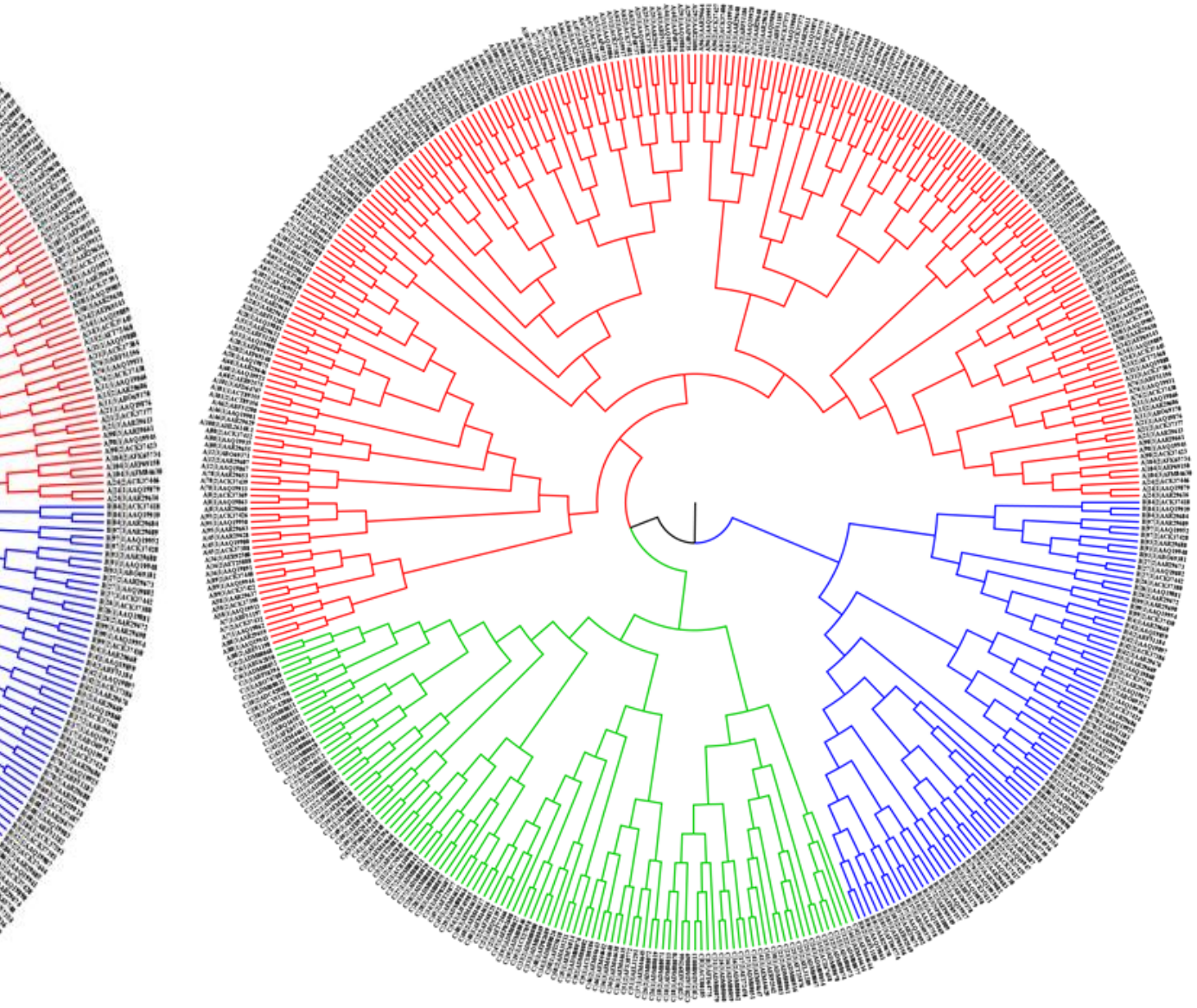
Introduction to Phylogenetic Methods

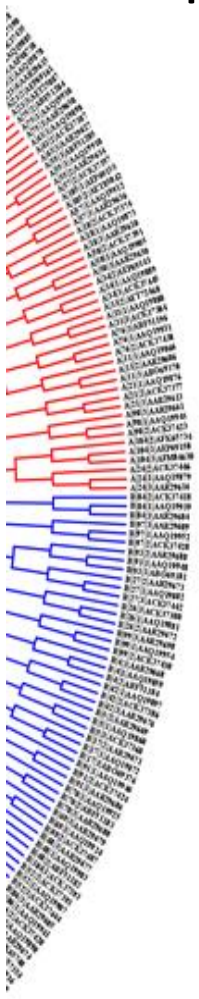
Dr. Urmila Kulkarni-Kale
S. P. Pune University| University of Southeastern Norway | Citadel Precision Medicine
urmila.Kulkarni.kale@gmail.com



