Escuela Profesional de
Ciencia de la Computación

ICC Fase 1

Bioinformatics

Phylogenetic trees

MSc. Vicente Machaca Arceda

Universidad Nacional de San Agustín de Arequipa

June 5, 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

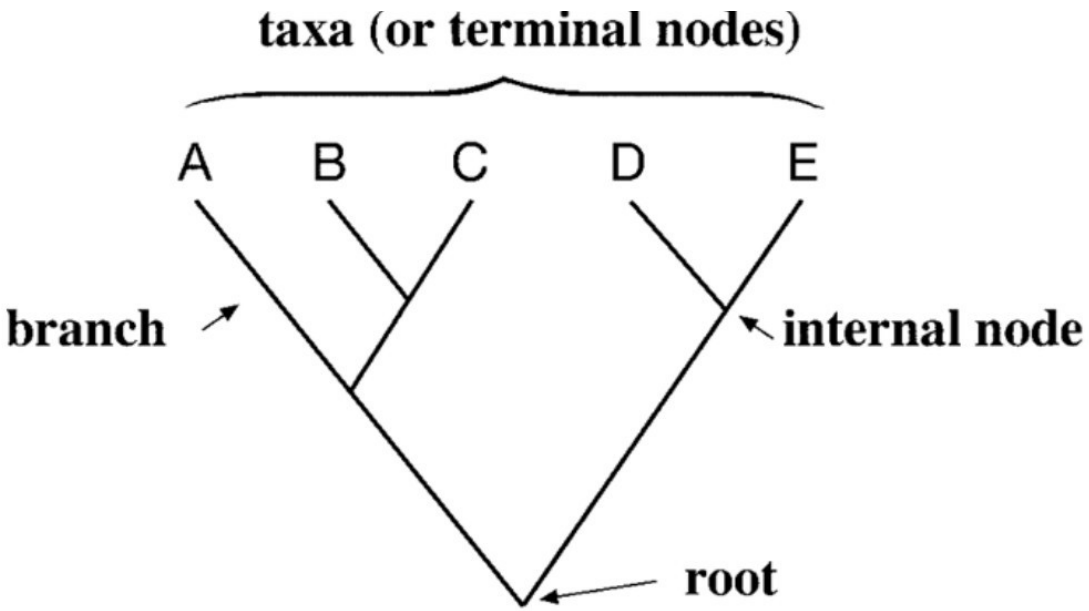


Figure: A typical bifurcating phylogenetic tree showing root, internal nodes, terminal nodes and branches. Source: [1]

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

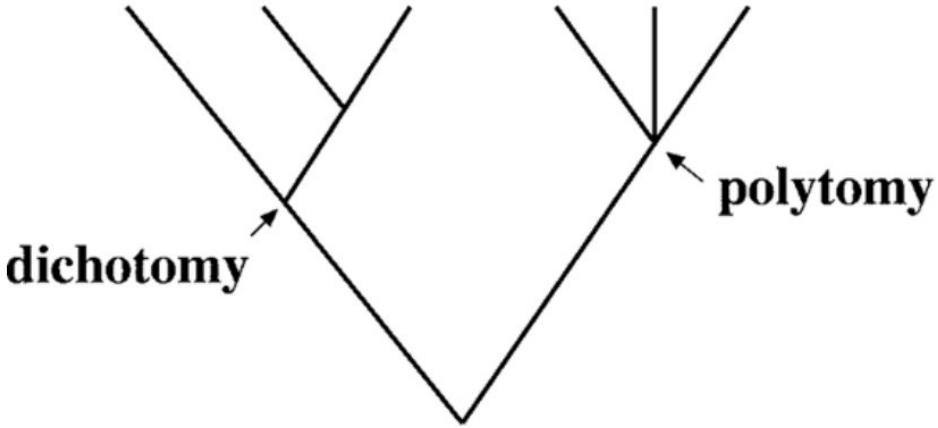


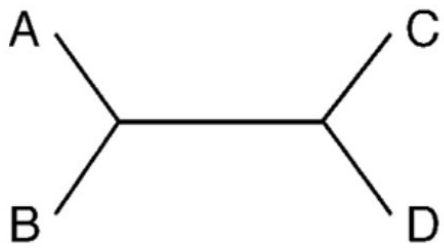
Figure: A phylogenetic tree showing an example of bifurcation and multifurcation. Source: [1]

