Escuela Profesional de
Ciencia de la Computación

Bioinformatics

Phylogenetic trees

MSc. Vicente Machaca Arceda

Universidad Nacional de San Agustín de Arequipa

June 6, 2021

Table of Contents

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Introduction

Objectives

- Understand the importance of phylogenetic trees.

vmachacaa@unsa.edu.pe

Introduction

Objectives

- Understand the importance of phylogenetic trees.
- Understand and implement UPGMA.

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Phylogenetics

What is evolution?

In the biological context, evolution can be defined as the development of a biological form from other preexisting forms or its origin to the current existing form through natural selections and modifications [1].

vmachacaa@unsa.edu.pe

Phylogenetics

Definition

Phylogenetics is the study of the evolutionary history of living organisms using tree like diagrams to represent pedigrees of these organisms [1].

Molecular phylogenetics is the study of evolutionary relationships of genes and other biological macromolecules by analyzing mutations at various positions in their sequences [1].

vmachacaa@unsa.edu.pe

Phylogenetics

Example

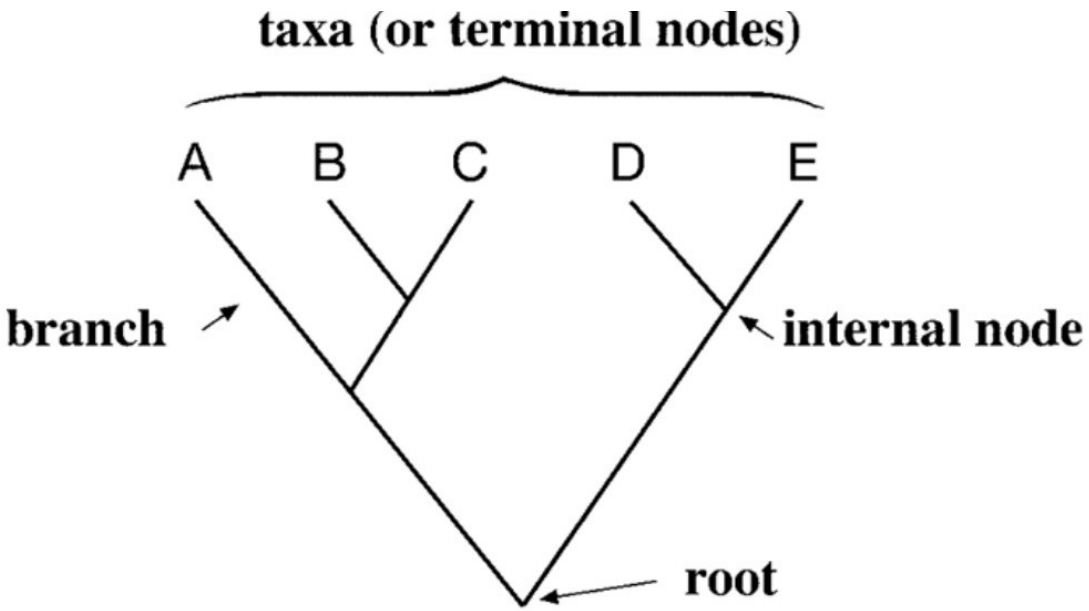


Figure: A typical bifurcating phylogenetic tree showing root, internal nodes, terminal nodes and branches. Source: [1]

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Major Assumptions

Major Assumptions:

- Molecular sequences used in phylogenetic construction are homologous, meaning that they share a common origin.
- Each position in a sequence evolved independently.

