Dr. M. Sathiyabama
Associate Professor
Department of Botany

Plant Molecular Systematics

Types of molecular data?

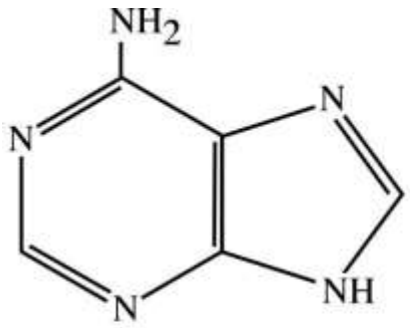
- DNA sequences
- DNA restriction sites: RFLPs
- Allozymes - different forms of proteins
- Microsatellites - DNA regions w/tandem repeats
- RAPDs - Random Amplification of Polymorphic DNA
- AFLPs - Amplified Fragment Length Polymorphism

How are Plant Molecular Data Acquired?

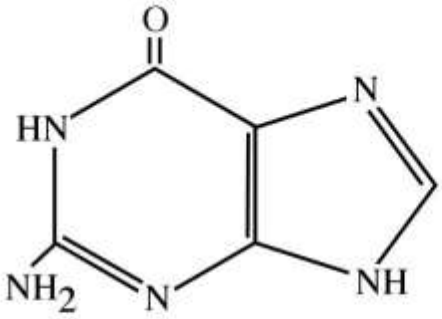
Plant collected: voucher prepared!

Live samples, e.g., allozyme analysis

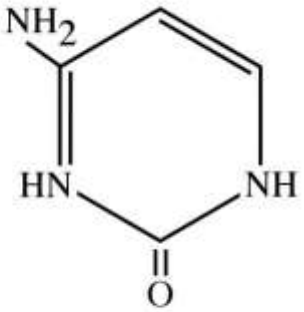
Dried or liquid-preserved samples, e.g., DNA analysis



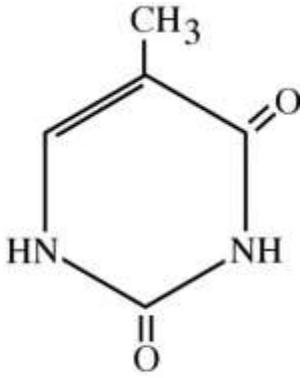
adenine



guanine



cytosine



thymine

A G C T

Purines Pyrimidines

What is DNA Sequence Data?

PCR: Polymerase Chain Reaction

What is it?

Process used to amplify DNA: replication into thousands of copies.

How does it work?

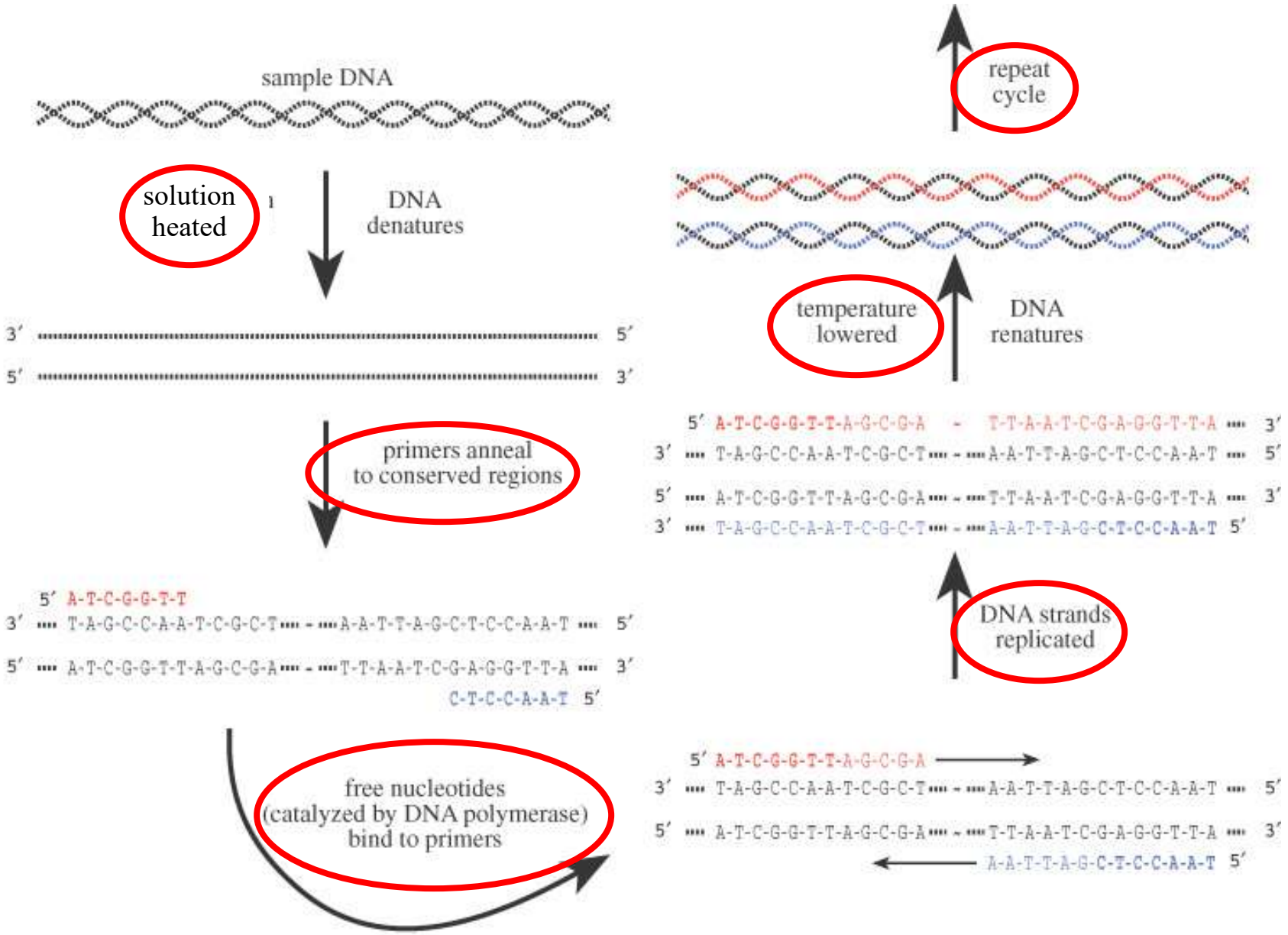
DNA isolated, purified, heated to denature

Primer used:

Primer = short, conserved DNA region

Complementary to ends of DNA to be amplified

+ Taq polymerase, nucleotides, buffer/salts



PCR: Polymerase Chain Reaction

DNA Sequencing

How does it work?

