

CGNPH Summer School

Sequence Alignment and SNP Analysis

I. A. Taiwo PhD

(Professor of Human Genetics and Bioinformatics)

Outline of Presentation

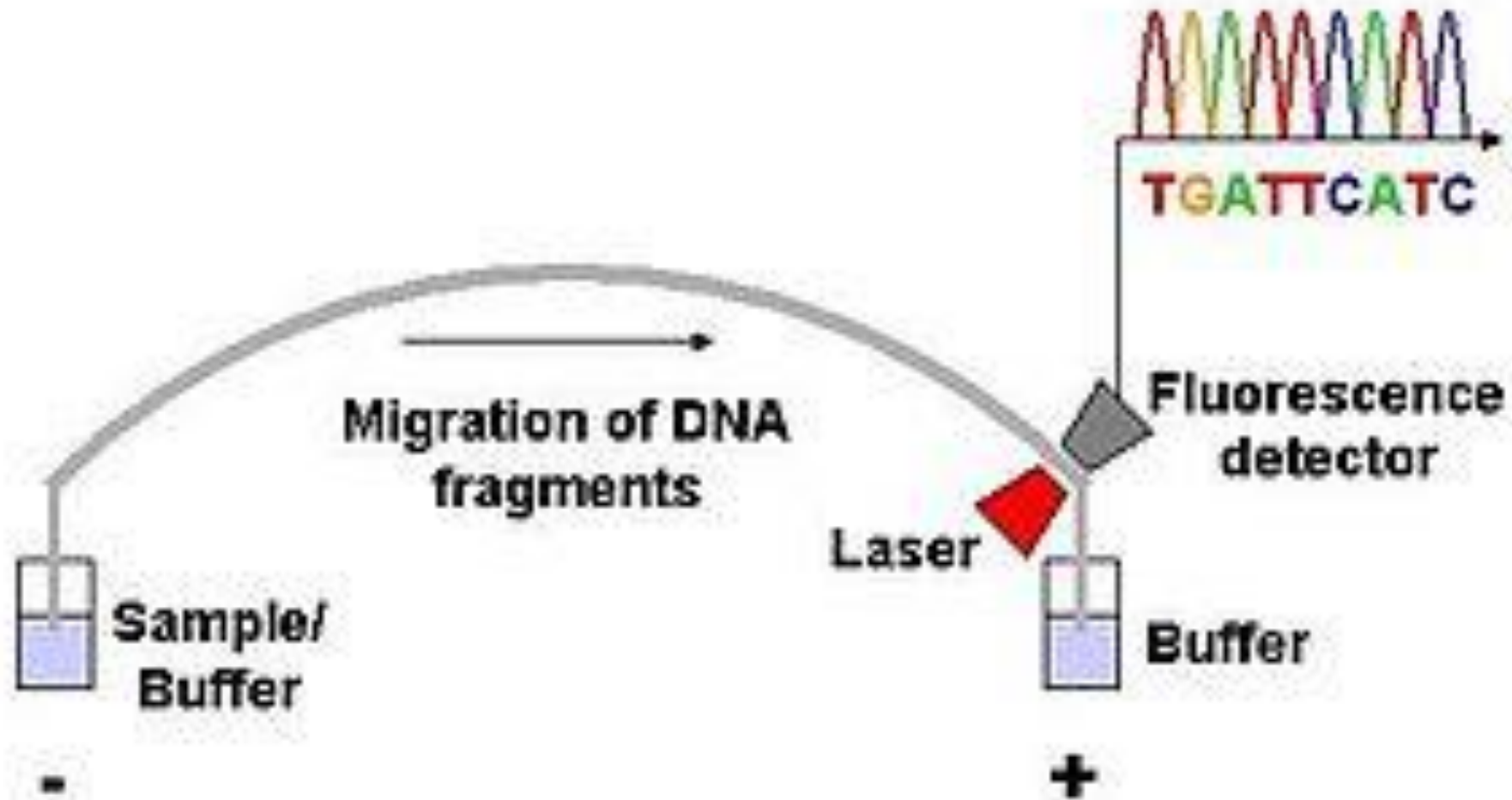
- **What is a sequence?** - An ordered list of items
- Two main types
 - DNA sequence: Order of nucleotides in a DNA molecule
 - Protein Sequence: Order of amino acids in a protein molecule

Bioinformatics Vs Computational Biology

- **Computational biology**
 - development of algorithms and statistical models to analyze biological data
- **Bioinformatics**
 - collection and storage of biological information
 - derives knowledge from computer analysis of biological data

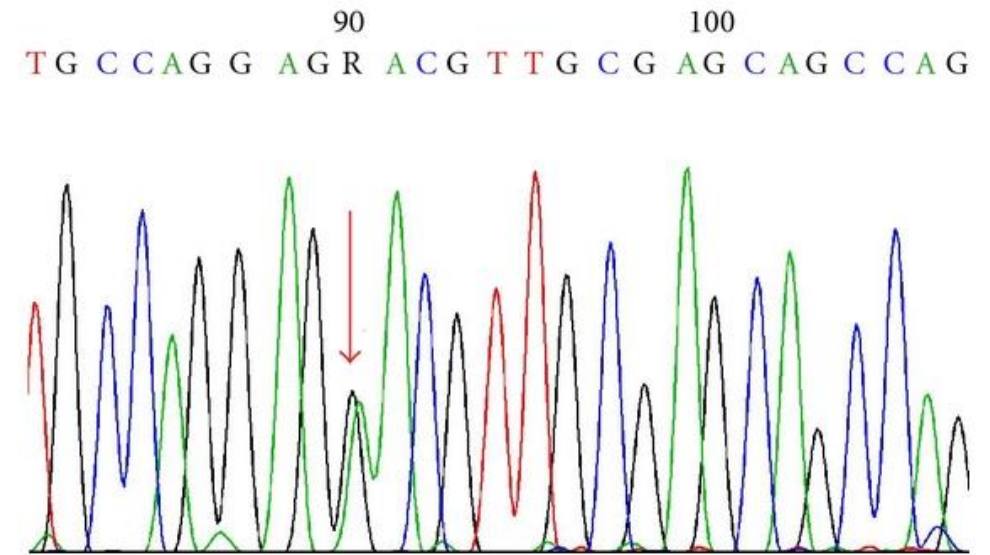
Sequence Data

Sanger Sequencing – 1st Generation Sequencing

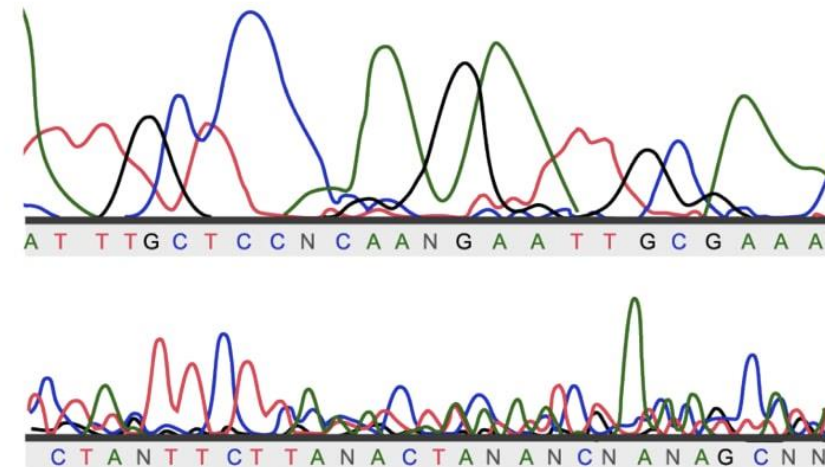


Analysis of sanger Sequence Data in Electropherogram

Mixed Signals: Heterozygosity



Mixed Signals: Unclear signal

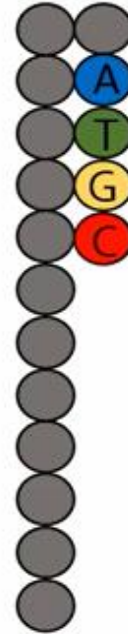


Next Generation Sequencing (NGS)

2nd Generation Sequencing or Massively Parallel Sequencing (MPS)



NGS Illumina



Sequencing
by synthesis



Fluorescence signal
Capture



A T G C
Conversion to
sequence format

.fastq

.bcl

Sequencing errors

```
*new 1 - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
new 1
1 @SRR031716.1 HWI-EAS299_4_30M2BAAXX:3:1:944:1798 length=37
2 GTGGATATGGATATCCAAATTATATTGCATAATTG
3 +
4 !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
5
```

Each symbol tells quality of respective base

! Lowest quality Score: Q -0
I Highest quality Score: Q - 40

Sequence Data

IUPAC degenerate base symbols^[2]