vmachacaa@unsa.edu.pe

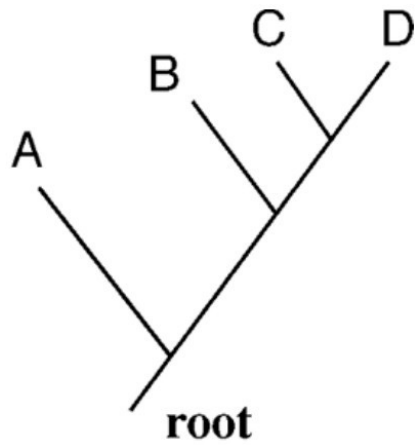
Table of Contents

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].

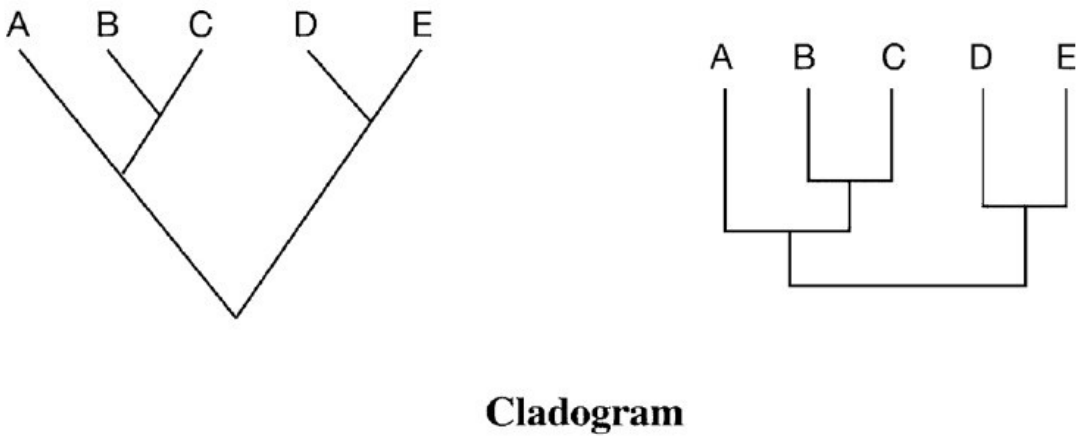
vmachacaa@unsa.edu.pe

Table of Contents

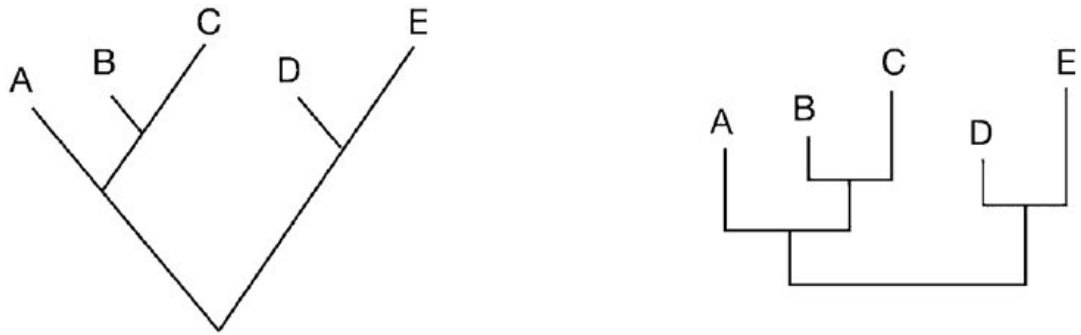
vmachacaa@unsa.edu.pe

