

Molecular Sampling and Sequencing for Phylogeny

Udayan Borthakur

Programme Coordinator

Wildlife Genetics

Aaranyak

udayan@aaranyak.org

About Aaranyak

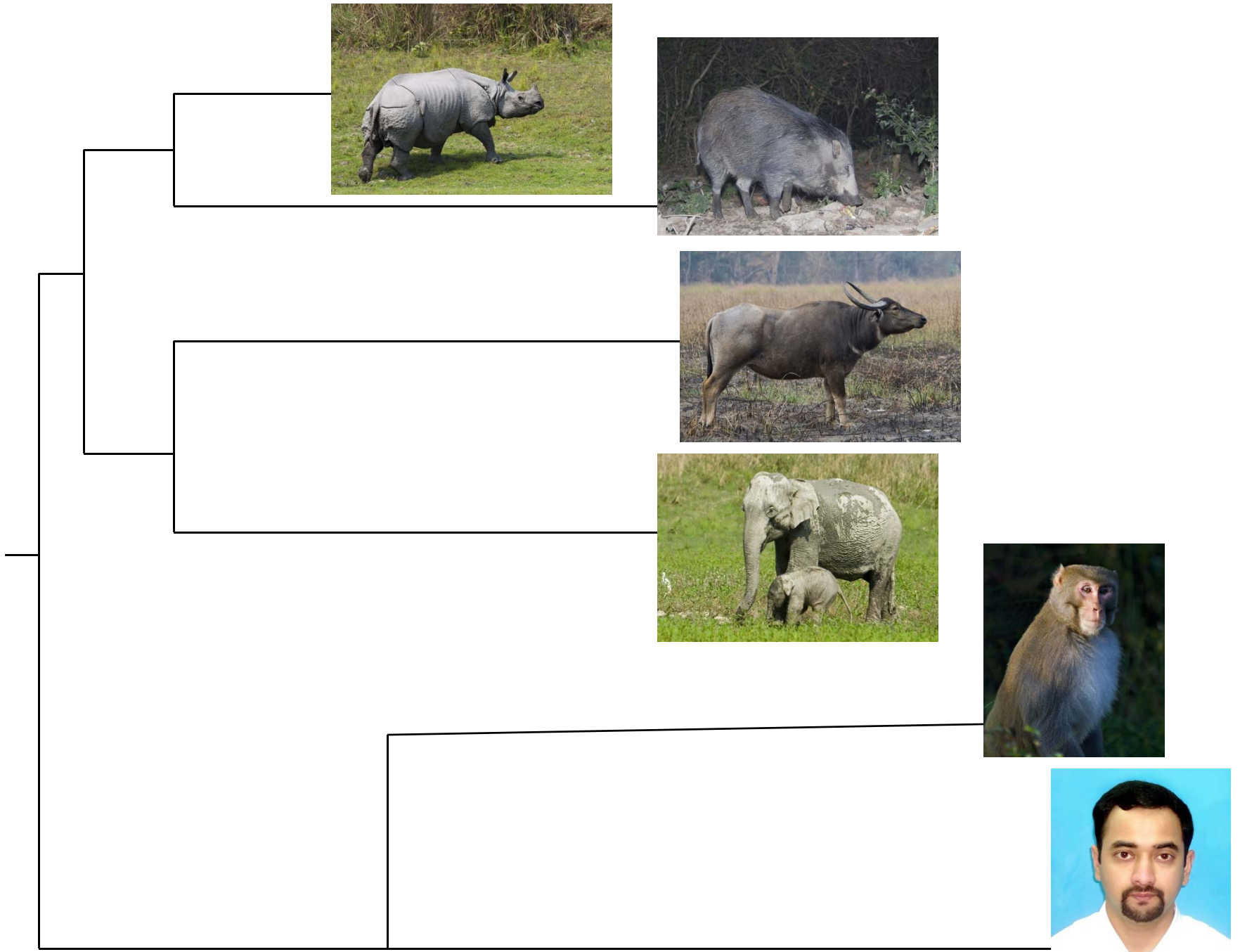
www.aaranyak.org

Aaranyak is a society for biodiversity research and conservation in Northeast India