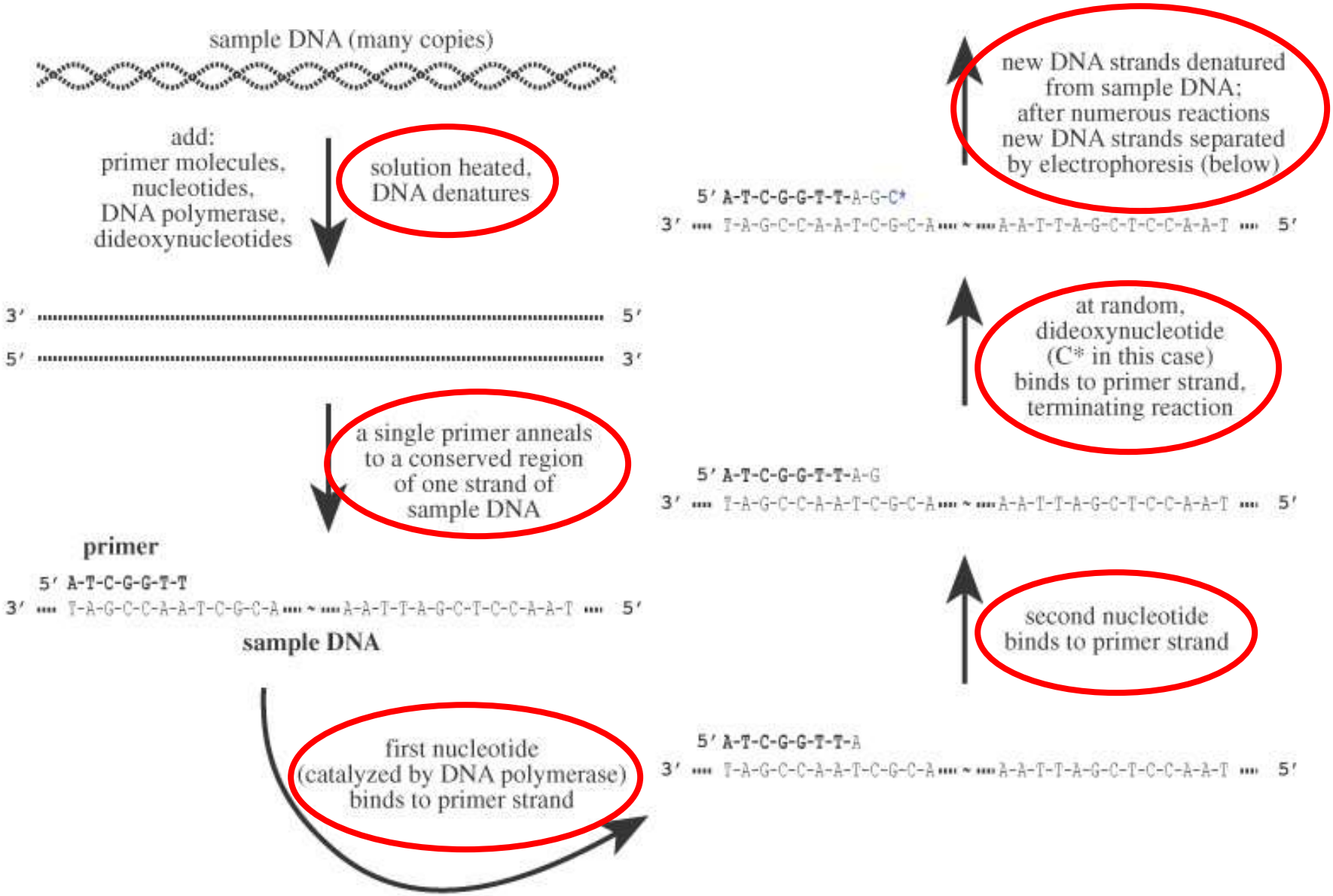
Similar to PCR amplification

But, small amount of **Dideoxynucleotides** used (along with higher conc. of nucleotides)

Dideoxynucleotides, once joined to new DNA strand, terminate polymerase reaction.

Dideoxynucleotides identified by fluorescence pattern.

Length of DNA strands determined by electrophoresis.



DNA Sequencing

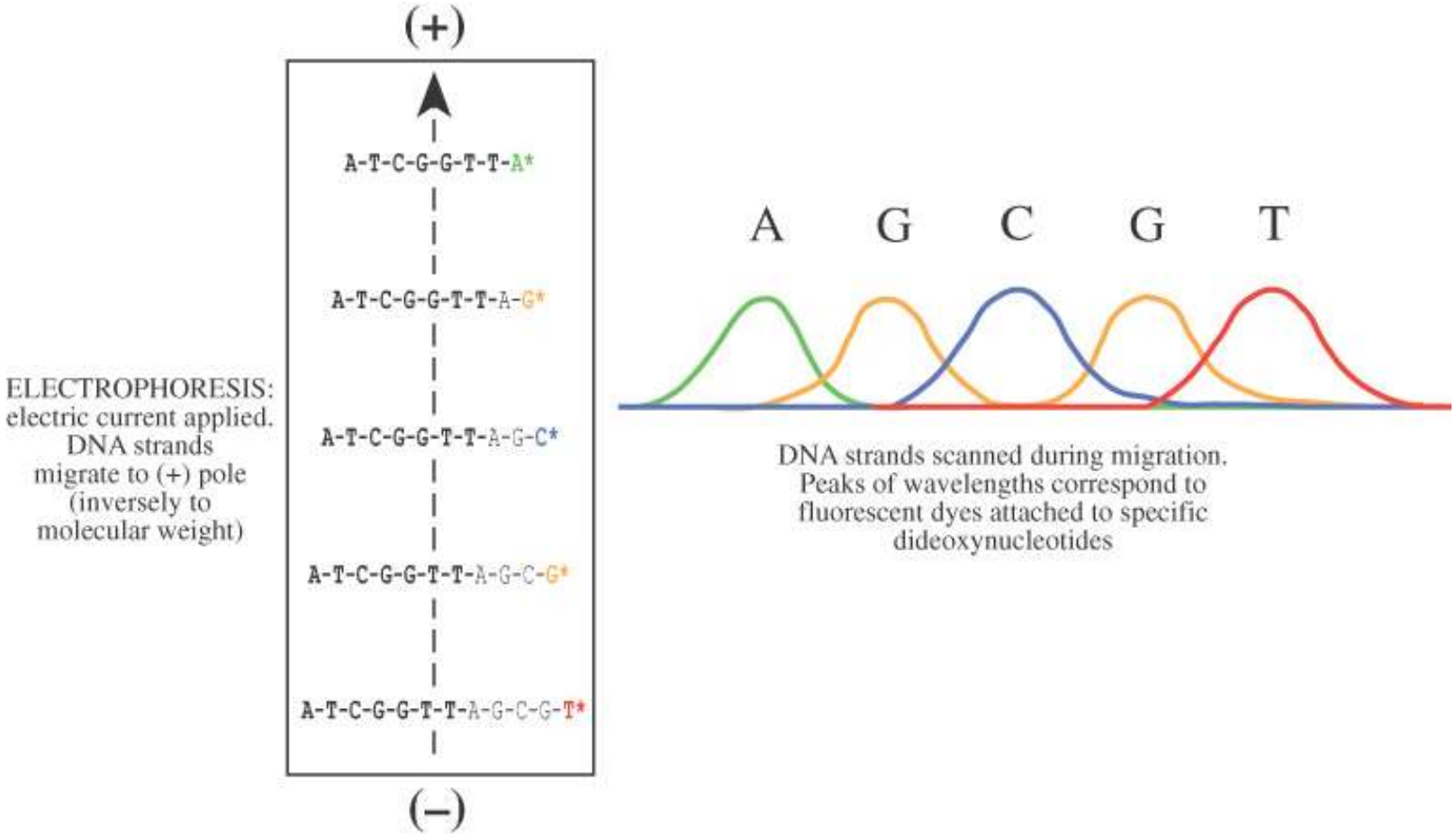


Figure 14.3 DNA sequencing reactions. A* = dideoxyadenine; C* = dideoxycytosine; G* = dideoxyguanine; T* = dideoxythymine.