Description	Symbol	Bases represented					Complementary bases
		No.	A	C	G	T	
Adenine	A	1	A				T
Cytosine	C			C			G
Guanine	G				G		C
Thymine	T					T	A
Uracil	U					U	A
Weak	W	2	A			T	W
Strong	S			C	G		S
Amino	M		A	C			K
Ketone	K				G	T	M
Purine	R		A		G		Y
Pyrimidine	Y			C		T	R
Not A	B	3		C	G	T	V
Not C	D		A		G	T	H
Not G	H		A	C		T	D
Not T ^[a]	V		A	C	G		B
Any one base	N	4	A	C	G	T	N
Gap	-	0					-
a. ^ Not U for RNA							

IUPAC 3-letter and 1-Letter Amino Acid Symbols

Amino Acid	3 Letter Abbreviation	IUPAC Notation	Translating Codons
Alanine	Ala	A	GCT, GCC, GCA, GCG
Arginine	Arg	R	CGT, CGC, CGA, CGG, AGA, AGG
Asparagine	Asn	N	AAT, AAC
Aspartic acid	Asp	D	GAT, GAC
Cysteine	Cys	C	TGT, TGC
Glutamine	Gln	Q	CAA, CAG
Glutamic acid	Glu	E	GAA, GAG
Glycine	Gly	G	GGT, GGC, GGA, GGG
Histidine	His	H	CAT, CAC
Isolucine	Ile	I	ATT, ATC, ATA
Methionine	Met	M	ATG
Leucine	Leu	L	TTA, TTG, CTT, CTC, CTA, CTG
Lysine	Lys	K	AAA, AAG
Phenylalanine	Phe	F	TTT, TTC
Proline	Pro	P	CCT, CCC, CCA, CCG
Serine	Ser	S	TCT, TCC, TCA, TCG, AGT, AGC
Threonine	Thr	T	ACT, ACC, ACA, ACG
Tryptophan	Trp	W	TGG
Tyrosine	Tyr	Y	TAT, TAC
Valine	Val	V	GTT, GTC, GTA, GTG
STOP	Stop	*	TAA, TGA, TAG

Table 2: Amino acids, their symbols, and codons that translate to the ac

Sequence File Formats

FASTA Format

Header —●>VIT_201s0011g03530.1
Sequence —●AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
●GACCATGAGAACAAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA

FASTq Format

Identifier —|@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Sequence —|TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNNTAGTTTCTTGAGA
+ sign & identifier —|+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Quality scores —|efcfffffcfeefffcfffffdddf`feed]`_]_Ba_^__[YBBBBBBBBBBBRTT\]][] dddd`

Base T
phred Quality] = 29

What is Annotation?

- The task of adding notes, explanation, comment, or extra text to something.

Note: Annotation gives meaning to something.

Unannotated Sequence Data

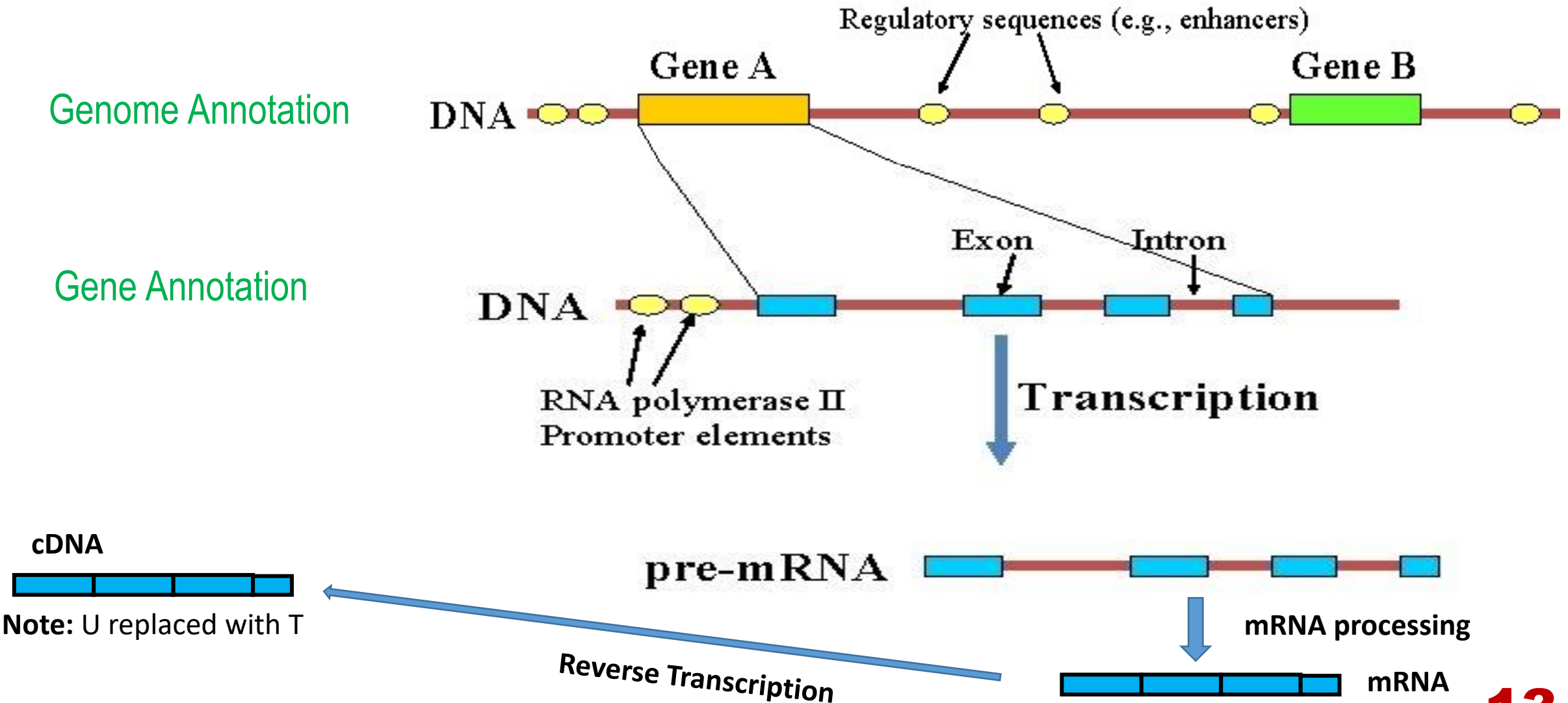
AAAGGGAAAGAGAGGAGTGGATTCCCAGTACTAGAAAGGGGAAAGAGAGGAGTGG
ATTCCCAGTACTAATTCATCAGGTGAGACGCGCAAAAGGGGAAAGAGAGGAGTGGAT
TCCCAGTACTAGAAAGGGGAAAGAGAGGAGTGGATTCCCAGTACTAATTCATCAGGTG
AGACGCGCAAAAGGGGAAAGAGAGGAGTGGATTCCCAGTACTAGAAAGGGGAAAGAG
AGGAGTGGATTCCCAGTACTAATTCATCAGGTGAGACGCGCAAAAGGGGAAAGAGAG
GAGTGGATTCCCAGTACTAGAAAGGGGAAAGAGAGGAGTGGATTCCCAGTACTAATTC
ATCAGGTGAGACGCGCAAAAGGGGAAAGAGAGGAGTGGATTCCCAGTACTAGAAAG
GGAAAGAGAGGAGTGGATTCCCAGTACTAATTCATCAGGTGAGACGCGCAAAAGGGG
AAAGAGAGGAGTGGATTCCCAGTACTAGAAAGGGGAAAGAGAGGAGTGGATTCCCA
GTACTAATTCATCAGGTGAGACGCGCAAAAGGGGAAAGAGAGGAGTGGATTCCCAGT
ACTAGAAAGGGGAAAGAGAGGAGTGGATTCCCAGTACTAATTCATCAGGTGAGACGC
GCAAAAGGGGAAAGAGAGGAGTGGATTCCCAGTACTAGAAAGGGGAAAGAGAGGAG
TGGATTCCCAGTACTAATTCATCAGGTGAGACGCGCACCCAAGGGGGTGTGAGACCAA

Sequence Annotation

Annotated Sequence Data

Genome Annotation

Gene Annotation



Sequence Alignment

What is sequence alignment?

Residue-to-residue comparison between two or more sequences.