"Nothing in biology makes sense except in the light of evolution"

Prof. Theodosius Dobzhansky, 1973

Phylogenetics & Viral typing:
The starting point for characterization of virus isolates



- Classification
- Evolution
- Diversification
- Epidemiology
- Diagnostics
- Vaccine design
- Drug design

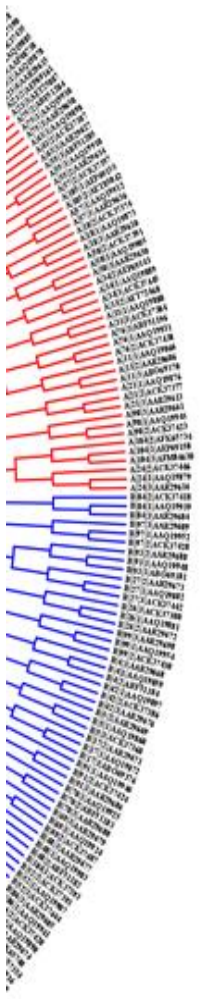
Methods used

- Alignment-based
- Alignment-free

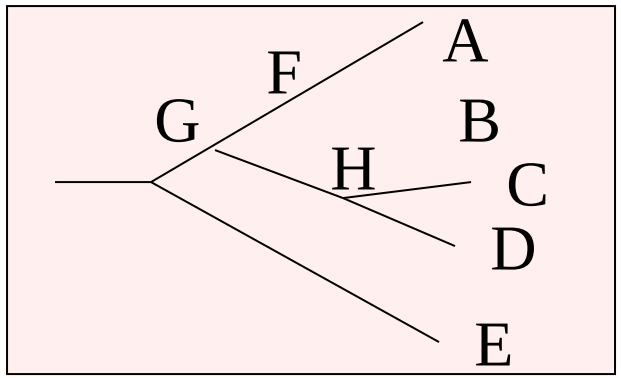
Regions used

- Individual genes
- Genomic regions
- Whole genomes

Phylogenetic tree: Concepts & Terminologies

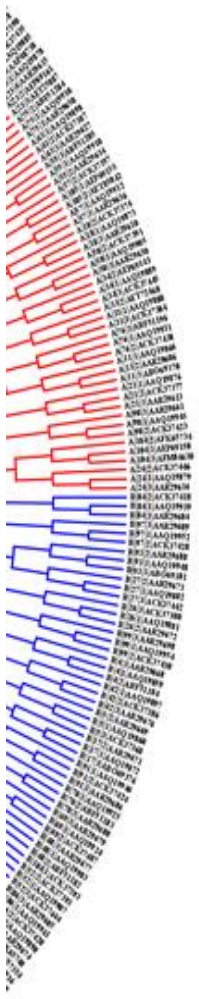


- Phylogenetic tree: a graph that illustrates evolutionary relationships using nodes and branches
- Nodes: represents OTUs
- OTUs: operational taxonomic units
- Branches: define relationships
- Topology: a branching pattern
- Branch length: represents number of changes that have occurred in that branch
- External/terminal nodes: OTUs
- Internal nodes: common ancestors
- Scaled branches: lengths proportional to number of changes
- Un-scaled branches: uniform branch lengths



- Monophyletic taxa: derived from a single common ancestors → A & B [derived from F]
- Polyphyletic taxa: derived from more than one common ancestor → A & C [derived from F & H]
- Paraphyletic taxa: derived from common ancestor but the group doesn't include all descendent taxa of the common ancestor → A, C & E
- Clad: A monophyletic group that includes all descendent species of the common ancestor

Types of trees



Cladogram vs. Phylogram

- Cladograms have uniform branch lengths and only represents relationships.
- Phylograms have length proportional to change or distance

Rooted vs. Unrooted

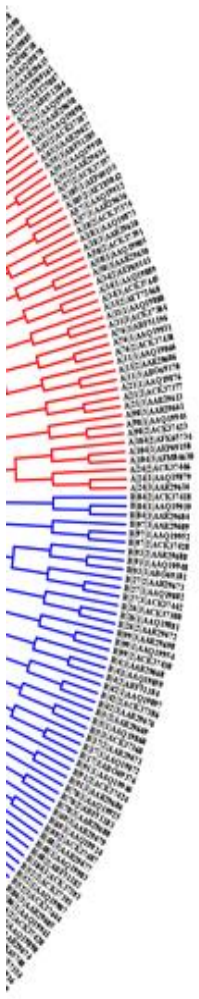
- Rooted tree has a defined origin as opposed to a network of relationships in unrooted tree.
- Most trees are unrooted