Recognized as Scientific Industrial Research Organization (SIRO), by Ministry of Scientific & Industrial Research, Govt. of India

Aaranyak currently has seven programmes or departments

Wildlife Genetics is the latest addition, starting from January 2008



Contents

1. Phylogeny (1)
2. Phylogenetic tree (1)
3. Types of phylogenetic tree (1)
4. Data required (1)
5. Molecular phylogenetics (13)
 - i. Evolutionary models
 - ii. Molecular clock
6. Methods of constructing phylogenetic tree (11)
 - i. Distance based
 - ii. Character based
7. Sampling for phylogeny (3)
8. Sequencing for phylogeny (3)
9. Suggested readings (1)
10. Practicals (2)

Phylogeny

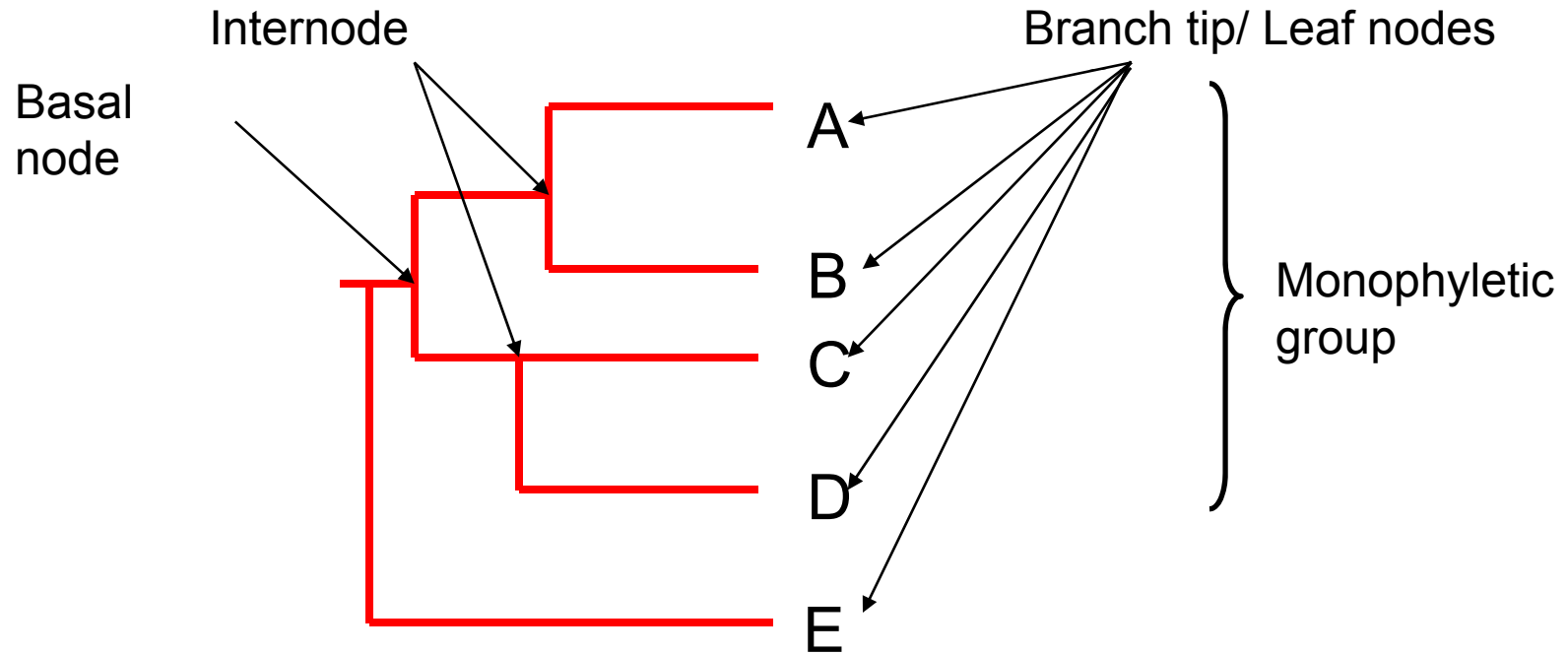
The study of evolutionary relatedness among various groups of organisms (Species, populations of a species)

