **root :** is the common ancestor of all taxa.
- **distance scale :** scale which represents the number of differences between sequences (e.g. 0.1 means 10 % differences between two sequences)
- **lineage :**Branch path depicting an ancestor-descendant relationship
- **leaf :** Is a node with degree one.In phylogeny it represents a single present day taxon.
- **Internal node:** It has a degree greater than one .It represents common ancestors.
- **Split:** Is a partition of the taxa into 2 non empty sets. Each edge in the tree represents a split
- **Subtree:** Is the subset of a tree

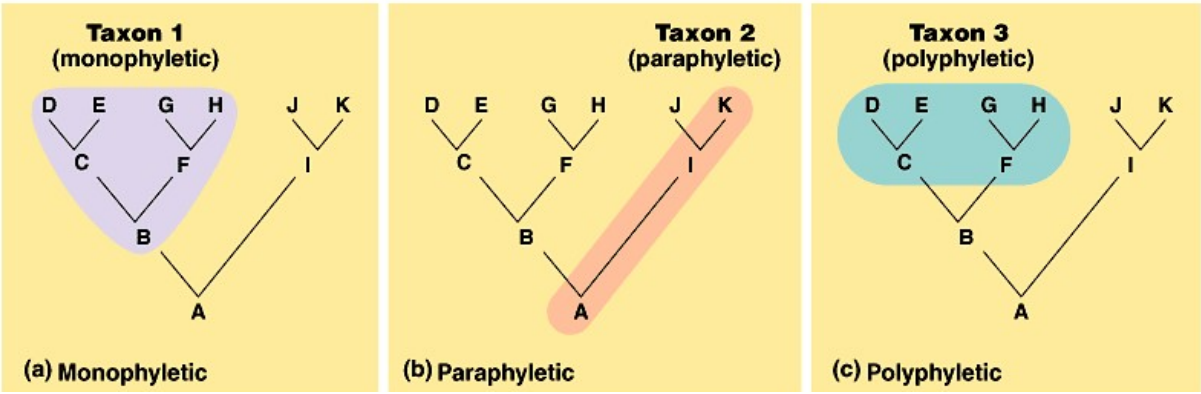
A Group (Taxa) consists of collection of organisms and is classified as

Monophyletic taxa : A group composed of a collection of organisms, including the most recent common ancestor of all those organisms and all the descendants of that most recent common ancestor. A monophyletic taxon is also called a clade.
Paraphyletic taxa : A group composed of a collection of organisms, including the most recent common ancestor of all those organisms. Unlike a monophyletic group, a paraphyletic taxon does not include all the descendants of the most recent common ancestor.
Polyphyletic taxa : A group composed of a collection of organisms in which the most recent common ancestor of all the included organisms is not included, usually because the common ancestor lacks the characteristics of the group.

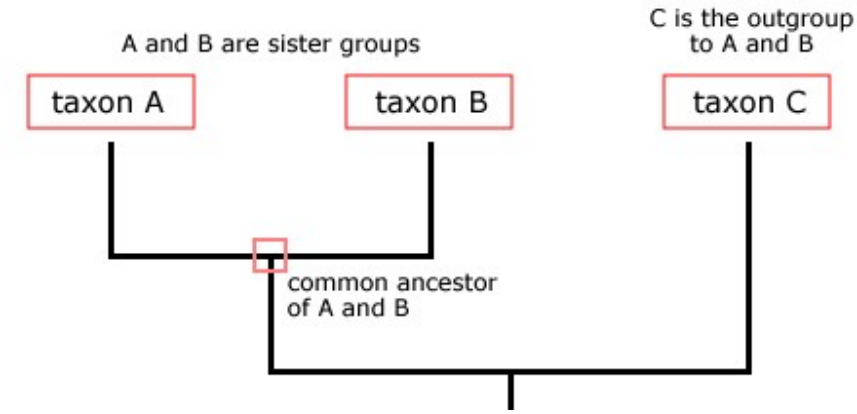
Polyphyletic taxa are considered "unnatural", and usually are reclassified once they are discovered to be polyphyletic.

Examples : marine mammals, bipedal mammals, flying vertebrates, trees, algae, etc.