Artistic

- Slanted, rectangular, circular

- Species tree: represents evolutionary pathways of a group of species
- Gene tree: constructed using single gene from each species
- GT can differ from ST in two ways:
  - Divergence of two genes sampled from two species may predate divergence of two species
  - The branching pattern (Topology) of gene tree may be different from that of the species

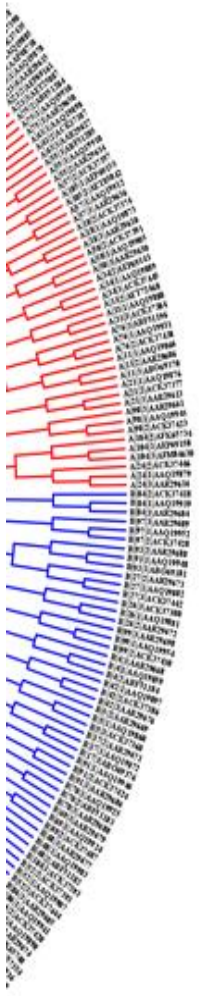
Workflow: Molecular Phylogeny Analysis (MPA) What, When, Why & How?



- Define the objective
- Curate a set of reference sequences (known types)
- Carry out Multiple Sequence Alignments (MSA)
- Undertake phylogenetic analysis:
  - distance-based (NJ: Neighbor-Joining)
  - character-based (MP: Maximum parsimony)
  - Bayesian-based (ML: Maximum likelihood)
- Generate datasets for bootstrapping
- Use clustering method
- Generate consensus tree
- Assess tree topology(ies)
- Analysis of inferred tree(s)

Molecular Phylogeny Analysis (MPA):

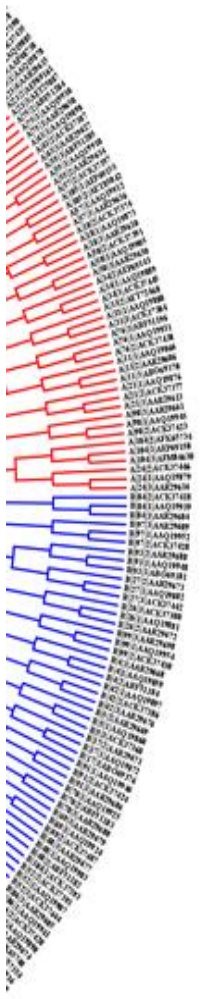
permits study of similarities within the group and differences between the groups



- Integral part of sequence-based bioinformatics analysis
- Applications:
  - Evolution of gene(s) in a group of species
  - Evolution of species
  - Assignment of species in the taxonomic hierarchy
  - Assignment of genotype/serotype, strains
  - Study of novel properties (drug resistance)

- Alignment-based methods
- Distance-based methods
- Character-based methods
- Bayesian methods

