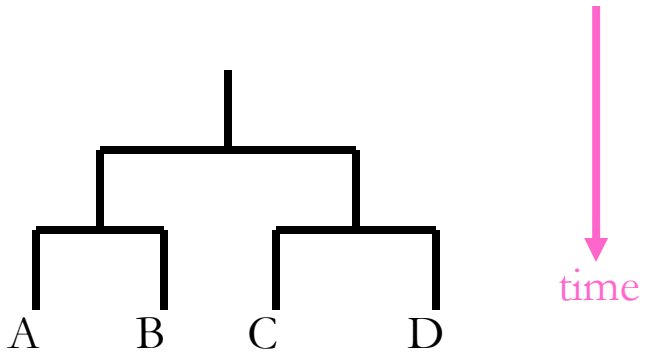
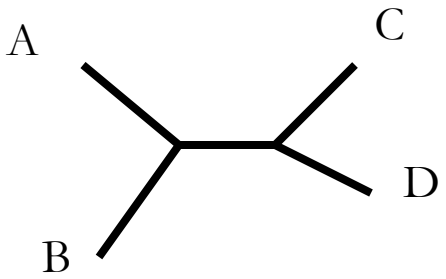
Phylogenetic Analysis

Dr Safdar Ali
Department of Biological Sciences
Aliah University, Kolkata
Date: 06/04/2020

Phylogenetic Trees



A rooted tree



An unrooted tree

time?

Bioinformatics & Functional Genomics 3rd Edition, 2015,
Jonathon Pevsner, Wily and Blackwell Publishing, Chapter 7

Kimura Neutral Theory

- Till 1960s, the model of molecular evolution was that most changes in genes are selected for or against. (Darwin).

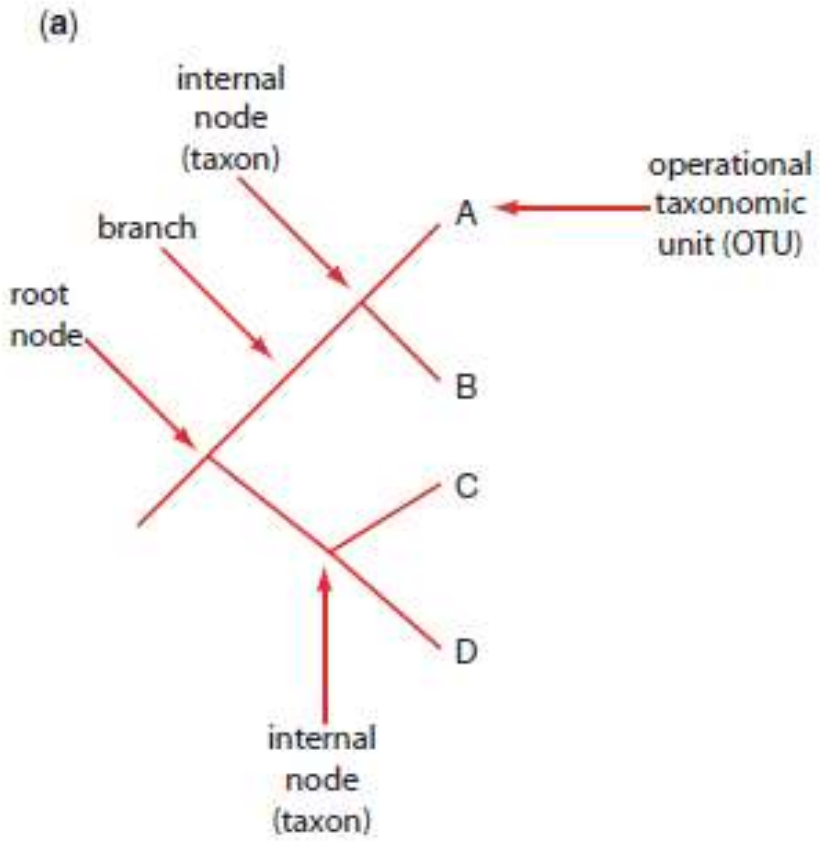
- Motoo Kimura (1968, 1983) proposed a different model to explain evolution at the DNA level. Kimura (1968) noted that the rate of amino acid substitution averages approximately one change per $28x10^6$ years for proteins of 100 residues.

- He further estimated that the corresponding rate of nucleotide substitution must be extremely high (one base pair of DNA replaced in the genome of a population every 2 years on average).

