E-RESOURCE

UDAI PRATAP COLLEGE, VARANASI

Programme/Class: Diploma in Plant Identification, Utilization & Ethnomedicine

UG, Year: II, Semester: III, Paper: I, Subject: Botany

Course code: B040301T (NPE-2020); **Course Title:** Flowering Plants Identification & Aesthetic Characteristics

Credits: 4, Course compulsory, Max Marks: 25+75

Class: B.Sc. –Botany

Year: II, Paper: I, UNIT: II

Topic: Introduction to Taxonomic Evidences fromMolecular biology data (Protein and Nucleic acid homology)

Name: Prof. Ajai Kumar Singh, Department of Botany, Faculty of Science,
Mobile No. 9450538149, E-mail: ajaiupcollege@gmail.com

MOLECULAR SYSTEMATICS

- *Uses macromolecular data such as DNA/RNA sequences for the study of evolution and phylogeny.*
- *A plant cell contains three genomes. The chloroplast genome is the smallest and nuclear genome is the largest among the three types of plant genomes [third one is the mitochondria genome].*
- *The homologous regions of DNA provide character and character states which are compared and used for inferring the phylogenetic relationships.*
- *Unweighted Pair-Group Method with Arithmetic Averages* ***[UPGMA]****.*
- *There are two general types of trees:* ***Cladogram*** *and* ***Phylogram****.*
- ***BOOTSTRAPPING*** *is a standard statistical technique for evaluating the robustness of the computed trees.*
- *The main aim of phylogenetic analysis is to reveal the evolutionary relationships between different taxa and obtain an understanding of the evolution of life.*
- *Phylogenetic trees provide a concise way to visualise evolution from common ancestors.*

The term '**molecular systematics**' is used to refer to macromolecular systematic, i.e., the use of DNA and RNA sequences to infer evolutionary basis and relationship among organisms.

Molecular techniques provide powerful tools for the study of evolution and phylogeny.

- The utilization of information on DNA and RNA for understanding the phylogenetic relations has received a great boost over the last two decades, meriting the establishment of a new field of study, referred to as Molecular Systematics.

  {DNA is responsible for building and maintaining organism structure, it gives a physical characteristic, that make organisms unique.

  GENE = Gr. Genos = 'birth'

  The word spawned 'genome' = DNA

  Denish botanist WILHELM JOHANNSEN coined the word gene to describe the Mendelian Units of heredity.**}**

  [GENE: The basic unit of heredity passed from parent to child. Gene are made up of sequences of DNA and are arranged one after the another, at specific location on chromosome. Basic unit of inheritance. Basic physical and functional unit of heredity.It's a segment of DNA that contains instructions for building and maintaining an organism. Genes serves as the blueprint for the synthesis of proteins and play a crucial role in the inheritance of traits from one generation to next.]
- As molecular data reflects *Gene Level Changes*, it was believed to reflect phylogeny better then morphological data.
- In many cases molecular data have allowed the placement of taxa whose relationships were known to be problematic.
- The closely related species are expected to have greater similarities in their genetic material than the distantly related species.
- Systematists use molecular data from three different locations with in a plant cell: *Chloroplast, Mitochondria and the Nucleus.*
- They yield three different types of genomes (DNA)
- Since mitochondrion genome (made up of 200-2500 kbp), undergoes a lot of rearrangements, many different forms may be found within the same cell, *hence is of little significance in phylogenetic systematics.*
- However, the chloroplast and nuclear DNA is highly stable not only within the same cell, but also within a species, and therefore serves as a useful taxonomic tool.
- The nuclear DNA, difficult to analyse, but has two great advantages in phylogenetic studies.

  (a). Certain nuclear sequences evolve more rapidly than chloroplast DNA (cpDNA) sequences, and thus allow finer level of discrimination at population level than cpDNA.

  (b). Also, whereas the nuclear genome is inherited Biparently (a combination of male and female genome), the chloroplast genome is inherited Maternally. Both mtDNA and cpDNA are Uniparental in origin. Thus, the hybrid plant will possess the nuclear component of both parents but the cpDNA and mtDNA will have only the maternal complement.

- Molecular data is being used to construct phylogenetic trees of green plants at almost all ranks.
- It principally helps in establishing affinities and relationship within and outside the ranks. For example, monophyly of families Apiaceae, Caprifoliaceae, Scrophulariaceae, Apocynaceae, Saxifragaceae etc. is well supported by molecular data analysis.

- Sequences of 18S and 26S genes have been used in phylogenetic studies, because they have some highly conservative regions which help in alignment, and other variable regions, which help to distinguish phylogenetic groups.
- In recent years ITS (internal transcribed spacer) region has also been used to determine relationships among species. The ITS region has also supported relationships inferred from chloroplast studies and morphology.
- By Gene Mapping, Jansen & Palmer (1987) found a unique order of genes in the large single-copy region of the chloroplast genome in Asteraceae. This unique order is explained by single inversion of the DNA, a feature lacking in all other angiosperms, strongly confirming that the family Asteraceae is MONOPHYLETIC.
- Similarly, the family Poaceae has three inversions in the chloroplast genome. Out of these three inversions, one is unique to the family and confirms its monophyletic status. Of the other two, one is shared with Joinvilleaceae and one with families Joinvilleaceae and Restionaceae, suggesting that these two are sister group of Poaceae.
- By using side-copy sequence analysis hybridization in *Atriplex* concluded that the division of this taxon is not correct.
- On the basis of sequence analyses of the plastid *atp*B and *rbcl* DNA, Bayer *et al*., (1999) found a support for an expanded order Malvales, including most of the taxa previously included in Sterculiaceae, Tiliaceae and Bombacaceae with the Malvaceae and sub-divide this enlarged family Malvaceae into nine subfamilies based on molecular, morphological and biogeographical data.
- Similarly, genome analysis of cereal grasses has provided useful information to systematists. Of the common cereal grasses, rice has the smallest genome (400 mb) (megabase is a unit of measurement used to designate the length of DNA/ 1mb = 1 million bases].
- Maize genome is 2500 mb, whereas the largest genome is found in Wheat (17,000mb).
- Genomic studies in genus *Gossypium* (Wendel *et al*., 1995) using isozymes, nuclear ITS sequences, and chloroplast restriction size analysis, indicated that New World diploids are monophyletic, as are the Old World diploids.
- Most common methods used over the recent years include studies on **chloroplast gene, ITS region of ribosome, phytochrome B, and granule bound starch synthase I.**
- An encouraging result of these diverse studies was met in tribe Stipeae of grasses.
- RbcL data has supported that Caryophyllaceae is monophyletic.
- It has also supported the union of family pairs viz. Asclepiadaceae-Apiaceae, Araliaceae-Apiaceae, and Brassicaceae-Capparidaceae.
- The data also supported the polyphyletic nature of Saxifragaceae and Caprofoliaceae.

PLANT GENOMES

DNA molecule contains the genes and the total DNA content of a cell is known as its **genome**.
Plants have three different genome:

a. Plastid;
b. Mitochondrial; and
c. Nuclear

As stated above, the chloroplast and mitochondrion are generally inherited uniparentally (maternally in flowering plants).
The nuclear genome is biparental.

Genome sizes are described in terms of the **number of pairs of DNA bases they contain**.

Genome	Inheritance	Genome Size (kbp)	Status
Chloroplast	Generally maternal	120-160	Smallest
Mitochondrion	-do-	200-2500	Medium
Nucleus	Biparental	$1.1x10^{6}$to $1.1x10^{11}$	Largest

[**A kilobase (kb) is a unit of measurement in molecular biology equal to 1000 base pairs of DNA/RNA. The total number of DNA base pairs on Earth is estimated at $5.0x10^{37}$ with a weight of 50 billion tonnes.**]

(bp = base pair; one bp corresponds to approximately 3.4 A (340 pm) of length along the strand, and to roughly 618 or 643 Daltons for DNA and RNA respectively. Kb (= kbp) – kilo-base-pair = 1,000 bp). [kb (- kbp) – kilo base pair – 1,000,000 bp. Gb – Giga base pairs – 1,000,000,000 bp].

MOLECULAR MARKERS

Chloroplast Genome

Chloroplast have circular genome (*Maurya et al.*, 2020). Typically plastid genome contains 120-160 kb. Chloroplast genome is distinguishable into a **small single-copy** region (**SSC**) and a **large single copy** region (**LSC**). In chloroplast genome most genes are in single copy.

The most characteristic feature of this chloroplast genome is the presence of two regions that encode the same gene, situated in opposite direction. These are known as **inverted repeats** (**IRs**). IRs separate the SSC and LSC regions. These genes tend to accumulate mutations more rapidly as compared to mitochondrial genes in plants.

cpDNA is being used more frequently in systematic and phylogenetic studies of plants.

ADVANTAGES OF **cpDNA** IN TAXONOMIC STUDIES

- **Relatively smaller size (120-200kb) and are present in many copies per cell, making it easy to isolate in sufficient quantities from even very small amounts of plant material.**
- **Most genes in chloroplasts are single copy and easy to examine the entire genome.**
- **Conserved Genome (minimal variation encountered within and among conspecific populations).**
- **The lack of frequent structural changes (inversions, transpositions, deletions, and insertions) in the chloroplast genome makes it relatively easy to work in comparative studies.**
- **The relatively slow rate of nucleotide substitution in cpDNA minimizes the problem of parallel and convergent evolution when comparing genomes of congeneric species.**
- **Uniparental inheritance;**
- **Large number of chloroplasts per cell;**
- **Relatively stable not only in an organism but also in a species;**
- **High nucleotide substitution rates;**
- **Easy to isolate and analyse;**
- **All cpDNA molecules carry basically same set of genes, however, arrangement varies in different species.**
- **Introns and spacers of cpDNA have been widely used but substitution rates are often too low to distinguish closely related species.**

The use of sequence data from the cpDNA has proven to be very useful in systematic interpretations and **it has been sequenced in many taxa and found very informative.**
Some commonly used cpDNA geneand genic regions employed in plant systematics include
i.*rbc*L, ii.*trn*, L-trnF, iii.*ndh*F, iv.*mat*K, v.*trn*H-*psb*A.

***rbc*L**
Among plant genes, *rbc*L (Ribulose-1,5-biphosphate carboxylase/oxygenase) was the first to be sequenced. It encodes the large subunit of ribulose-1,5-biphosphate carboxylase (RUBISCO), located in the LSC region of the chloroplast genome and is a critical enzyme in photosynthesis. The length of the genome is 1430bp.
In most of the phylogenetic studies, this locus has proved to be useful for reconstructing phylogenies at the generic level and above. Good quality sequences can be easily obtained.
The *rbc*L sequences are phylogenetically conserved and have served well across all green plants.

***ndh*F**
It is located in the SSC region of the chloroplast genome. This region codes for NADH dehydrogenase F gene. It is involved in various reactions of respiration having 2235 base pairs and its 5´ region is very different from the 3´ region. It is larger than *rbc*L and has a greater number of variable sites. Some indels **[An insertion/deletion polymorphism, commonly abbreviated "indel", is a type of genetic variation in which a specific nucleotide sequence is present (insertion) or absent (deletion)].** Some indels of this molecular marker have also been shown to be of phylogenetic significance. Its utility has been shown at the familial, subfamilial, tribal and generic levels due to the longer size and high sequence divergence.

***mat*K**
*mat*K (Maturase K) is a group II intron present in the gene *trn*K that codes transfer RNA for lysine. It has a length of 1550 bp and encodes for an enzyme known as maturase, which splices type II introns from RNA transcripts. It is located in the LSC region of the chloroplast genome. It is a fast-evolving region with high substitution rates. It is widely used in systematic due to its low transition/transversion ratio, presence of conserved regions, and the ability to discriminate between species.
***mat*K has provided useful comparative data among genera within a family and even among species within the genus.**

***atp*B**
Located in the LSC region of the chloroplast genome and codes for beta subunit of ATP Synthetase. It is involved in the synthesis of ATP through proton translocation. Due to the absence of introns [**Introns** are noncoding sections of an RNA transcript, or the DNA encoding it, that are spliced out before the RNA molecule is translated into protein. The section of DNA (or RNA) that code for proteins are called exons.], this region is easily aligned during phylogenetic analysis. It is conservative in nature and has also been used to analyze the phylogenetic relationships in ferns.

***rp*/16**
This is noncoding intron region is of about 1059 bp. It has a lower transition/transversion ratio but shows higher nucleotide divergence and genetic distance.

***trn*H – psbA [tRNA-Histidine-Photosystem II protein D1] intergenic spacer**

It is one of the most variable intergenic spacers present in the chloroplast genome of angiosperms. It varies from 296-1120 bp in length with an average of approximately 450 bp, located in LSC region of the chloroplast genome. This region is useful due to its high variability and the ability to discriminate between species but the small length limits its usefulness. As the length is short, the number of variable sites provided by this locus is less. The ease of amplification makes it a highly used locus.

***trn*L-*trn*F: Intron and intergenic spacer**

The*trn*L and *trn*F code for tRNAs for leucine and phenylalanine, respectively. The *trn*L-*trn*F sequence consists of an intron in the transfer RNA gene *trn*L (UAA) and the adjacent intergenic spacer between *trn*L and *trn*F (GAA). It is located in the LSC region of the chloroplast genome. It is highly used in systematic studies due to its high amplification rate, universality and good species discrimination ability. This marker is very useful for analyzing relationships at lower levels of the hierarchy particularly between genera and species, however, it has been used up to tribal level.