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Terminology

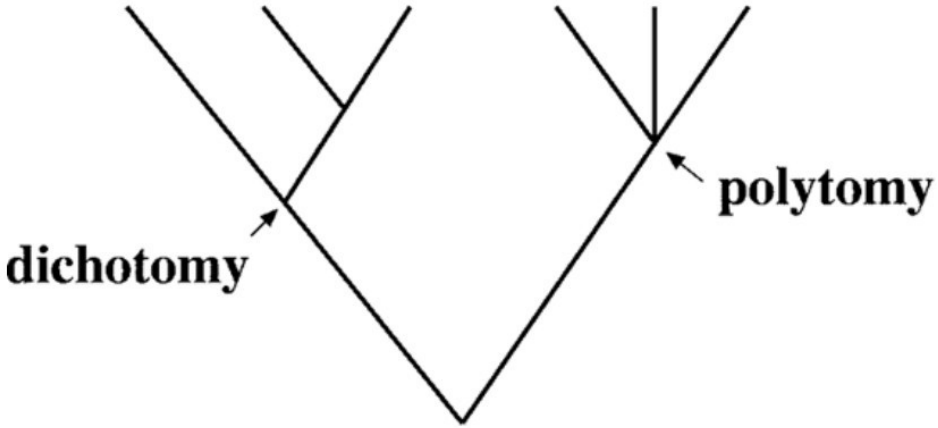


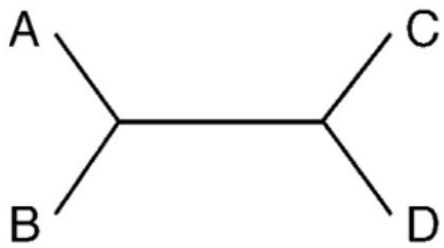
Figure: A phylogenetic tree showing an example of bifurcation and multifurcation. Source: [1]

vmachacaa@unsa.edu.pe

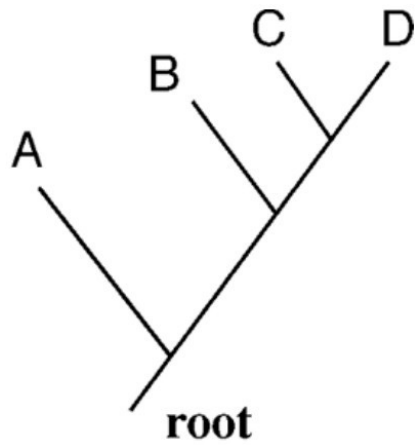
Table of Contents

vmachacaa@unsa.edu.pe

Rooted and unrooted



Unrooted



Rooted

Figure: An illustration of rooted versus unrooted trees. Source: [1]

vmachacaa@unsa.edu.pe

Rooted and unrooted

The root of the tree is not known; the common ancestor is already extinct [1].

There are two ways to define the root of a tree:

- **Outgroup**.- Which is a sequence that is homologous to the sequences under consideration, but separated from those sequences at an early evolutionary time.
- **Midpoint rooting approach**.- The midpoint of the two most divergent groups judged by overall branch lengths is assigned as the root.

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Gene versus species phylogeny

Gene phylogeny

Describes the evolution of that particular gene/protein. This sequence may evolve more or less rapidly than other genes or may have a different evolutionary history from the rest of the genome [1].

Species phylogeny

The species evolution is the combined result of evolution by multiple genes in a genome [1].

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Forms of representation

cladograms and phylograms

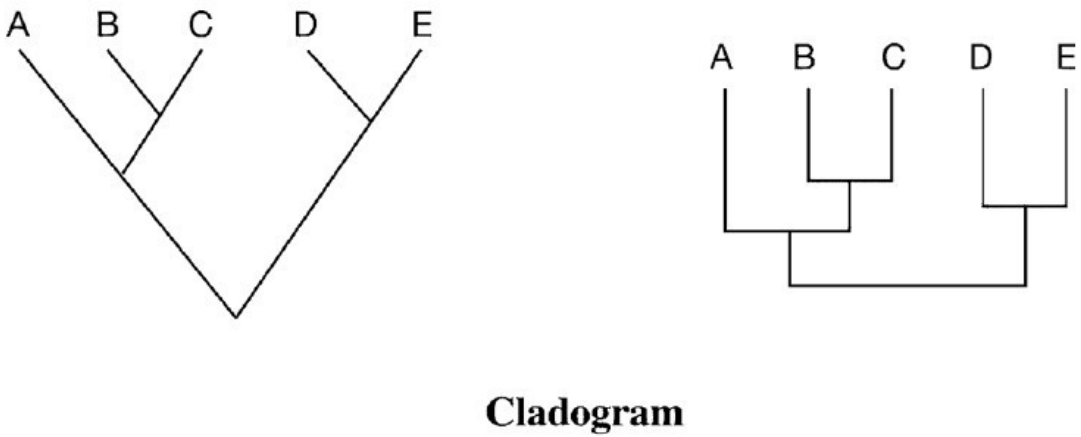
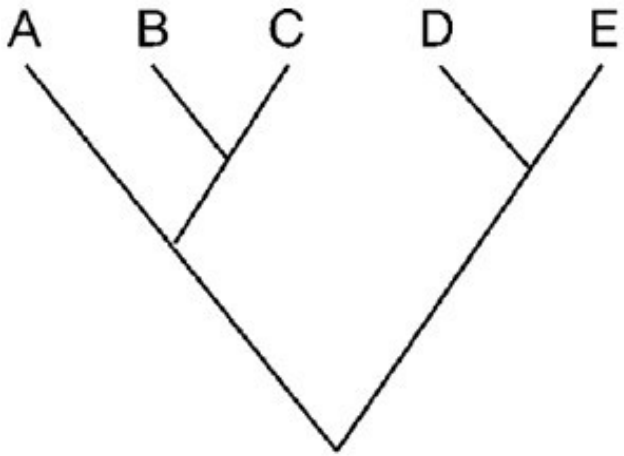