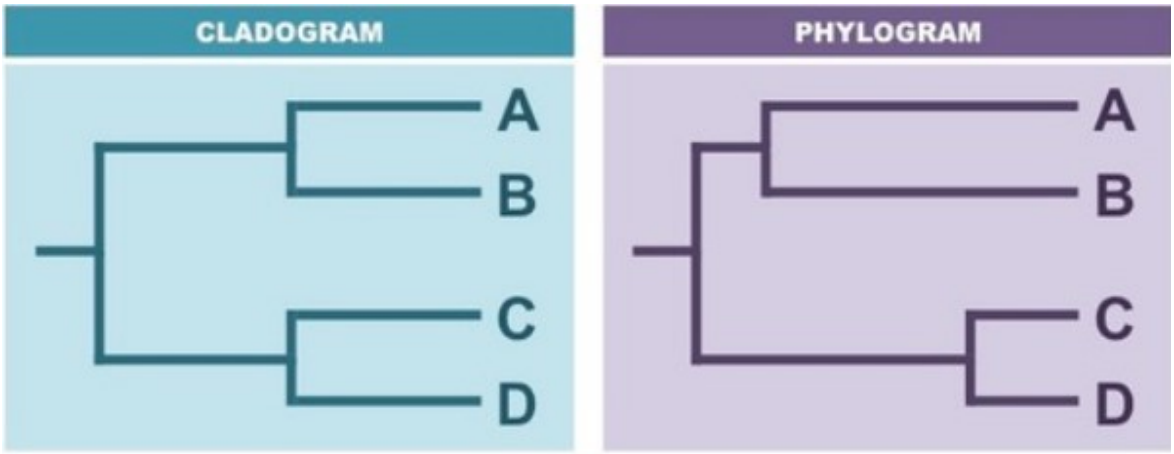
Many phylogenies also include an outgroup — a taxon outside the group of interest. All the members of the group of interest are more closely related to each other than they are to the outgroup. In the following tree taxon A and B are called sister taxas since both share the common ancestor



Tree topologies

phylogram : A phylogram is a branching diagram (tree) that is assumed to be an estimate of a phylogeny. The branch lengths are proportional to the amount of inferred evolutionary change.

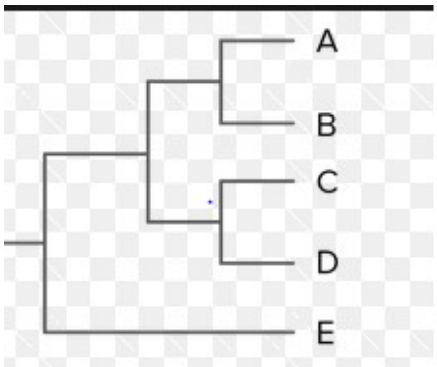
cladogram :A cladogram is a branching diagram (tree) assumed to be an estimate of a phylogeny where the branches are of equal length. Therefore, cladograms show common ancestry, but do not indicate the amount of evolutionary "time" separating taxa. It is possible to see the tree distances by clicking on the diagram to get a menu of options.



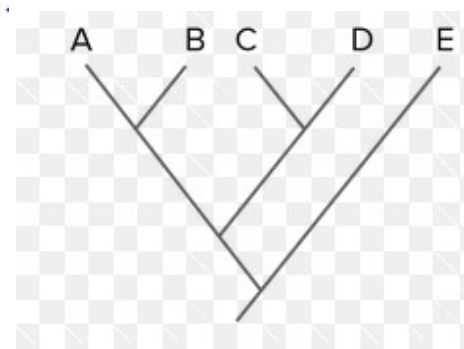
Phenogram: a diagram depicting taxonomic relationships among organisms based on overall similarity of many characteristics without regard to evolutionary history or assumed significance of specific characters: usually generated by computer.

Dendrogram: general term for a branching diagram

Shape of phylogenic tree : Can be rectagular or slanted



Rectangular cladogram



slanted cladogram

Phylogenetic trees take several forms:

They can be *rooted* or *unrooted*,.

rooted *tree:* A *rooted* tree is a tree in which one of the nodes is stipulated to be the root, and thus the direction of ancestral relationships is determined.

unrooted *tree* :An *unrooted* tree, as could be imagined, has no pre-determined root and therefore induces no hierarchy. Therefore, in this case, the distance between the nodes should be symmetric (since the tree edges are not directed).

Additive tree: contain additional information, namely branch lengths. These are numbers associated with each branch that correspond to some attribute of the sequences, such as amount of evolutionary change. In the example shown in Figure 5, sequence A has acquired 4 substitutions since it shared a common ancestor with sequence B. Other commonly used terms for additive trees include "metric trees" and "phylograms."