Mitochondrial Genome

An important component of eukaryotic cell and it is bound by double layered membrane. **It is a true organelle.**

Mitochondrion contains its own genetic information in the form of **double stranded circular DNA.**The DNA found in mitochondrion is called **mtDNA**. Mitochondrial genome carries several genes and is **maternally inherited**. In contrast to cpDNA sequences, mtDNA has been **employed much less frequently in plant systematic due to high degree of intramolecular recombination and a low rate of base pair substitution.**

mtDNA in plants is **large, variable** and is also **less abundant** in plant leavesas compared to cpDNA.

Presently, very few sequences of mtDNA are availablefor study in plant systematics.

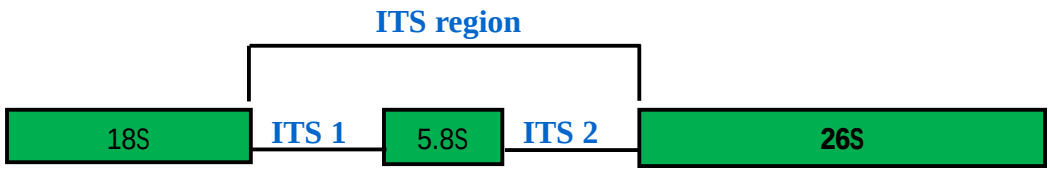
Mitochondrial DNA has been the molecule of choice in the animal kingdom.

Nuclear Genome

Plant nuclear genomes vary greatly in size.

The model organism *Arabidopsis thaliana*has one of the smallest known (1.35×10^8bp) plant genomes. The genome of the lily (*Fritillaria assyriaca*) is one of the largest (1×10^{11}bp).

Ribosomal genes are arranged in tandem repeats and are subjected to concerted evolution. It is **biparentally inherited**. The nuclear ribosomal genes encode the small subunit (18S) and the large subunit (26S) which are separated by a smaller (5.8S) gene.



Structure of the ITS region[WORKHORSEof plant molecular systematics]

Between the three genes, there are short internal transcribed spacers (**ITS**). Each set of the three genes (18S, 26S, 5.8S and ITS regions) is separated by a larger spacer, called intergenic spacer (**IGS**).

Sequences of the 18S gene (about 1800 bp) and 26S gene (about 3300bp) have been used for inferring relationships among large group of plants. Both have some regions that are highly conserved and others that are quite variable, which helps in distinguishing phylogenetic groups.

The nuclear genome sequences of **nrDNA** (Nuclear Ribosomal DNA), particularly the ITS region, have been **widely used in phylogeny reconstructionat both the generic and specific levels because of their higher substitution rate compared to commonly used regions of cpDNA.**

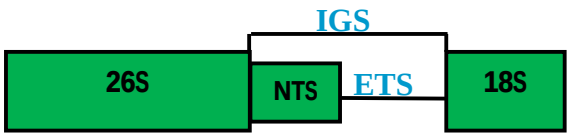
ITS is a highly suitable and powerful region for resolving plant taxonomic and phylogenetic problems in most plant lineages, especially between closely related species.

WORKHORSE OF PLANT MOLECULAR SYSTEMATICS

The entire ITS region [ITS1+5.8S+ITS2] can be easily amplified using universal primers. Total length of ITS regions plus intervening 5.8S gene is fairly short and relatively uniform [600-700 bp]. ITS is also called **Workhorse of plant molecular systematics.** It provides sufficient number ofparsimony (A rule used to choose among possible cladograms, which states that the cladogram implying the least number of changes in character states is the best.) informative characters, with good resolution and strong clade support.

Another region of the nuclear rDNA, i.e., **ETS** [External Transcribed Spacer]. It is extensively in use to supplement the ITS data in the phylogenetic analysis. It is longer than ITS 1 and ITS 2 put together. **It is assumed that the restrictions posed by ITS in resolving systematic relationships might be overcome by the use of ETS.**

However, the use of ETS is limited due to the presence of highly variable non-transcribed spacer (NTS) region flanking its 5´end [Fig.]. Because of this no universal primers are available and thus sequencing becomes challenging. To overcome this limitation, the entire IGS region (that includes NTS+ETS) is amplified using universal primers flanking 18S and 26S region.



Position of the ETS region

GENERATING DNA SEQUENCE DATA

DNA sequence data basically refers to the sequence of nucleotide: Adenine (A), Thymine (T), Cytocine (C), Guanine (G), in a particular stretch of DNA of a given taxon. **The homologous regions of DNA (i.e., regions having similarity due to ancestry) among the taxa under study provide character and character states which are used for inferring the phylogenetic relationships. DNA obtained from the samples are compared.**

DNA sequence data are generated in two ways:

1. A gene-by gene approach in which a gene of interest is selected, isolated from a large number of plants and sequenced.
2. A genomic approach, in which an entire chloroplast or nuclear genome is sequenced and the sequences of many genes from the genome are analyzed.

Steps in Acquiring DNA Sequence Data

Extraction of DNA either by CTAB method or DNA Extraction kit
[From live leaf material or silica dried leaf]

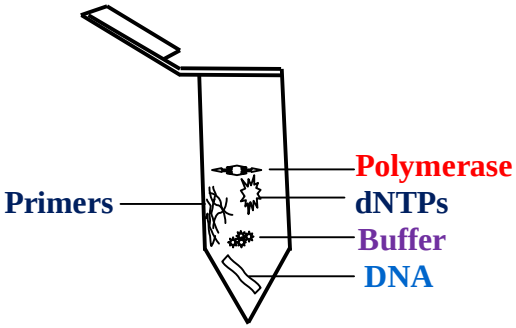
↓

Genomic DNA is amplified by using polymerase chain reaction [PCR]

↓

Product is cleaned and sequenced in automated sequencer

Polymerase Chain Reaction [PCR]



The technique is used to amplify specific, target DNA fragments from low quantities of source DNA. Different components of PCR include:

1. **DNA template:** It is the sample DNA that contains the target sequence.
2. **DNA polymerase:**It is a type of enzyme that synthesizes new strands of DNA complementary to the target sequence. PCR requires a DNA polymerase enzyme that makes new strands of DNA, using existing strands as templates. The DNA polymerase typically used in PCR is called *Taq* polymerase, after the heat-tolerant bacterium (*Thermus aqaticus*) from it was isolated.
3. ***Taq* buffer:** Taq buffer with $MgCl_2$provides an optimal and stable chemical environment for the DNA polymerase to work adequately.
4. **Primers:**PCR primers areshort pieces of single stranded DNA (15-30 nucleotides in length) which bind to certain nucleotide sequences along the DNA strand.
5. **dNTPs:**Deoxynucleotide triphosphate are single units of the bases A, T, G, and C, which are essentially “building blocks” for new DNA strands.

STEPS IN PCR

There are three steps involved in polymerase chain reaction.

Denaturation: The initialdenaturation step is carried out at the beginning of PCR to separate the double-stranded template DNA into single strands so that the primers can bind to the target and initiate extension.

Annealing: After denaturation of the double stranded DNA, each primer hybridizes to one of the two separated strands.

Extension: After primer annealing, the next step in PCR is to extend the 3′ end of primers, complementary to the tempelate. In this step, 5′ to 3′ polymerase activity of the DNA polymerase incorporates dNTPs and synthesizes the daughter strands.

PCR steps of denaturation, annealing, and extension are repeated (or "cycled") many times to amplify the target DNA.

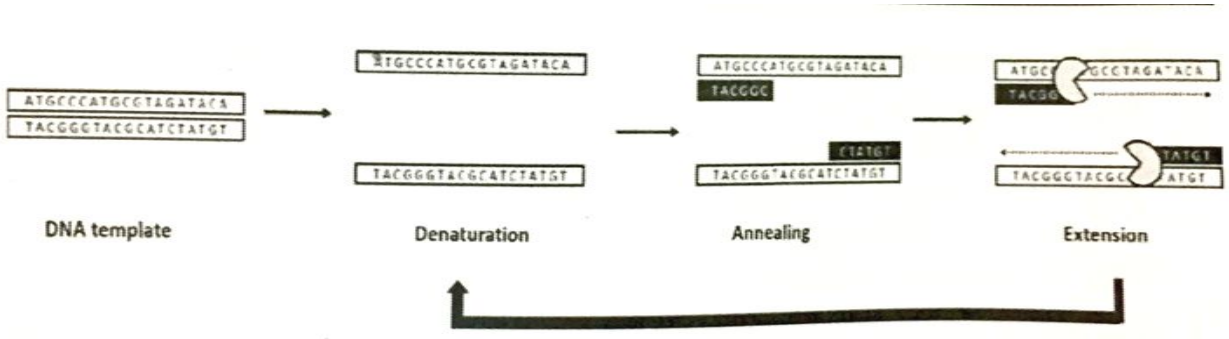
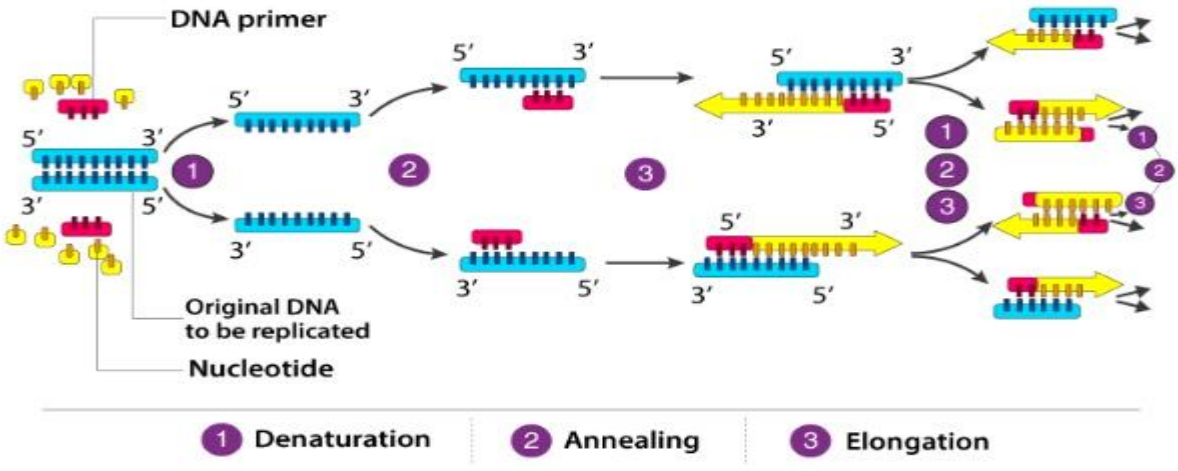


Fig. Steps involved in Polymerase Chain Reaction (PCR)



PCR

Gel Electrophoresis: Gel electrophoresis is the technique that is used to separate proteins as well as nucleic acids on the basis of charge in the presence of an electric field. In molecular systematics, once the DNA molecules have been amplified by PCR, they are loaded onto a gel made of agarose and subjected to electric charge. The negatively charged DNA molecule moves from negative pole to the positive one. The rate of migration is dependent on the size of the DNA molecule. **The larger the size, slower is the rate of migration.** The bands on the gel are not visible with the naked eye and hence the gels are stained (usually with **ethidium bromide**) before loading the samples. A DNA ladder is loaded next to the DNA sample. The DNA ladder produces fragments of known size or base pairs. It can thus be used to determine the size of the corresponding bands on the gel. The gel is visualized under UV light.

DNA Sequencing Reaction:

PCR product is cleaned and processed for sequencing. The machine which reads sequences is known as SEQUENCER.

The replicated DNA is placed in a tube with DNA polymerase, nucleotides, primers, and dNTPs. The sequencing reaction is present in following figure.

After sequencing, electropherograms obtained from the sequencer (Fig.) are aligned for phylogenetic analysis.

Sequences of both individual genes and whole genomes are available in the GenBank, at the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

GenBank is the repository of a vast amount of publicly accessible data.

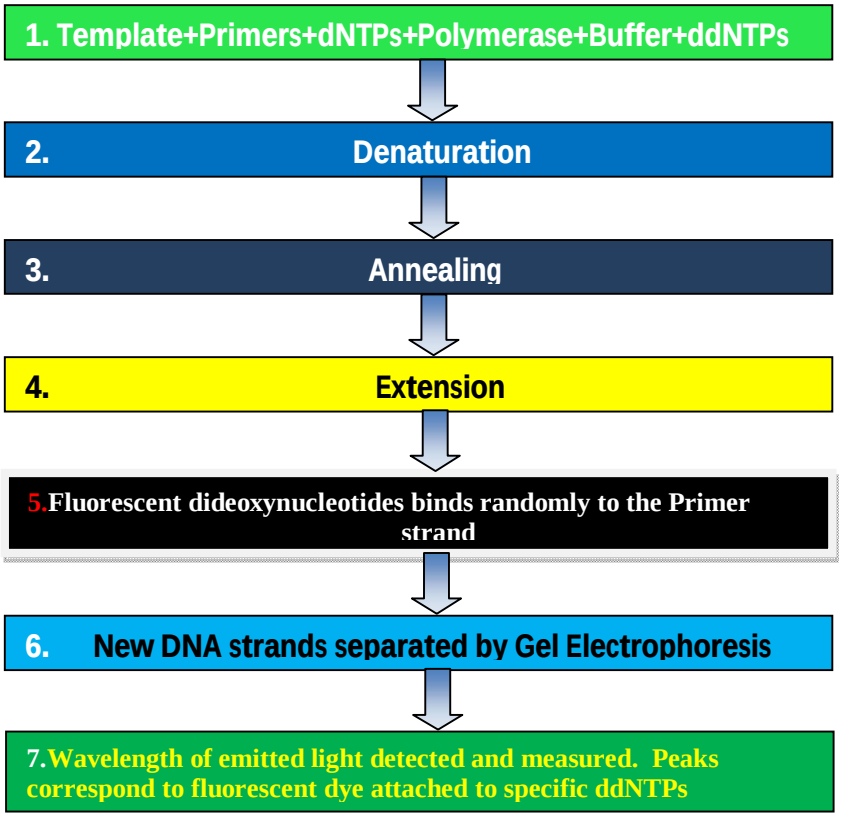


Fig.: Steps in a typical sequencing reaction

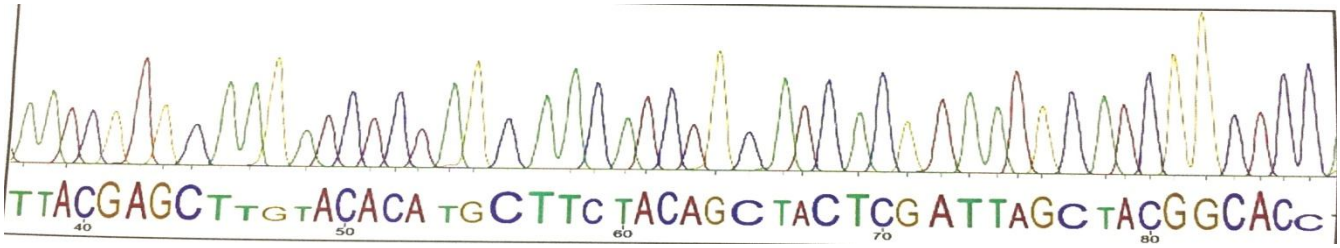


Fig. Sequence of nucleotides of the DNA strand

Sequence Alignment:

After the sequences have been received from the sequencer, they need to be aligned. Accurate alignment is crucial to any phylogenetic analysis.
The alignment is necessary so that identical nucleotides can be determined.
In sequence alignment four steps are involved:

IDENTICAL NUCLEOTIDES ARE ALIGNED

↓

TRANSITION

↓

TRANSVERSIONS

↓

Gaps needed in order to accommodate alignments that are separated by one or more nucleotides Compared to the other sequence being aligned

Phylogenetic trees are generated from aligned DNA sequences. **Two DNA sequences that show high similarity are generally presumed to be homologous (evolved from a common ancestral sequence).**

PHYLOGENETIC ANALYSIS

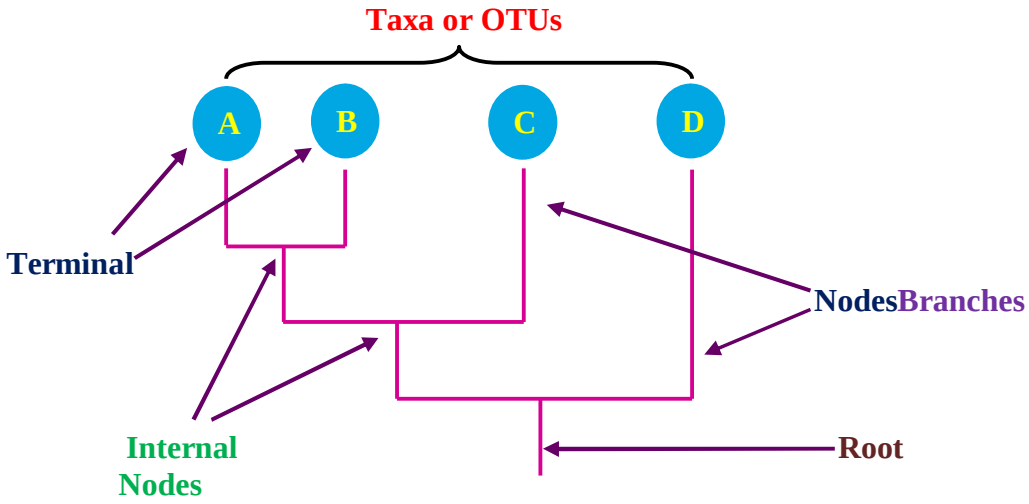
Phylogenetic analysis aims at uncovering the evolutionary relationships between different species or taxa, to obtain an understanding of the evolution of life on earth.

The purpose of phylogenetic analysis is to place species in the context of their common ancestry relationship.

Phylogenetic trees are generally generated from molecular sequences.

STRICT CONSESUS TREES: Contain only monophyletic groups that are common to all trees.

A phylogenetic tree is a graphic representation of the historical course of evolution (See fig. below). Phylogenetic trees provide a concise way to visualize evolution as descent from commonancestors and are well suited to represent evolutionary histories. **The phylogenetic tree and corresponding classification predict properties of newly discovered or poorly known organisms.**



A typical phylogenetic tree showing Root, Nodes and Branches

A tree is a simple structure that is composed of two elements-nodes and branches (also called edges). The point of divergence of one clade into two is termed a node. Nodes are the points where branches meet.
The nodes may be external or internal.
The **External Nodes**are located at the tip of the tree and represent taxa.
The tips of a tree, sometimes referred to as leaves, are the external nodes and biologically they represent existing taxa.
The Operational Taxonomic Trees (**OTUs**) are the names of the sequences, genera or species. The OTUs are the ends (terminal nodes) of the terminal branches. The branches joining them to other parts of the tree are called terminal branches, and all the other branches are internal branches. They represent the real data (morphological, nucleotide or amino acid characters) from which everything else in the tree is inferred. In case of sequence data, the external nodes are represented by sequences.
The region between two nodes is called an **Internode.**

Types Of Trees

There are two general types of trees: **CLADOGRAM** and **PHYLOGRAM.**
Trees can be produced that are Rectangular or Slanted in form.
For Rectangular Phylogenetic Trees, only the horizontal branches indicate the phylogenetic distance. The vertical branches are included only to join together the horizontal branches. The trees may be circular which are simply rectangular trees bent into a circle.

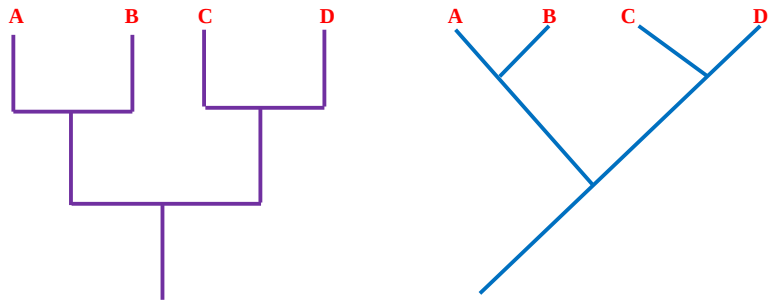


Fig. A phylogenetic tree in rectangular and slanted form (The tree can be drawn as angled form [right]or squared form [left].

Cladogram: In a cladogram, the branches are not always proportional to the distances between the OTUs. The branching is meant only to indicate the relationship between the OTUs. Cladograms show branching patterns to indicate evolutionary relationships, but usually do not indicate phylogenetic distances.

Fig. Phylogenetic tree drawn as cladogram (left) and phylogram (right). The branch lengths are unscaled in the cladogram and scaled in the phylogram.

Phylogram: A diagrammaticrepresentation in which branch lengths are proportional to the amount of change attributed to a particular branch.

In phylogram, the edges are drawn to scale and represent distances or the number of mutations along an edge.

These trees convey both the branching pattern and the distance among the OTUs.

The term Phylogram is often used for a Cladogram that has an absolute time scale.

Phylogram show evolutionary relationships as well as phylogenetic distances.
For rectangular phylograms, the horizontal branches convey the distances, while the vertical branches are only used to join adjacent branches together.
They contain no distance information.
Short branches indicate more recent changes, while long branches indicate more distant changes.

Rooted/Unrooted Trees
Trees may be rooted/unrooted. The branching pattern of a tree, whether rooted or unrooted, called a TOPOLOGY.

Rooted trees have a root that denotes common ancestry.

Unrooted Trees

An unrooted tree is a tree without a defined root. [Fig. below]
In an unrooted tree, all nodes of degree 1 are called **leaves** and all other nodes are called **internal nodes**.
In an unrooted tree the branches represent evolutionary lineages.
An unrooted tree is usually drawn to scale as a radial diagram.
An unrooted tree only depicts the relationships among the taxa, and does not represent any evolutionarypathway.

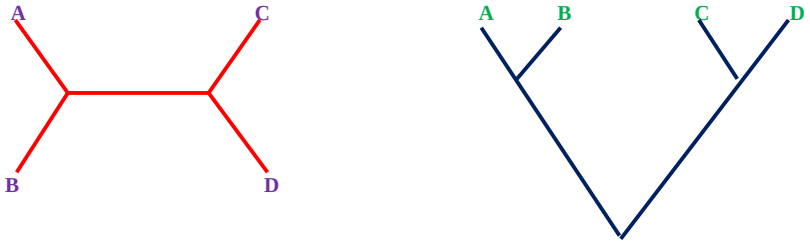
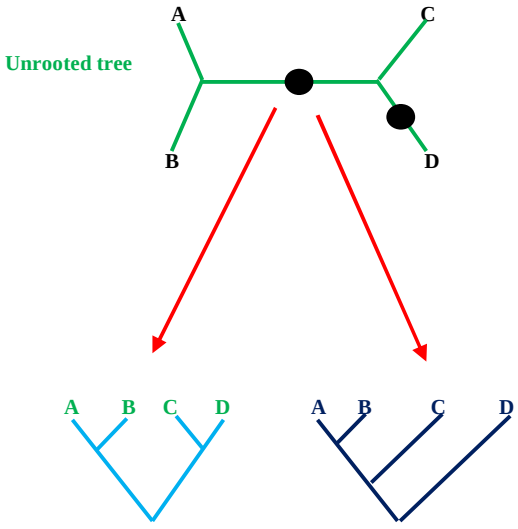


Fig. Unrooted and Rooted tree

Rooted Trees

The evolutionary of a set of species is usually described by a rooted phylogenetic tree [Fig]. A rooted tree can be depicted as a phylogram or cladogram. A tree is said to be rooted if there is a particular node, the root, from which a unique directional path leads to each extant [taxa that are still in existence, i.e., are still alive] taxon.



Rooted trees

Fig. Rooted tree from an Unrooted tree

Polytomy

Polytomies in a tree diagram represent complete uncertainty [Fig]. The unresolved complex of lineages is known as **polytomy.** In fig. below, A is sister to B which is represented by the dichotomous branching. However, the relationship between C, D, and E is uncertain and hence shown by polytomy.

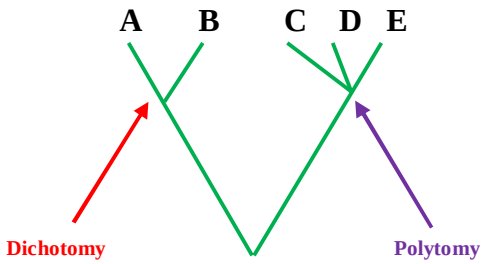


Fig. A phylogenetic tree showing an example of bifurcation and polytomy

Phylogeny

Phylogeny describes the evolutionary history of a set of taxa.
A group of organisms is called a **clade**or a **monophyletic group,** if it contains all the descendants of a common ancestor itself.
Phylogenetic trees are generally computed from aligned DNA or protein sequences.

Outgroup or Sister Group

A **sister group** is defined as an OTU that is closely related, but clearly outside the group being analyzed. In this case, the sister group is termed as outgroup and the group being analyzed is ingroup [Fig. below]

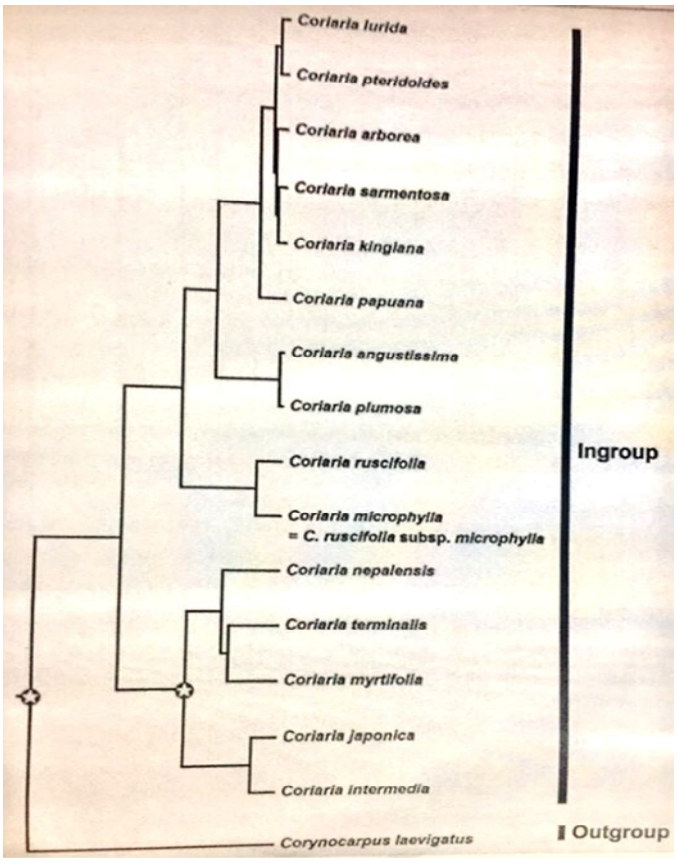


Fig. A phylogenetic tree showing outgroup and ingroup (After Renner *et al.*, 2020)

Rotation of trees

Trees can be rotated at any of the nodes because this does not alter the relationships or the distances [Fig.]. Free rotation is possible around each of the horizontal branches at their nodes. The relationships and distances (branch lengths) remain unchanged by each of these rotations.