DNA Sequencing

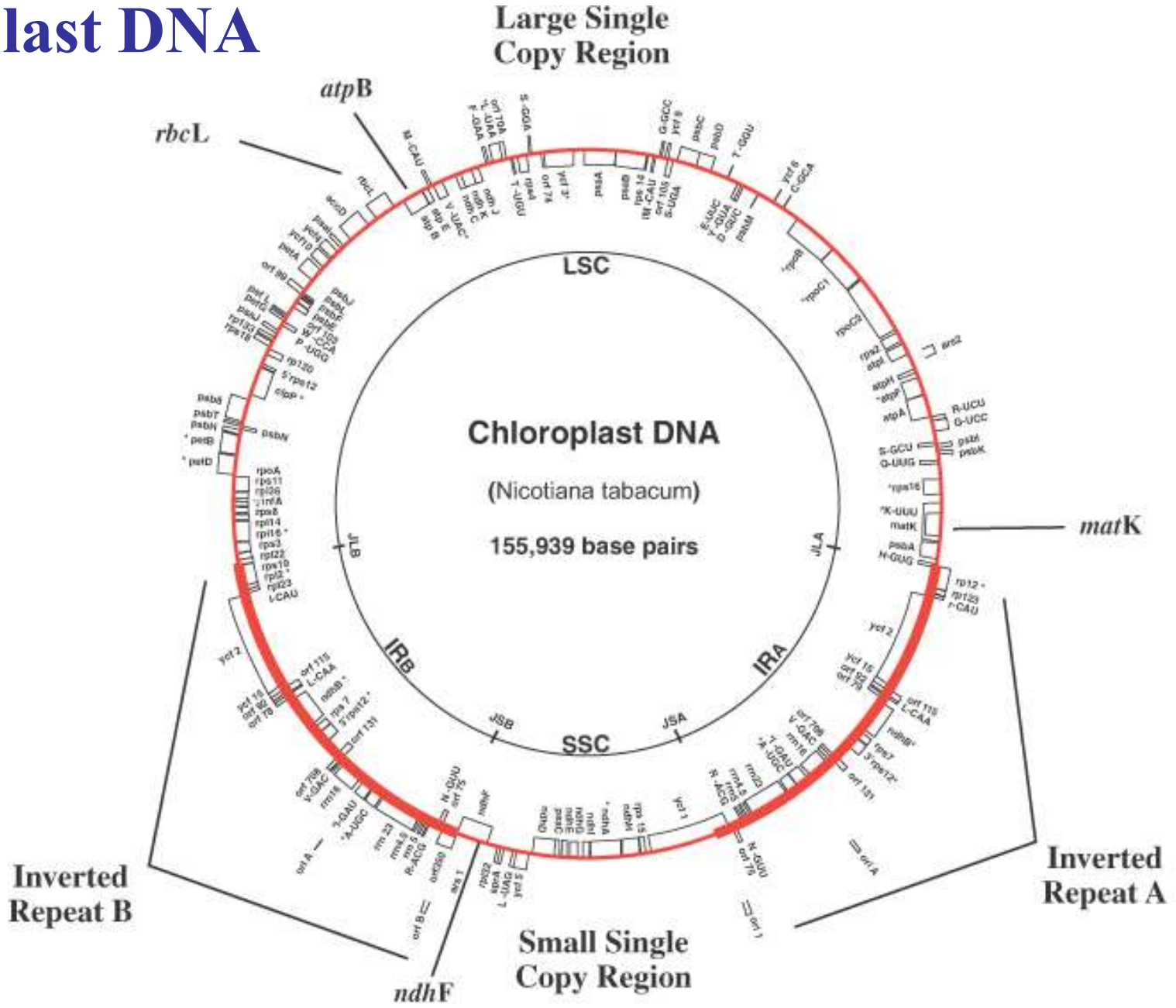
Types of DNA sequence data

Chloroplast cpDNA

Nuclear nDNA

Mitochondrial mtDNA (not used much with plants; used w/ animals)

Chloroplast DNA



Some Chloroplast coding genes:

Gene	Location	Function
atpB	Large single-copy region of chloroplast	Beta subunit of ATP synthethase, which functions in the synthesis of ATP via proton translocation
rbcL	Large single-copy region of chloroplast	Large subunit of ribulose 1,5-bisphosphate carboxylase/oxygenase (RUBISCO), which functions in the initial fixation of carbon dioxide in the dark reactions
matK	Large single-copy region of chloroplast	Maturase, which functions in splicing type II introns from RNA transcripts
*ndh*F	Small single-copy region of chloroplast	Subunit of chloroplast NADH dehydrogenase, which functions in converting NADH to NAD + H^+, driving various reactions of respiration

CHLOROPLAST INERGENIC SPACER REGIONS			
3'rps16-5'trnK	*petL-psbE*	*rpl32-trnL*	*trnL* intron
3'trnK-matK intron	*psaI-accD*	*rpoB-trnC*	*trnL-trnF*
3'trnV-ndhC	*psbA-3'trnK*	*rps16* intron	*trnQ-5'rps16*
5'rpS12-rpL20	*psbB-psbH*	*rps4-trnT*	*trnS-rps4*
atpI-atpH	*psbD-trnT*	*trnC-ycf6*	*trnS-trnfM*
matK-5'trnK intron	*psbJ-petA*	*trnD-trnT*	*trnS-trnG*
ndhA intron	*psbM-trnD*	*trnG* intron	*trnT-trnL*
ndhF-rpl32	*rpl14-rps8-infA-rpl36*	*trnH-psbA*	*ycf6-psbM*
ndhJ-trnF	*rpl16* intron		

Chloroplast inter-genic spacer regions used!