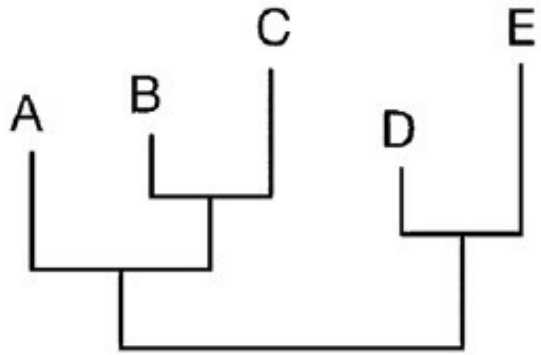
Figure: Phylogenetic trees drawn as cladograms (top) and phylograms (bottom). Source: [1]

vmachacaa@unsa.edu.pe

Forms of representation

Newick





(((B,C),A),(D,E))

(((B:1,C:2),A:2),(D:1.2,E:2.5))

Newick format

Figure: Newick format of tree representation. Source: [1]

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

The true tree

The search for a correct tree topology can sometimes be extremely difficult and computationally demanding. The number of potential tree topologies can be enormously large even with a moderate number of taxa [1].

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{1}$$

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{2}$$

where N_R and N_U are the number of rooted and unrooted trees, n is the number of taxa.

vmachacaa@unsa.edu.pe

The true tree

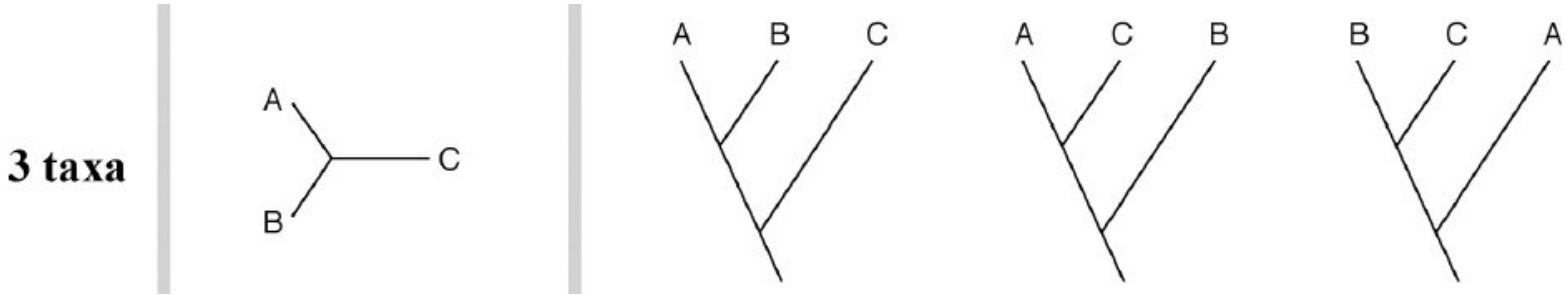


Figure: Unrooted and rooted trees for 3 taxa. Source: [1]