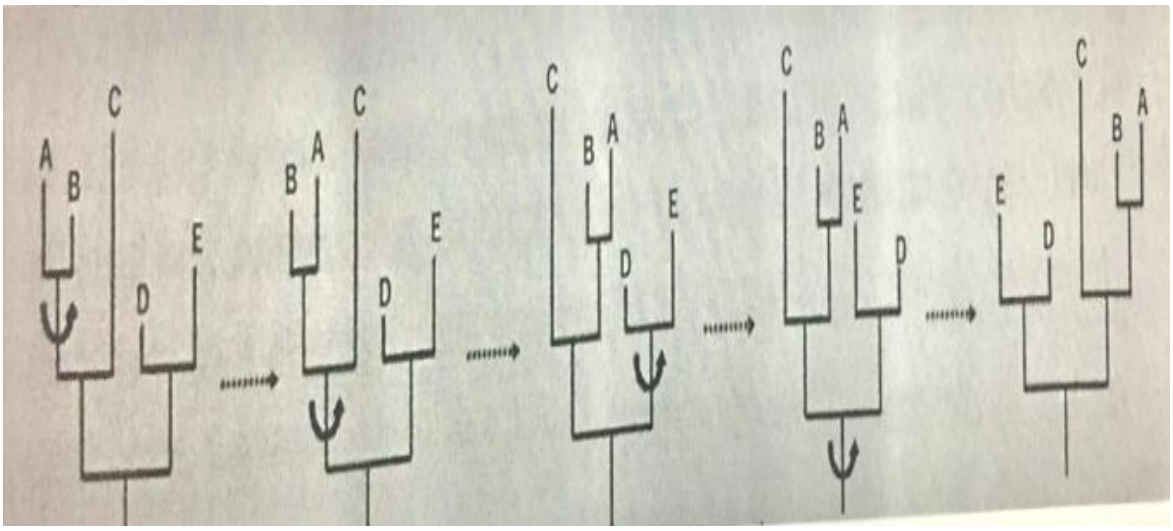


Fig. Rotation of branches

TREE TERMINOLOGY

Character trees and Sequence Trees: Trees based on characters are actually character trees, and trees based on sequences are sequence trees. The former shows the evolution of characters and latter evolution of sequences respectively.

Gene Trees: It is essential to remember that gene trees are not necessarily the same as species tree. The branching pattern of a tree constructed by using genes is called a gene tree. This may be different from that of a species tree, which is the best estimate of the cladogram depicting taxa relationship.

Species Tree: Species tree is the tree like history of population branching. Species tree is a tree that represents the evolutionary history of a group of species. In a species tree, the time of divergence between two species refers to the time when the two species were reproductively isolated.

Consensus Tree: Consensus tree are graphs that summarize the common knowledge claims of different phylogenetic trees. It represents those parts of the evolutionary history on which the different phylogenetic trees agree.

A consensus tree that includes only clades present in all input trees is called strict consensus tree.

Super Tree: It isa single phylogenetic tree that compiles many trees with overlapping taxa into a single comprehensive tree. It is helpful in reconstructing the phylogeny of large groups.

Diagonal Tree: A tree diagram format in which the nodes are connected by straight diagonal lines.

Phylogenetic network: A phylogenetic network is any graph used to represent evolutionary relationships between a set of taxa that labels some of its nodes [usually the leaves].

Evolutionary Models: Models determine the way in which a programme calculates branch lengths. Branch lengths are supposed to indicate the amount of genetic change between an ancestor and its descendants [Hall, 2008].

MAJOR METHODS FOR ESTIMATING PHYLOGENETIC TREES

Phylogenetic trees are reconstructions of evolutionary history based on data collected from sampled organisms. Majority of the phylogenetic reconstructions are based on molecular sequence data.

The genomic regions used in phylognetics must be carefully chosen to reflect the time of evolutionary events and/or the taxonomic group being evaluated.

There are two primary approaches to tree estimation:

1. **Distance-based approach:**This approach uses an algorithm to estimate a tree from the data. The algorithmic approach is fast and it yields only a single tree from any given data set.

   The two algorithmic methods used in phylogenetic analyses are:

   a. **Neighbour-joining [NJ],**
   b. **Unweighted Pair-Group Method with Arithmetic Mean [UPGMA].**

The two algorithmic methods, NJ and UPGMA are both distance methods.

2. **Tree-searching approach:**

The tree-searching method estimates many trees, then uses some criteria to decide which is the best tree or best set of trees.

The tree searching methods are generally slower, and some will produce several equally good trees.

Methods included in tree-searching approach are:Parsimony, Maximum Likelihood and Bayesian analysis.

Distance-based Approach

UPGMA

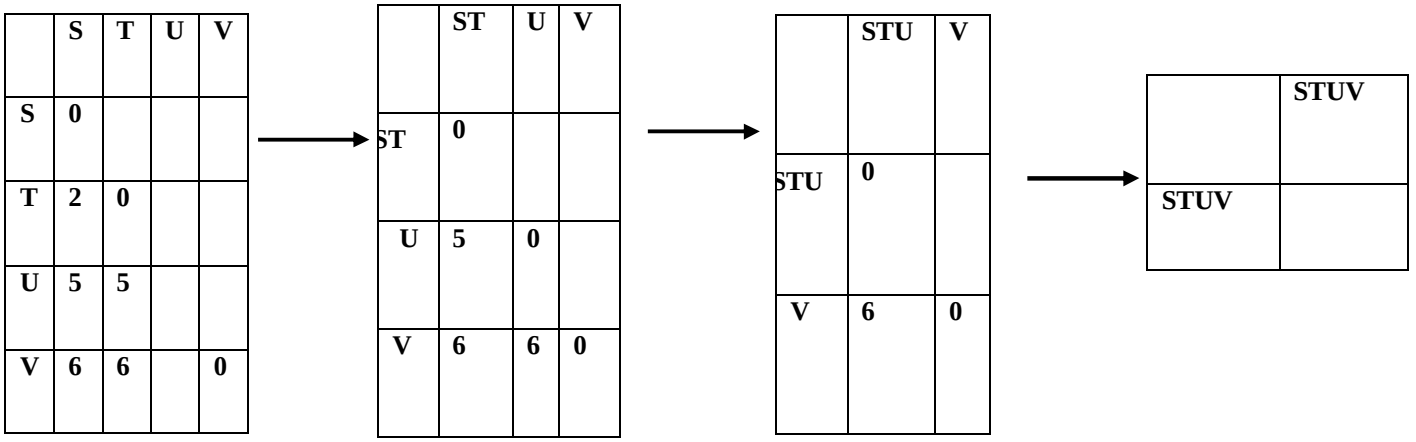
UPGMA stands for **Unweighted Pair-Group Method with Arithmetic Averages**, developed by Sokal & Michener (1958). It is an example of a clustering method.

A tree constructed by UPGMA method is called a PHENOGRAM.

It is an 'ultrametric' method, i.e., it assumes that all taxa are equidistant from the root and it hence depicts a rooted tree with branch lengths not corresponding to the evolutionary changes.

UPGMA produces good trees when gene frequency data are used for phylogenetic reconstruction .

In UPGMA matrix values are presented in the matrix form [Fig.].



Calculation of distance is done as: d(ST, U)=[d (S, U) + d(T, U)]/2 = (5+5)/2 = 5

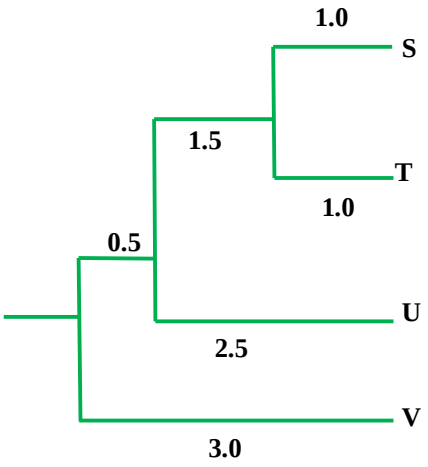


Fig. Sequential clustering in the UPGMA method and the resultant phylogenetic tree

Neighbour Joining (NJ)

It is the most widely used and powerful system of the distance methods, developed by Saitou and Nei (1987). An efficient tree building method based on the minimum evolution principles. In distance method, the evolutionary distance between every pair of taxa is calculated. This distance is then used to generate the evolutionary tree as the distance between any two taxa is the sum of all branches between them.

This method does not examine all possible topologies, but at each stage of taxon clustering a minimum evolution principle is used. This method is called the **neighbour joining (NJ) method.**

In NJ method, neighbours are defined as two taxa that are connected by a single node in an unrooted tree. NJ produces a single, bifurcating tree i.e., each internal node has exactly two branches descending from it. NJ method produces a phylogram, rather than a cladogram [Fig.]. the tree generated by this method is an unrooted tree. NJ is used frequently for moderate and large datasets because it is rapid due to its relatively simple calculations. This method does not involve any measure of the quality of tree obtained.

NJ differs from UPGMA in that it does not construct but directly calculates distances to internal nodes.

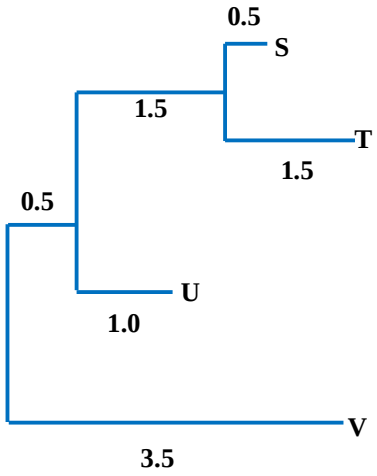


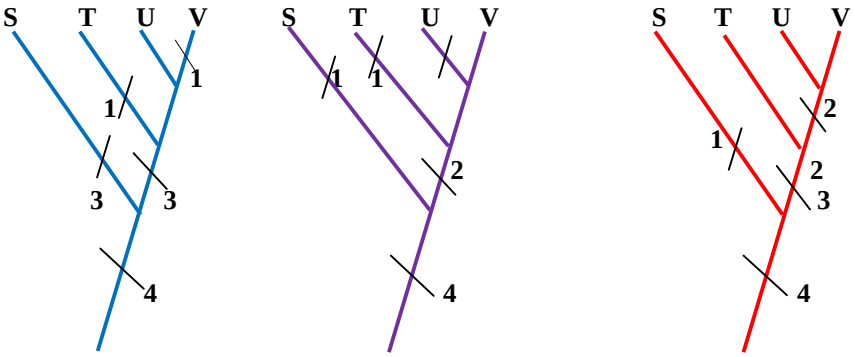
Fig. A NJ tree showing evolutionary distances on the branches

Tree-searching Approach

Maximum Parsimony (MP)

One of the most widely used sequence-based tree reconstruction methods [Fig.]. Parsimony simply means the simplest explanation.

The basic premise of Parsimony is that taxa sharing a common characteristic do so because they inherited that characteristic from a common ancestor.



Number of steps: 5 **Number of steps: 5** **Number of steps: 4**

Fig. The most parsimonious tree from the three possible rooted trees of four taxa is the one having least number ofevolutionary steps.

Parsimony looks for the tree or trees with the minimum number of changes. In other words, maximum parsimony is a method in which the best hypothesis is taken to be that which minimizes the total cost of the implied character state changes. A MP method searches for phylogenetic tree that explains the given dataset using a minimum number of observable mutations. Parsimony operates by selecting the tree or tree(s) that minimize the number of evolutionary steps.

There are three main search strategies that can be used in MP.

(1) **Exhaustive search:** In this search, the length (number of total nucleotide in the tree) of every possible tree is calculated, and one or more MP trees is/are reported.

(2) **Heuristic search:** In heuristic search, a random number is generated from among the number of possible trees, and the length of the tree corresponds to the selected random number. Adjacent trees are then analyzed until a minimum length tree is found, which is reported as the MP tree.

(3) **Branch-and-bound search:** It is also called **semi-exhaustive method** because it does find the MP tree, but does not examine every single tree. It begins by joining each OTU one by one, and creating a "search tree". In this method, large groups of trees are eliminated from the search because they all are longer than one being evaluated, and, therefore, longer than the current shorted trees.

MP differs significantly from NJ and UPGMA, which are based on the initial distance calculations.

Maximum Likelihood (ML)

Maximum likelihood looks for the tree that, under some model of evolution, maximizes the likelihood of observing the data. ML almost always recovers a single tree. Some programmes can be used to save multiple trees. A ML method aims at determining a tree that maximizes the likelihood of generating the given dataset, under a given model of evolution. There are several advantages of ML method. It provides a systematic framework for explicitly incorporating assumptions and knowledge about the process that generated the given data.

This method is based on the likelihood of sequence changes as determined by the frequencies of changes of each of the types of bases (or amino acids). ML method begins by calculating the base frequencies in the entire dataset and then determines the probabilities of change among each of the possible base changes. The ML begins to join the OTUs together, beginning with the two that are closest, and proceeding stepwise by adding each successive OTU onto the tree in order of its distance to the other OTUs. It then begins to calculate each of the branch lengths based on the likelihood of base changes along that branch. Once it has calculated all of the 1nL (natural log of the likelihood) of the branches, it calculates a value of the entire tree by summing all of the 1nL branch lengths. This will be the ML tree [Fig.].

.....

Bayesian Analysis

Yang *et al.* (1955) developed a Bayesian approach of deducing a Bayesian inference of phylogenetic trees. The goal of Bayesian analysis is to estimate the distribution of a

quantity called the posterior probability of a phylogenetic tree (Huson *et al.*, 2010). This method is similar to ML because it relies on the evaluation of probabilities of sequence change along the branches of the trees but the calculations involved are different than those for ML.

Bayesian inference is based on the concept of **posterior probabilities**, i.e., on the probabilities that are estimated following some model (i.e., prior expectations). Bayesian methods attempt to compute the posterior distribution of trees, based on given input data, a specified model of evolution and a presumed prior distribution of phylogenetic trees [Fig.].