Ultrametric trees:
sometimes also called "dendrograms") are a special kind of additive tree in which the tips of the trees are all equidistant from the root of the tree. This kind of tree can be used to depict evolutionary time, expressed either directly as years or indirectly as amount of sequence divergence using a molecular clock.

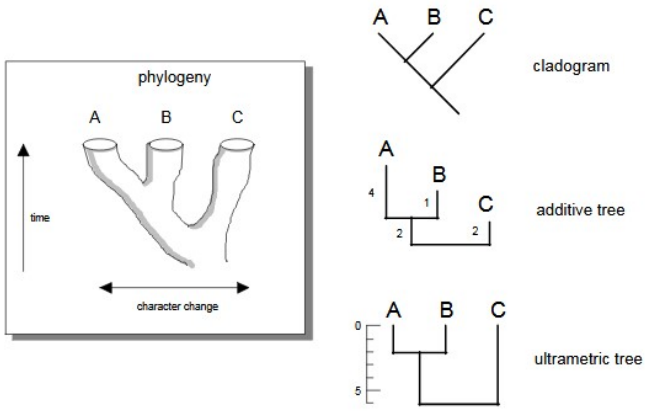


Figure 5. A phylogeny and the three basic kinds of tree used to depict that phylogeny. The cladogram represents relative recency of common ancestry; the additive tree depicts the amount of evolutionary change that has occurred along the different branches, and the ultrametric tree depicts times of divergence.

Consensus Tree: Is a tree that represents common featurs among two or more primary trees

Phylogenetic Analysis

Phylogenetic analysis is concerned with construction of phylogenetic trees and analyse them to get meaningful inferences.There are two types of method to construct trees

1.Distance based methods

2.Character based methods

Distance based methods

- Requires a distance matrix created by calculating pairwise distance between 2 sequences.Distance matrix are derived in such a way that each mismatch beween two sequence add to the distance and each identity subtracts from the matrix.Insertions and deletions are given a larger weight than replacements(substitution). Substitution can be of 2 types transitions and translation and give weight .it is possible to correct for multiple substitutions at a single site.
- From the obtained distance matrix, a phylogenetic tree is calculated with clustering algorithms. These cluster methods construct a tree by linking the least distant pair of taxa, followed by successively more distant taxa.

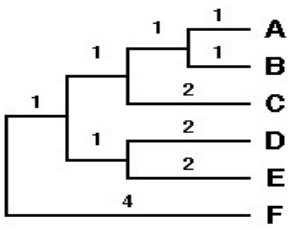
Following are the clustering method used by distance based methods

UPGMA method

Unweighted Pair Group Mehod Using Assymmetric Average is the simplest clustering method.This build the tree sequentially by the following steps

1. Assume that initially each species is a cluster on its own. It starts by grouping each taxa with smallest pair wise distance in the distance matrix
2. Then joins closest 2 clusters and recalculates distance of the joint pair by taking the average. A node is hence placed at the midpoint distance between them.
3. It then creates a reduced matrix by creating a new cluster as a single taxon.
4. Repeat this process until all species are connected in a single cluster
5. The last taxon added is considered the out group producing a rooted tree.

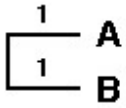
Suppose we have the following tree consisting of 6 OTU(operational taxonomic unit)s:



The pair wise evolutionary distances are given by the following distance matrix:

	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

We now cluster the pair of OTUs with the smallest distance, being A and B, that are separated a distance of 2. The branching point is positioned at a distance of 2 / 2 = 1 substitution. We thus constuct a subtree as follows:



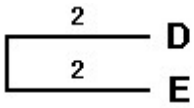
Following the first clustering A and B are considered as a single composite OTU(A,B) and we now calculate the new distance matrix as follows:

```
dist(A,B),C = (distAC + distBC) / 2 = 4
dist(A,B),D = (distAD + distBD) / 2 = 6
dist(A,B),E = (distAE + distBE) / 2 = 6
dist(A,B),F = (distAF + distBF) / 2 = 8
```

In other words the distance between a simple OTU and a composite OTU is the average of the distances between the simple OTU and the constituent simple OTUs of the composite OTU. Then a new distance matrix is recalculated using the newly calculated distances and the whole cycle is being repeated:

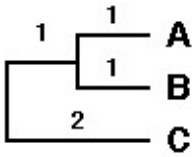
Second cycle

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



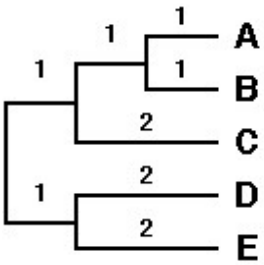
Third cycle

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



Fourth cycle

	AB,C	D,E
D,E	6	
F	8	8



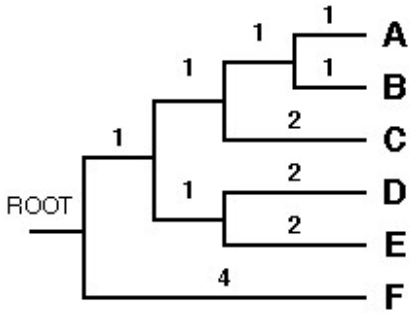
Fifth cycle

The final step consists of clustering the last OTU, F, with the composite OTU.

	ABC,DE
F	8

Although this method leads essentially to an unrooted tree, UPGMA assumes equal rates of mutation along all the branches, as the model of evolution used. The theoretical root, therefore, must be equidistant from all OTUs. We can here thus apply the method of mid-point rooting. The root of the entire tree is then positioned at dist (ABCDE),F / 2 = 4.

The final tree as inferred by using the UPGMA method is shown below.



So now we have reconstructed the phylogenetic tree using the UPGMA method. As you can see we have obtained the original phylogenetic tree we started with.

Neighbour –Joining method

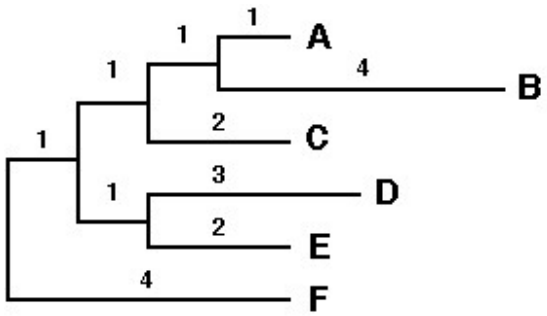
Algorithm

Neighbor-joining is a recursive algorithm. Each step in the recursion consists of the following steps:

1. Based on the current distance matrix calculate a modified distance matrix Q (see below).
2. Find the least distant pair of nodes in Q (= the closest neighbors = the pair with the lowest distance value). Create a new node on the tree joining the two closest nodes: the two nodes are linked by their common ancestral node.
3. Calculate the distance of each of the nodes in the pair to their ancestral node.
4. Calculate the distance of all nodes outside of this pair to their ancestral node.
5. Start the algorithm again, considering the pair of joined neighbors as a single taxon (the terminal nodes are replaced by their ancestral node and the ancestral node is then treated as a terminal node) and using the distances calculated in the previous step.

Example of the method

Suppose we have the following tree:



Since B and D have accumulated mutations at a higher rate than A. The Three-point criterion is violated and the UPGMA method cannot be used since this would group together A and C rather than A and B. In such a case the neighbor-joining method is one of the recommended methods.

The raw data of the tree are represented by the following distance matrix:

First we need a distance matrix

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

We have in total 6 OTUs (N=6).

	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5
Distance matrix					

Cycle 1

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

Cycle 2

	U_1	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8

Cycle 3

	U_1	C	U_2
C	3		
U_2	3	4	
F	7	8	6

Cycle 4

	U_2	U_3
U_3	2	
F	6	6

Cycle 5

	U_4
F	5

Step 1

S calculations

S_x = (sum all D_{xi})/($N-2$), where N is the # of OTUs in the set.

Cycle 1:
$S_A = (5+4+7+6+8)/4 = 7.5$
$S_B = (5+7+10+9+11)/4 = 10.5$
$S_C = (4+7+7+6+8)/4 = 8$
$S_D = (7+10+7+5+9)/4 = 9.5$
$S_E = (6+9+6+5+8)/4 = 8.5$
$S_F = (8+11+8+9+8)/4 = 11$

Cycle 2:
$S_{U_1} = (3+6+5+7)/3 = 7$
$S_C = (3+7+6=8)/3 = 8$
$S_D = (6+7+5+9)/3 = 9$
$S_E = (5+6+5+8)/3 = 8$
$S_F = (7+8+9+8)/3 = 10.6$

Cycle 3:
$S_{U_1} = (3+3+7)/2 = 6.5$
$S_C = (3+4+8)/2 = 7.5$
$S_{U_2} = (3+4+6)/2 = 6.5$
$S_F = (7+8+6)/2 = 10.5$

Cycle 4:
$S_{U_2} = (2+6)/1 = 8$
$S_{U_3} = (2+6)/1 = 8$
$S_F = (6+6)/1 = 12$

Cycle 5: Because $N-2 = 0$, we cannot do this calculation.