[*phyle/phylon* meaning "tribe, race," and *genetikos* meaning "relative to birth"]

The basic aim of reconstructing a phylogeny is to obtain evolutionary relationship among organisms of interest in the form of a **phylogenetic tree**

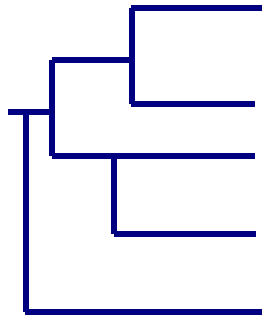
Phylogenetic Tree

A **phylogenetic tree** is a tree diagram showing the evolutionary relationships among various biological species or other entities that are believed to have a common ancestor



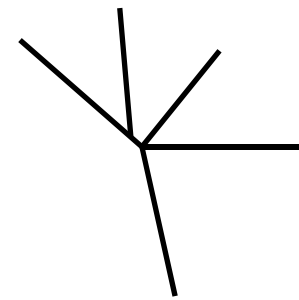
The branching pattern of a tree is called tree topology

Types of phylogenetic tree



Rooted Tree

A direct tree with unique node corresponding to the most recent common ancestor of all the organisms at the leaf nodes of the tree



Unrooted Tree

Illustrates the relatedness of the organisms at the leaf nodes without making assumptions about the common ancestry

A rooted tree can be inferred from an unrooted tree by including an **outgroup** to the input data or including additional assumptions about the relative rates of evolution on each branch, such as **molecular clock hypothesis**

Data requirement for obtaining a phylogenetic tree:

- *Morphological and physiological data*

(e.g., matrices of various body part measurements)

- *Molecular sequence data* – **Molecular phylogenetics**

Sequences of single or multiple fragments of DNA (e.g., particular genes)

Organism A: ATGATCTGATCGGCCCAATATATTTC

Organism B: ATGATGTCACCGCCCCAATATATTTC

Organism C: ATGATCTGACCGCCCTATATATTTC

Molecular phylogenetics

The primary cause of evolution is the mutational changes in the DNA sequence

- i. Nucleotide substitution
- ii. Insertion/ deletion
- iii. Recombination etc.

These mutant sequences may be spread through the population by

- i. Genetic drift and/or
- ii. Natural selection

Nucleotide substitutions

Terms to know before going further:

- **Synonymous** and **nonsynonymous** substitutions
- **Transition** and **transversion**

16 different types of nucleotide pairs between a pair of sequences:

Class	Nucleotide pair			
Identical nucleotide	AA	TT	CC	GG
Transition type pair	AG	GA	TC	CT
Transversion type pair	AT	TA	AC	CA
	TG	GT	CG	GC

Analyzing nucleotide differences between sequences

Evolutionary models

Models determines the way in which the evolutionary distances between the taxa under consideration are calculated.

The simplest model of analyzing the differences between two sequences:

P distance = proportion of nucleotide sites at which two sequences are different

$$\hat{p} = n_d/n$$

When p is large (i.e., greater differences between the sequences) it gives an underestimate of the number of nucleotide substitutions per site

Evolutionary models – Nucleotide substitution models

One parameter model – Jukes-Cantor model

(Jukes and Cantor 1969)

- Only one parameter – *nucleotide substitution rate*
- Probabilities of any nucleotide changing to any other nucleotide is equal

Kimura 2-parameter model

- Extends the Juke-Cantor correction by taking into account the possibility that the rates at which transitions and transversions occur might well be different
- Thus, if all changes were equal, we should see twice as many transversions as transitions
- However, in reality, we observe more transitions than transversions, as most transversions are nonsynonymous