Forms of representation

cladograms and phylograms



Cladogram



Phylogram

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

vmachacaa@unsa.edu.pe

Forms of representation

Newick

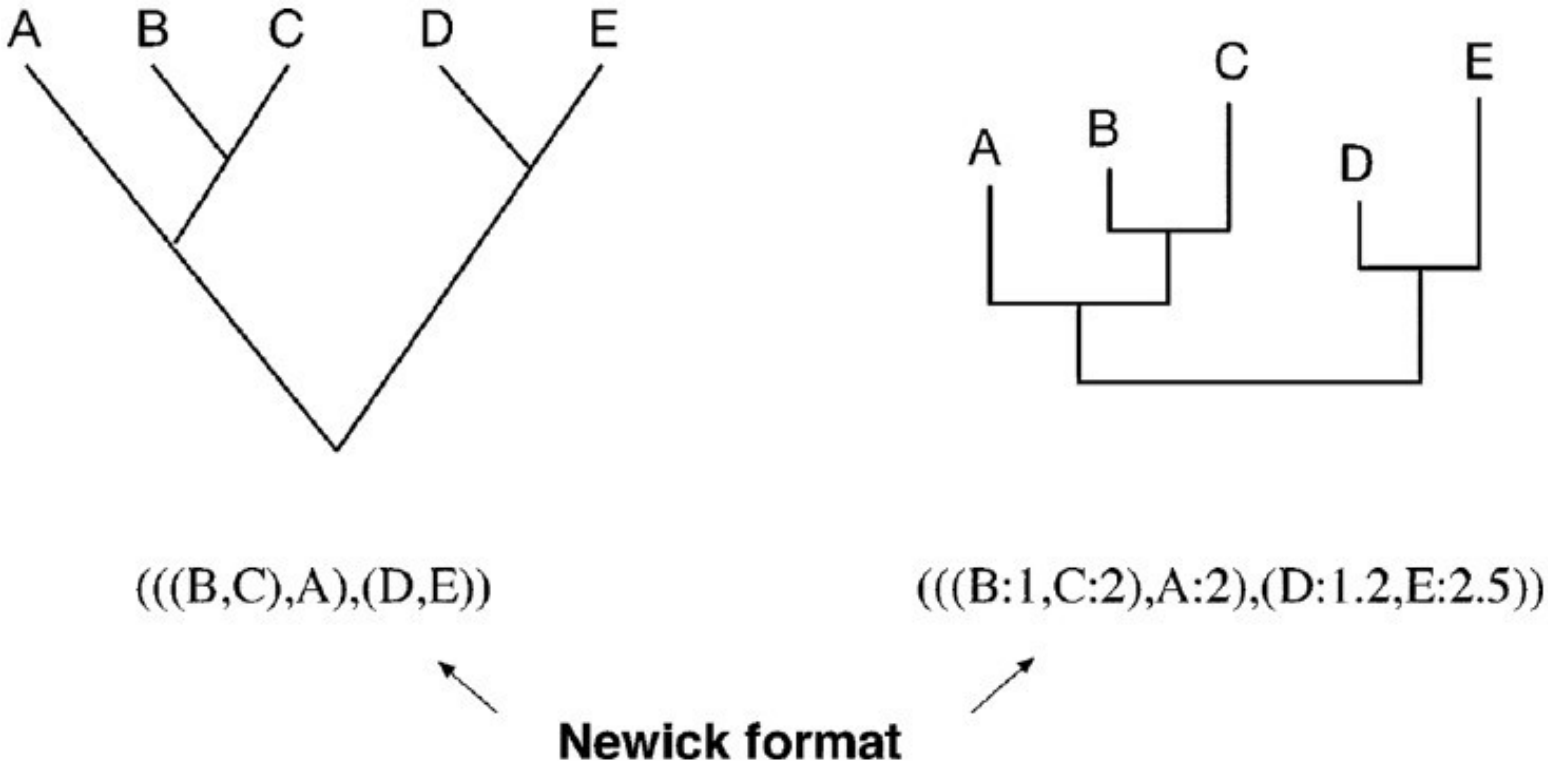


Figure: Newick format of tree representation. Source: [1]