Step 2

Calculate pair with smallest (M), where $M_{ij} = D_{ij} - S_i - S_j$.

Cycle 1: Smallest are
$M_{AB} = 5 - 7.5 - 10.5 = -13$
$M_{DE} = 5 - 9.5 - 8.5 = -13$
Choose one of these (AB here).

Cycle 2: Smallest is
$M_{CU_1} = 3 - 7 - 8 = -12$
$M_{DE} = 5 - 9 - 8 = -12$
Choose one of these (DE here).

Cycle 3: Smallest is
$M_{CU_1} = 3 - 6.5 - 7.5 = -11$

Cycle 4: Smallest is
$M_{U_2F} = 6 - 8 - 12 = -14$
$M_{U_3F} = 6 - 8 - 12 = -14$
$M_{U_2U_3} = 2 - 8 - 8 = -14$
Choose one of these ($M_{U_2U_3}$ here).

Step 3

Create a node (U) that joins pair with lowest M_{ij} such that $S_{iU} = D_{ij}/2 + (S_i - S_j)/2$.

Cycle 1: U_1 joins A and B:
$S_{AU_1} = D_{AB}/2 + (S_A - S_B)/2 = 1$
$S_{BU_1} = D_{AB}/2 + (S_B - S_A)/2 = 4$

Cycle 2: U_2 joins D and E:
$S_{DU_2} = D_{DE}/2 + (S_D - S_E)/2 = 3$
$S_{EU_2} = D_{DE}/2 + (S_E - S_D)/2 = 2$

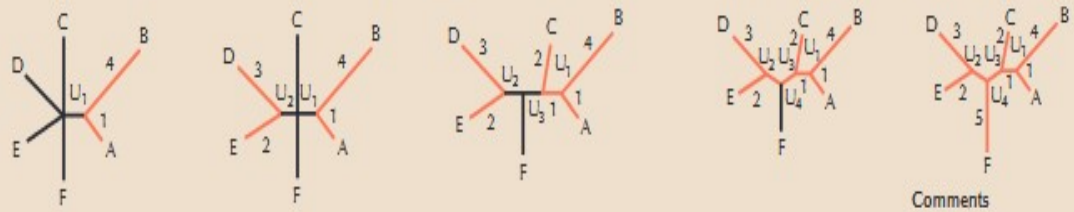
Cycle 3: U_3 joins C and U_1:
$S_{CU_3} = D_{CU_1}/2 + (S_C - S_{U_1})/2 = 2$
$S_{U_1U_3} = D_{CU_1}/2 + (S_{U_1} - S_C)/2 = 1$

Cycle 4: U_4 joins U_2 and U_3:
$S_{U_2U_4} = D_{U_2U_3}/2 + (S_{U_2} - S_{U_3})/2 = 1$
$S_{U_3U_4} = D_{U_2U_3}/2 + (S_{U_3} - S_{U_2})/2 = 1$.

Cycle 5: For last pair, connect U_4 and F with branch length = 5.

Step 4

Join *i* and *j* according to *S* above and make all other taxa in form of a star. Branches in black are of unknown length. Branches in red are of known length.



Step 5

Calculate new distance matrix of all other taxa to U with $D_{xU} = D_{ix} + D_{jx} - D_{ij}$, where *i* and *j* are those selected from above.

Comments

Note this is the same tree we started with (drawn in unrooted form here).

- Advantages
  - is fast and thus suited for large datasets and for bootstrap analysis
  - permist lineages with largely different branch lengths
  - permits correction for multiple substitutions
- Disadvantages
  - sequence information is reduced
  - gives only one possible tree
  - strongly dependent on the model of evolution used.

Character Based methods

There are 2 types of cladistic methods they are

- Maximum Parsimony
- Maximum Likelyhood

Maximum Parsimony

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 sequence changes. The **most parsimonious tree** is the one with the **fewest evolutionary changes** for all sequences to derive from a common ancestor. This is a more time-consuming method than the distance methods.