vmachacaa@unsa.edu.pe

The true tree

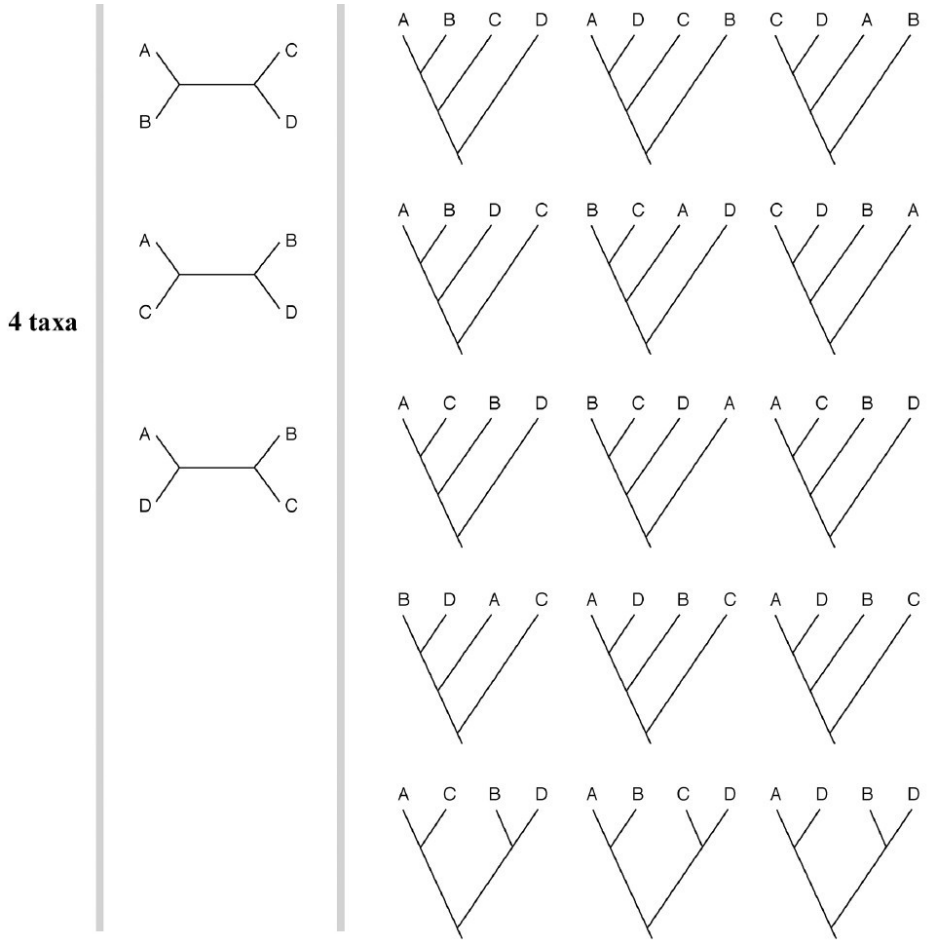


Figure: Unrooted and rooted trees for 4 taxa. Source: [1]