Tamura 3-parameter model

- Adds a correction for compositional bias, i.e., if the base frequencies differ greatly from equal, perhaps because of mutational bias, then that difference needs to be accounted for

Tamura-Nei model

- Extends from Tamura 3-parameter model by distinguishing between transitional substitution rates among purines and transversional substitution rates among pyrimidines

Other models of this general family:

Felsenstein 84 and HYK model

General time-reversible (GTR) model

(Tavare 1986)

Consists of:

i) equilibrium base-frequency vector,

$$\Pi = (\Pi_A + \Pi_C + \Pi_G + \Pi_T)$$

ii) 6 substitution probabilities: a, b, c, d, e, f

e.g., $a = \text{Prob [A to C]} = \text{Prob [C to A]}$

GTR has a rate matrix as below:

$$Q = \begin{pmatrix} -\left(\frac{\pi_1 x_1}{\pi_2} + \frac{\pi_1 x_2}{\pi_3} + \frac{\pi_1 x_3}{\pi_4}\right) & \frac{\pi_1 x_1}{\pi_2} & \frac{\pi_1 x_2}{\pi_3} & \frac{\pi_1 x_3}{\pi_4} \\ x_1 & -\left(x_1 + \frac{\pi_2 x_4}{\pi_3} + \frac{\pi_2 x_5}{\pi_4}\right) & \frac{\pi_2 x_4}{\pi_3} & \frac{\pi_2 x_5}{\pi_4} \\ x_2 & x_4 & -\left(x_2 + x_4 + \frac{\pi_3 x_6}{\pi_4}\right) & \frac{\pi_3 x_6}{\pi_4} \\ x_3 & x_5 & x_6 & -\left(x_3 + x_5 + x_6\right) \end{pmatrix}$$

Rate variation among sites

- It is possible to assume that evolutionary rates might be different at different sites

e.g., the start codon of a gene is almost always **ATG**, but sometimes **GTG**

Thus at site 2 and 3 the substitution rate will be zero

- In general, positions on the interior of proteins, near active sites evolve more slowly than do positions near the surfaces