Advantages

- Reconstruct ancestoral nodes there by using all the evolutionary data
- Perform better than distance methods
- Provides numerous parsimonious trees

Disadvantages

- Branch length canot be determined,only topology
- Slower than distance methods
- Sensitive to the order in which sequences are added to the tree

Maximum Likelyhood

This method also uses **each position** in an alignment, evaluates all possible trees, and calculates the **likelihood for each tree** using an explicit **model of evolution** (<-> Parsimony just looks for the fewest evolutionary changes). The likelihood's for each aligned position are then multiplied to provide a likelihood for each tree. The tree with the maximum likelihood is the most probable tree. This is the slowest method of all but seems to give the best result and the most information about the tree.

Advantages

- Reconstructs ancestral nodes so using all evolutionary data
- Generate branch length
- Generates statistical estimate of significance of each branch
- Perform better than distance methods

Disadvantages

- It is very slow

Hidden Markov Model

A Markov process is a random process in which the future is independent of the past, A trivial example is throwing a die the change is the transition from one state to next. A markov process is characterized by the property that the change is depend only on the current state, the previous states are immaterial.

The HMM can be visualized as a finite state machine, that move through a series of states and produce some kind of output either when the machine has reached at a particular state or when it is moving from one state to another..HMM working as a finite state machine generates protein sequence by emitting amino acid as it proceeds through a sequence of states.There are also transition probabilities for moving from one state to another .The sum of all transition probabilities =1

For modelling biological sequences, you have to follow these steps:

1. You need to collect a set of sequences of your interest.
2. You then need to define a vocabulary $V = \{X_1, X_2, X_3, \ldots, X_N\}$. For example, for DNA sequences, $N = 4$ and $V = \{A, T, C, G\}$.
3. You then need to develop a model, which is then used for generating typical sequences from the class of training sequences (e.g.) protein families.

Following figure shows the markov model for DNA sequence

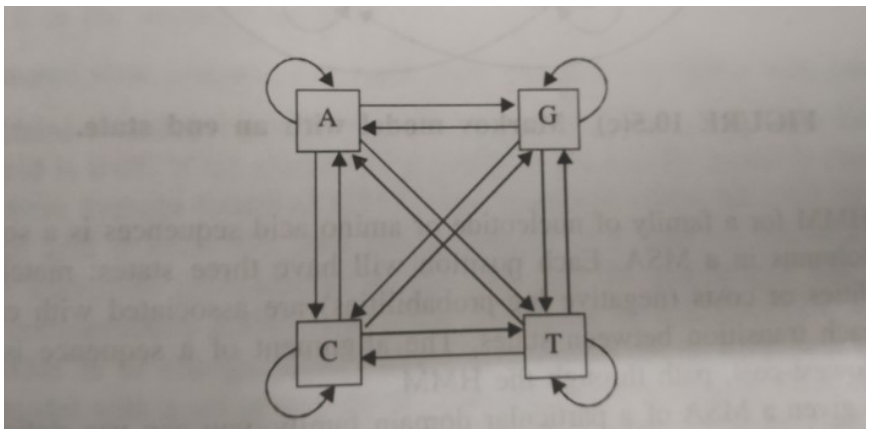


FIGURE 10.5(a) Markov model for DNA sequence.