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

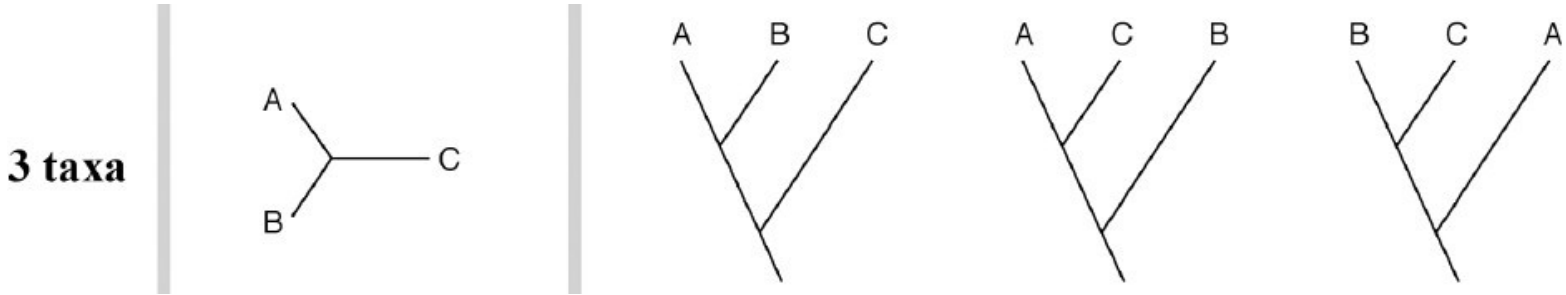


Figure: Unrooted and rooted trees for 3 taxa. Source: [1]