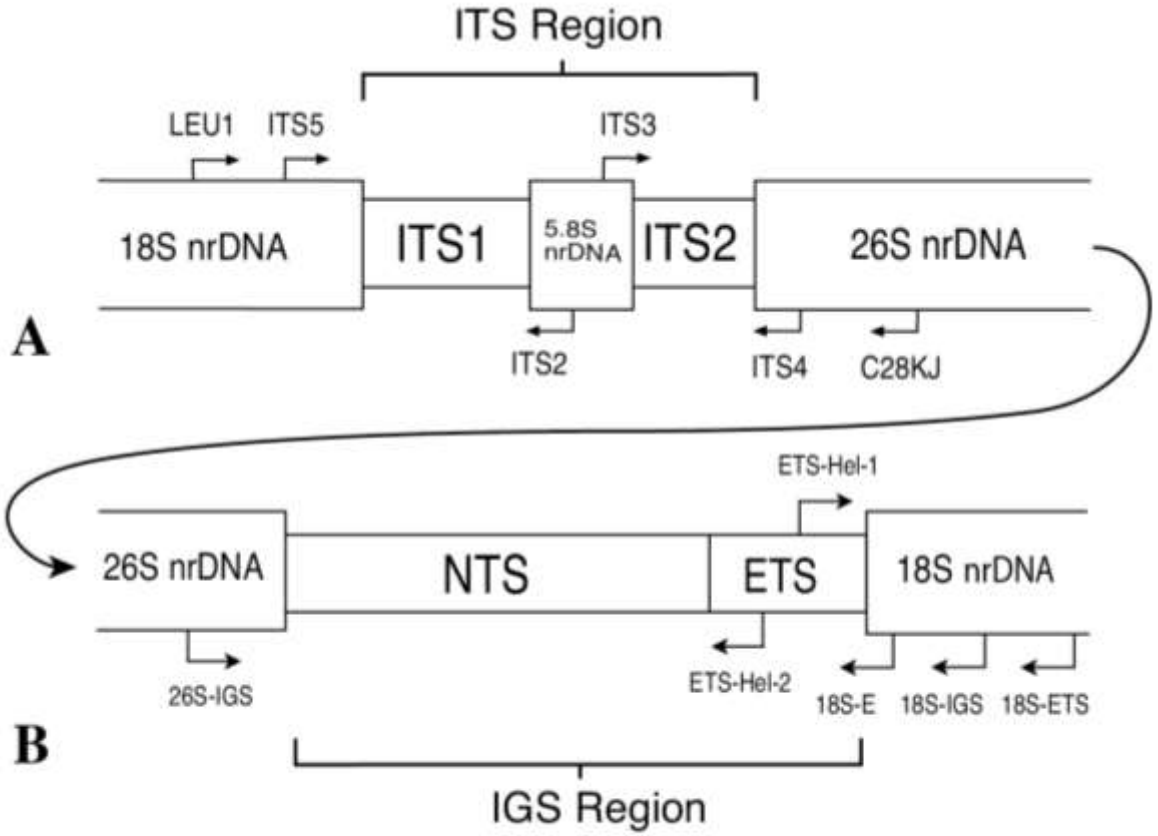


Figure 14.5 **A.** Internal transcribed spacers (ITSs) of nuclear ribosomal DNA, illustrating the ITS region and flanking subunits, and showing the orientations and locations of primer sites. After Baldwin et al. (1995). **B.** External transcribed spacer (ETS) of the intergenic spacer (IGS) region, also showing orientations and locations of primer sites. After Baldwin and Markos (1998).

Nuclear DNA: ITS/ETS Sequence Data: used in lower-level analyses

DNA alignment

	DNA Alignment	*Character Coding*
	0000000000000000000111111111111111111111	
	8888888889999999999000000000011111111122	
	1234567890123456789012345678901234567890 1	1 2 3 4 5 6
Taxon 1	GCCTAGCC**A**AAGCTC**T**TCCAAG**G**TG**A**CTCTCA**GTTCAA**GCT	2 0 3 2 0 4
Taxon 2	GCCTAGCC**A**AAGCTC**T**TCCAAG**C**TG**A**CTCTCA------GCT	2 0 3 1 0 5
Taxon 3	GCCTAGCC**T**AAGCTC**A**ACCAAG**G**TG**T**CTCTCA**GTTCAA**GCT	2 3 0 2 3 4
Taxon 4	GCCTAGCC**T**AAGCTC**T**TCCAAG**G**TG**T**CTCTCA**GTTCAA**GCT	2 3 3 2 3 4
Taxon 5	GCCTAGCC**A**AAGCTC**T**TCCAAG**C**TG**A**CTCTCA------GCT	2 0 3 1 0 5
Taxon 6	CCCTAGCC**A**AAGCTC**T**TCCAAG**C**TG**A**CTCTCA**GTTCAA**GCT	1 0 3 1 0 4
Taxon 7	CCCTAGCC**A**AAGCTC**T**TCCAAG**C**TG**A**CTCTCA**GTTCAA**GCT	1 0 3 1 0 4
Taxon 8	GCCTAGCC**T**AAGCTC**T**TCCAAG**C**TG**A**CTCTCA**GTTCAA**GCT	2 3 3 1 0 4

Figure 14.6 Example of alignment of DNA sequences of 41 nucleotide sites (positions 81–121) from eight taxa. Variable nucleotide sites are in **bold**. Note deletion of six bases in taxon 2 and taxon 5. Possible character coding of variable sites is seen at right. Coding of nucleotides is as follows: A = state 0; C = state 1; G = state 2; T = state 3. In this example, the deletion is coded as a single binary character (character 6), coded differently from nucleotides, as state 4 = deletion absent and state 5 = deletion present.

Character Coding

Weighting of DNA sequence data

	A	G	C	T
A	0	1	5	5
G	1	0	5	5
C	5	5	0	1
T	5	5	1	0

Step Matrix:

Transition: PY <–> PY or PU <–> PU

Transversion: PY <–> PU or PU <–> PY

Models of Molecular Evolution

A

	A	C	G	T
A	$-\mu(a\pi_C + b\pi_G + c\pi_T)$	$\mu a\pi_C$	$\mu b\pi_G$	$\mu c\pi_T$
C	$\mu a\pi_A$	$-\mu(a\pi_A + d\pi_G + e\pi_T)$	$\mu d\pi_G$	$\mu e\pi_T$
G	$\mu b\pi_A$	$\mu d\pi_C$	$-\mu(b\pi_A + d\pi_C + f\pi_T)$	$\mu f\pi_T$
T	$\mu c\pi_A$	$\mu e\pi_C$	$\mu f\pi_G$	$-\mu(c\pi_A + d\pi_C + e\pi_G)$

B

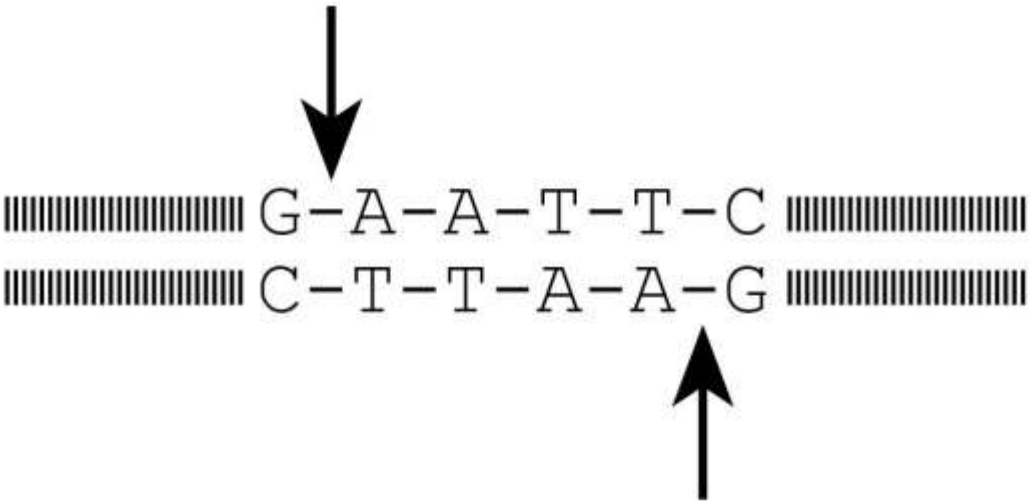
	A	C	G	T
A	$-3/4\mu$	$1/4\mu$	$1/4\mu$	$1/4\mu$
C	$1/4\mu$	$-3/4\mu$	$1/4\mu$	$1/4\mu$
G	$1/4\mu$	$1/4\mu$	$-3/4\mu$	$1/4\mu$
T	$1/4\mu$	$1/4\mu$	$1/4\mu$	$-3/4\mu$

