Gene Technology

Lecture-2

For all computational purposes, DNA is represented as a string of 4-letter alphabets - A, T, C, G:

attgctacgttacatcgctgca

How do we get this string representation from a dynamic double-stranded molecule?

DNA Sequencing - determine the precise sequence of nucleotides in a sample of DNA

To carry out this task we need to be able to chop the DNA, store it, make copies of it.

Let's consider the example of detecting if a person is infected by the novel coronavirus SARS-CoV-2

- **uses Real Time RT-PCR Nucleic Acid Detection Kit based on the PCR method which uses a fluorescent probe and a specific primer to detect three specific regions within the SARS-CoV-2 nucleocapsid protein N gene.**

- **How is the SARS-CoV-2 genome sequenced?**
- **How does one identify the coordinates of N gene on it? i.e., how to construct a physical map of a genome?**
- **How does one select which regions in this gene would give specificity for the presence of SARS-CoV-2?***
- **How are the specific probe regions extracted and amplified for detection?**
- **Is it possible to store the DNA sample for re-testing? How?**

To sequence a gene, we need to

- **Identifying the region of interest**
- **Isolate it from the organism – DNA fragmentation**
- **moving it to another easily manageable organism such as a bacterium for obtaining multiple copies – cloning**

Such manipulations are conducted by a toolkit of enzymes:

Restriction endonucleases - used as molecular scissors

DNA ligase - to bond pieces of DNA together

- a variety of additional enzymes that modify DNA are used to facilitate the process.

Restriction endonucleases are enzymes that make **site-specific** cuts in the DNA – **chemical scissors**

Ability to cut DNA into discrete fragments allows to understand

- how genetic material of an organism is **organized**
- how expression of genetic information is **controlled**
- how **alteration** of genetic information can give rise to genetically inherited disorders, etc.
- in **bulk production** of pharmaceutically important proteins

First restriction enzyme was isolated from H. influenzae in 1970 by Daniel Nathans and Kathleen Danna

- awarded the Nobel Prize for Medicine in 1978

Restriction endonucleases are enzymes that make site-specific cuts in the DNA – chemical scissors

First restriction enzyme was isolated from H. influenzae in 1970 by Daniel Nathans and Kathleen Danna

- awarded the Nobel Prize for Medicine in 1978

Radioautogram of ^{14}C-labelled SV40 DNA (a tumor virus) cleaved with endonuclease R.

- 11 distinct DNA bands were visible on polyacrylamide gel electrophoresis, indicating that the enzyme always cut SV40 resulting in the same 11 pieces

Background

How were these restriction endonucleases identified?

Bacteria are under constant attack by bacteriophages – a virus that infect and replicates within a bacterium

To protect themselves, bacteria have developed a method to chop up any foreign DNA - such as that of an attacking phage

These bacteria build an endonuclease - an enzyme that cuts DNA - it circulates in the bacterial cytoplasm, waiting for phage DNA.

These endonucleases are termed "restriction enzymes" because they restrict the infection of bacteriophages.

Why the restriction enzymes do not chew up the genomic DNA of their host?

Background

A bacterium that makes a particular restriction endonuclease, also synthesizes a companion DNA methyltransferase,

- which methylates the DNA target sequence for that restriction enzyme, thereby protecting it from cleavage.

DNA from an attacking bacteriophage will not have these protective methyl groups and will be destroyed.

Methyl groups block the binding of restriction enzymes, but do not block the normal reading and replication of the genomic information stored in the host DNA.