Aligned and Unaligned Sequences

Unaligned Sequences

Sequence 1: A T G A C T

Sequence 2 : A G A C C

Aligned Sequences

Sequence 1: A T G A C T

Sequence 2 : A - G A C C

Types of Alignment

Based on Number of Sequences

Pairwise Sequence Alignment (PSA) - Alignment of two sequences

Sequence 1: A T G A C T

Sequence 2 : A - G A C C

Multiple Sequence Alignment (MSA) - Alignment of more than two sequences

Sequence 1: A T G A C T

Sequence 2 : A - G A C C

Sequence 3 : A T G A T C

Types of Alignment

Based on Length of Sequences

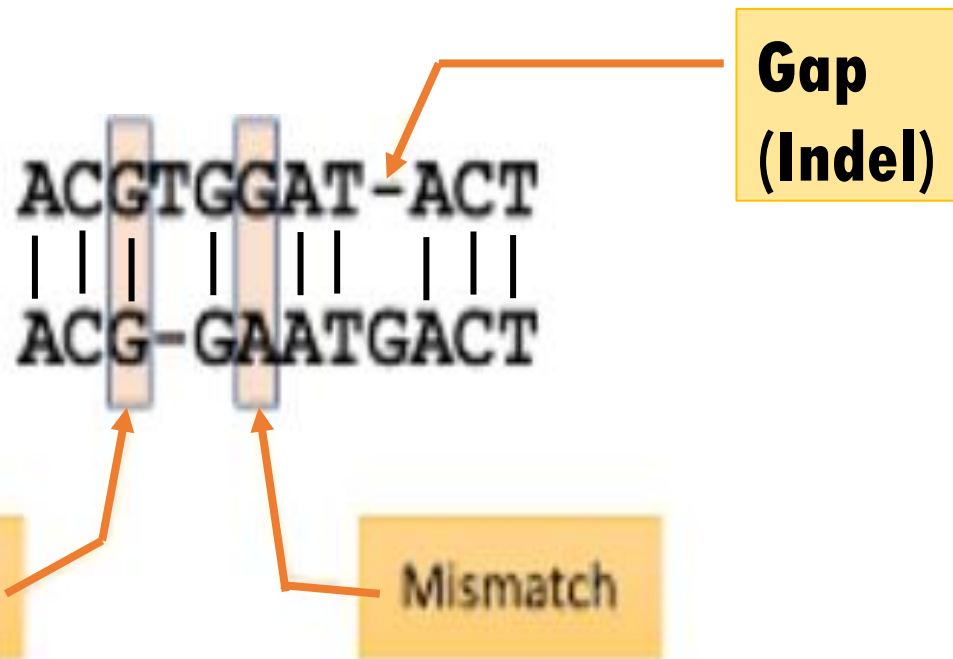
- **Global Alignment-** Alignment of two sequences of similar length over their entire length
Sequence 1: A T G A C T
Sequence 2 : A - G A C T
- **Local Alignment-** Alignment of two sequences of different length at their region of similarity. (**Note:** Used to identify domain, motifs or other protein signatures).
Sequence 1: A T G A C T
Sequence 2 : A - G A C C

Needleman-Wunsch is an algorithm for optimal global alignment developed by Saul B. Needleman and Christian D. Wunsch and published in 1970.

Smith–Waterman algorithm is also a rigorous dynamic programming method for finding optimal local alignments.

Determine the % Similarity (or ident) of PSA

- Identity – matches;
- similarity – similar characteristics
- For Nucleotide sequence: identity = similarity
- For Protein Sequence: identity $\neq$ Similarity



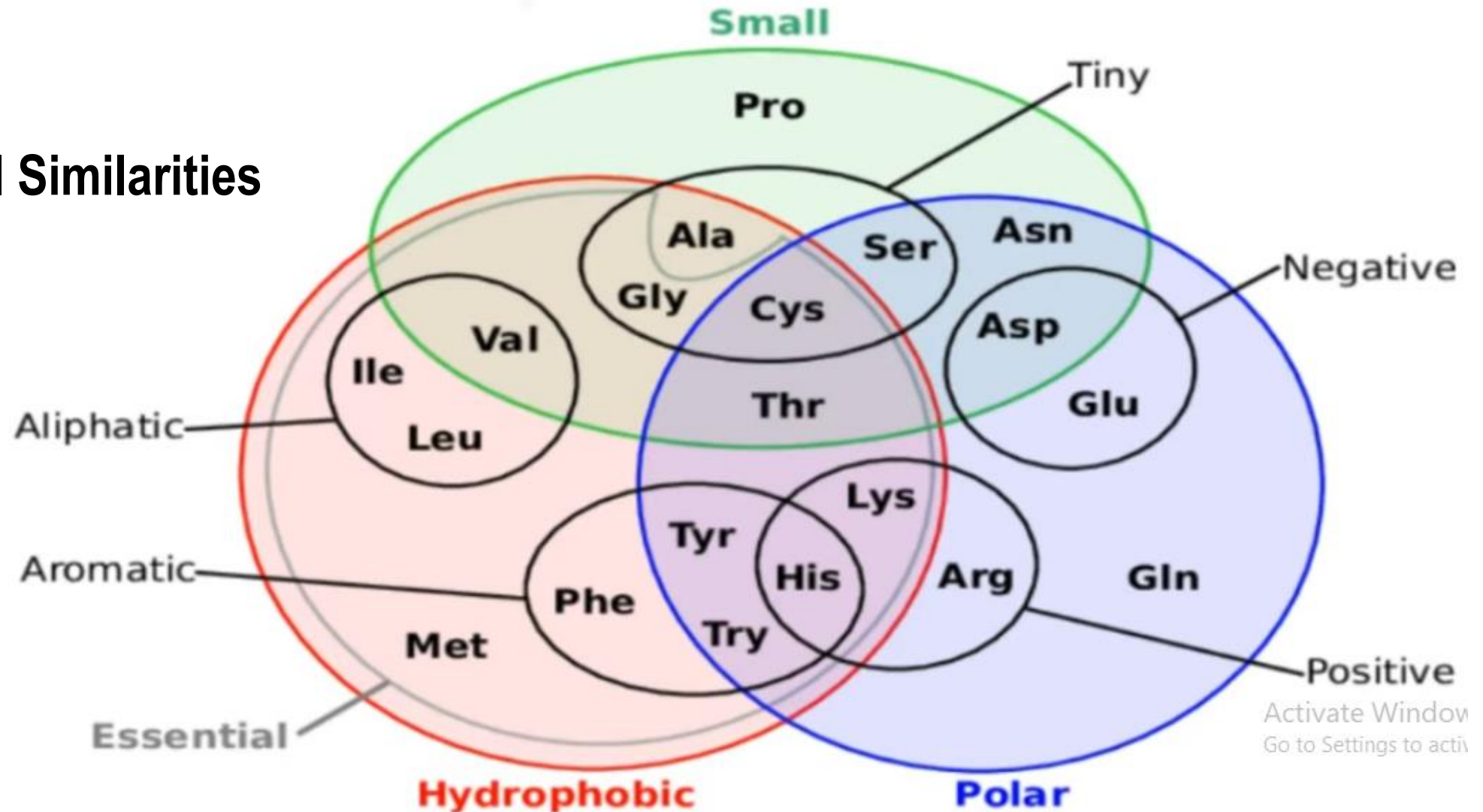
$$\text{Identity or Similarity (\%)} = \frac{\text{Match}}{\text{Match} + \text{Mismatch}} \times 100$$
$$= (9/10) \times 100 = \mathbf{90\%}$$

Note: Gaps were not considered in determining similarity or identity of the alignment

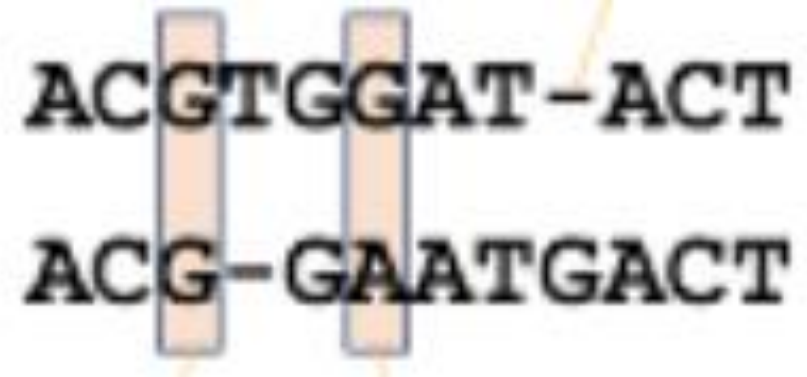
Identity and Similarity of Protein Sequence Alignment

- For protein sequence alignment, ident is not necessarily equal to sim because different amino acids are not identical but may be similar.