vmachacaa@unsa.edu.pe

The true tree

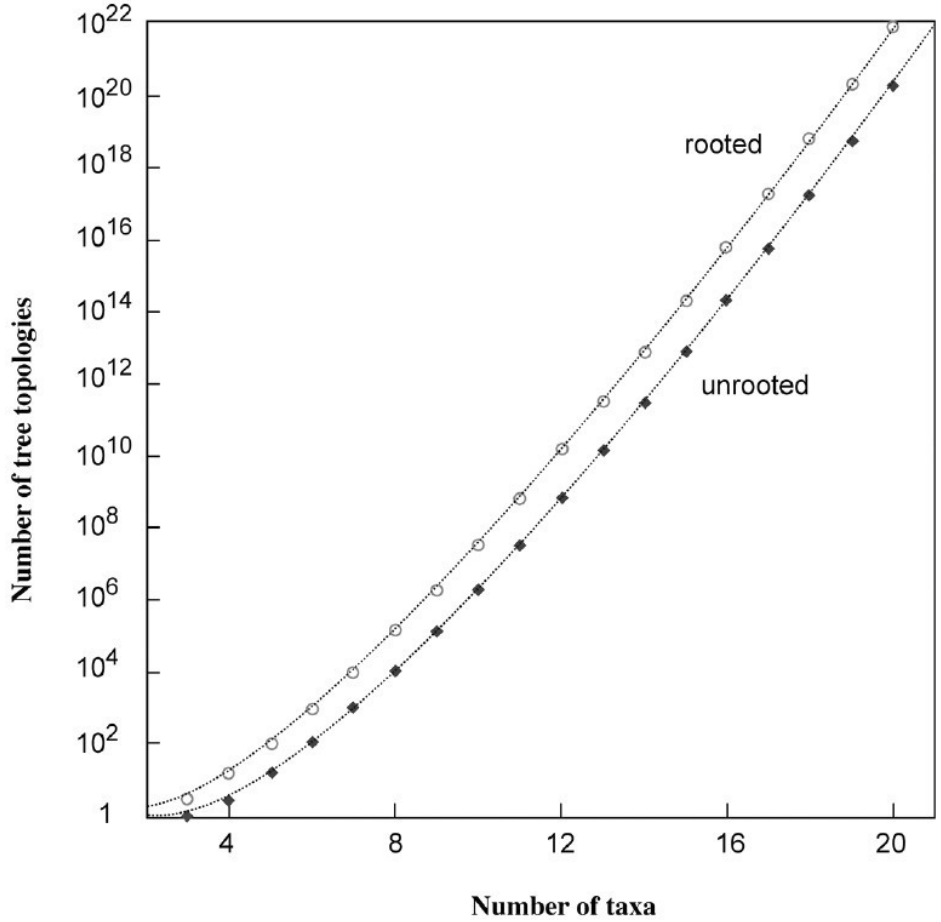


Figure: Total number of rooted (○) and unrooted (⧫) tree topologies as a function of the number of taxa. The values in the y-axis are plotted in the log scale. Source: [1]

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

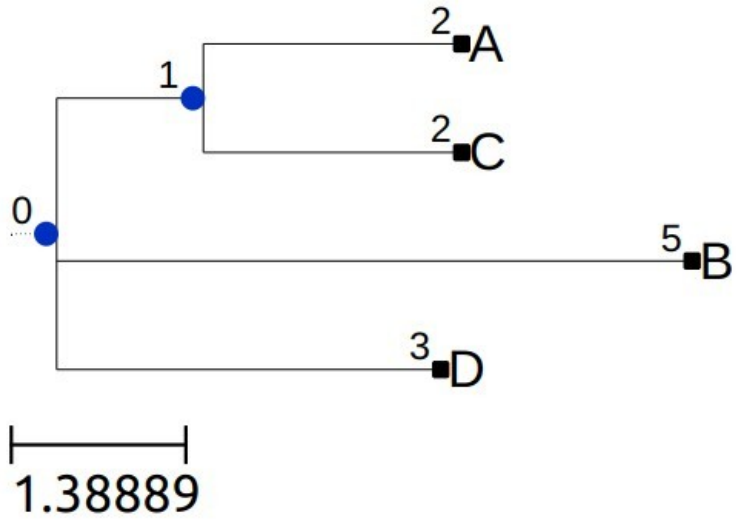
Methodology

Input-Output

Distances between sequences [A, B, C, D]:

$$\begin{bmatrix} 0 & 8 & 4 & 6 \\ 8 & 0 & 8 & 8 \\ 4 & 8 & 0 & 6 \\ 6 & 8 & 6 & 0 \end{bmatrix}$$

Output:



vmachacaa@unsa.edu.pe

Methodology

- Choice a molecular marker.
- Alignment.
- Multiple substitution (distances).
- Phylogenetics building.

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Choice of Molecular Markers

Nucleotide or protein sequence data?

vmachacaa@unsa.edu.pe

Choice of Molecular Markers

Use nucleotides for:

- Studying very closely related organisms, nucleotide sequences, which evolve more rapidly than proteins.

Use proteins because:

- Protein sequences are relatively more conserved as a result of the degeneracy of the genetic code.
- Sixty-one codons encode for twenty amino acids, meaning thereby a change in a codon may not result in a change in amino acid.

vmachacaa@unsa.edu.pe

Choice of Molecular Markers

Moreover:

- Protein sequences allow more sensitive alignment than DNA sequences (20 vs. 4 characters).

vmachacaa@unsa.edu.pe

Choice of Molecular Markers

Moreover:

- Protein sequences allow more sensitive alignment than DNA sequences (20 vs. 4 characters).
- Two randomly related DNA sequences can result in up to 50% sequence identity, compared to 10% for protein sequences.

vmachacaa@unsa.edu.pe

Choice of Molecular Markers

Moreover:

- Protein sequences allow more sensitive alignment than DNA sequences (20 vs. 4 characters).
- Two randomly related DNA sequences can result in up to 50% sequence identity, compared to 10% for protein sequences.
- In the alignment of DNA, gaps almost always cause frameshift errors. Protein sequences have a higher signal-to-noise ratio.

vmachacaa@unsa.edu.pe

Choice of Molecular Markers

Moreover:

- Protein sequences allow more sensitive alignment than DNA sequences (20 vs. 4 characters).
- Two randomly related DNA sequences can result in up to 50% sequence identity, compared to 10% for protein sequences.
- In the alignment of DNA, gaps almost always cause frameshift errors. Protein sequences have a higher signal-to-noise ratio.
- DNA is informative for closely related sequences. Moreover, if we take into account that sequences evolve faster at the DNA level.