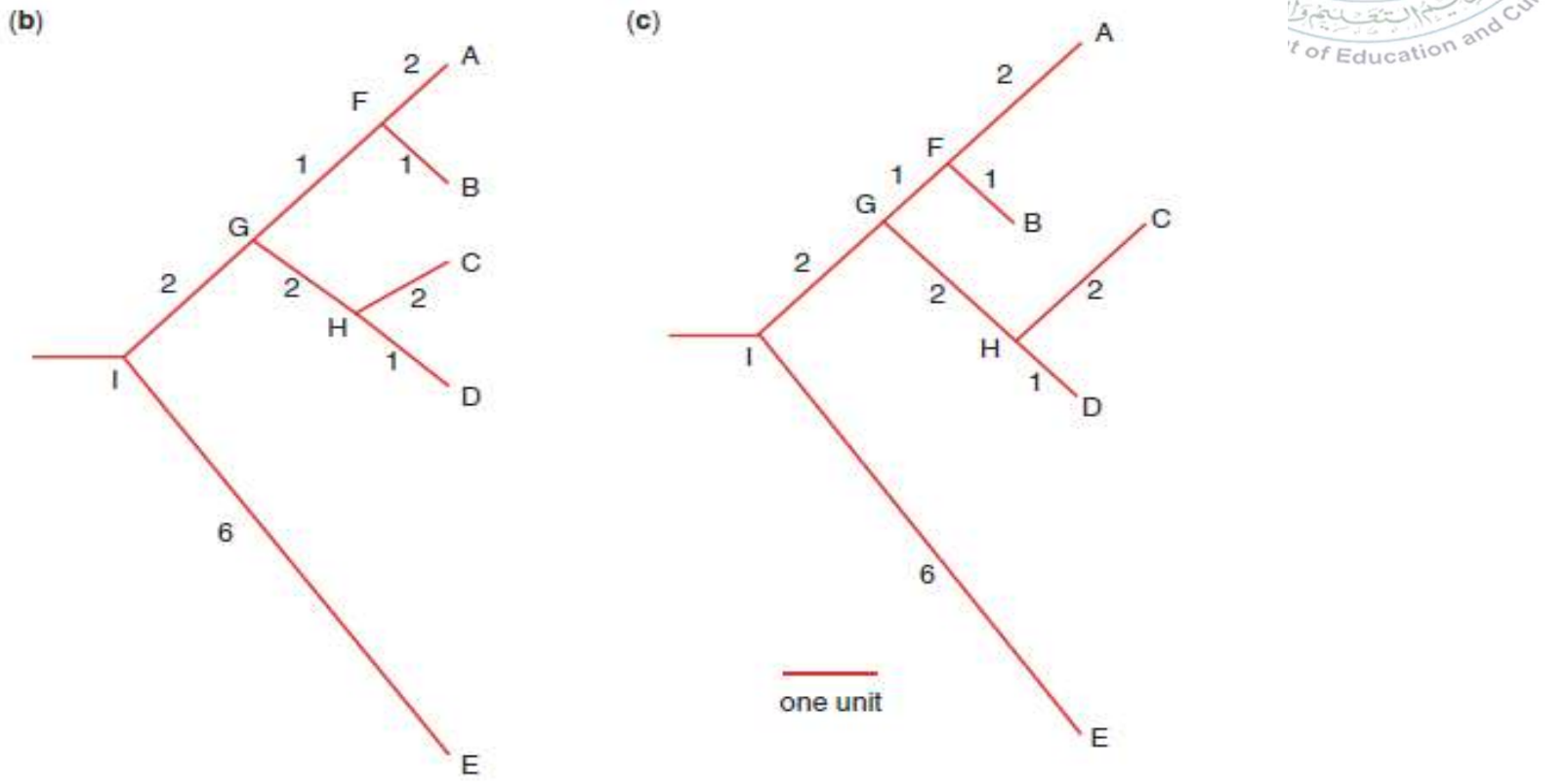
- Kimura's conclusion was that most observed DNA substitutions must be neutral or nearly neutral, and that the main cause of evolutionary change (or variability) at the molecular level is random drift of mutant alleles.

- Most nonsynonymous mutations are deleterious, and thus are not observed as substitutions in the population. Under this model, called the neutral theory of evolution, **positive Darwinian selection plays an extremely limited role**.