C

	A	C	G	T
A	$-1/4\mu(\kappa+2)$	$1/4\mu$	$1/4\mu\kappa$	$1/4\mu$
C	$1/4\mu$	$-1/4\mu(\kappa+2)$	$1/4\mu$	$1/4\mu\kappa$
G	$1/4\mu\kappa$	$1/4\mu$	$-1/4\mu(\kappa+2)$	$1/4\mu$
T	$1/4\mu$	$1/4\mu\kappa$	$1/4\mu$	$-1/4\mu(\kappa+2)$

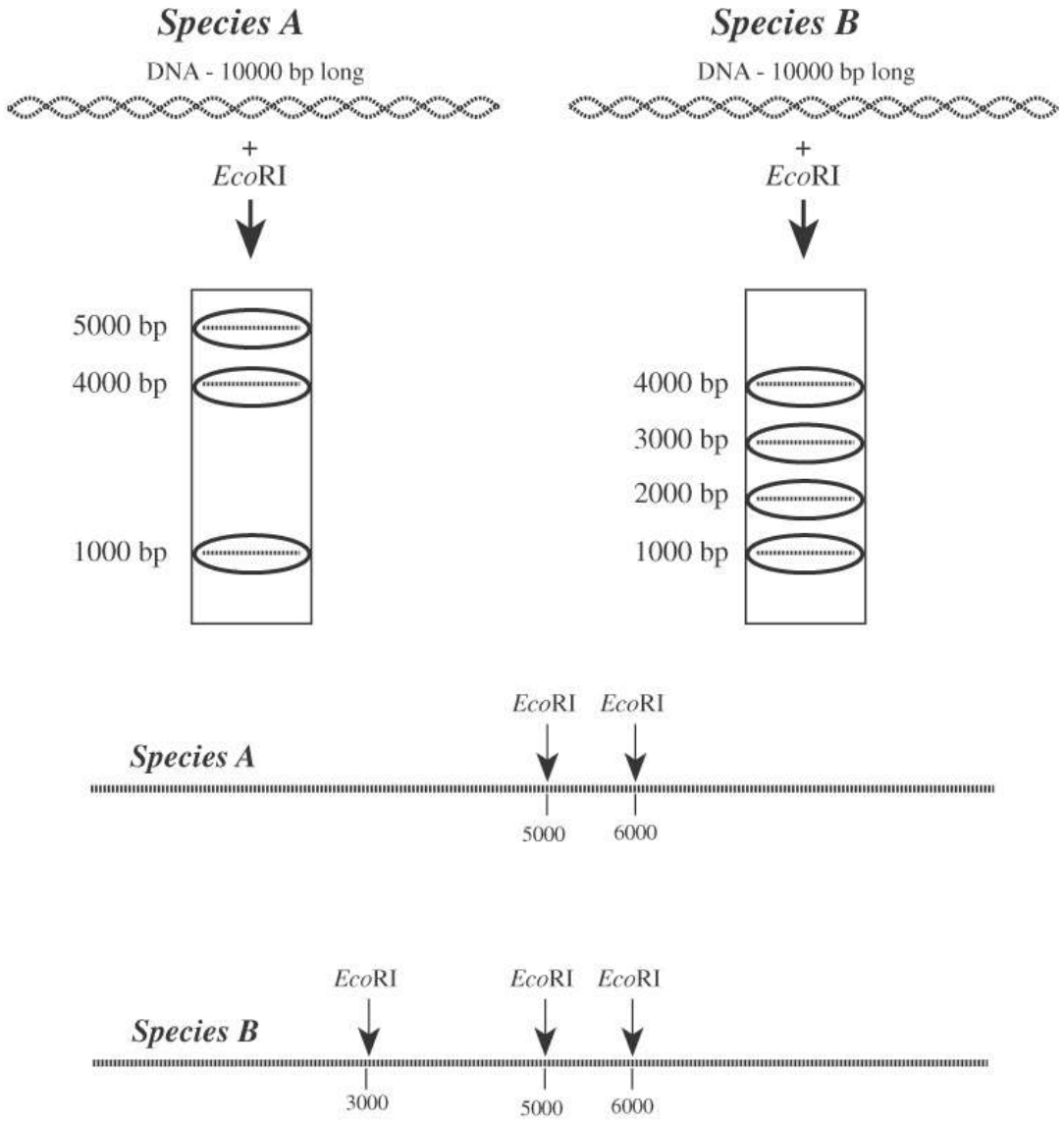
Figure 2.17 Models of base substitution. **A.** General time reversable model, in which probabilities of change from one base to another are a function of mean instantaneous base substitution rate (μ), relative rate parameters (a,b,c,d,e,f), and base frequencies (π_A, π_C, π_G, π_T). **B.** Jukes-Cantor (**JC**) model, in which substitution rates are the same. **C.** Kimura's two-parameter model (**K2P**), in which base frequencies are the same but transitions (in red) and transversions (in blue) occur at different rates.

Restriction site?