Physicochemical Similarities of Amino Acid



Score of Pairwise Sequence Alignment (PSA)



Scoring Scheme

Match = +4

Mismatch = -2

Indel (Gap) = -3

Score of the alignment = $4+4+4-3+4-2+4+4-3+4+4+4 = 28$

OR

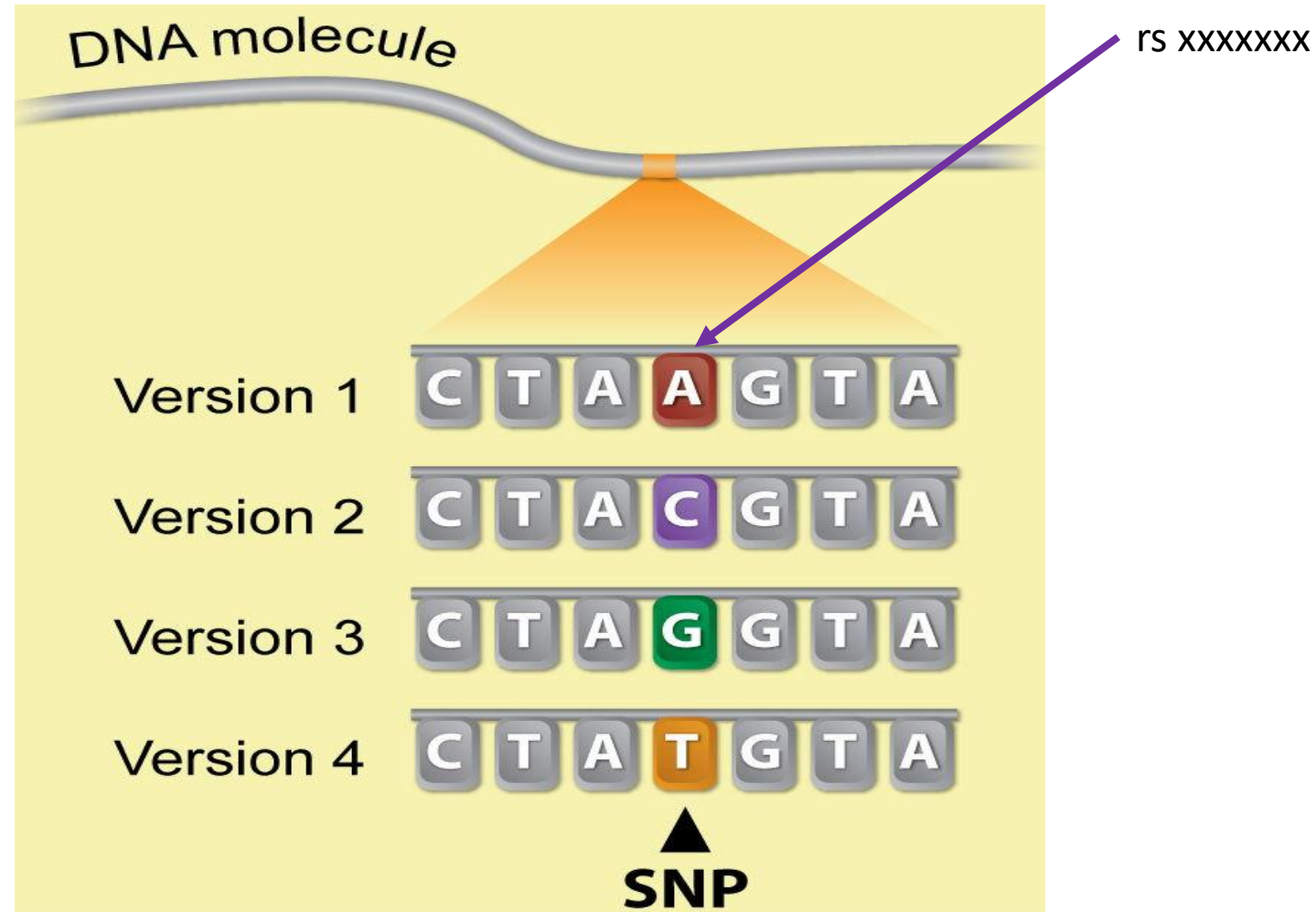
= 9 matches + 1 mismatch + 2 indels

= $9(4)+1(-2)+2(-3) = 28$

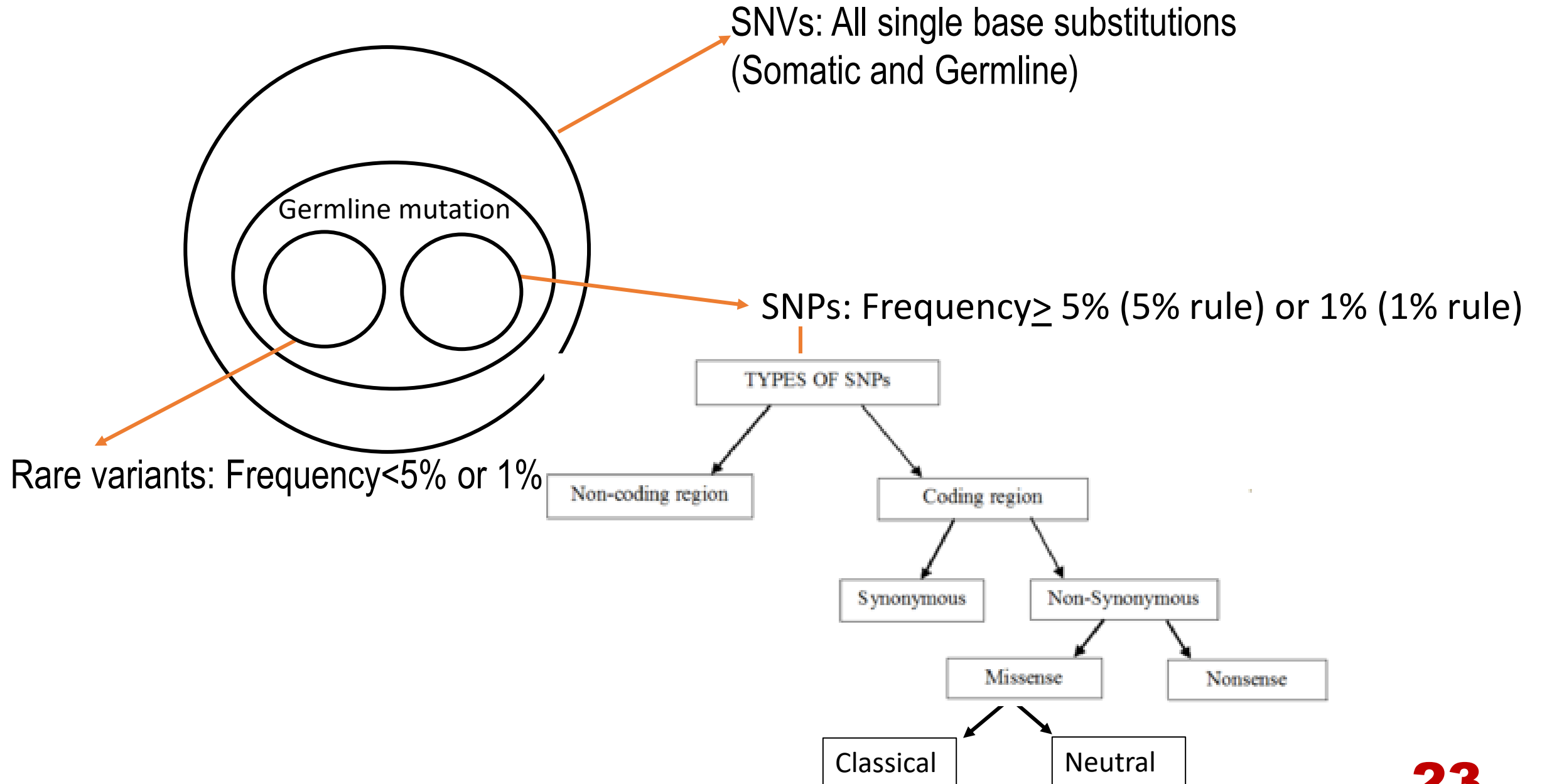
Multiple Sequence Alignment and SNPs

SNP

- a genomic variant at a single base position in the DNA
- The most common type of genetic variation in human populations