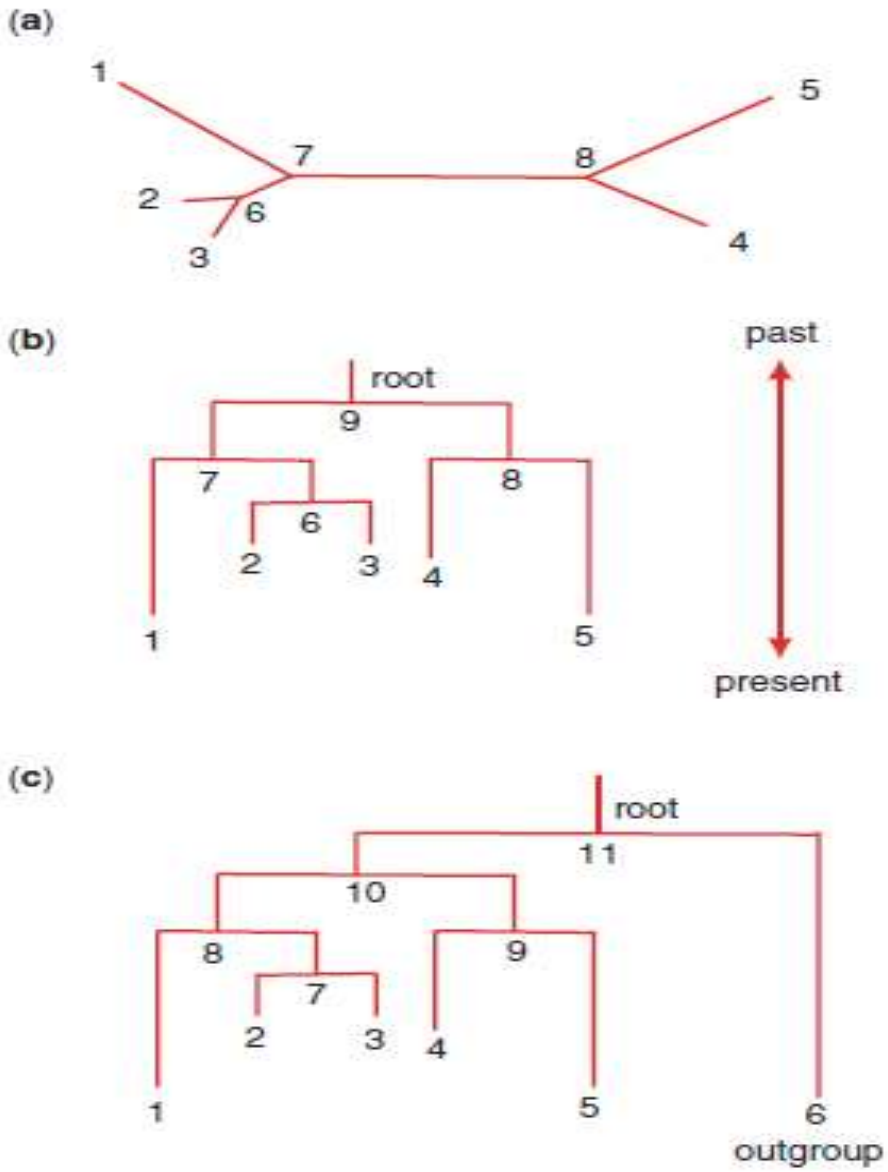
- Indeed, the existence of a molecular clock makes sense in the context of the neutral hypothesis because most amino acid substitutions are neutral. (If substitutions occurred primarily in the context of positive or negative selection, it is unlikely that they could account for clock-like evolution and hence substitutions are tolerated)



Bioinformatics & Functional Genomics 3rd Edition, 2015, Jonathon Pevsner, Wily and Blackwell Publishing, Chapter 7



Bioinformatics & Functional Genomics 3rd Edition, 2015,
Jonathon Pevsner, Wily and Blackwell Publishing, Chapter 7



Bioinformatics & Functional Genomics 3rd Edition, 2015, Jonathon Pevsner, Wily and Blackwell Publishing, Chapter 7

(a)

	A	C	T	G
A	0	1	1	1
C	1	0	1	1
T	1	1	0	1
G	1	1	1	0

(b)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	2	1	1	2	1	2	2	2	2	2	2	1	2	2	1	1	1	2	2
C		0	2	3	1	1	2	2	3	2	3	2	2	3	1	1	2	2	1	1
D			0	1	2	1	1	2	2	2	3	1	2	2	2	2	2	1	3	1
E				0	3	1	2	2	1	2	2	2	2	1	2	2	2	1	2	2
F					0	2	2	1	3	1	2	2	2	3	2	1	2	1	2	1
G						0	2	2	2	2	2	2	2	2	1	1	2	1	1	2
H							0	2	2	1	3	1	1	1	1	2	2	2	3	1
I								0	1	1	1	1	2	2	1	1	1	1	3	2
K									0	2	1	1	2	1	1	2	1	2	2	2
L										0	1	2	1	1	1	1	2	1	1	2
M											0	2	2	2	1	2	1	1	2	3
N												0	2	2	2	1	1	2	3	1
P													0	1	1	1	1	2	2	2
Q														0	1	2	2	2	2	2
R															0	1	1	2	1	2
S																0	1	2	1	1
T																	0	2	2	2
V																		0	2	2
W																			0	2
Y																				0