Evolutionary models

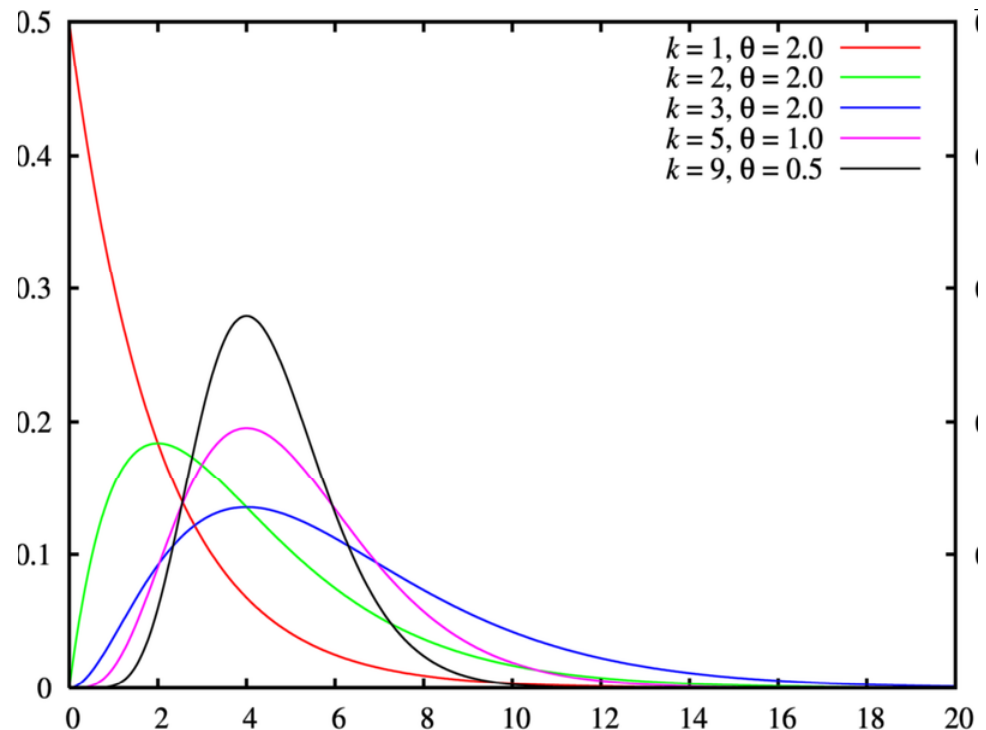
Rate variation among sites

One commonly used distribution – **Gamma distribution**

requires a scale parameter θ and a shape parameter k

The larger the θ , the more spread out the distribution

k affect the *shape* of a distribution rather than simply shifting it (as a location parameter does) or stretching/shrinking it (as a scale parameter does)



Invariant sites

- Some sites may not free to vary at all, such as initiation codons.
- However all the *identical sites* along a sequence alignment are not necessarily *invariant sites*
- Identical sites may simply mean that the sequences are too closely related for any substitution to have occurred
- Invariant sites can be estimated as per selected model specified

(available in some software such as MrBayes, BEAST)

Evolutionary models

Both **rate variation among sites** and **Invariant sites** can be incorporated into an evolutionary model while constructing a phylogenetic tree

e.g., GTR + Gamma + I

Homoplasy:

- i. Character reversal – a character changed but then reverted back to its original state (e.g., A to G to A)
- ii. Convergence – unrelated taxa evolved the same taxa independently
- iii. Parallelism – different taxa may have similar properties that predispose a character to evolve in a similar way

Molecular clock hypothesis

- Assumes that the rate of substitution is approximately constant over evolutionary time
- In reality, the actual number of substitutions is subject to stochastic errors and no gene would evolve at a constant rate for longer evolutionary time
- Still may be useful for certain groups of organisms, in estimating the time of divergence between the organisms
- Molecular clock can also be calibrated using fossil data for certain groups of organisms,

e.g., in birds it is estimated that there is a sequence divergence of approx. 2% in 1 million years (Weir 2006)

Methods of constructing a phylogenetic tree

Distance based methods

These methods convert aligned sequences into a distance matrix of pairwise differences

- i. Unweighed Pair-Group Method with Arithmetic Mean (UPGMA)
- ii. Neighbour joining (NJ)

Character based method

Use multiple sequence alignments directly by comparing characters within each column (each site of the DNA sequence)

- i. Parsimony
- ii. Maximum likelihood
- iii. Bayesian methods

Distance matrix

P-distance among 12 taxa:

Taxa1	0.000											
Taxa2	0.075	0.075										
Taxa3	0.066	0.066	0.038									
Taxa4	0.047	0.047	0.075	0.085								
Taxa5	0.057	0.057	0.085	0.094	0.009							
Taxa6	0.057	0.057	0.085	0.075	0.038	0.047						
Taxa7	0.075	0.075	0.104	0.094	0.057	0.066	0.019					
Taxa8	0.113	0.113	0.142	0.123	0.132	0.123	0.142	0.160				
Taxa9	0.132	0.132	0.160	0.160	0.113	0.104	0.132	0.151	0.142			
Taxa10	0.057	0.057	0.085	0.085	0.104	0.113	0.113	0.132	0.104	0.160		
Taxa11	0.066	0.066	0.085	0.075	0.085	0.094	0.085	0.104	0.085	0.132	0.075	
Taxa12	0.066	0.066	0.085	0.075	0.085	0.094	0.085	0.104	0.085	0.132	0.075	0.000

Characters:

each site or column of a sequence alignment is considered as a single character.