Premise for distance-based methods
Sequence-alignment → Distance calculations



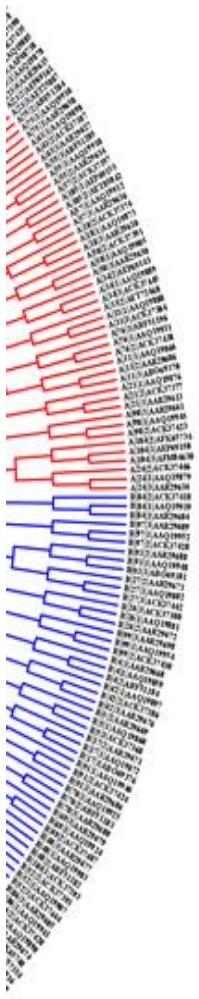
- SEQ1 ACGTACGTAA
- SEQ2 ACGTTCGTAT
- SEQ3 TCCATCGTAA

Similarity	Distance
(1-2) 80%	1-0.8=0.2
(1-3) 60%	1-0.6=0.4
(2-3) 60%	1-0.6=0.4

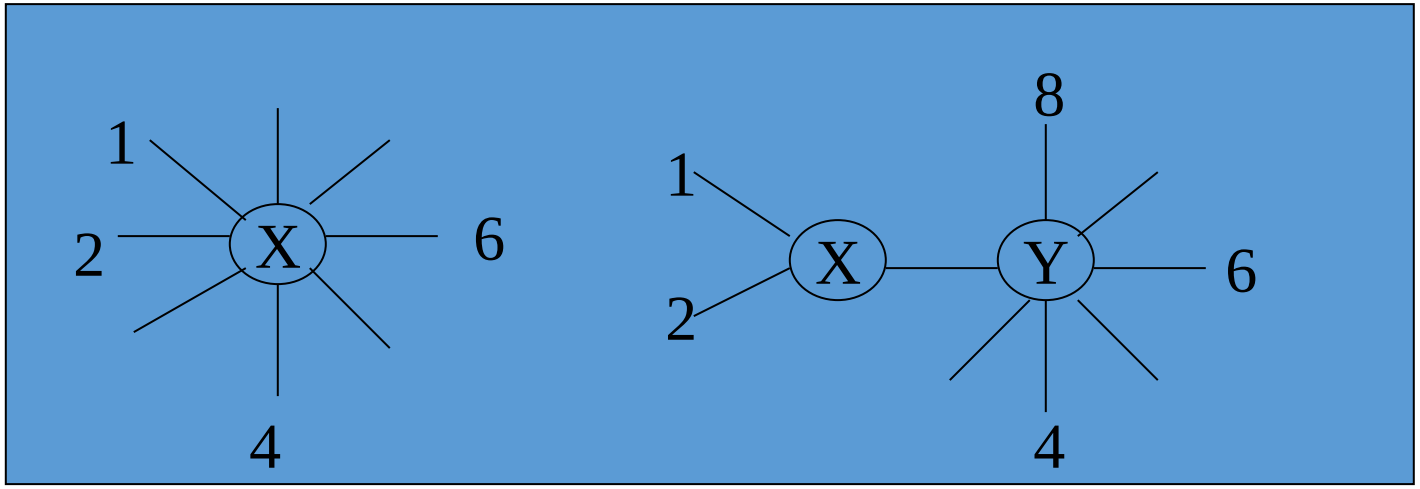
Distance matrix

	S1	S2	S3
S1	0		
S2	0.2	0	
S3	0.4	0.4	0

Neighbors-joining Method

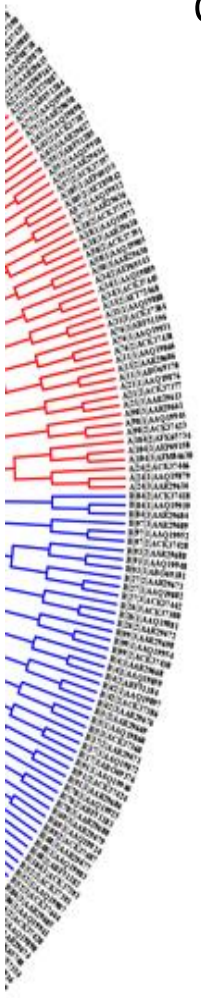


- Alignment-based & Distance-based method
- Find neighbors sequentially that may minimize the total length of the tree
- This method starts with a star-like tree with absence OTUs clusters
- Separate most similar pair of OTUs (1 & 2) from others by deriving one interior branch X that connects nodes X & Y (Y: common node for remaining OTUs



Premise for character-based method:

A nucleotide site is phylogenetically informative if it favors a subset of trees over other possible trees