Marcov Chain Monte Carlo (MCMC)is a computationalmethod used ina Bayesian framework for estimating the posterior distribution of trees and other parameters. It estimates the divergence time on a fixed phylogenetic tree.

Bayesian majority-rule consensus tree: A consensus tree composed of clades with a posterior probability higher than some threshold (usually 0.5) that is used to summarize the results of a Bayesian MCMC phylogenetic analysis.

Bootstrap Analysis

Bootstrapping is a standard statistical technique for evaluating the robustness of the computed trees [Fig.].

Bootstrapping can be used to test the strengths of each of the branches on a phylogenetic tree.

A method for assessing strength (confidence) in a statistical inference by generating many random, pseudo-replicate to see how frequently a particular result is obtained. Both maximum parsimony and maximum likelihood methods aim at producing an optimal phylogenetic tree.

If bootstrap value is higher than 95%, the branch is considered well supported, those at 70-90% are moderately supported, those that are below 50% are unsupported by the data. Generally, the unsupported branches are based on only a few nucleotide changes, while well-supported branches are based on many nucleotide changes.

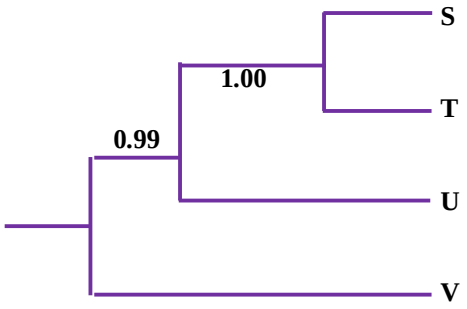


Fig. Bootstrap majority rule consensus tree

TYPES OF MOLECULAR DATA

Types of molecular data acquired include DNA sequences, DNA restriction sites, microsatellites, RAPDs and AFLPs. It may be divided into two broad groups:

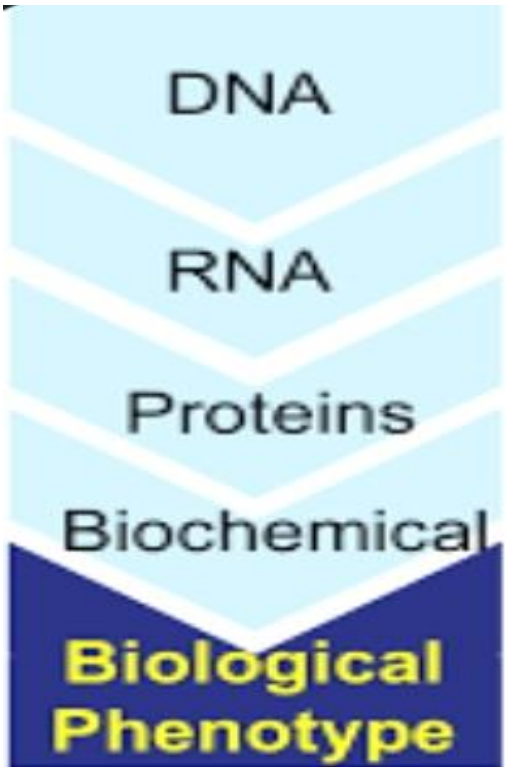
1. **PCR independent** (RFLP-Restriction Fragment Length Polymorphism)
2. **PCR dependent** (RAPD-Random Amplified Polymorphic DNA; PCR-RFLPs; AFLP-Amplified Fragment Length Polymorphism; Microsatellites, (also called SSRs, simple sequence repeats/STRs, Short tandem repeats)-DNA that contains tandem repeats, which are short (usually 2-5) repeats of nucleotides.

Random Amplified Polymorphic DNA (RAPD)

RAPD is used to study DNA variation within and between populations by using PCR (polymerase chain reaction). In this method, the total DNA from an organism is subjected to PCR by using short primers (10bp). The PCR technique amplifies intervening DNA segments of various lengths. When many such PCRs with random primers are performed, fragments can be found that distinguish plants or populations.

Amplified Fragment Length Polymorphism (AFLP)

This technique was developed by Vos *et al*., (1995). AFLP is a selective PCR amplification of restriction fragments from a total digest of genomic DNA. In this method, the entire DNA is first digested with two restriction enzymes. Those restriction fragments that perfectly match the primer sequences are amplified. Usually, 50-100 restriction fragments are amplified and detected on polyacrylamide electrophoretic gels. The polymorphism can be used to measure the intraspecific as well as interspecific variation.