vmachacaa@unsa.edu.pe

The true tree

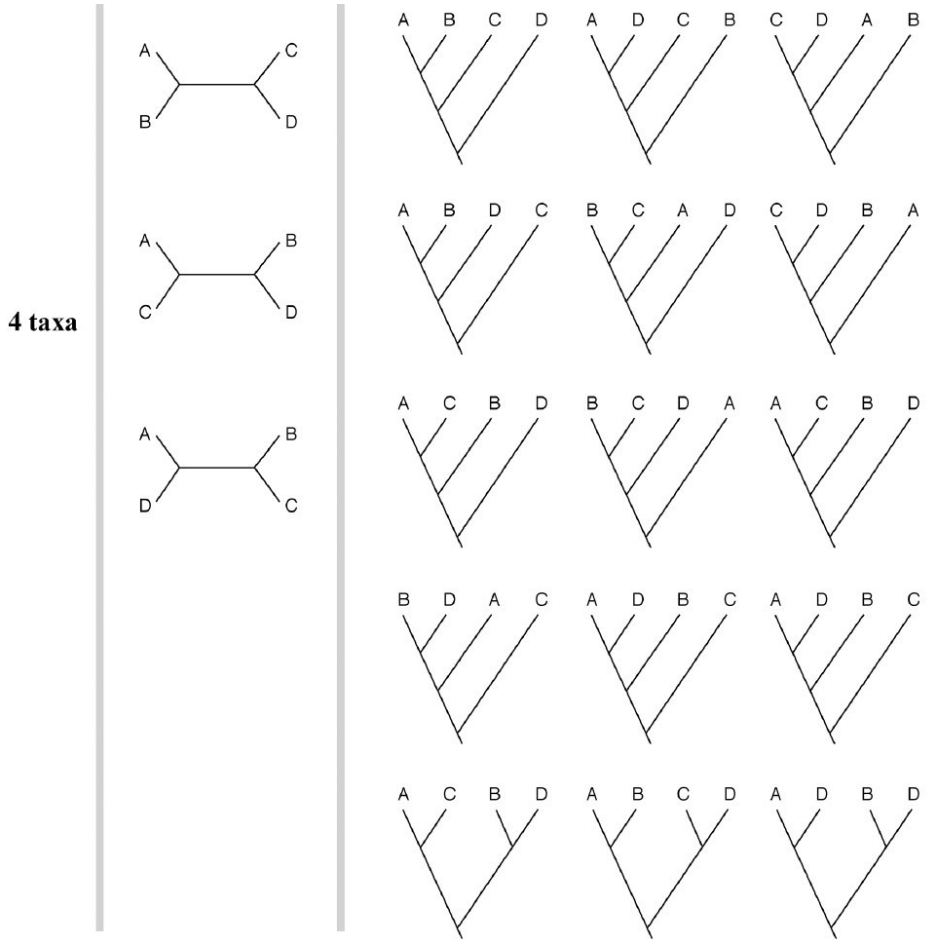


Figure: Unrooted and rooted trees for 4 taxa. Source: [1]

vmachacaa@unsa.edu.pe

The true tree

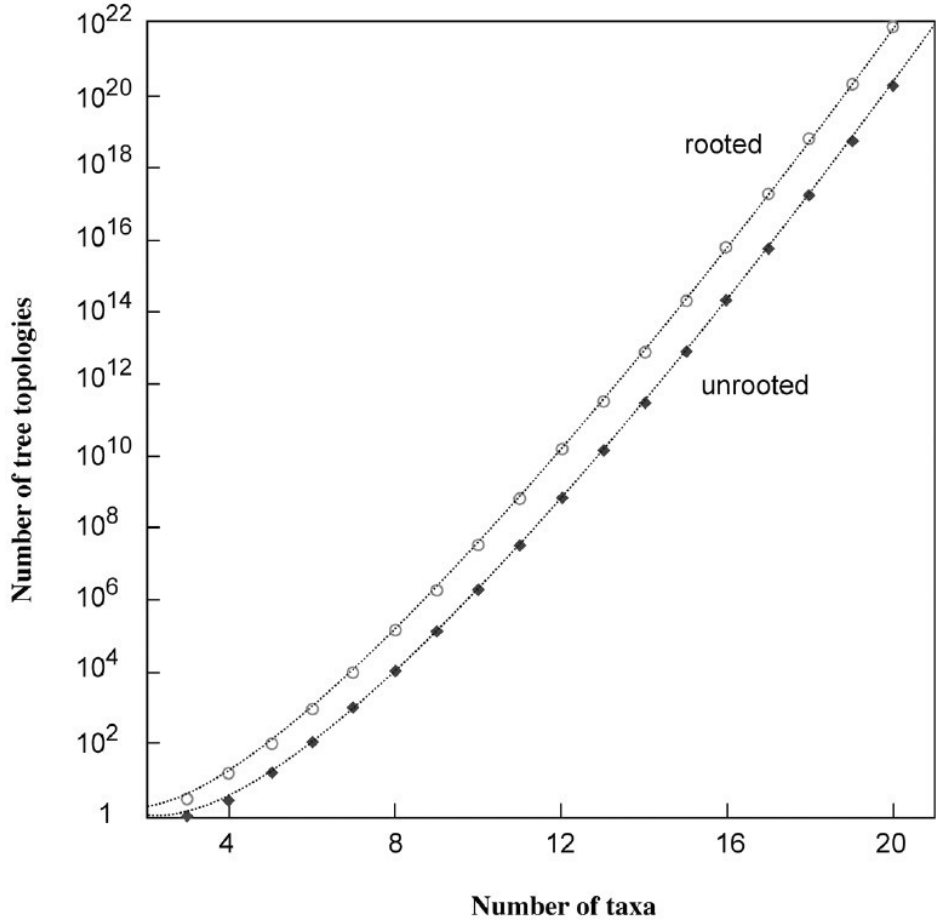


Figure: Total number of rooted (○) and unrooted (⧫) tree topologies as a function of the number of taxa. The values in the y-axis are plotted in the log scale. Source: [1]

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Methodology

- Choice a molecular marker.
- Alignment.
- Multiple substitution.
- 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

vmachacaa@unsa.edu.pe

Questions?

vmachacaa@unsa.edu.pe

References I


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

vmachacaa@unsa.edu.pe