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Alignment

Only the correct alignment produces correct phylogenetic.

- In some cases, researchers like to remove all insertions and deletions and only use positions that are shared by all sequences in the dataset. As a consequence, many phylogenetic signals are lost.
- There is an automatic approach in improving alignment quality. For example: Rascal and NorMD

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Multiple Substitutions

After alignment, we need to measure the distance between sequences. A simple measure could be the number of substitutions in a alignment.

Figure: This alignment has 8 substitutions, so the distance between these sequences is 8. Source: [1]

vmachacaa@unsa.edu.pe

Multiple Substitutions

Homoplasy

Substitutions are complex, for example:

When A is replace by C. The nucleotide may have actually undergone a number of intermediate steps to become C, such as $A \rightarrow T \rightarrow G \rightarrow C$

Such multiple substitutions obscure the estimation of the true evolutionary distances between sequences. This effect is known as **homoplasy** [1].

vmachacaa@unsa.edu.pe

Multiple Substitutions

Substitution Models

The statistical models used to correct **homoplasy** are called substitution models or evolutionary models [1].

vmachacaa@unsa.edu.pe

Substitution Models

Jukes–Cantor Model

This model assumes that all nucleotides are substituted with equal probability.

$$d_{AB} = -\frac{3}{4} ln \left[1 - \left(\frac{4}{3} \right) p_{AB} \right] \tag{3}$$

where, d_{AB} is the evolutionary distance between A and B. p_{AB}, is the observed sequence distance measured by the proportion of substitutions over the entire lenght.

vmachacaa@unsa.edu.pe

Substitution Models

Jukes–Cantor Model

Example: If an alignment of sequences A and B is twenty nucleotides long and six pairs are found to be different, the sequences differ by 30%, or have an observed distance 0.3. What is the distances using Jukes–Cantor Model?

vmachacaa@unsa.edu.pe

Substitution Models

Jukes–Cantor Model

Solution:

$$d_{AB} = -\frac{3}{4} ln \left[1 - \left(\frac{4}{3} \right) p_{AB} \right]$$

$p_{AB} = 0.3$

$$d_{AB} = -\frac{3}{4} ln \left[1 - \left(\frac{4}{3} \right) 0.3 \right] = 0.38$$

vmachacaa@unsa.edu.pe

Substitution Models

Kimura Model

This is a more sophisticated model in which mutation rates for transitions and transversion are assumed to be different.

$$d_{AB} = -\frac{1}{2} ln(1 - 2p_{ti} - p_{tv}) - \frac{1}{4} ln(1 - 2p_{tv}) \tag{4}$$

where, d_{AB} is the evolutionary distance between A and B. p_{ti}, is the observed frequency for transition and p_{tv} the frequency of transversion[1].

[1] Transitions are interchanges of two-ring purines (A G) or of one-ring pyrimidines (C T): they therefore involve bases of similar shape. Transversions are interchanges of purine for pyrimidine bases, which therefore involve exchange of one-ring and two-ring structures.

vmachacaa@unsa.edu.pe

Substitution Models

Kimura Model

Example: The comparison of sequences A and B differ by 30%. If 20% of changes are a result of transitions and 10% of changes are a result of transversions. What is the distances using Kimura Model?

vmachacaa@unsa.edu.pe

Substitution Models

Kimura Model

Solution:

$$d_{AB} = -\frac{1}{2} ln(1 - 2p_{ti} - p_{tv}) - \frac{1}{4} ln(1 - 2p_{tv})$$

$p_{ti} = 0.2$
$p_{tv} = 0.1$

$$d_{AB} = -\frac{1}{2} ln(1 - 2 * 0.2 - 0.1) - \frac{1}{4} ln(1 - 2 * 0.1) = 0.40$$

vmachacaa@unsa.edu.pe

Substitution Models

Protein sequences

For protein sequences, the evolutionary distances from an alignment can be corrected using a PAM or JTT amino acid substitution matrix.