Bioinformatics & Functional Genomics 3rd Edition, 2015,
Jonathon Pevsner, Wily and Blackwell Publishing, Chapter 7

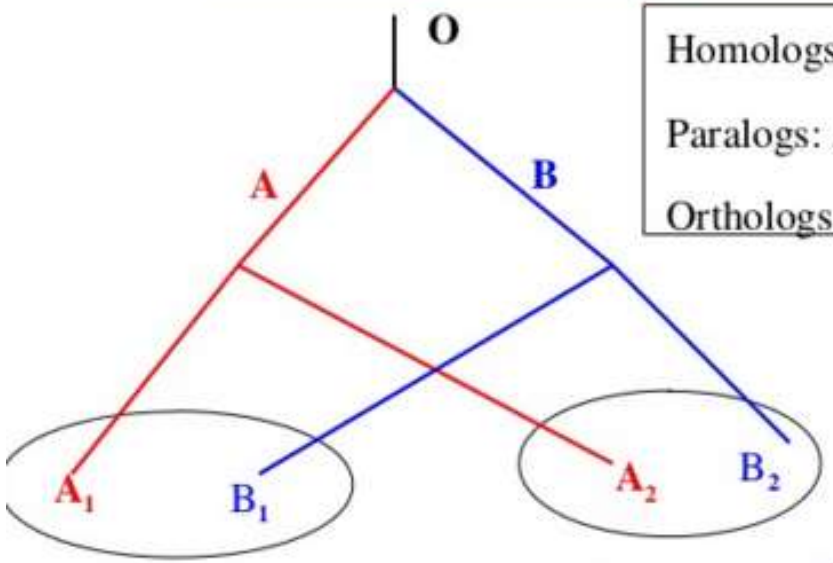
Homolog, Ortholog, Paralog

- Homologous genes that share a common ancestry and function in the absence of any evidence of gene duplication are called orthologs.

- When there is evidence for gene duplication, the genes in an evolutionary lineage derived from one of the copies and with the same function are also referred to as orthologs.

- The two copies of the duplicated gene and their progeny in the evolutionary lineage are referred to as paralogs.

Homolog - Paralog - Ortholog

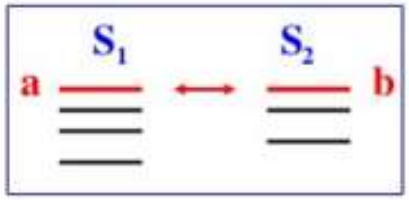


Homologs: A_1, B_1, A_2, B_2

Paralogs: A_1 vs B_1 and A_2 vs B_2

Orthologs: A_1 vs A_2 and B_1 vs B_2

Sequence analysis



Assumptions for Phylogenetic Analysis

- Sequences in the sample evolved according to a single stochastic process
- All positions in the sequence evolved according to the same stochastic process
- Each position in the sequence evolved independently

Steps

- Multiple Sequence Alignment
- Substitution model
- Tree building
- Tree evaluation

Distance Based Methods

- Distance-based methods begin the construction of a tree by calculating the pairwise distances between molecular sequences (Felsenstein, 1984; Desper and Gascuel,2006).

- A matrix of pairwise scores for all the aligned proteins (or nucleic acid sequences) is used to generate a tree. The goal is to find a tree in which the branch lengths correspond as closely as possible to the observed distances.

- The main distance-based methods include the unweighted pair group method with arithmetic mean (UPGMA) and neighbor joining (NJ).

- Distance-based methods of phylogeny are computationally fast, and thus they are particularly useful for analyses of a larger number of sequences (e.g., .50 or 100).

Bioinformatics & Functional Genomics 3rd Edition, 2015,
Jonathon Pevsner, Wily and Blackwell Publishing, Chapter 7

Character Based Methods

- The main idea behind maximum parsimony is that the best tree is that with the shortest branch lengths possible (Czelusniak et al., 1990).

- According to maximum parsimony theory, having fewer changes to account for the way a group of sequences evolved is preferable to more complicated explanations of molecular evolution.

- The assumption of phylogenetic systematics is that genes exist in a nested hierarchy of relatedness, and this is reflected in a hierarchical distribution of shared characters in the sequences.

- The most parsimonious tree is supposed to best describe the relationships of proteins (or genes) that are derived from common ancestors.

Bioinformatics & Functional Genomics 3rd Edition, 2015,
Jonathon Pevsner, Wily and Blackwell Publishing, Chapter 7

Thank You