- S1: A A G A G T T C A
- S2: A G C C G T T C T
- S3: A G A T A T C C A
- S4: A G A G A T C C T

Invariant sites

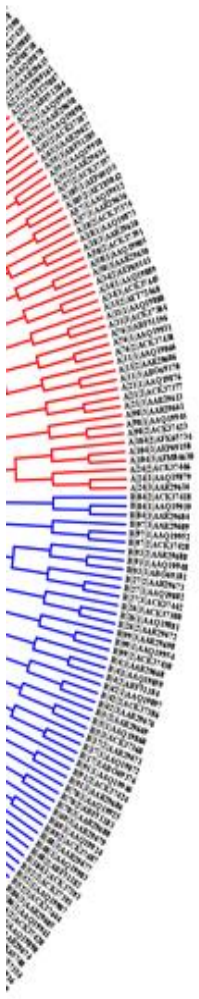
Informative variable sites

Variable sites

Uninformative variable sites

Singleton

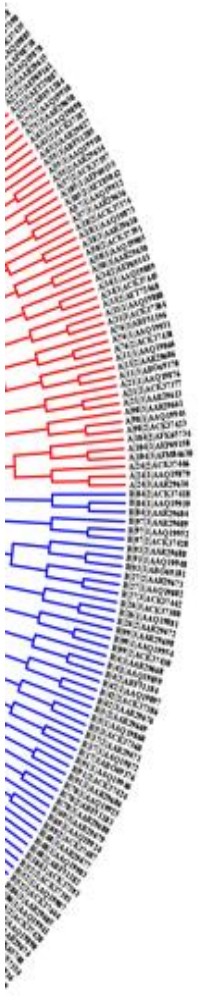
Maximum Parsimony Methods

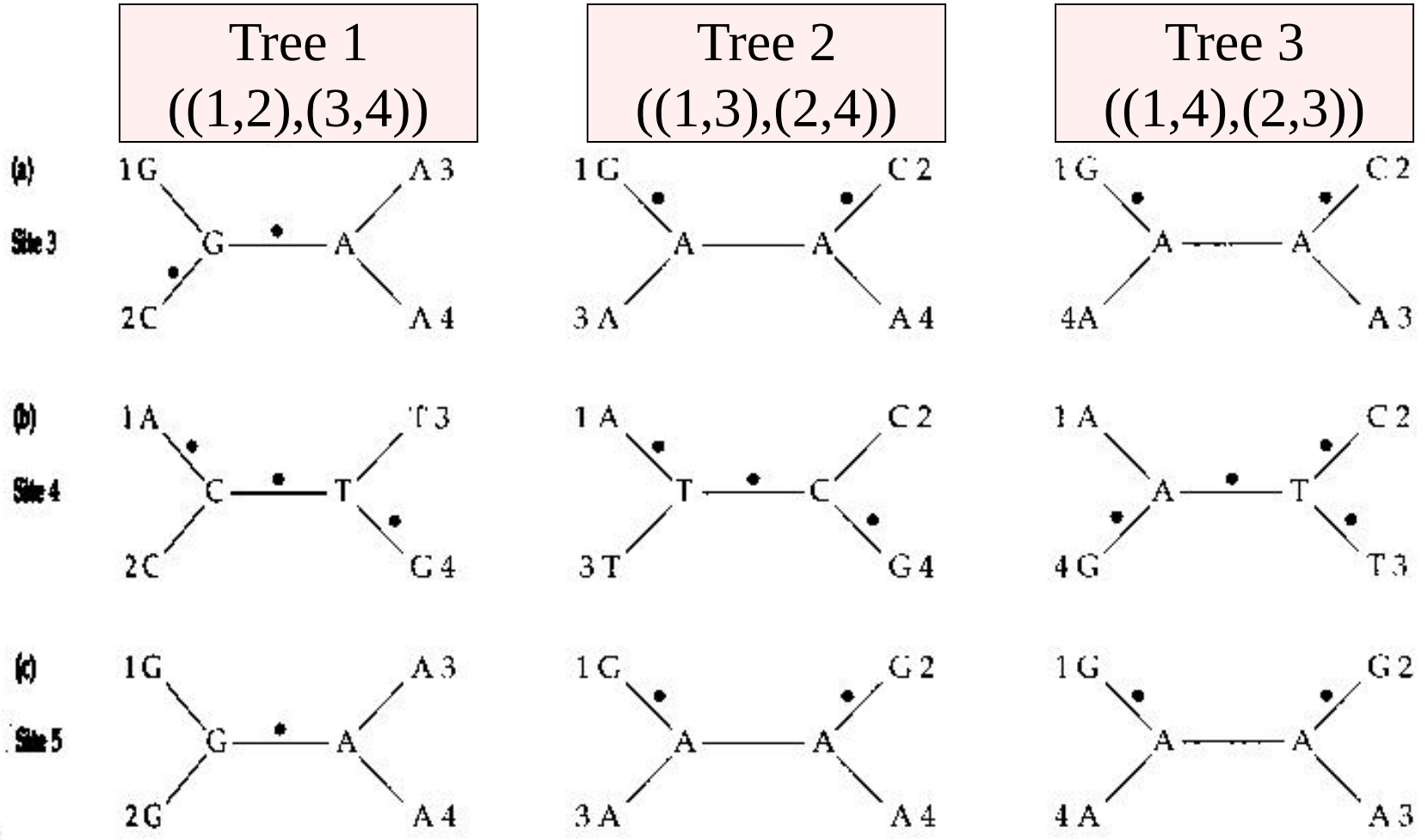


- Originally developed for protein sequences
- Principle: Identify of the Topology that requires the smallest number of evolutionary changes (substitutions) to explain the observed differences amongst the OTUs
- The tree that uses discrete character sets and shortest pathway leading to these character sets is the best tree and called a maximum parsimony tree
- If 2 or more trees are found and no unique tree can be inferred, trees are said to be equally parsimonious