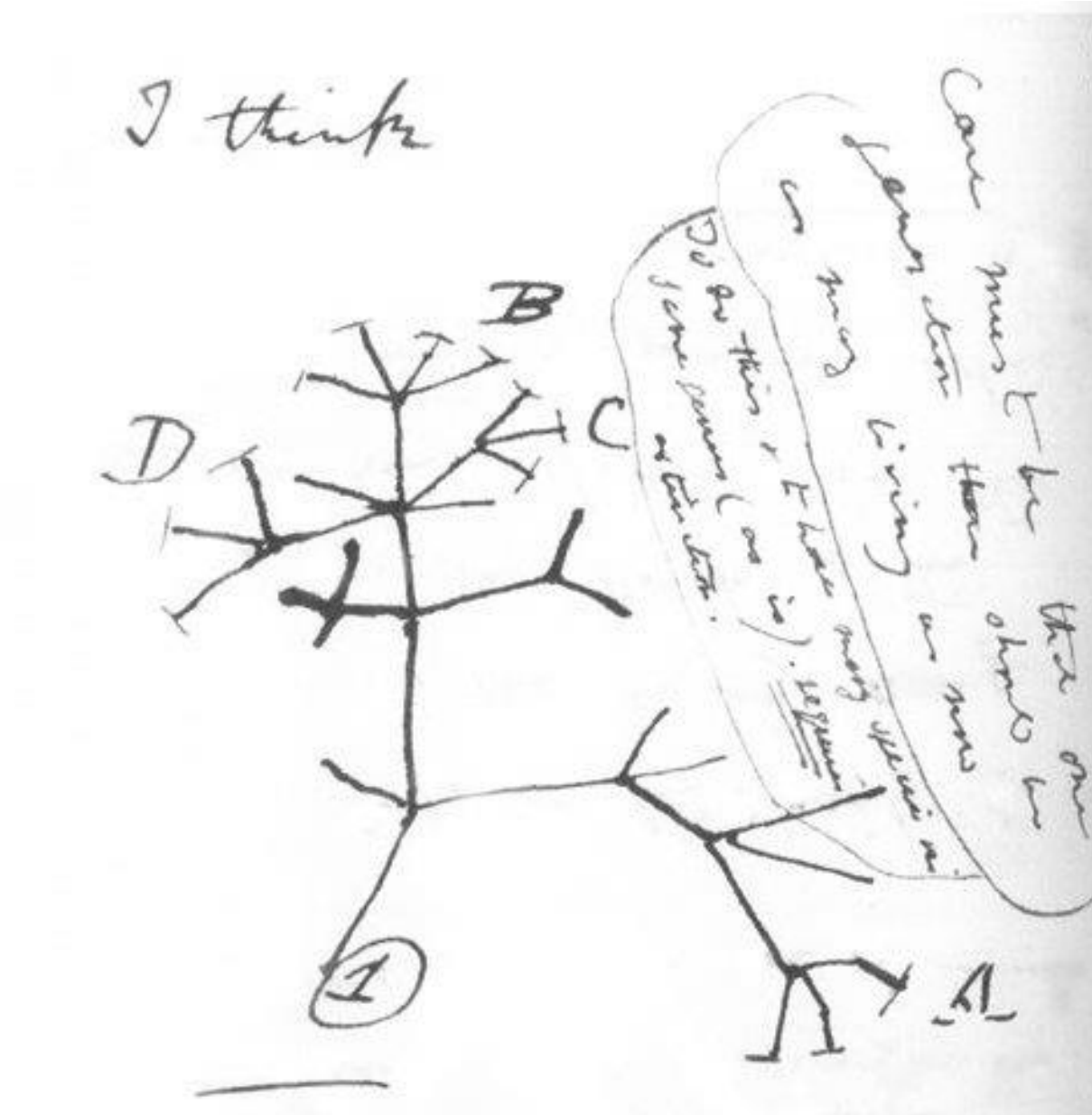
Single Nucleotide Variations (SNVs) and Single Nucleotide Polymorphisms (SNPs)



Phylogenetics

The First Tree: Darwin's Tree

History of Phylogenetics



What is phylogeny/Phylogenetics

- **Phylogeny** - Evolutionary relationship
- **Phylogenetic Tree** – A diagram that describes evolutionary relationship

Phenetics and Cladistics

- Phenetics (Numerical taxonomy) – Grouping based on similarity regardless of evolutionary relationship (phylogeny).
 - The drawing depicting phenetic relationship is phenogram.

Note: Linnaeus was a pheneticist
- Cladistics – Grouping based on shared ancestral characters or evolutionary relationship.
 - The drawing depicting cladistic relationship is a cladogram

Phenogram Vs Cladogram for the Following:



Dolphin



Shark



Eel



Snake



Bat



Bird



Mouse



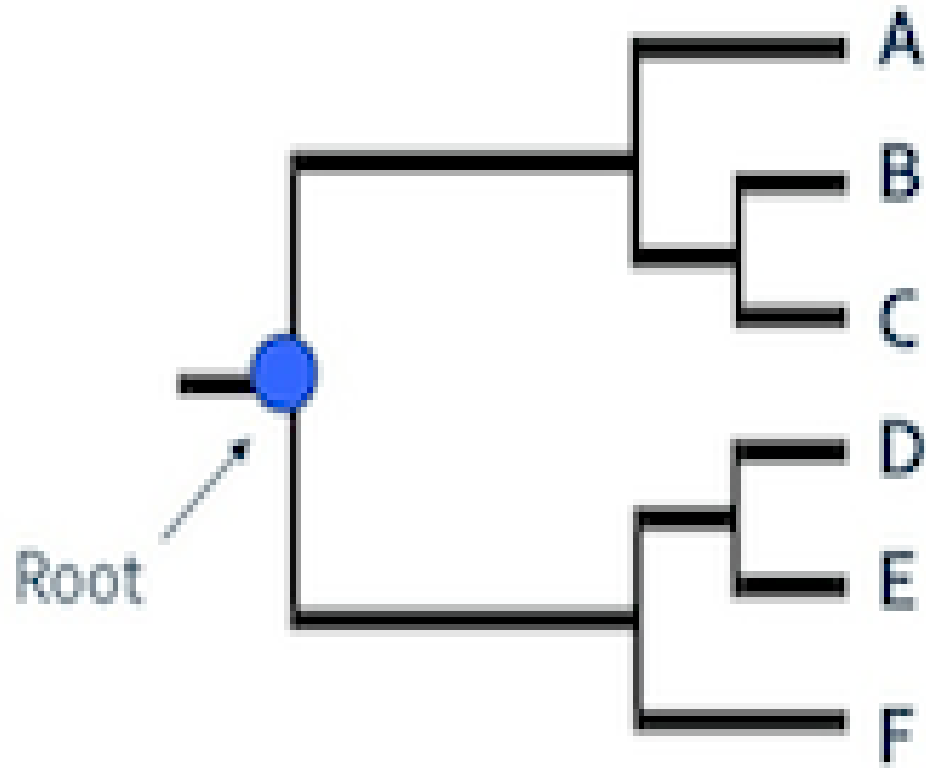
Lizard

Types of Phylogenetic Trees with respect to the Taxa on the Tree

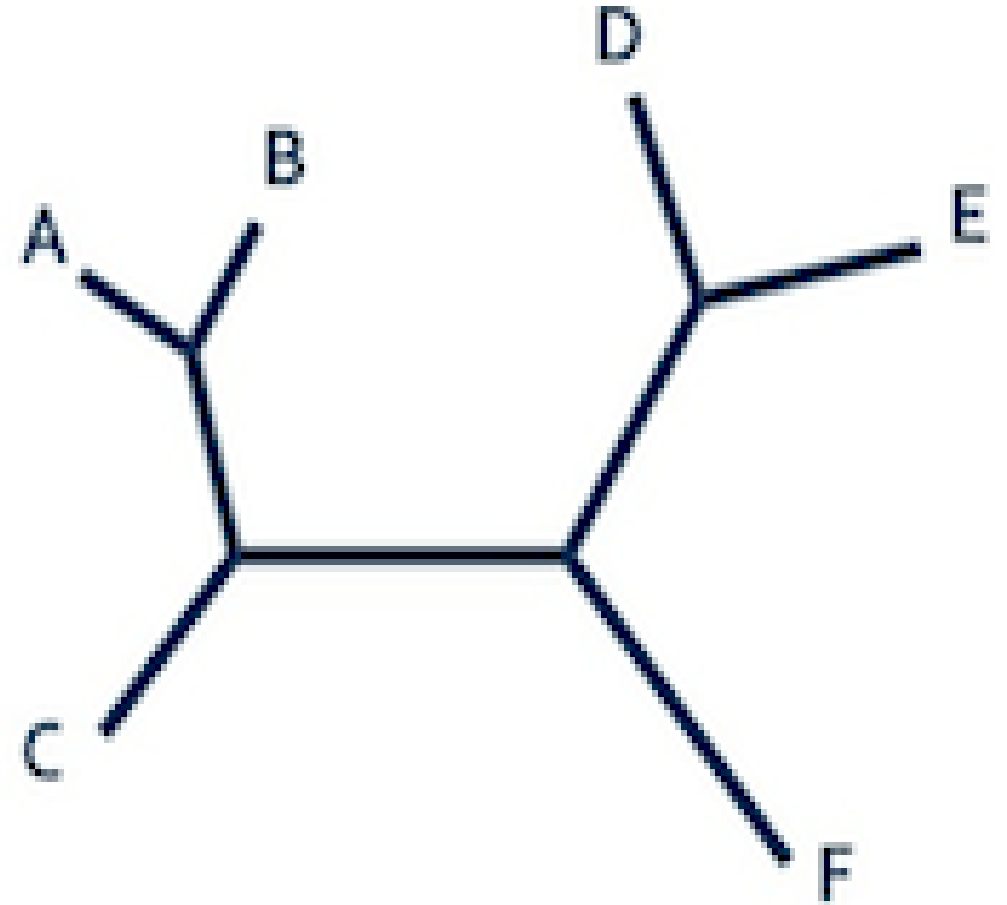
- 1. Rooted Tree** – Has one or more ancestral outgroup when compared to the ingroup
- 2. Unrooted Tree** – Has no ancestral outgroup among the ingroup