For example, the Kimura model for correcting multiple substitutions in protein distances is:

$$d = -ln(1 - p - 0.2p^2) \tag{5}$$

where, p is the observed pairwise distance between two sequences.

vmachacaa@unsa.edu.pe

Multiple Substitutions

Among-Site Variations

In all these calculations, different positions in a sequence are assumed to be evolving at the same rate. However, this assumption may not hold up in reality.

Nevertheless:

- In DNA sequences, the rates of substitution differ for different codon positions. The third codon mutates much faster than the other two.
- For protein sequences, some amino acids change much more rarely than others.

vmachacaa@unsa.edu.pe

Multiple Substitutions

Among-Site Variations

It has been shown that there are always a proportion of positions in a sequence dataset that have invariant rates and a proportion that have more variable rates. **The distribution of variant sites follows a γ distribution pattern.**

vmachacaa@unsa.edu.pe

Multiple Substitutions

Among-Site Variations

Probability curves of γ distribution. The mathematical function of the distribution is $f(x) + (x^{\gamma-1}e^{-x})/\Gamma(\gamma)$. The curves assume different shapes depending on the γ-shape parameter (γ).

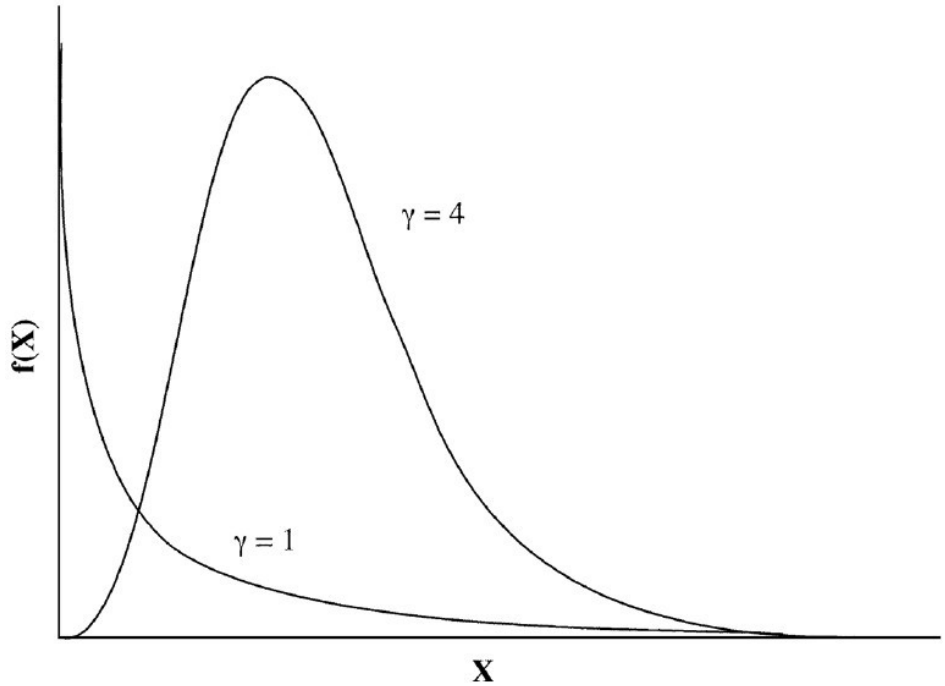


Figure: Probability curves of γ distribution. Source: [1]

vmachacaa@unsa.edu.pe

Substitution Models

Among-Site Variations

For the Jukes-Cantor model, the evolution distance can be adjusted:

$$d_{AB} = \frac{3}{4}\alpha\left[1 - \left(\frac{4}{3}p_{AB}\right)^{-1/\alpha} - 1\right]$$

For the Kimura model, the evolutionary distance with γ correction factor becomes:

$$d_{AB} = \frac{\alpha}{2}\left[(1 - 2p_{ti} - p_{tv})^{-1/\alpha} - \frac{1}{2}(1 - 2p_{tv})^{-1/\alpha} - \frac{1}{2}\right]$$

where α is the γ correction factor (default = 1).

vmachacaa@unsa.edu.pe

Questions?

vmachacaa@unsa.edu.pe

References I


J. Xiong, *Essential bioinformatics*. Cambridge University Press, 2006.

vmachacaa@unsa.edu.pe