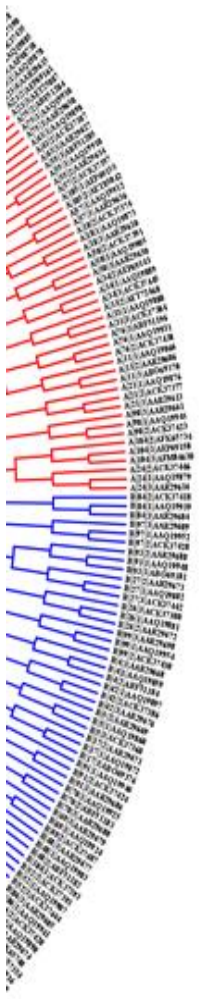
3 possible unrooted trees for 4 DNA sequences

• : inferred substitution on the branch



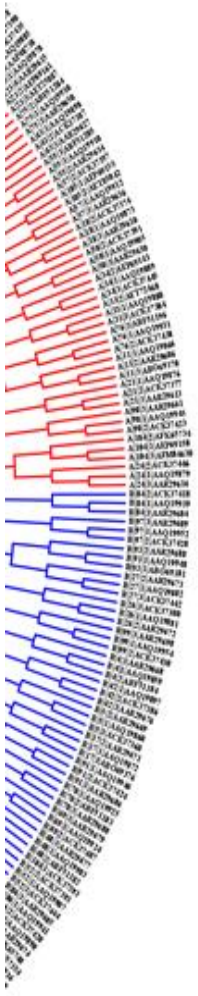


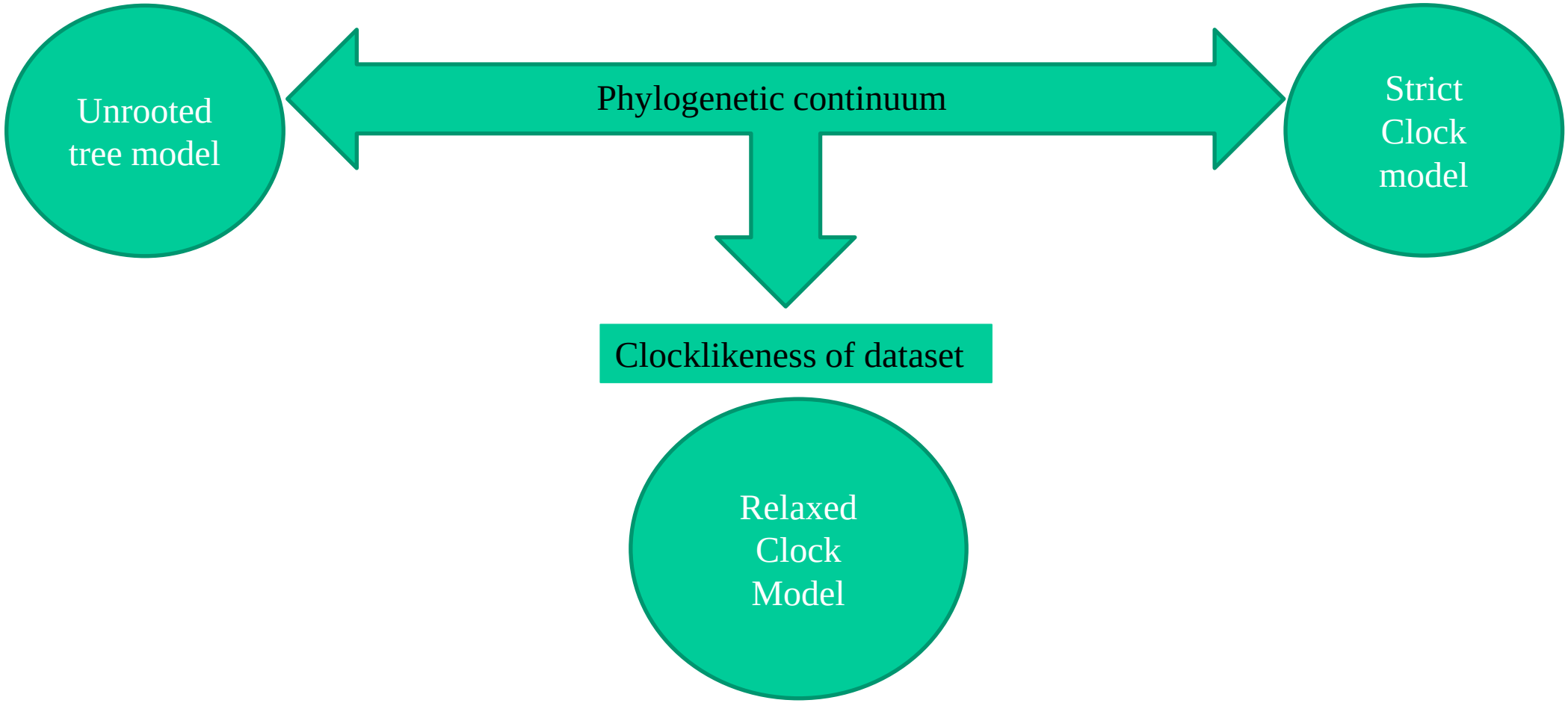
Maximum Likelihood Method

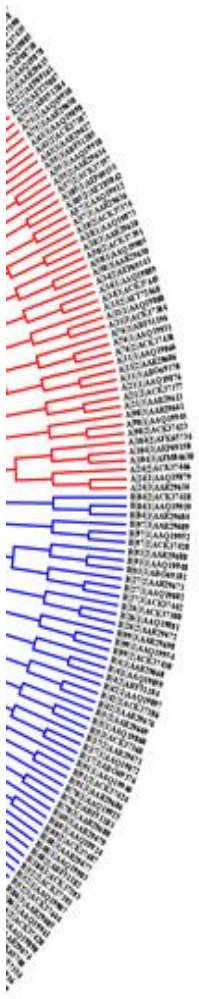


- The likelihood (L) of a phylogenetic tree is the probability of observing the data under a given tree and specified model of character state changes.
- L = P(data|Tree)
- Find the tree (amongst available) with highest L value

Bayesians models:
Account for variations in substitution rate



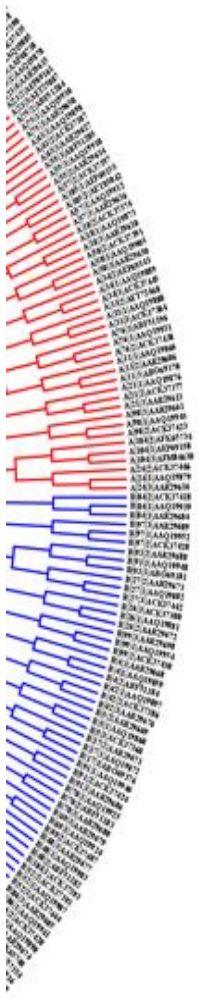




NSRs estimation using BEAST

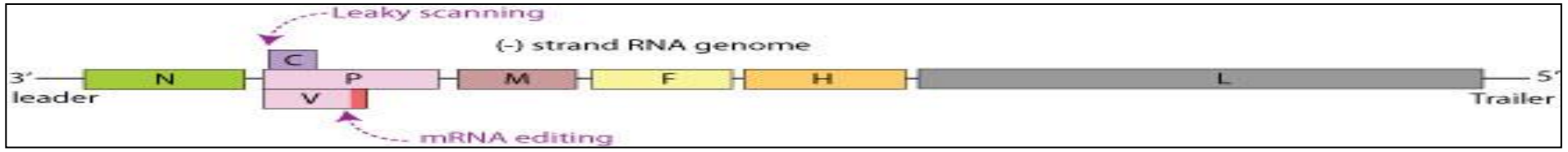
- RNA viruses are prone to errors due to replication machinery
- Derived using Maximum Likelihood (ML) method in molecular phylogeny
- NSR is a function of rate of mutation & rate of replication of virus
- Mutation rates indicate the rate at which errors are made during genome replication
- NSR indicate the rate at which evolution proceeds at the molecular level, replacing pre-existing alleles by new mutations
  - Influenza A viruses (-ve sense ssRNA): 1.8 × 10-3 substitutions per site per year
  - Human enterovirus 71 (+ ve sense ssRNA): 3.4 x 10-3 s/s/y
  - SARS-CoV-2 virus (+ve sense ssRNA): 7.8×10^{-4} s/s/y (range, 1.1×10^{-4} - 15×10^{-4})