Note: The ingroup is made up of the taxa of interest

Relationship between Type and Shape of a Tree

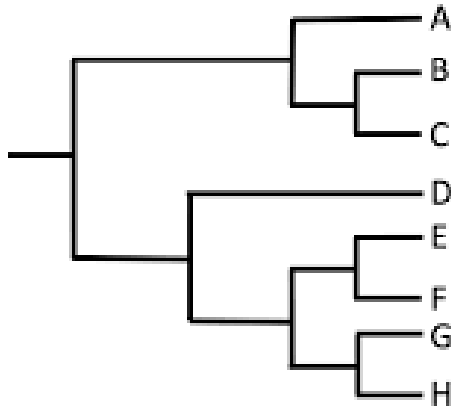


Rooted Tree - Rectangular

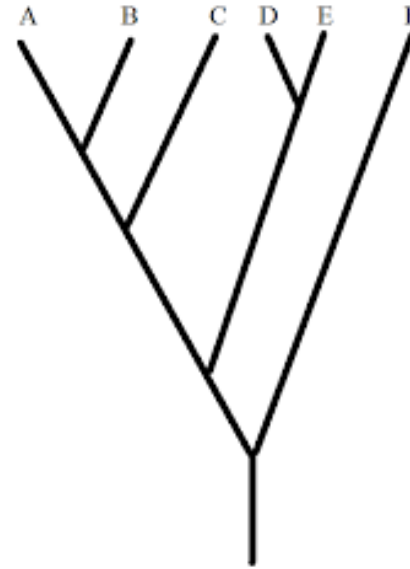


Unrooted Tree - Radial

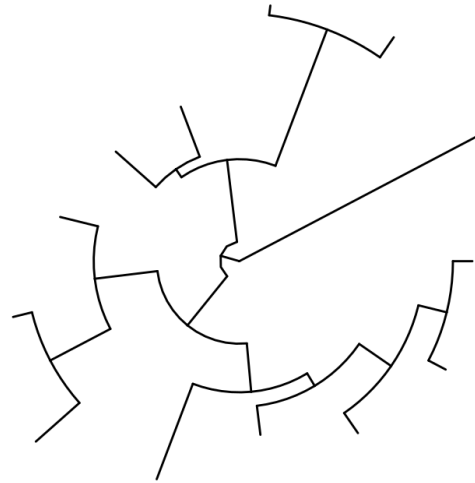
Types of Phylogenetic Trees with respect to the Tree Shape or Layout



Rectangular



Triangular



Circular

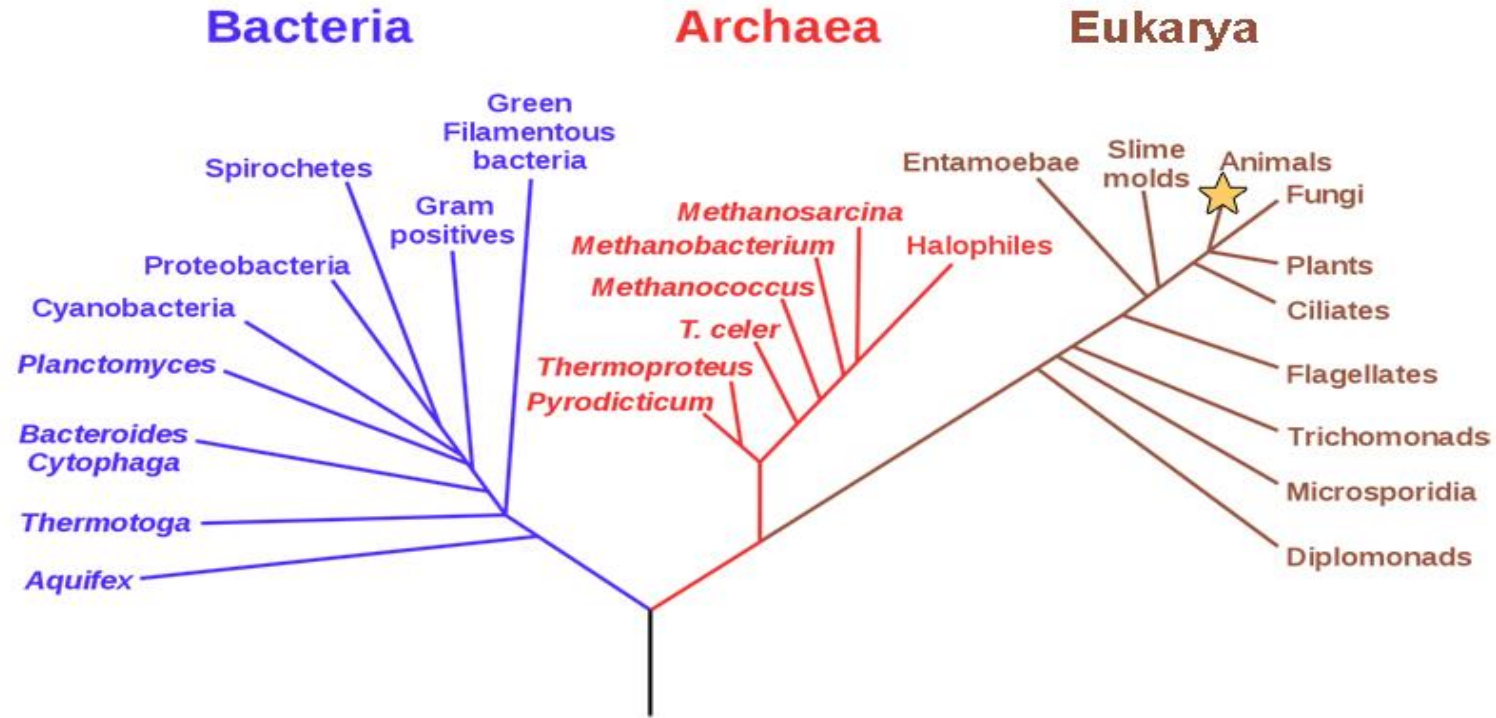


Radial

What type of tree is this with respect to shape/layout?