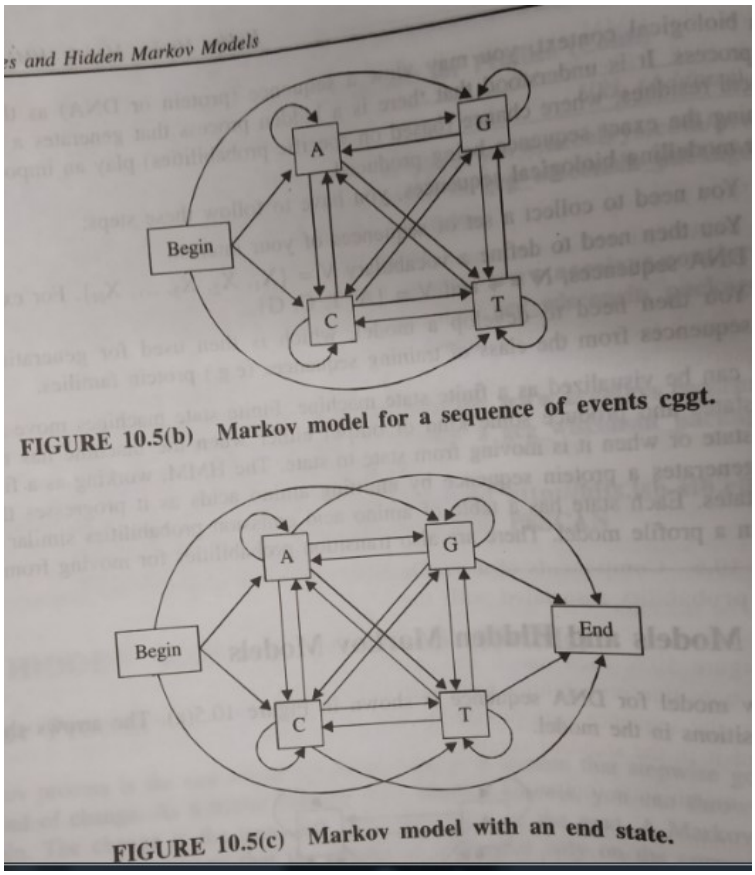
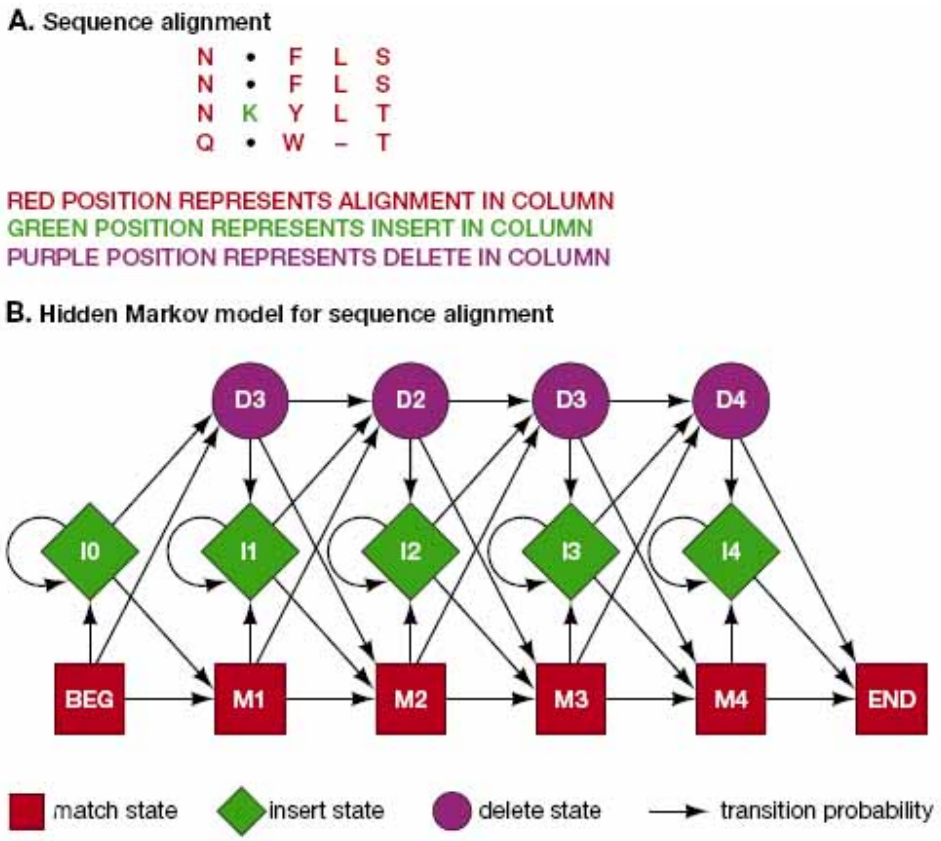


FIGURE 10.5(b) Markov model for a sequence of events cggt.

FIGURE 10.5(c) Markov model with an end state.

A linear HMM fro a family of nucleotide or amino acid sequence is a set of positions that relaed to columns inMSA.Each position will have 3 states :match ,insert or