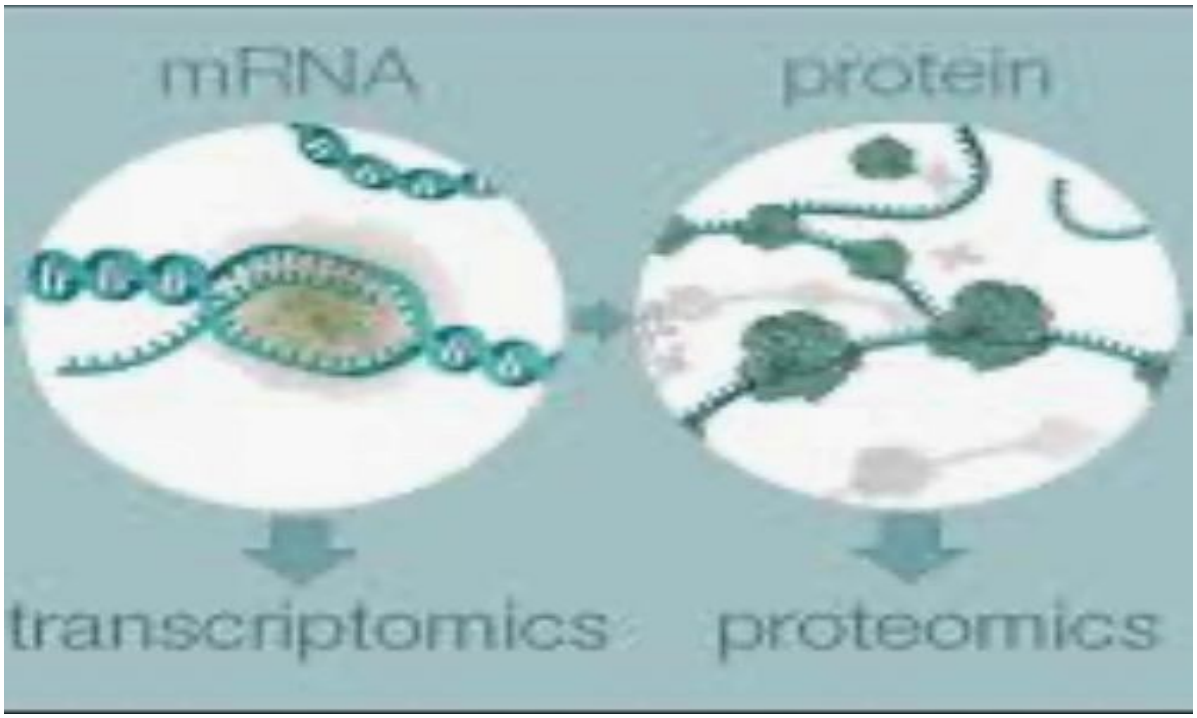


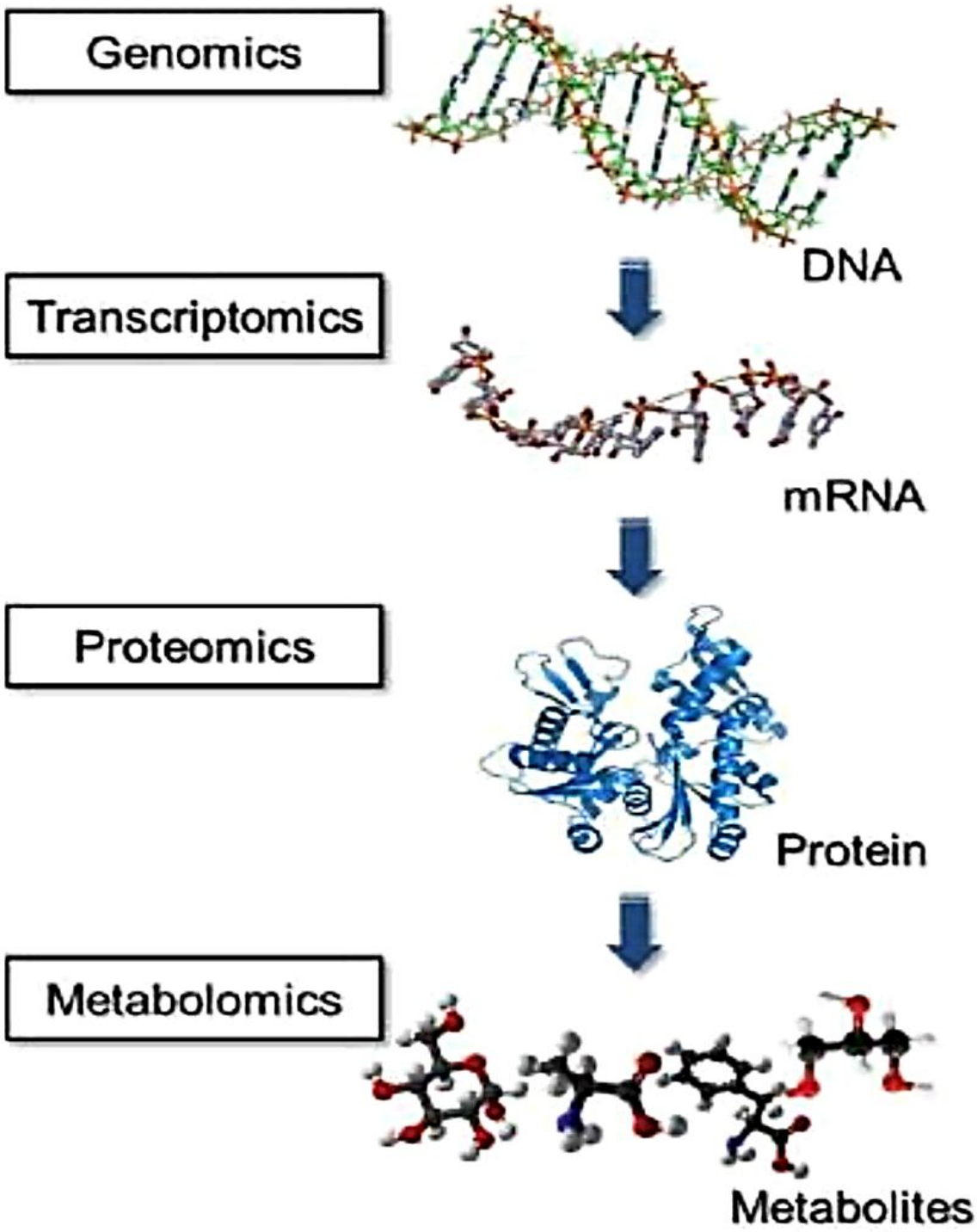
Genomics

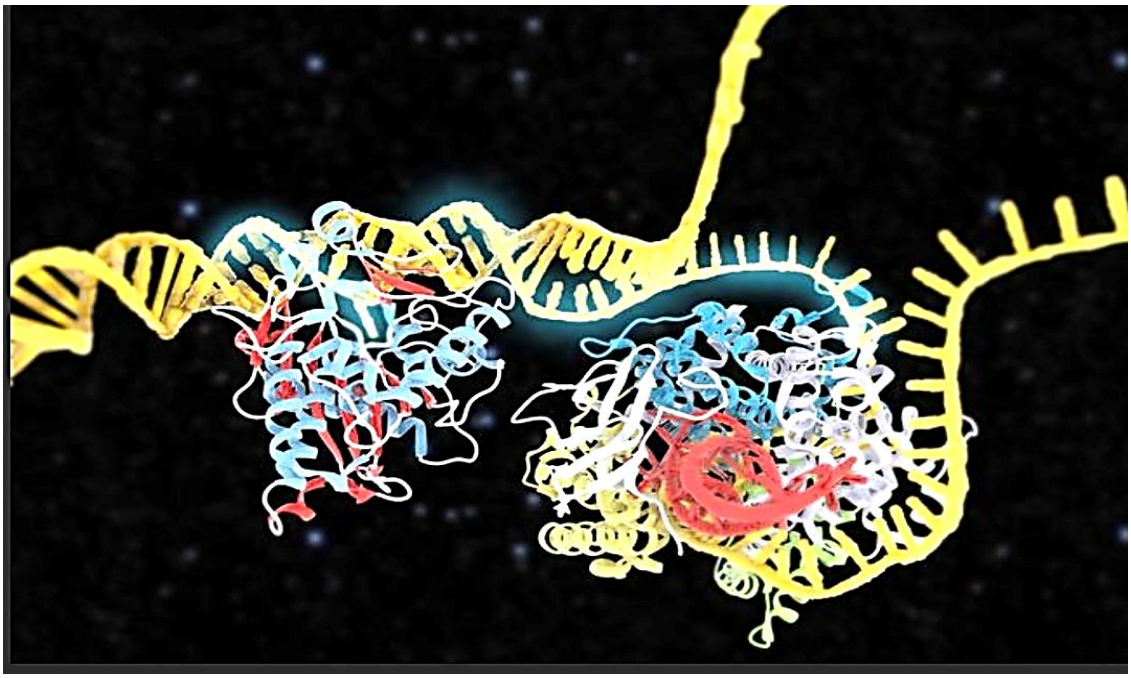
Transcriptomics

Proteomics

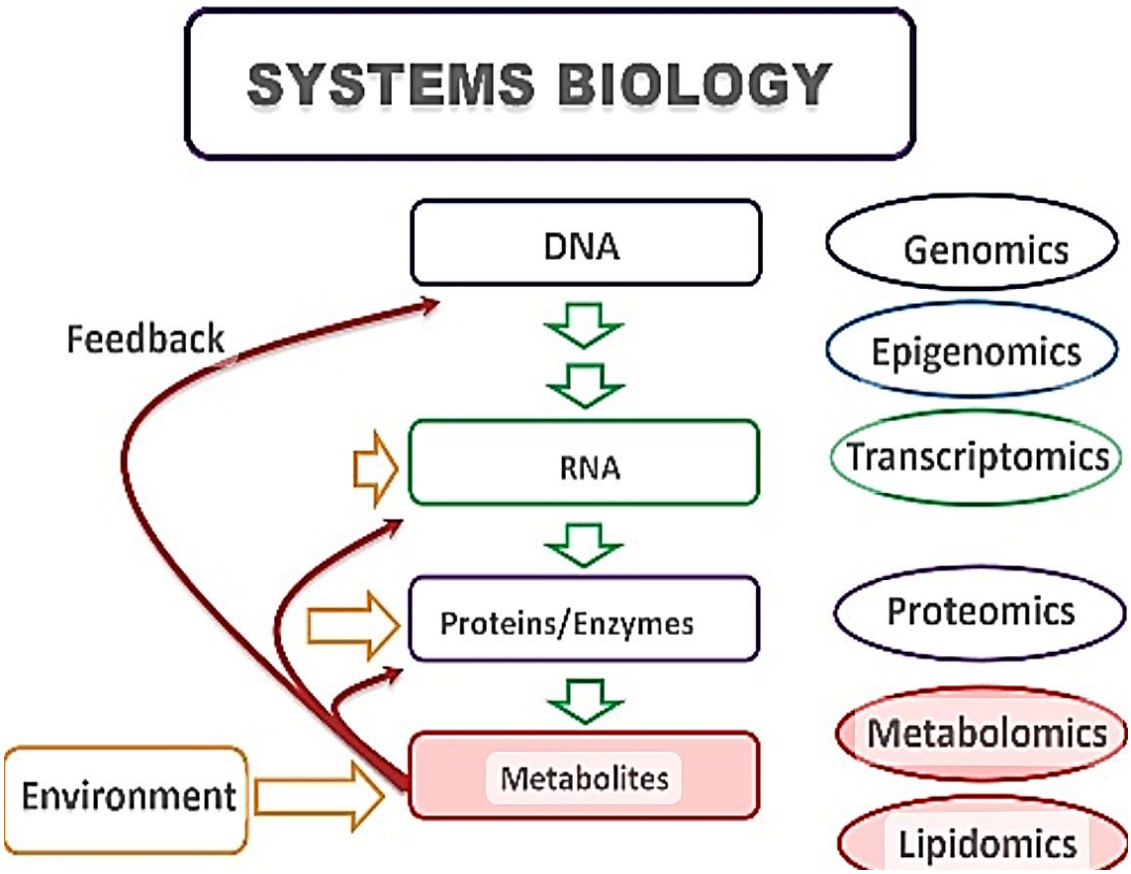
Metabolomics

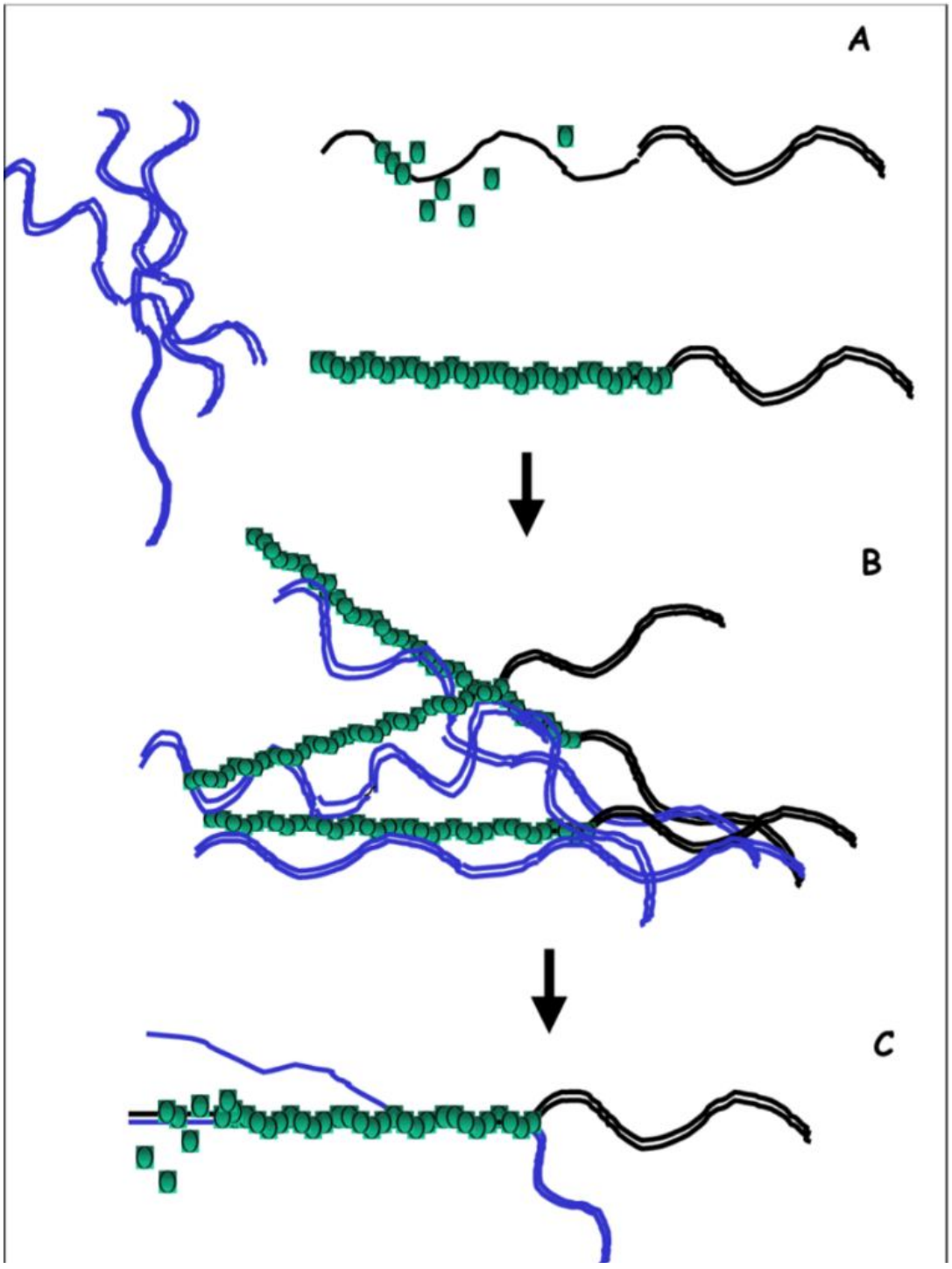



Genomics
DNA
Transcriptomics
mRNA
Proteomics
Protein
Metabolomics
Metabolites



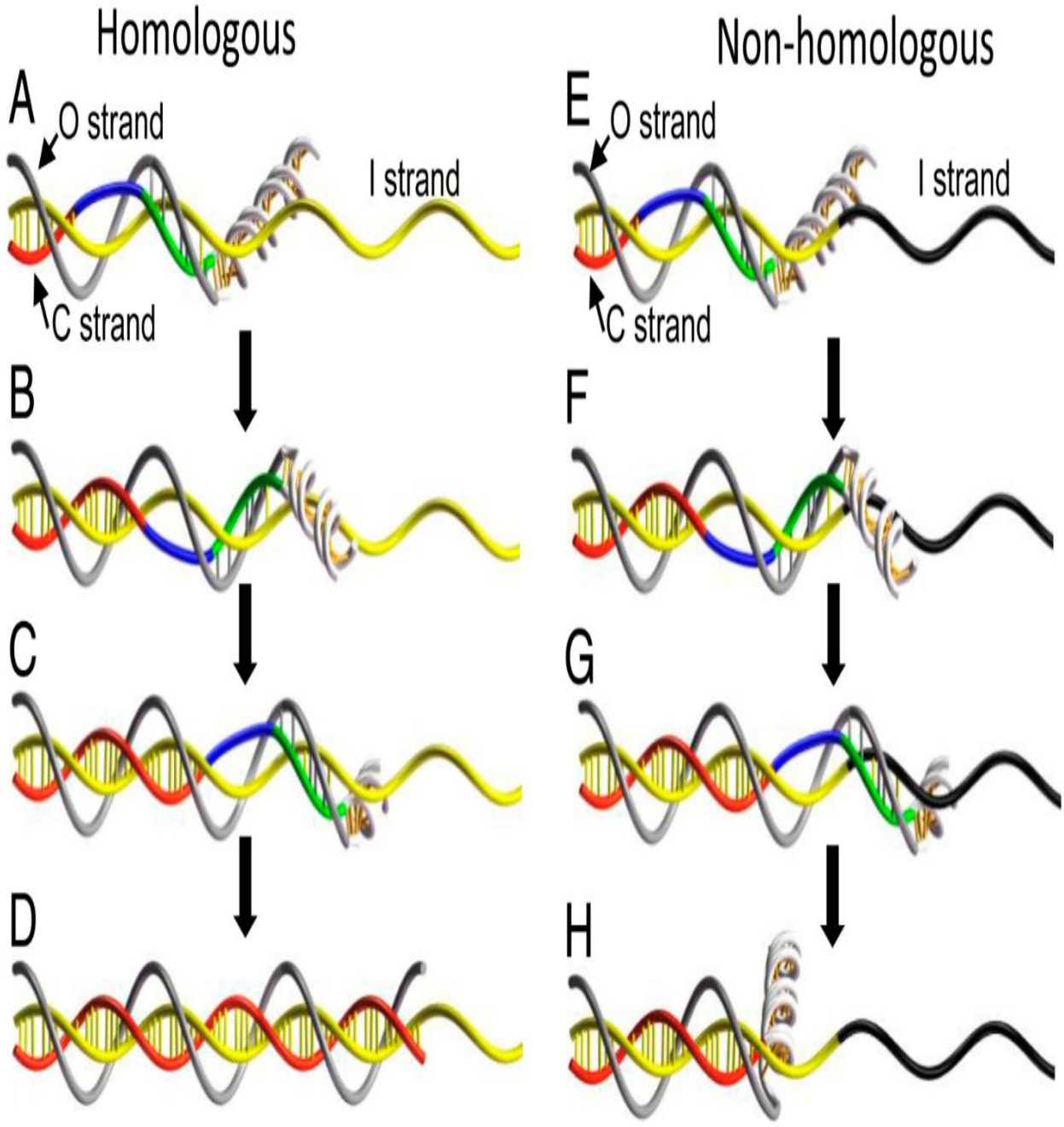
GENE ???


SYSTEMS BIOLOGY
DNA
RNA
Proteins/Enzymes
Metabolites
Feedback
Environment
Genomics
Epigenomics
Transcriptomics
Proteomics
Metabolomics
Lipidomics



An overview of the early steps of homologousRecombination promoted by RecA protein.

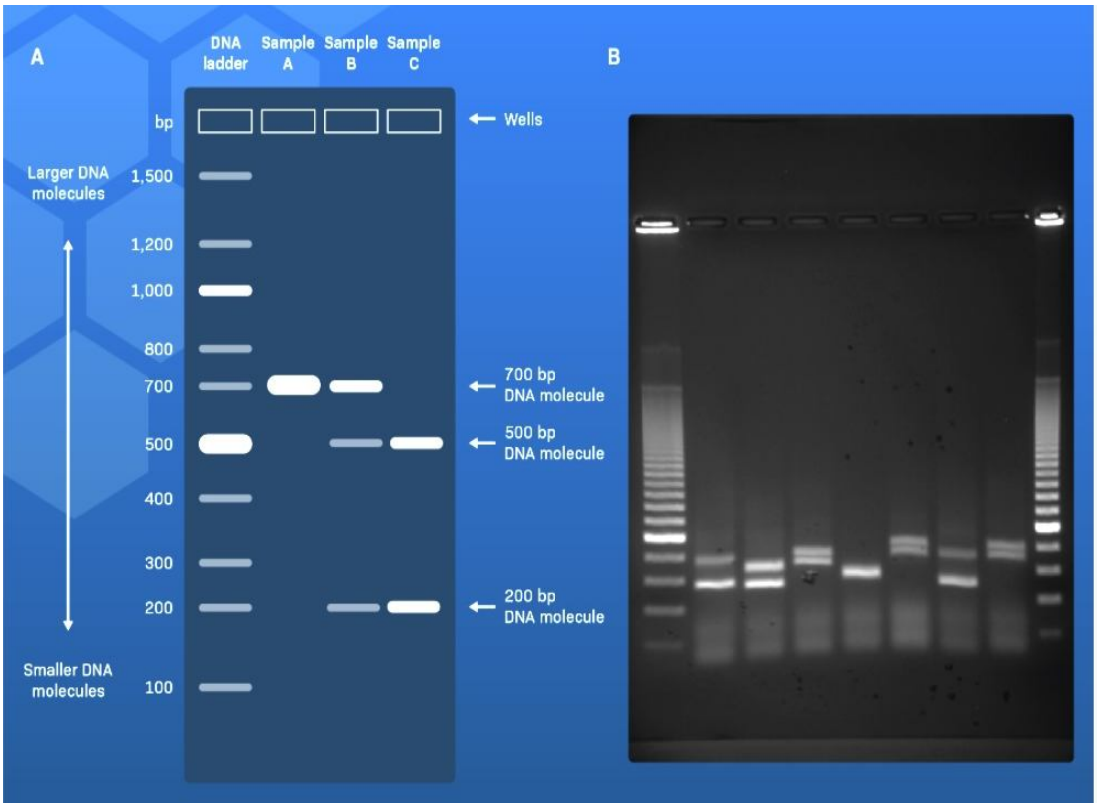
A. Formation of the nucleoprotein filaments.
B. Search for homology.
C. Homologous pairing and strand exchange.