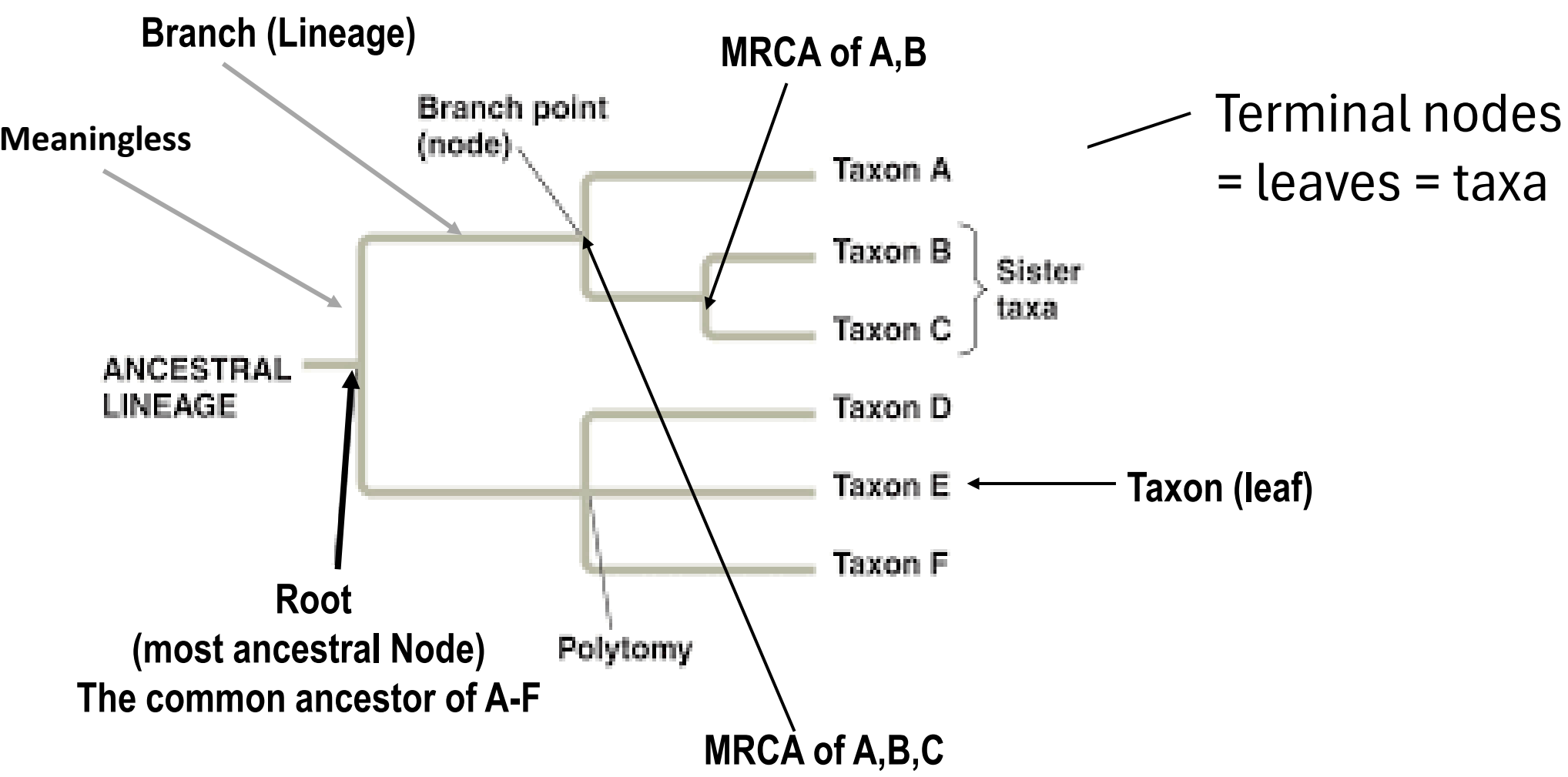
Phylogenetic Tree of Life

★ = You are here



Answer: Triangular tree

Parts of a Phylogenetic Tree



Tree Topology and Equivalent Trees

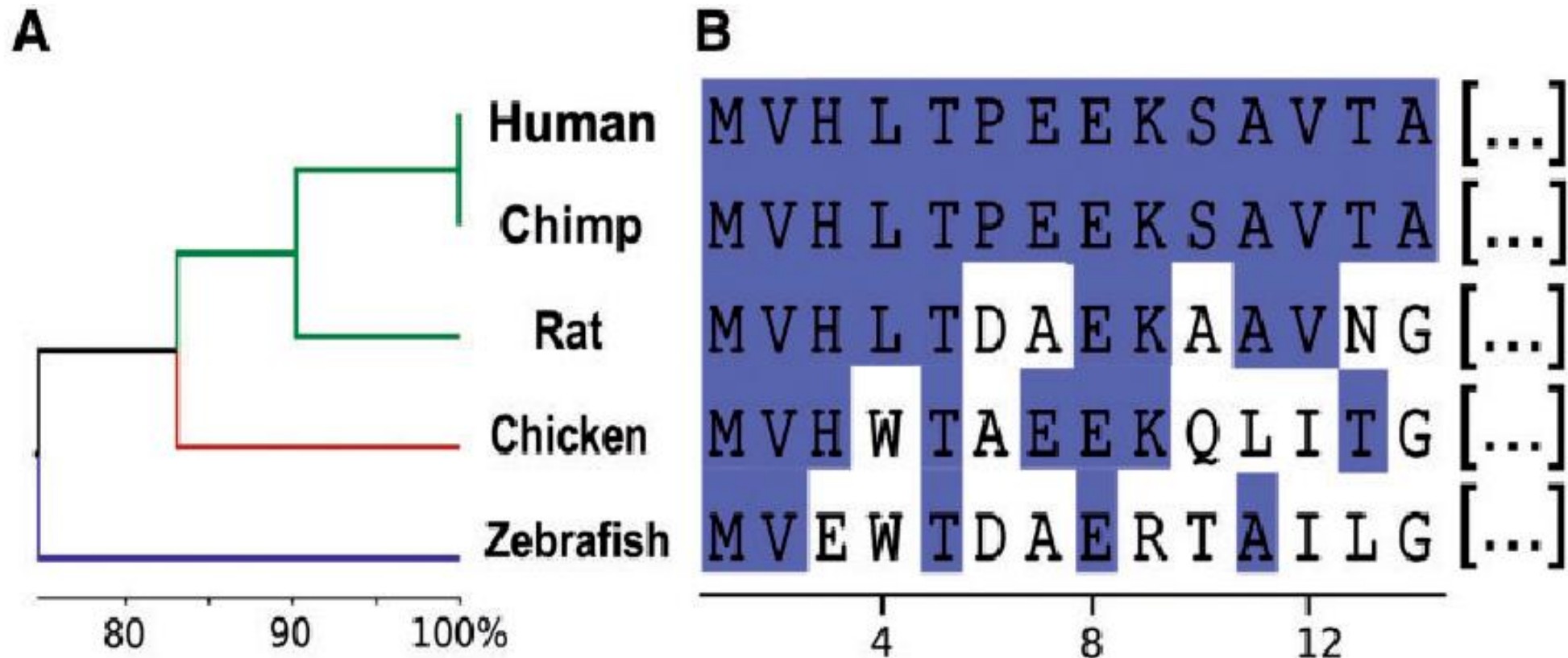
- Tree Topology – Branching pattern of a tree
- Trees are the same when rotated at a node



Note: These trees are equivalent

Molecular Phylogeny and Sequence Alignment

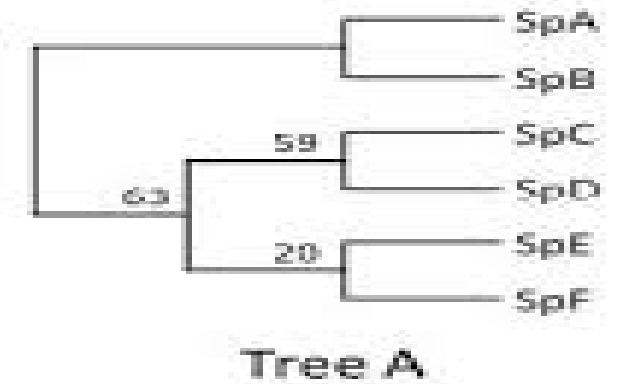
- Phylogeny is derived from multiple sequence alignment.



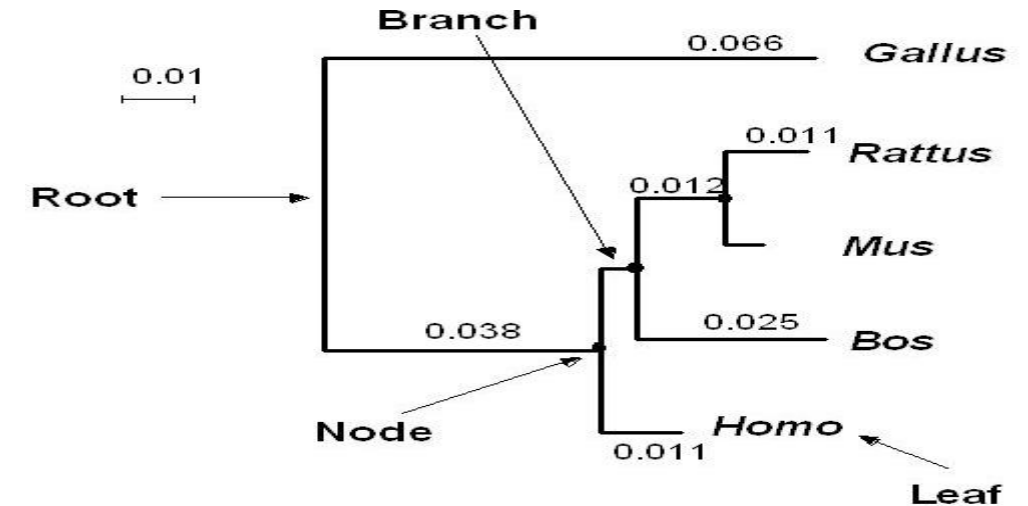
Bootstrap or Branch length?

- **Bootstrap Support** - The percentage of times the branch occur when resampled

Note: Bootstrap support >70 is well supported.



- **Branch Length**— Genetic distance measured as number of substitutions per site.

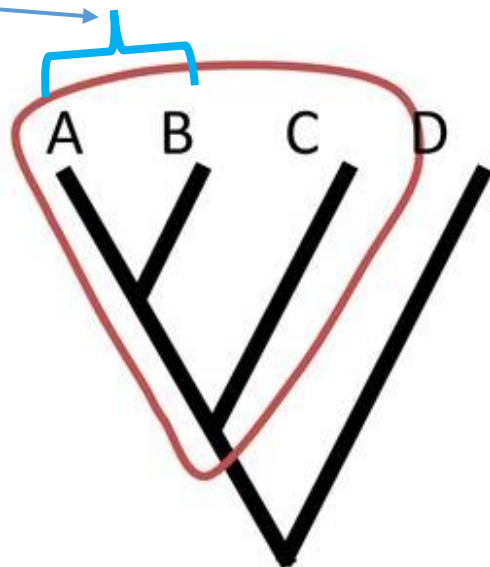


Evolutionary Groups/Relationships)

Monophyletic group

Includes an ancestor
all of its descendants

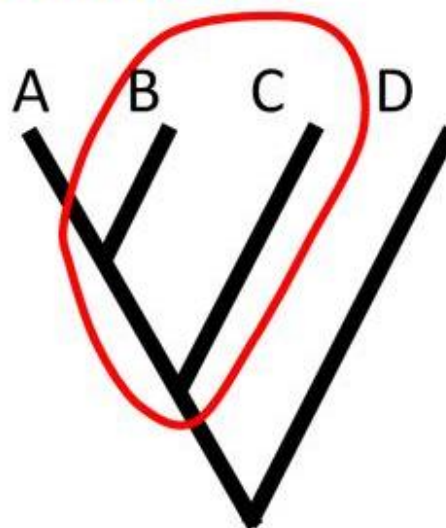
Sister taxa



How could this happen?