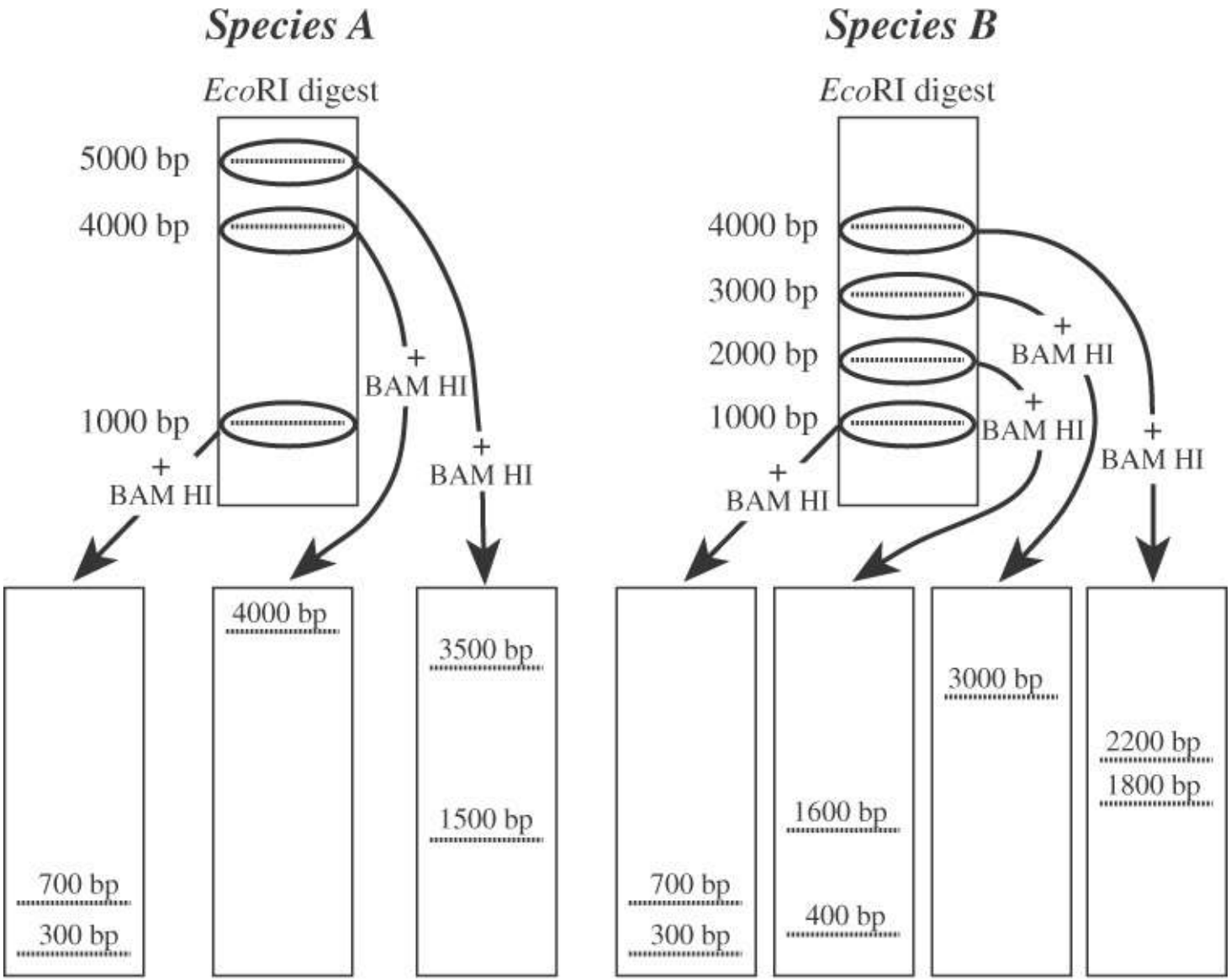


Restriction Enzymes: E.g., EcoR1

RFLP: Restriction Fragment Length Polymorphism



RFLP: Restriction Fragment Length Polymorphism



RFLP: Restriction Fragment Length Polymorphism

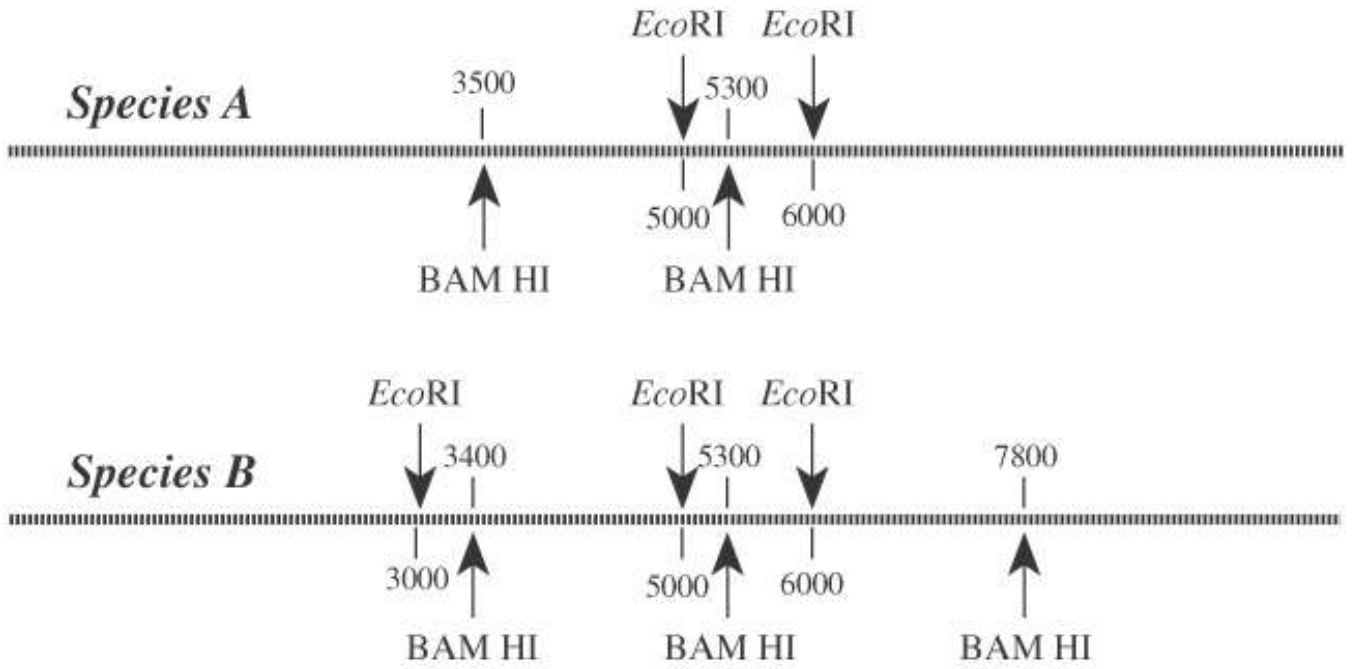


Figure 14.10 Example of restriction site analysis of species A and B, using restriction site enzyme *Eco*RI, followed by restriction site enzyme BAM HI. Possible restriction site maps of species A and B are shown in the lower portion of the figure.

	CHARACTERS						
	*Eco*RI	BAM	BAM	*Eco*RI	BAM	*Eco*RI	BAM
TAXA	3000	3400	3500	5000	5300	6000	7800
Species A	−	−	+	+	+	+	−
Species B	+	+	−	+	+	+	+

Figure 14.11 Character coding of restriction site map data of Figure 14.10, derived by presence or absence of *Eco*RI or BAM sites at specific locations along DNA.