G	C	C	T	G	A	T	G	A	A	A	C	T	T	C	G	G	A	T	C	C	C	T	A	C	T	A	G	G	C	A	T	C
G	C	C	T	G	A	T	G	A	A	A	C	T	T	C	G	G	G	T	C	C	C	T	A	C	T	A	G	G	C	A	T	C
G	C	T	T	G	A	T	G	A	A	A	T	T	T	C	G	G	A	T	C	T	C	T	A	T	T	A	G	G	C	A	T	C
A	C	T	T	G	A	T	G	A	A	A	N	T	T	C	G	G	A	T	C	T	C	T	A	N	T	A	G	G	C	A	T	C
G	C	T	T	G	A	T	G	A	A	A	C	T	T	T	G	G	A	T	C	C	C	T	A	C	T	A	G	G	C	A	T	C
G	C	T	T	G	A	T	G	A	A	A	C	T	T	T	G	G	A	T	C	C	C	T	A	C	T	A	G	G	C	A	T	C
A	C	T	T	G	A	T	G	A	A	A	C	T	T	T	G	G	A	T	C	A	C	T	A	C	T	A	G	G	T	A	T	T
A	C	N	T	G	A	T	G	A	A	A	C	T	T	N	G	G	N	T	C	T	C	T	N	C	T	G	G	G	N	A	T	T
G	C	T	T	G	A	T	G	A	A	A	C	T	T	C	G	G	A	T	C	A	C	T	A	C	T	A	G	G	A	A	T	T
G	C	T	T	G	A	T	G	A	A	A	T	T	T	C	G	G	A	T	C	C	C	T	C	C	T	A	G	G	A	A	T	T

Methods of constructing a phylogenetic tree

Unweighed Pair-Group Method with Arithmetic Mean (UPGMA)

1. A clustering method in which a pair of taxa with the smallest distance between them is clustered together
2. Node is placed at the midpoint of the branch length (genetic distance) between the pair of taxa
3. The distance matrix is rewritten with the distance from the 1st cluster to rest of the taxa
4. The process is repeated unless all the taxa in the matrix is clustered

UPGMA is a rarely used method in current phylogenetics

Methods of constructing a phylogenetic tree

Neighbour-joining (NJ)

- Similar to UPGMA in using a distance matrix and reducing it in size in each step, then reconstructs the tree from that series of matrices
 - Differs from UPGMA in that it does not construct clusters but directly calculates distance to inter nodes
1. Net divergence of each taxon from all other taxa is calculated
 2. This net divergence is used to calculate a corrected distance matrix
 3. The pair of taxa with corrected distance is taken and calculates the distance from each of the those taxa to the node that joins them
 4. A new matrix is created in which the new node is substituted for those two taxa

NJ does not assume that all the taxa are equidistant from the root

Methods of constructing a phylogenetic tree

Parsimony

- Based on the assumption that the most likely tree is the one that requires the fewest number of changes (maximum parsimony) to explain the data in the sequence alignment
 - Parsimony assumes that a character is more likely to be common to two taxa because it was inherited from a **common ancestor** than it is to be common because of **homoplasy**
1. An algorithm is used to determine the minimum number of steps necessary for any given tree (i.e., any given branching order) to be consistent with the data
 2. This minimum number of steps is the score for the tree, and the tree or trees with the lowest score are the most parsimonious trees

Parsimony method may recover more than one tree with the lowest score

Methods of constructing a phylogenetic tree

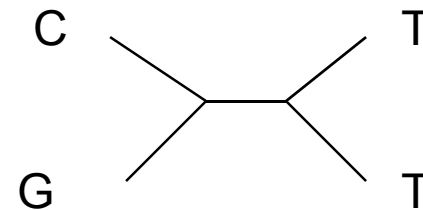
Maximum likelihood (ML)

- ML tries to infer an evolutionary tree by finding that tree which maximizes the probability of observing the data
- The process begins with a ***model of sequence evolution***
(An evolutionary model gives the instantaneous rates at which each of the 4 possible nucleotides changes to each of the other three nucleotides)

e.g.: 4 taxa with C, T, T and G at a certain position

Taxa 1: CGCATT**C**GCGTA
Taxa 2: CGCATT**T**GCGTA
Taxa 3: CGCATT**T**GCGTA
Taxa 4: CGCATT**G**GCGTA

One possible unrooted tree for these 4 taxa:

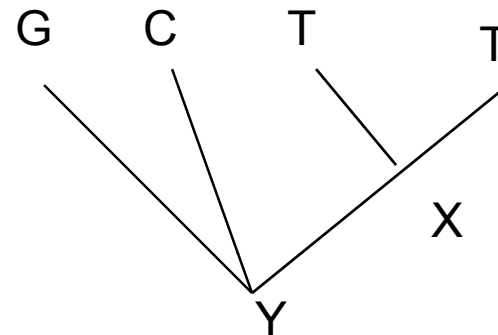


Methods of constructing a phylogenetic tree

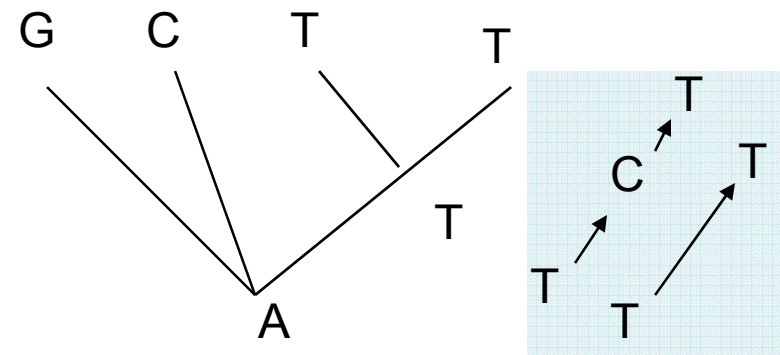
Maximum likelihood (ML)

Rooting the tree at any node:

There would be 4 possibilities for each X and Y nodes, i.e., 16 total scenario



One of these 16 scenarios would be:



Thus,

$$\text{Probability, } P_{\text{scenario1}} = P_A \times P_{AG} \times P_{AC} \times P_{AT} \times P_{TT} \times P_{TT}$$

Methods of constructing a phylogenetic tree

Maximum likelihood (ML)

Probability of each scenario must be determined and added to obtain the probability of the tree for a particular site,

$$P_{site1} = P_{scenario1} + P_{scenario2} + \dots + P_{scenario16}$$

The probability of observing all of the data at all of the sites would be the product of probabilities of each of each of the site i from 1 to N

$$P_{tree} = \prod_{i=1}^N P_i$$

As these numbers are much smaller, probability (or likelihood) for each site i is expressed as log likelihood, $\ln L_i$

$$\text{Thus, } \ln L_{tree} = \sum_{i=1}^N \ln L_i$$

Methods of constructing a phylogenetic tree

Bayesian analysis

- Based on the notion of ***posterior probabilities***: probabilities that are estimated, based on some model (***Prior expectations***), after learning something about the data
- Similar to ML in that in both, a model of evolution is postulated, and the best trees are searched that is consistent with both the model and the data (sequence alignment)
- Differ from ML in that while the ML seeks for tree that maximizes the possibility of observing the data given that tree, Bayesian analysis seeks the tree that maximises the probability of the tree given the data and the model of evolution.
- Unlike ML which searches for the single most likely tree, Bayesian analysis searches for the *best set* of trees.

Estimating the reliability of phylogenetic tree

“Reliability” often refers to the tree topology

It is the probability that the members of a given clade are always the members of that clade

Bootstrapping – a method of pseudorepeating the data collection for testing the reliability of a tree

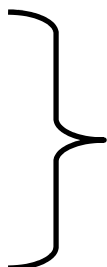
- A random site (i.e., a column) is taken from the sequence alignment and used as the first site, another random site is used as the second site and so on till the same number of sites as the original alignment is achieved – i.e., *sampling with replacement*
- A tree is constructed from the pseudo-alignment
- The original tree is compared with the new tree
- Every clade is scored on the basis of the presence/absence in the two trees compared and the process is repeated

Sampling for Phylogeny – Choice of marker

Protein?...DNA...?