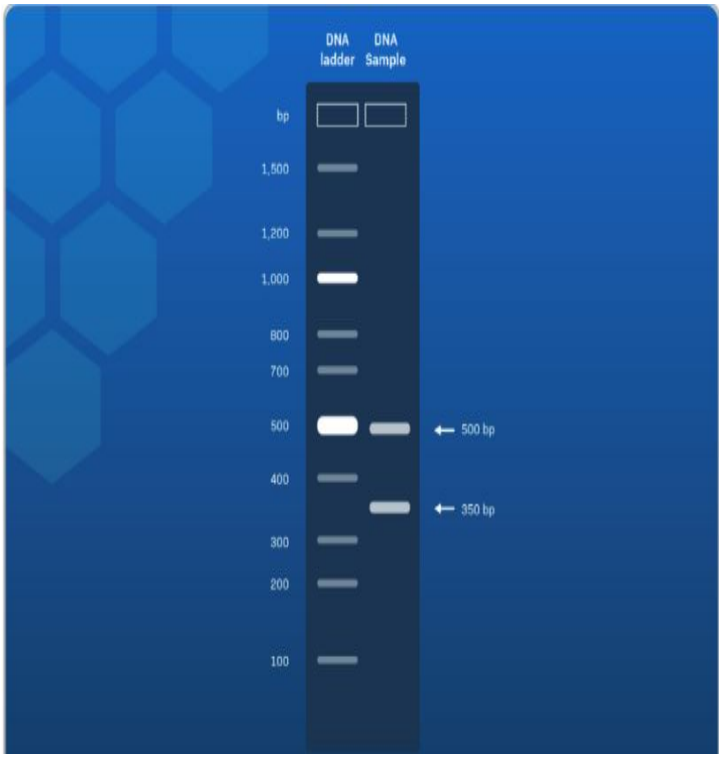


***Stepwise progression in Homologous Recombination* (HR).**

A-D: HR with fully homologous DNA. The C strand is divided into three parts.
(i). Post-strand exchanged (red),
(ii). homology testing (green), and
(iii). Transitional connection (blue).

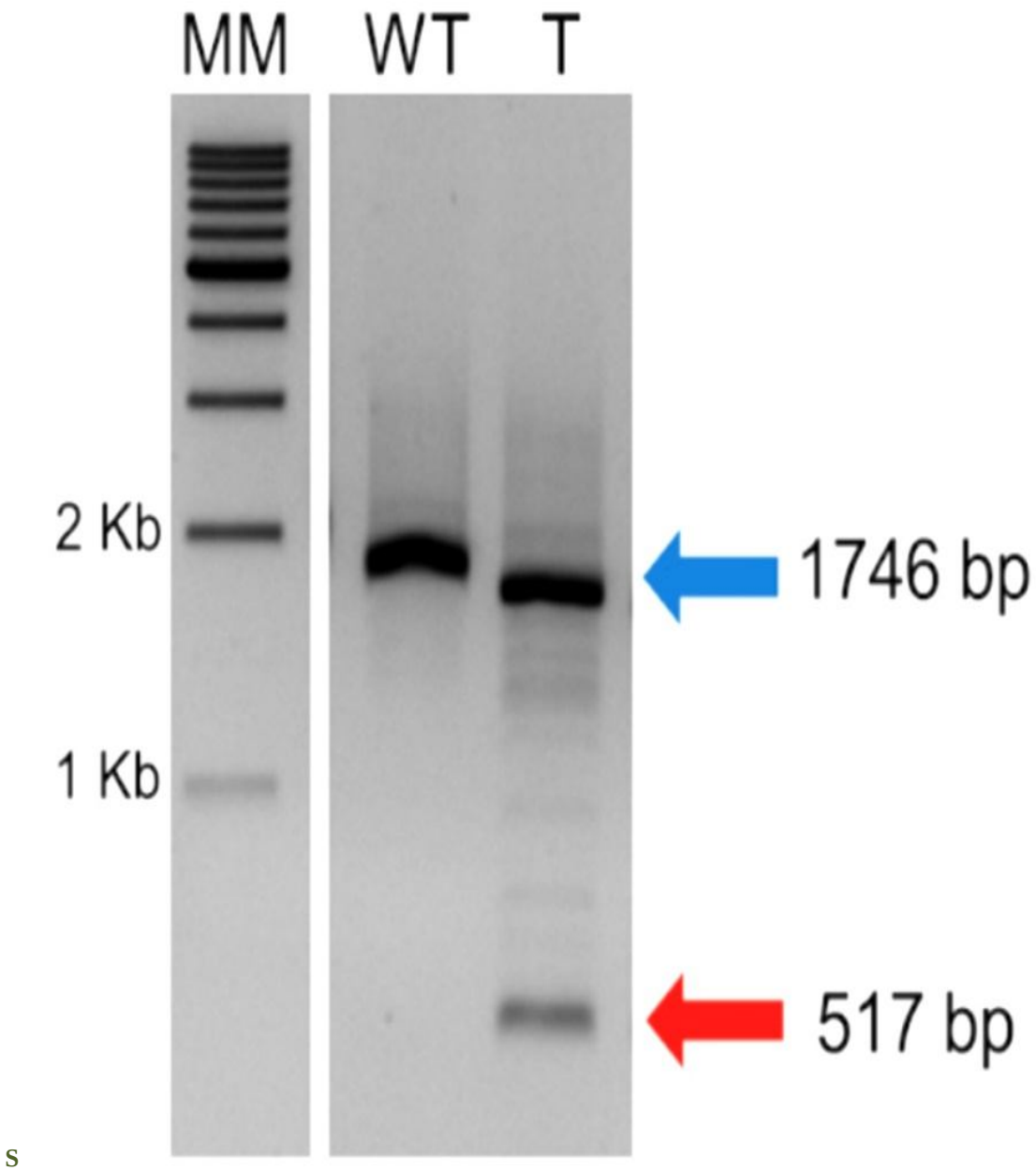
E-H: HR blocked by the nonhomologous segment.The homologous I strand is colored yellow, and the nonhomologous I strand is colored black.





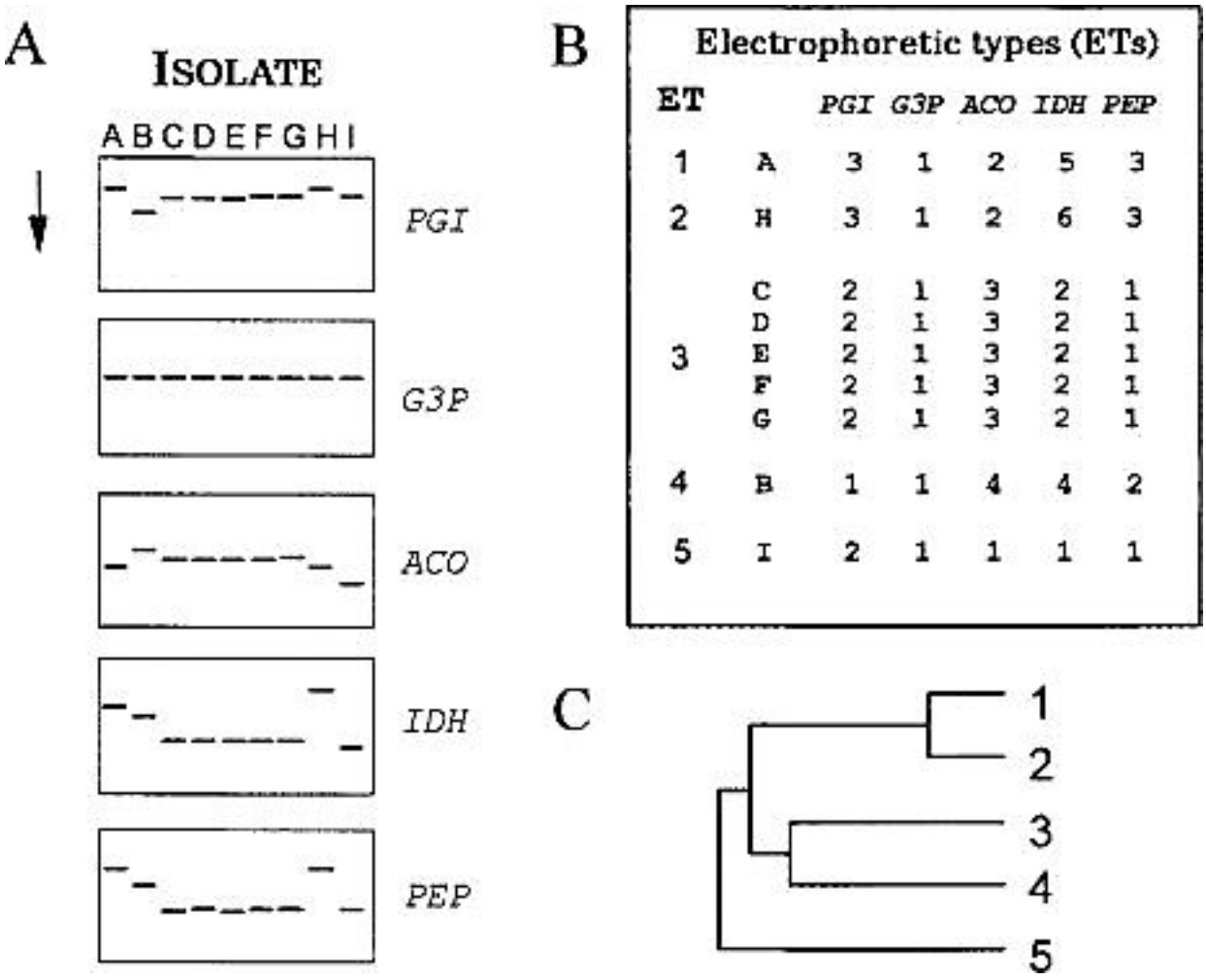
GEL ELECTROPHORESIS

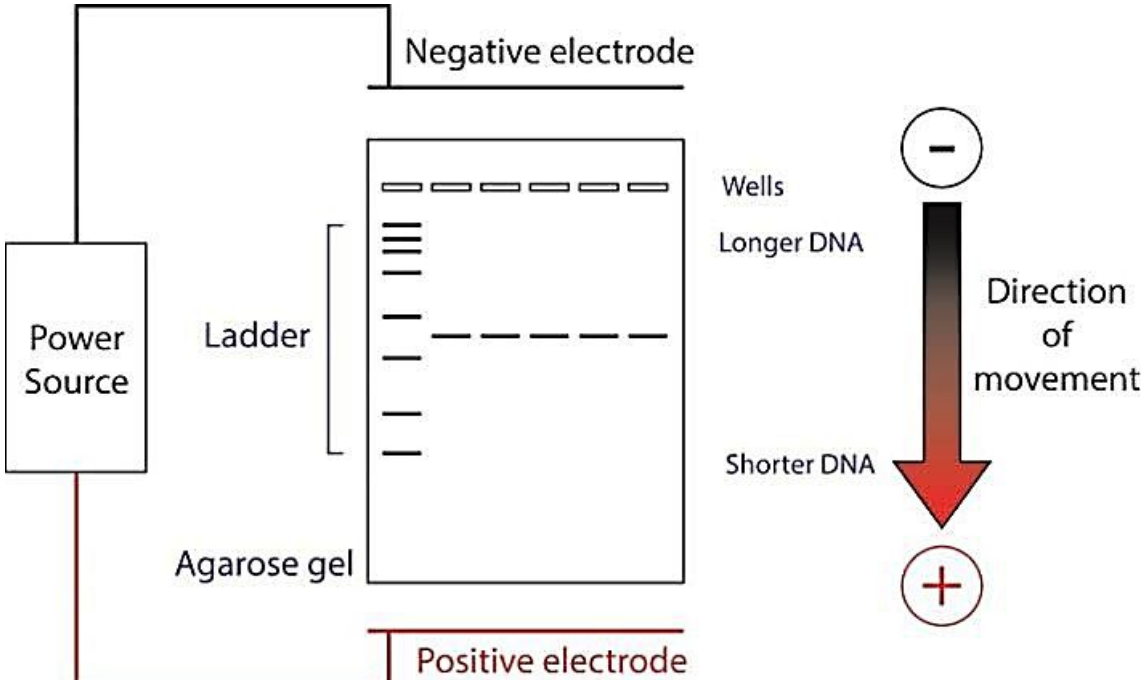
A stained agarose gel can be used to estimate the molecular weight of a DNA fragment. An illustration of a DNA sample containing 2 fragments (right lane) is presented. The molecular weight of each fragment can be approximated by comparing it to the closest band in the DNA ladder (left lane). The top band in lane 1 is closest to the 500 bp marker. The bottom band falls halfway between the 300 and 400 bp markers, and would be approximated as being 350 bp.