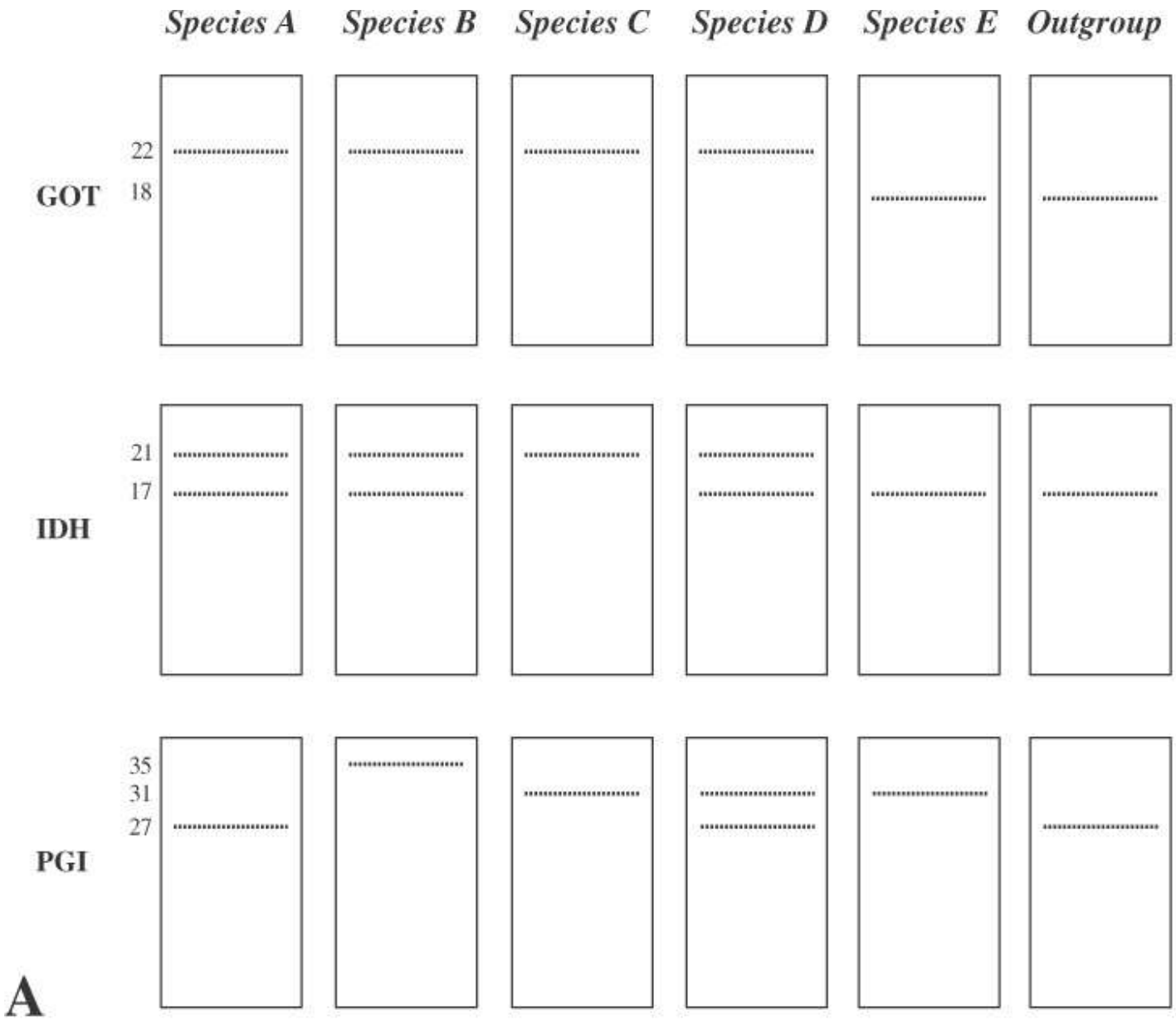
Allozymes-different forms of an enzyme

Used in the past frequently, rarely today.

Gives presence/absence of enzyme types.

Can have 2 allozymes per sample (2 alleles of a gene=heterozygous).

More difficult to code for phylogenetic study.



Allozyme Data

THE FUTURE: Next Generation Sequencing

Also called High Throughput Sequencing

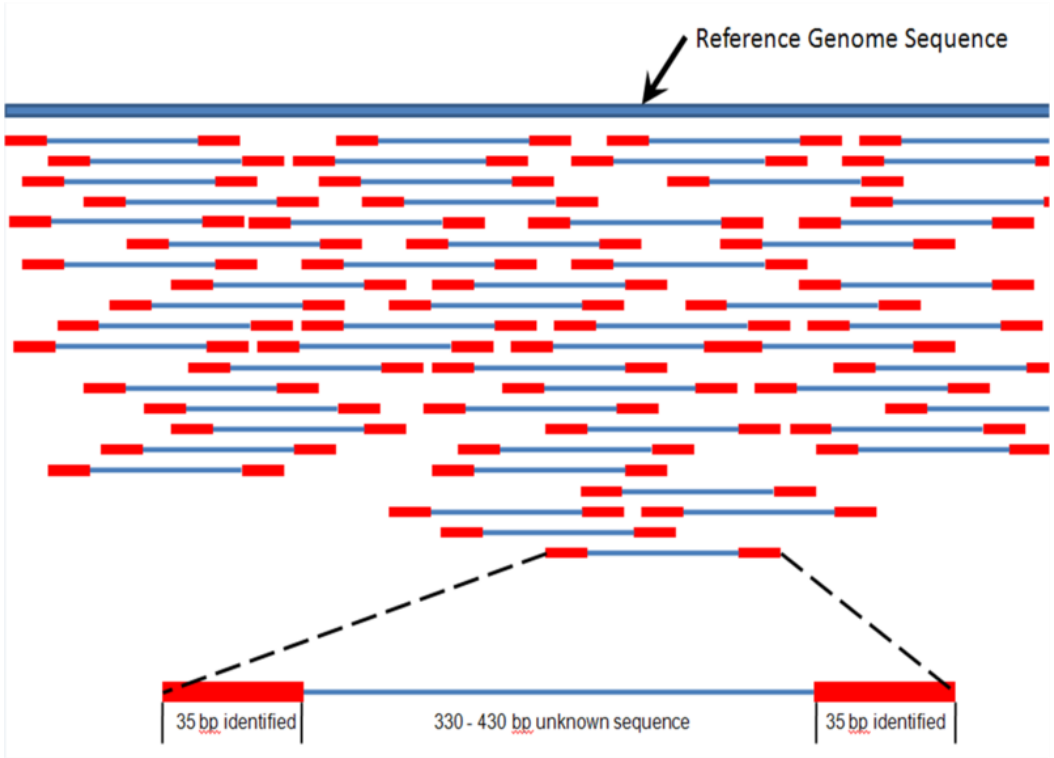
DNA is fragmented.

Fragments amplified producing as many as millions of sequences.

Many different techniques for amplifying and sequencing DNA fragments.

THE FUTURE: Next Generation Sequencing

Sequences are then read and overlapping sequence data aligned, using a reference.



THE FUTURE: Next Generation Sequencing

ADVANTAGE: Can get much more DNA sequence data (on order of 100-1,000x more than traditional studies).

Cost is much lower (per bp).

Can also sequence **transcriptomes**: sequences of translated mRNA, i.e., what is expressed.

DISADVANTAGE: Requires massive computer capabilities to assemble data.

Usually must have a reference genome (closely related taxon).

STOP HERE

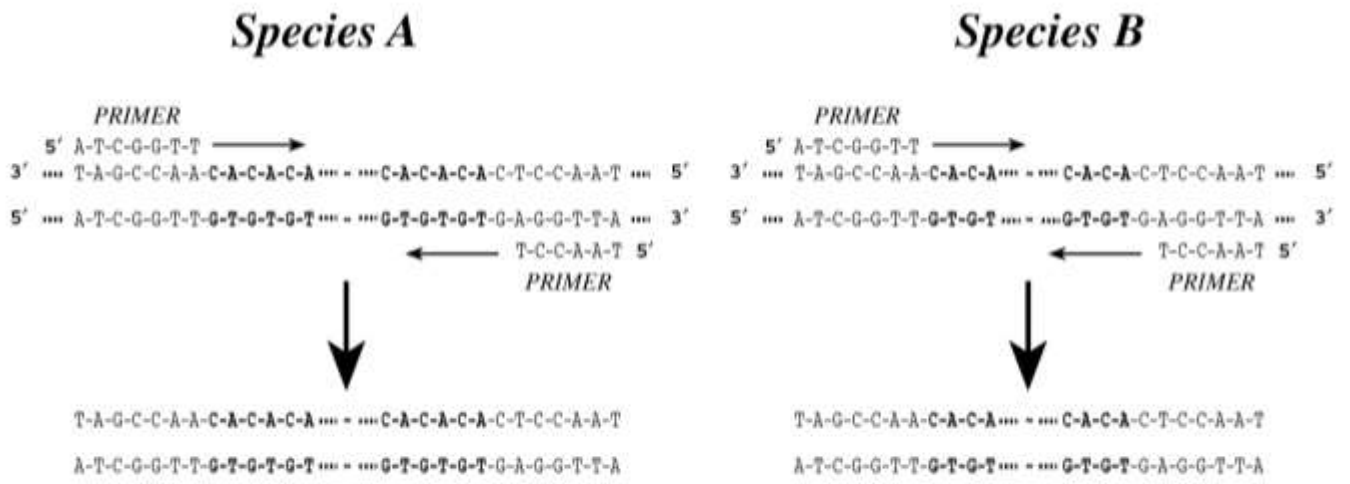


Figure 14.13 Microsatellite data. Primers were constructed to flank regions of tandem repeats. Note that tandem repeat region of species *A* is longer than that of species *B* and is thus a genetic difference between the two.