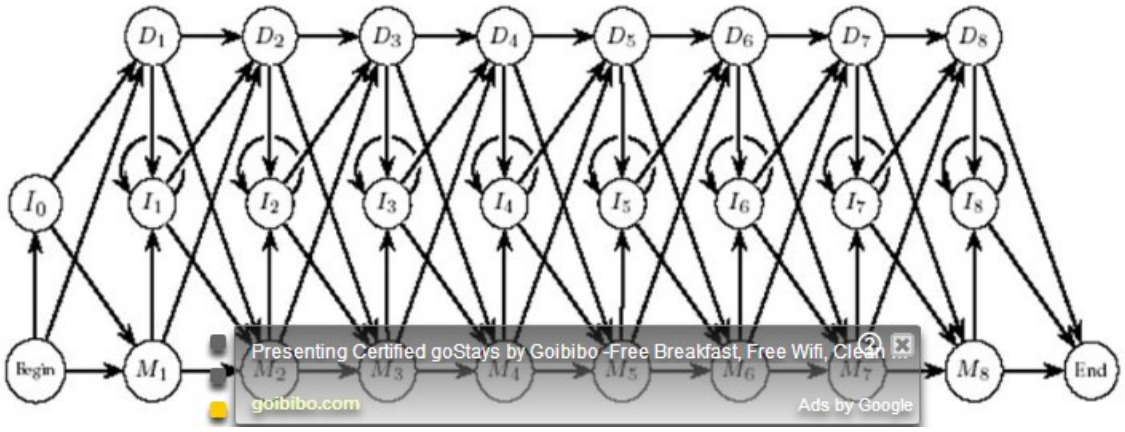
delete

- A **profile HMM** is a probabilistic representation of a multiple alignment.

- A given multiple alignment (of a protein family) is used to build a profile HMM.

- This model then may be used to find and score less obvious potential matches of new protein sequences.

- A profile HMM has three sets of states:
  - **Match states**: $M_1, \ldots, M_n$ (plus *begin/end* states)
  - **Insertion states**: $I_0, I_1, \ldots, I_n$
  - **Deletion states**: $D_1, \ldots, D_n$



Building a Profile HMM

1. Multiple alignment is used to construct the HMM model.

2. Assign each column to a *Match* state in HMM. Add *Insertion* and *Deletion* state.

3. Estimate the emission probabilities according to amino acid counts in column. Different positions in the protein will have different emission probabilities.

4. Estimate the transition probabilities between *Match, Deletion* and *Insertion* states.

HMM scoring algorithms

Once an HMM has been constructed we need to score the HMM and arrive at an alignment.

Viterbi algorithm

- Is a dynamic programming algorithm that can find the best alignment and its probability without going to all possible alignment
- Is used to produce multiple alignment of a set of sequences
- Each node in the HMM has a match state ,delete state and an insert state

In addition to a transition probability, the match state also has position-specific probabilities for seeing (or, more formally, emitting) a particular residue. Likewise, the insert state has probabilities for inserting a residue at the position given by the node. There is also a chance that no residue is associated with a node. That probability is indicated by the probability of transitioning to the delete state.

Both transition and emission probabilities can be generated from a multiple alignment of a family of sequences. An HMM can be compared (that is, aligned) with a new sequence to determine the probability that the sequence belongs to the modeled family. The most probable path through the HMM (i.e. which transitions were taken and which residues were emitted at match and insert states) taken to generate a sequence similar to the new sequence determines the similarity score.

The Viterbi algorithm can be defined as a matrix, where the columns of the matrix are indexed by the states in the model, and the rows are indexed by the sequence. Deletion states are not shown, since, by definition, they have a zero probability of emitting amino acid. The elements of the matrix are initialized to zero and then computed with these steps:

1. The probability that the amino acid A was generated by state I_0 is computed and entered as the first element of the matrix.
2. The probabilities that C is emitted in state M_1 (multiplied by the probability of the most likely transition to state M_1 from state I_0) and in state I_1 (multiplied by the most likely transition to state I_1 from state I_0) are entered into the matrix element indexed by C and I_1/M_1.
3. The maximum probability, max (I_1, M_1), is calculated.
4. A pointer is set from the winner back to state I_0.
5. Steps 2–4 are repeated until the matrix is filled.

Forward algorithm

Forward algorithm is used to calculate the sum over all paths inductively. The forward algorithm is similar to Viterbi. However, in step 3 of Viterbi algorithm, a sum rather than a maximum is computed, and no back pointers are needed. The probability of the sequence is found by summing the probabilities in the last column.

Application of HMM

1. Identification of G-protein coupled receptors
   G-protein-coupled receptors (GPCRs) mediate most of our physiological responses to hormones, neurotransmitters and environmental stimulants, and so have great potential as therapeutic targets for a broad spectrum of diseases
2. Clustering of paths for a sub group

3. Gene prediction

In computational biology, **gene prediction** or **gene finding** refers to the process of identifying the regions of genomic DNA that encode genes

4. Modelling protein domains
   Hidden Markov Models (HMMs) are a powerful tool for protein domain identification and moddelling