tMRCA estimation using BEAST



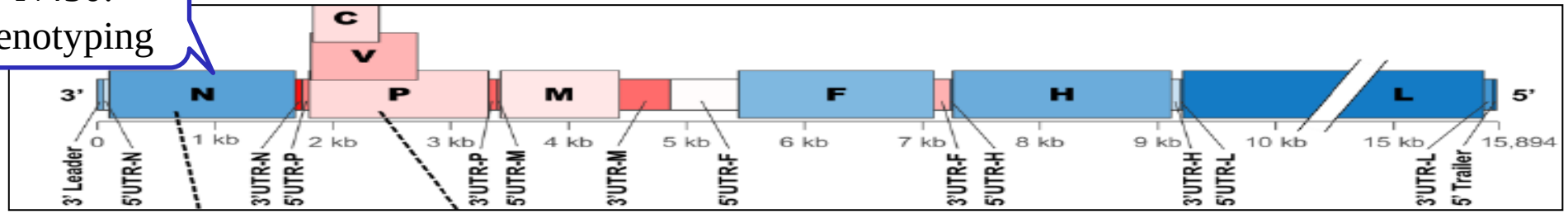
- Evolutionary rates, time of the most recent common ancestor (tMRCA) and demographic growth
- Bayesian framework using a Markov chain Monte Carlo (MCMC) method implemented in v.1.8.4 of the BEAST package
- Different coalescent priors and molecular clock models (Strict & Relaxed)
  - Constant population size;
  - Exponential growth &
  - Bayesian skyline plot (BSP)
- SARS-CoV-2 virus:
  - A mean tMRCA of the tree root of 73 days
  - Estimated *R* value was 2.6 (range, 2.1-5.1)
  - The estimated mean doubling time of the epidemic: 3.6 and 4.1 days

Nucleotide Substitution Rate (NSR) estimation: Measles virus



Genome Organization of MeV

N450: genotyping



Tolerance for insertional mutagenesis in different regions of the MeV genome (Beaty & Lee, 2016)

Genomic region: Measles virus of global isolates	Nucleotide substitution rate (substitutions per site per year)
N-450	7.11×10^{-4}
M-F intergenic region	**1.68×10^{-3}**
H-gene	5.94×10^{-4}
F-gene	2.63×10^{-4}
Complete genome	4.77×10^{-4}

Better candidate for genotyping

Vaidya et. al., 2020

Tracing the origin of the outbreak

Emerg Infect Dis. 2022 Apr; 28(4): 725–733.
doi: 10.3201/eid2804.211845

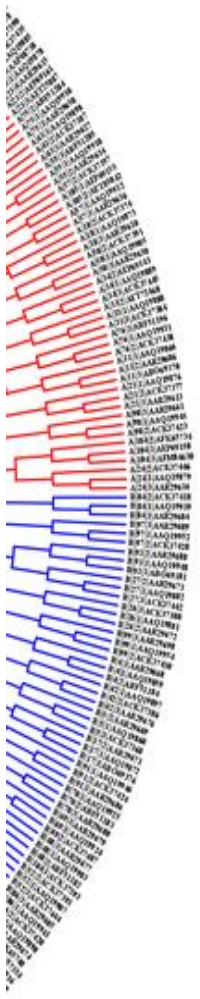
PMCID: PMC8962895
PMID: 35318918

Phylogenetic Analysis of Spread of Hepatitis C Virus Identified during HIV Outbreak Investigation, Unnao, India

Arati Mane, Sunitha Manjari Kasibhatla, Pallavi Vidhate, Vandana Saxena, Sandip Patil, Amrita Rao, Amit Nirmalkar, Urmila Kulkarni-Kale, and Samiran Panda

Approach: Sequencing & phylogenetic analysis of NS5 and Core regions of HCV isolates

Exploring the unknown of (un)known
Use curated data & right tool set(s)



Recombination analysis

- Detection of recombination events
- Identify major & minor parent
- RDP4 package (RDP, ENCONV, BOOTSCAN , MAXICHI , CHIMAERA , SiScan & 3SEQ)
- Input: MSA of genes/genomes

Population structure

- Identify genetically distinct lineages
- Map emerging lineages
- Identify admixed individuals
- Bayesian-based clustering approach
- STRUCTURE software
- Input: Gene/Complete genomes; PI sites
- LD (LIAN & DNASP)

Phylogenetic analysis

Nucleotide Substitution Rate

Divergence Time Estimation

- Evolutionary analysis
- Gene based
- Genome based
- Alignment-based (NJ, ML)
- Alignment-free (RTD: Return Time Distribution; developed in house)
- Geno- and Sero-Typing (RTD based)
- Input: MSA/Sequences

Selection pressure

Across all lineages

Subset of lineages

- Pervasive positive selection (SLAC, FEL, IFEL methods)
- Episodic positive selection (MEME method)
- Codon-based alignment
- Input: CDS of gene/genome
- Antigenic variation (experimental & predicted B- and T-cell epitopes