Microsatellites: Tandem repeats of nucleotides

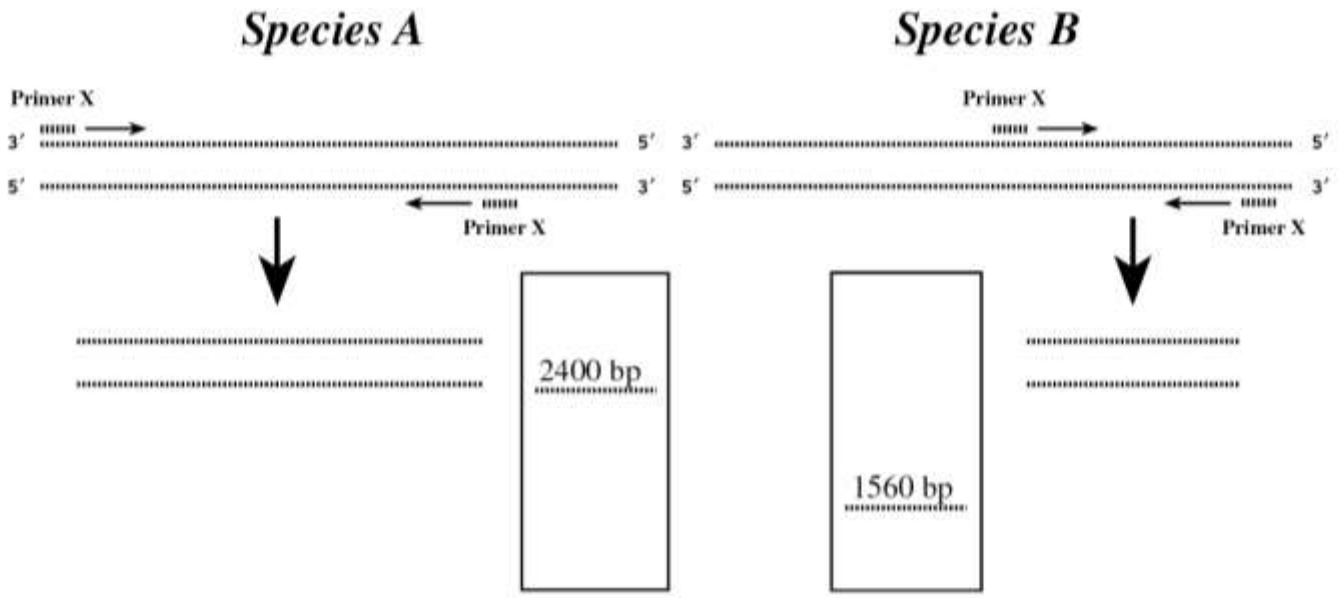


Figure 14.14 RAPDs data. In this example the same DNA regions for species *A* and *B* anneal to different randomly generated primers, resulting in amplified DNA of different lengths, a genetic difference between the two taxa.

RAPD's:

Random Amplified Polymorphic DNA

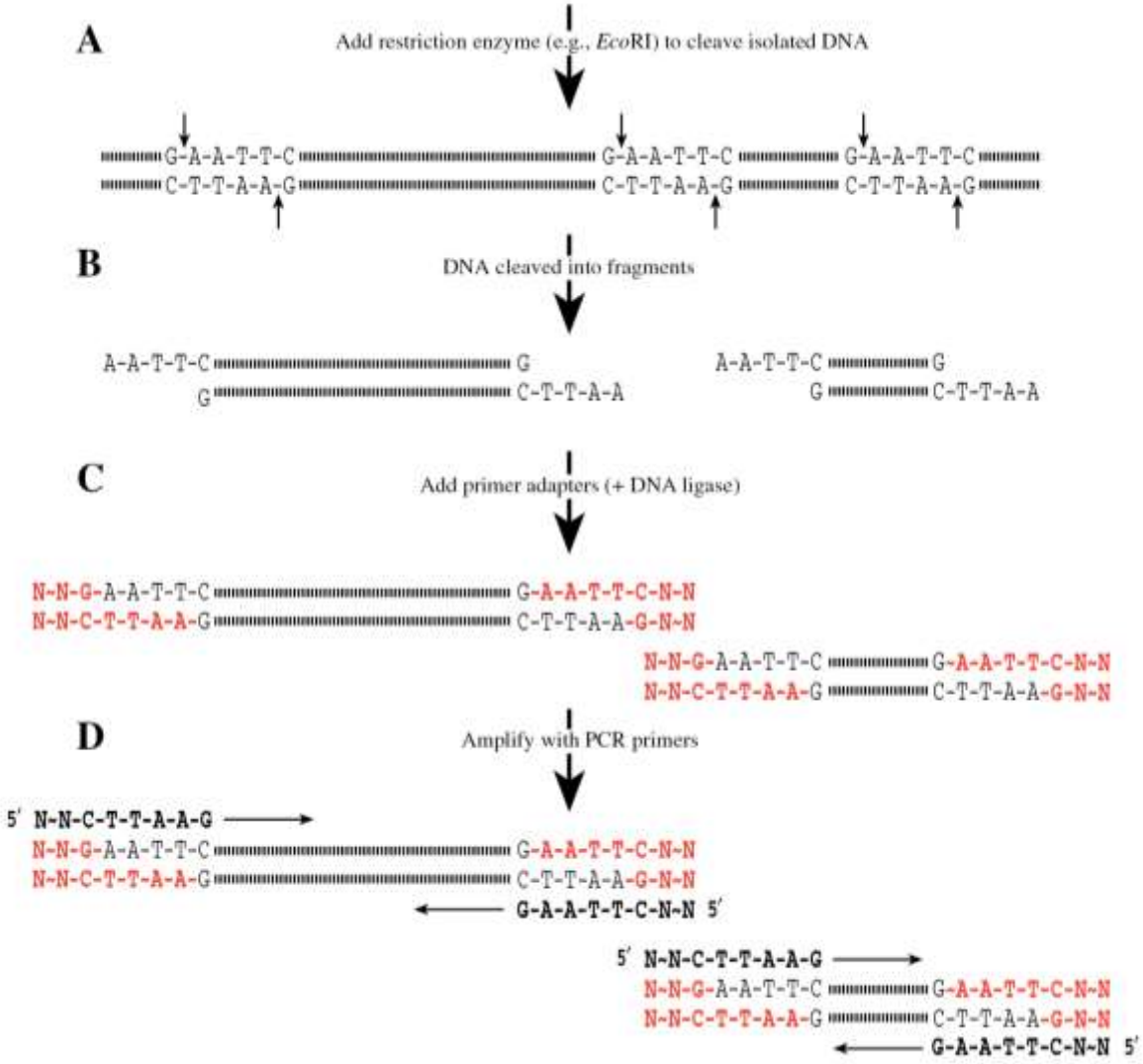


Figure 14.15 AFLP technique. The letters "N-N-" represent a length of nucleotides.

Amplified Fragment Length Polymorphism =AFLP's

THE FUTURE: Next Generation Sequencing

Also called High Throughput Sequencing

DNA is fragmented.

Fragments amplified producing as many as millions of sequences.

Sequences are then re
sequence data align