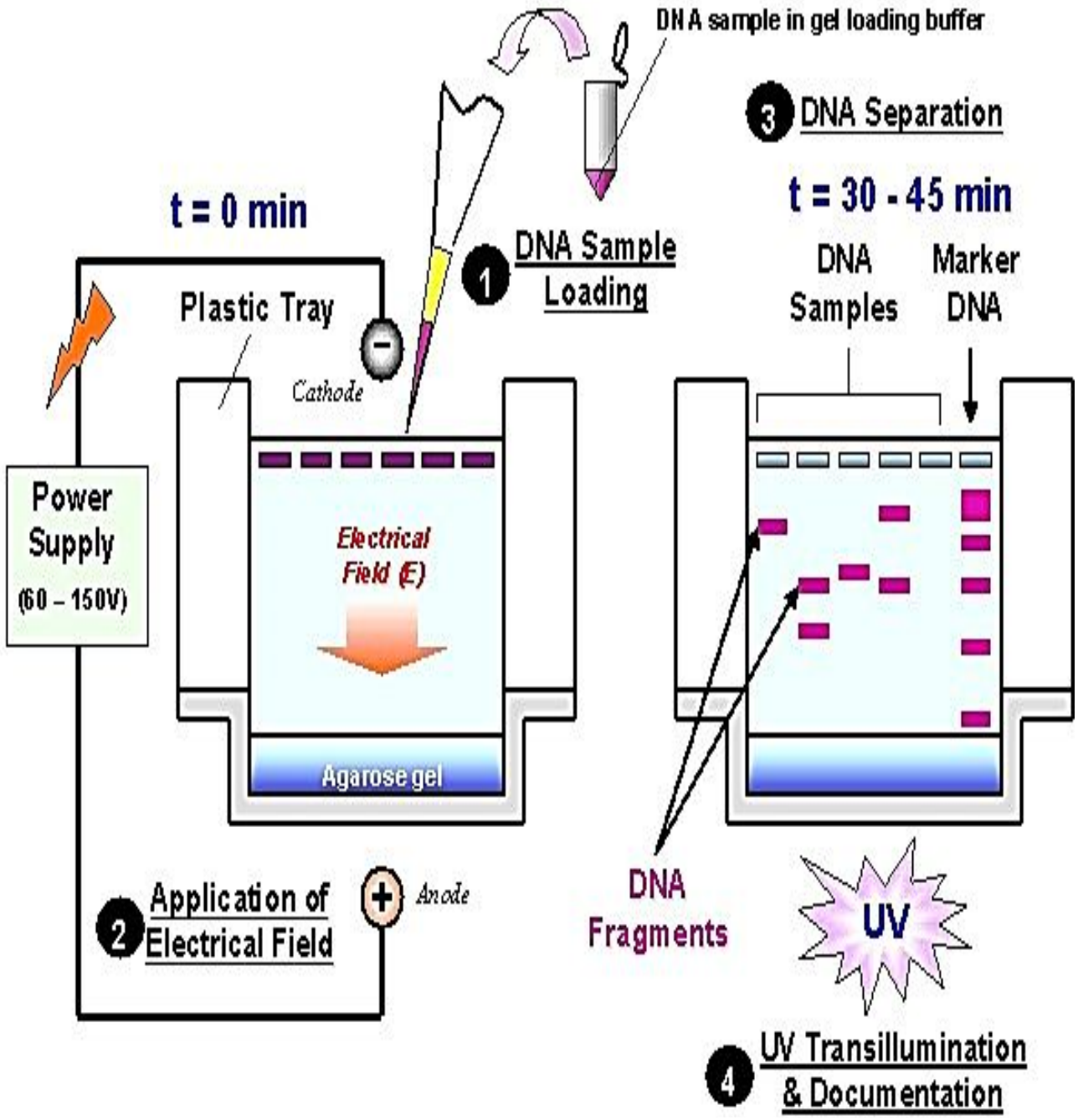


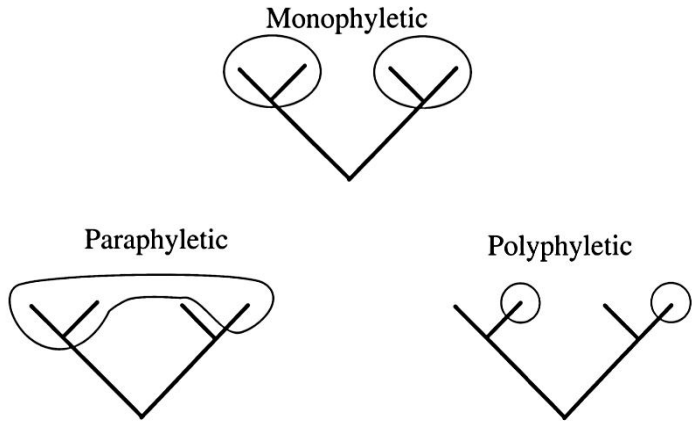
s

A stained agarose gel used to analyze CRISPR/Cas9-mediated editing of the TP53 gene. PCR was used to evaluate CRISPR/Cas9-mediated targeting of the TP53 gene designed to delete a 1229 bp region. The blue arrow denotes a 1746 bp PCR amplicon that represents the non-edited DNA sequence. The red arrow denotes a 517 bp PCR amplicon that results from an allele generated by CRISPR/Cas9 editing to delete the 1229 bp region. MM, 1 kb ladder; WT, wild type cells; T, gene-edited cells (Teixeira *et al*., 2020)

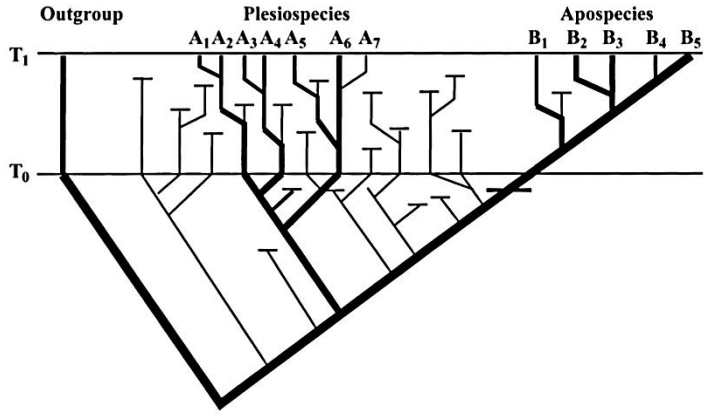

A
ISOLATE
A B C D E F G H I
PGI
G3P
ACO
IDH
PEP
B
Electrophoretic types (ETs)
ET
PGI G3P ACO IDH PEP
1 A 3 1 2 5 3
2 H 3 1 2 6 3
3 C 2 1 3 2 1
D 2 1 3 2 1
E 2 1 3 2 1
F 2 1 3 2 1
G 2 1 3 2 1
4 B 1 1 4 4 2
5 I 2 1 1 1 1
C
1
2
3
4
5


Negative electrode
Power Source
Ladder
Agarose gel
Positive electrode
Wells
Longer DNA
Shorter DNA
Direction of movement
−
+

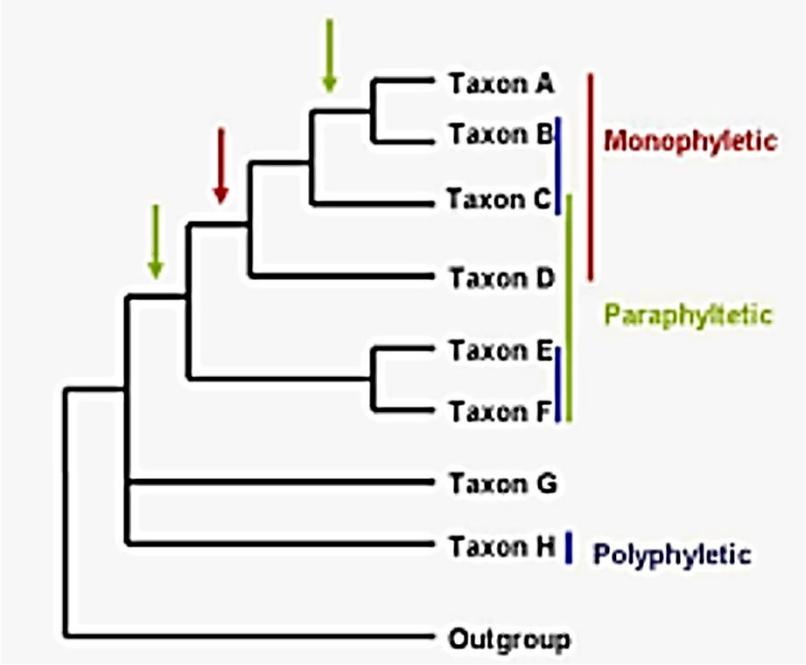

DNA sample in gel loading buffer
3 DNA Separation
t = 0 min
t = 30 - 45 min
1 DNA Sample Loading
DNA Samples
Marker DNA
Plastic Tray
Cathode
Power Supply
(60 – 150V)
Electrical Field (E)
Agarose gel
2 Application of Electrical Field
Anode
DNA Fragments
UV
4 UV Transillumination & Documentation

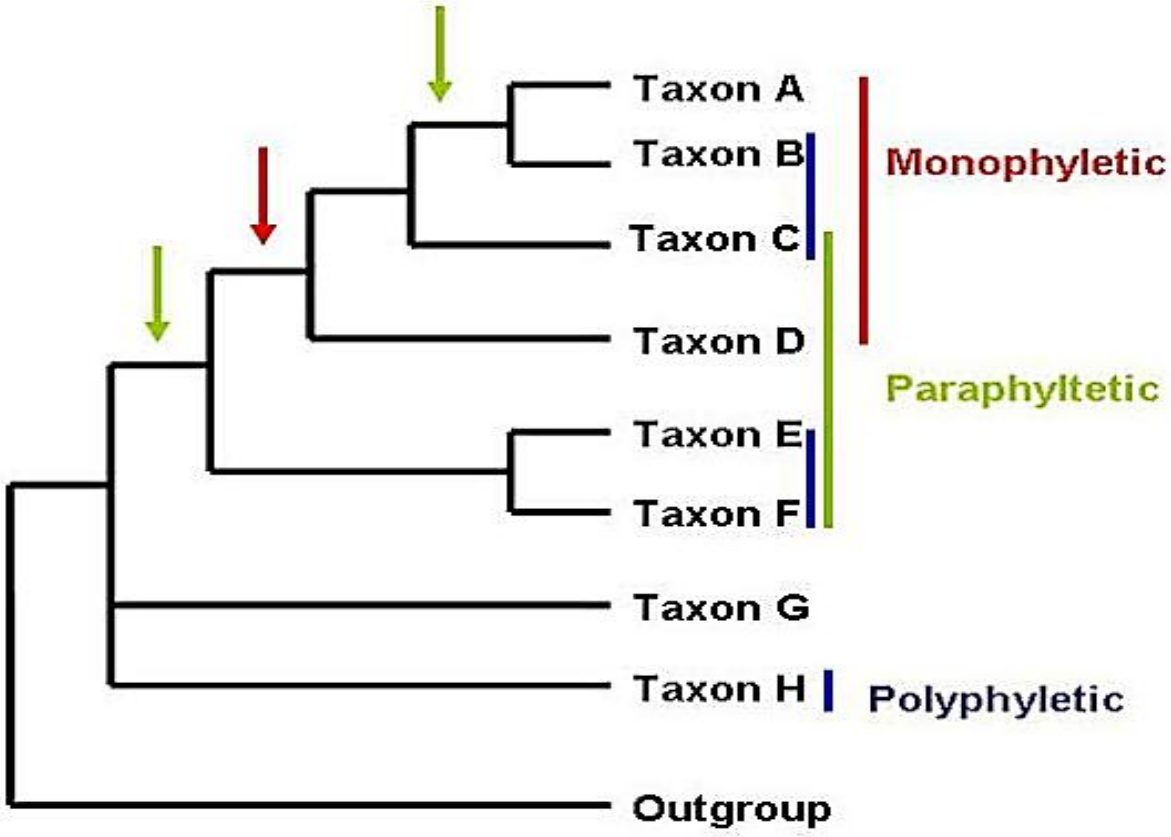


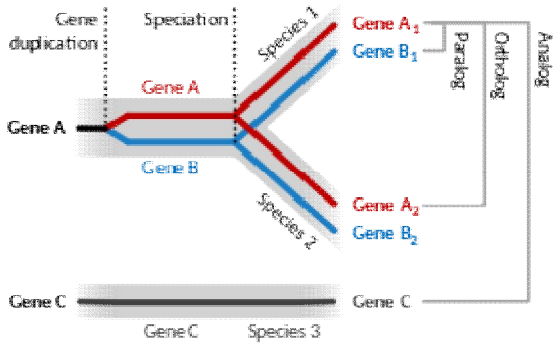
Cladistic relationship relative to cladograms



Apospecies and plesiospecies (Olmsstead, 1995). Under this evolutionary model, a set of populations is shown at the initial time T_0 when a speciation event occurs, as depicted by the thick horizontal line designating a synapmorphy forming the species. At initial time T_0, the new apospecies leaves a remnant set of populations that are now paraphyletic (plesiospecies). Later (at time T_1) extinction of populations leads to monophyly of both species. Bold lines designate populations surviving to time T_1. This shows the theoretical need to withdraw the strict criterion for monophyly in cladistic species concepts.


Taxon A
Taxon B
Taxon C
Taxon D
Taxon E
Taxon F
Taxon G
Taxon H
Outgroup
Monophyletic
Paraphyletic
Polyphyletic


Taxon A
Taxon B
Taxon C
Taxon D
Taxon E
Taxon F
Taxon G
Taxon H
Outgroup
Monophyletic
Paraphyletic
Polyphyletic



Homology among DNA, RNA, or proteins is typically inferred from their nucleotide or amino acid sequence similarity. Significant similarity is strong evidence that two sequences are related by evolutionary changes from a common ancestral sequence.

DECLARATION

This E-resource is exclusively meant for academic purposes and for enhancing teaching and learning only. Any other use for economic/commercial purpose is strictly prohibited. The users of the content shall not distribute, disseminate or share it with anyone else and its use is restricted to advancement of individual knowledge. The information provided in this e-content is authentic and best as per knowledge.

THANX