DNA as the marker of Choice due to the advent of modern high-throughput **Polymerase Chain Reaction (PCR)** and **DNA sequencing** methods

Selecting a DNA fragment from the genome for phylogeny reconstruction:

- Coding or non-coding sequence?
 - Nuclear or mitochondrial sequence?
 - Nuclear + mitochondrial sequence?
- 
- Depend on whether we need to reconstruct phylogenies of closely or distantly related taxa

Sampling for Phylogeny – What & How many samples

Taxon sampling – which taxa to include or exclude in a phylogenetic analysis

- This is an important aspect, as given a sample of taxa and a method of analysis, certain properties of the data could render the reconstruction of the "true" tree difficult or even impossible
- There is a risk of an investigator bias when deciding which terminal taxa to include or exclude in an analysis

Sampling for Phylogeny – What & How many samples

Adding more taxa or more samples to infer correct tree?

- In general science, assumption is that more data leads to a stronger hypothesis
- However, this may not always be true for a phylogenetic tree, given a particular model of evolution and taxa under consideration
- *For a phylogenetic tree, what should we add more— **Character** or **taxa**?*

The answer depends on two aspects:

- i. The **signal to noise ratio** in the sequence used
- ii. **Branch length heterogeneity** of the tree

Sequencing for phylogeny

1st step: selecting one/ multiple suitable DNA fragment

- Mitochondrial Protein-coding genes, such as a Cytochrome b, 16s rRNA gene, NADH2 etc. are often useful in constructing phylogeny for closely related taxa, such as **intra-species** or **inter-species** phylogeny
- However, to resolve a tree at the deeper nodes, such as at family level, nuclear protein-coding genes may be useful
- This is because, mitochondrial genes evolve faster than the nuclear genes, accumulating greater variation in the sequences
- Mitochondrial gene sequences get saturated in a shorter time span compared to the nuclear genes

Sequencing for phylogeny

2nd step: Laboratory strategy for obtaining sequences from the available samples

1. An optimized method for DNA extraction from the samples (e.g., variety of samples such as blood, muscle tissue, bone, samples from ancient museum specimen etc.)
2. Suitable PCR strategy
primer designing
single or multiple PCR reaction amplification (primer walking)
3. Sequencing strategy – direct sequencing or vector cloning

Sequencing for phylogeny

3rd step: detecting pseudogene copies in a mitochondrial DNA sequence

1. Analyzing sequences “in frame” for any intermediate stop codons – characteristics of nuclear copy of a mitochondrial gene (numt)
2. Discarding any numt sequences from the dataset or re-sequencing, may be a different PCR strategy

Suggested readings:

1. **Molecular Evolution and Phylogenetics** by Masatoshi Nei and Sudhir Kumar
2. **Phylogenetic trees made easy – a how to manual** by Barry G. Hall
3. **Evolution** by Futuyama
4. **Population Genetics – A concise Guide** by John H. Gillespie

Web-resources:

1. Phylogeny programs –
<http://evolution.genetics.washington.edu/phylip/software.html>

Questions before we proceed??

Practicals to be covered

1. Sequence search in NCBI database for taxa of interest
2. Multiple sequence alignment via **ClustalW** (in **Bioedit**) and **MUSCLE** (web-based)
3. Searching for numt in the practice dataset (to be provided) using **Transeq**(Web-based)
4. Using MEGA to compute pairwise distances
 - i. p-distance
 - ii. Using Kimura 2-parameter model
 - iii. Using Tamura-Nei model
 - iv. Maximum likelihood distance

Practicals to be covered

5. Using **MEGA** to construct phylogenetic tree by
 - i. UPGMA
 - ii. NJ
 - iii. Maximum parsimony
6. Bootstrap test of phylogeny for above three methods in **MEGA**
7. Introduction to Bayesian tree construction using **BEAST** software (Optional)