Paraphyletic group

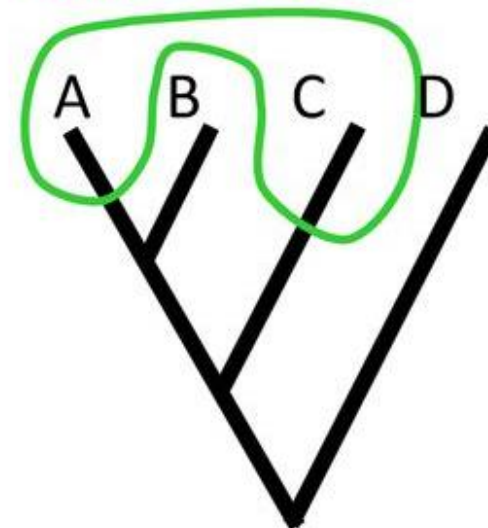
Includes ancestor and
some, but not all of its
descendants



Taxon A is highly derived
and looks very different
from B, C, and ancestor

Polyphyletic group

Includes two convergent
descendants but not their
common ancestor

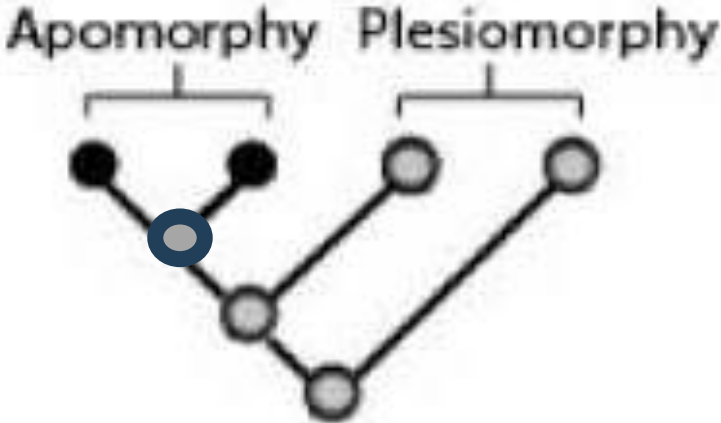


Taxon A and C share
similar traits through
convergent evolution

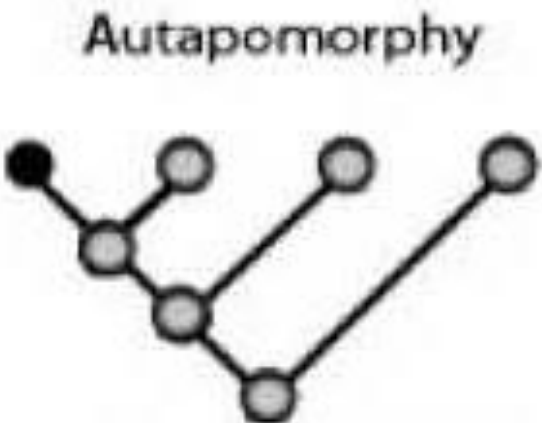
Only monophyletic groups (**clades**) are recognized in cladistic classification

Evolutionary Characters

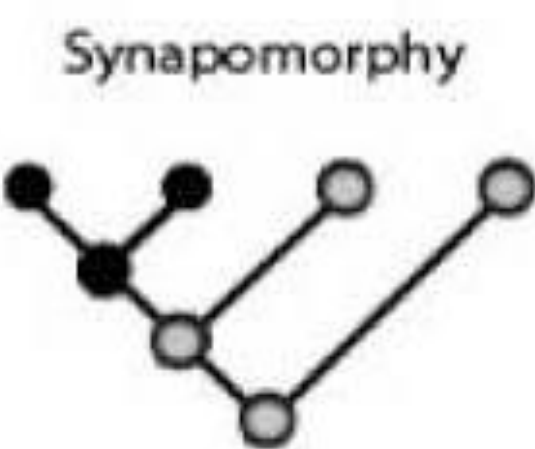
Derived trait
(Evolutionary novelty) Ancestral trait



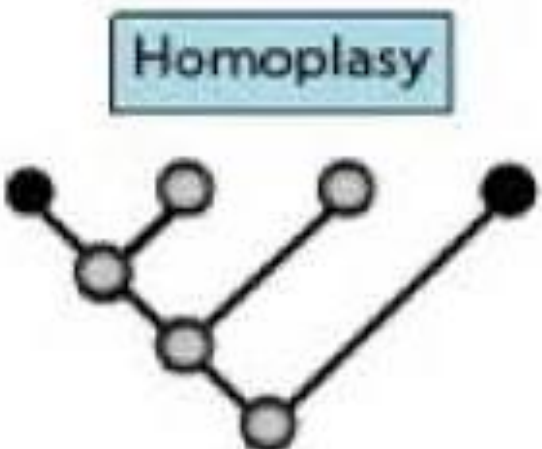
Derived trait unique
to only one taxon



Shared derived trait
present in a particular
clade



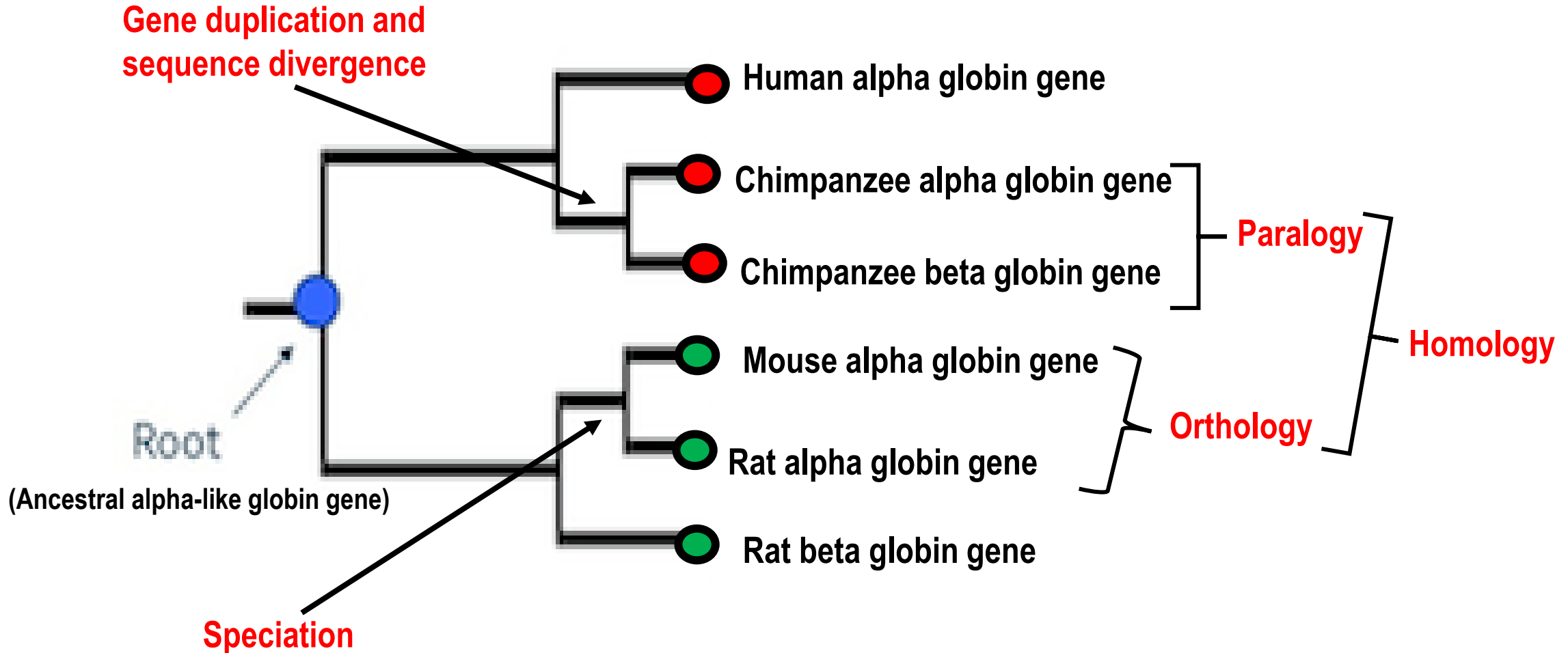
Shared trait not due to
common ancestor



Ancestral trait (○)
Derived trait (●)

Homology: Orthology and Paralogy

Homology: Similarity by common ancestry



Thank You!