DNA Fragmentation

Different endonucleases present in different bacteria recognize different nucleotide sequences

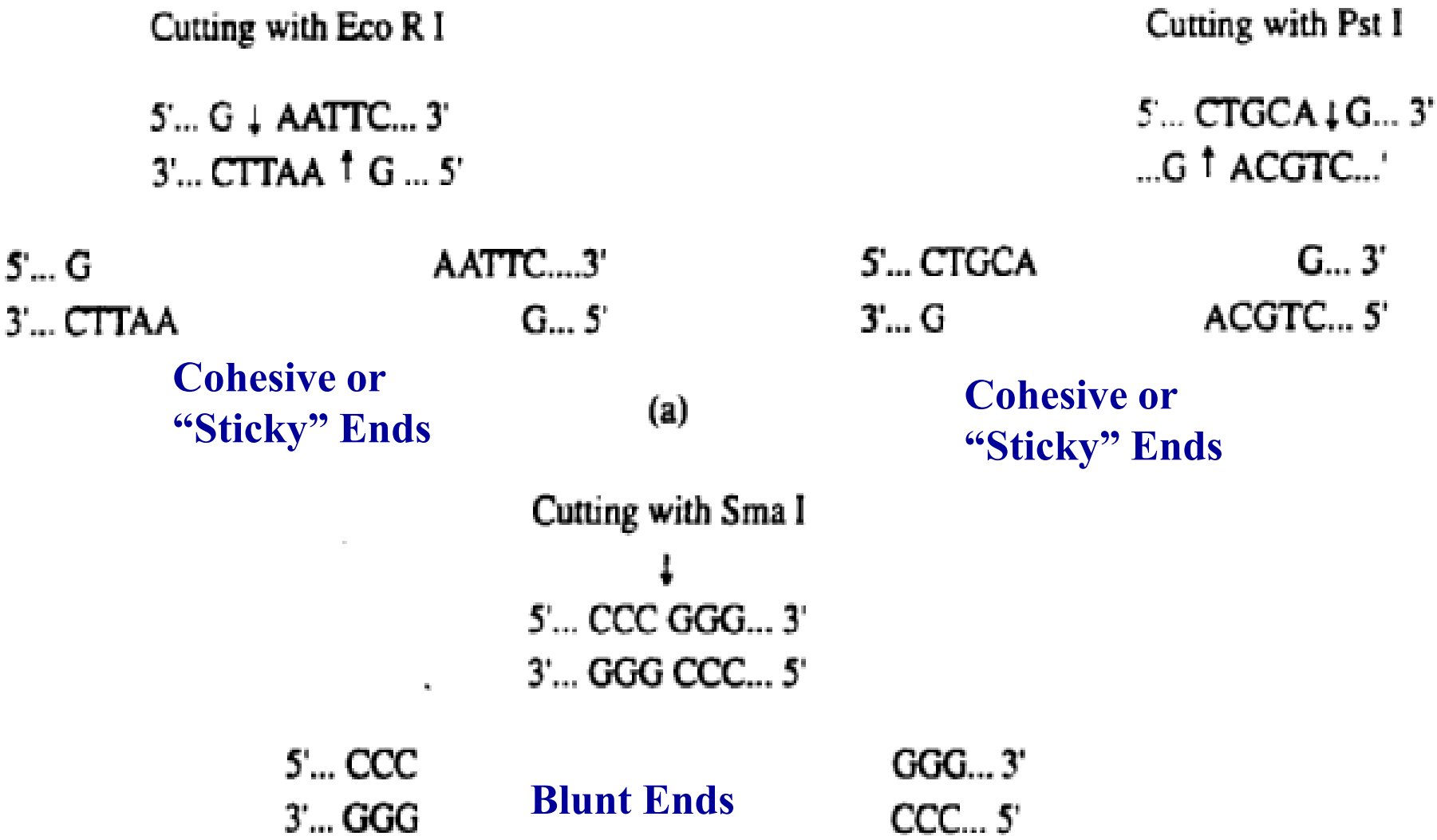
Naming of restriction enzymes - after their host of origin, e.g.,

- **EcoRI -** ***Escherichia coli***
- **Hind II & Hind III -** ***Haemophilus influenzae***
- **XhoI -** ***Xanthomonas holcicola***

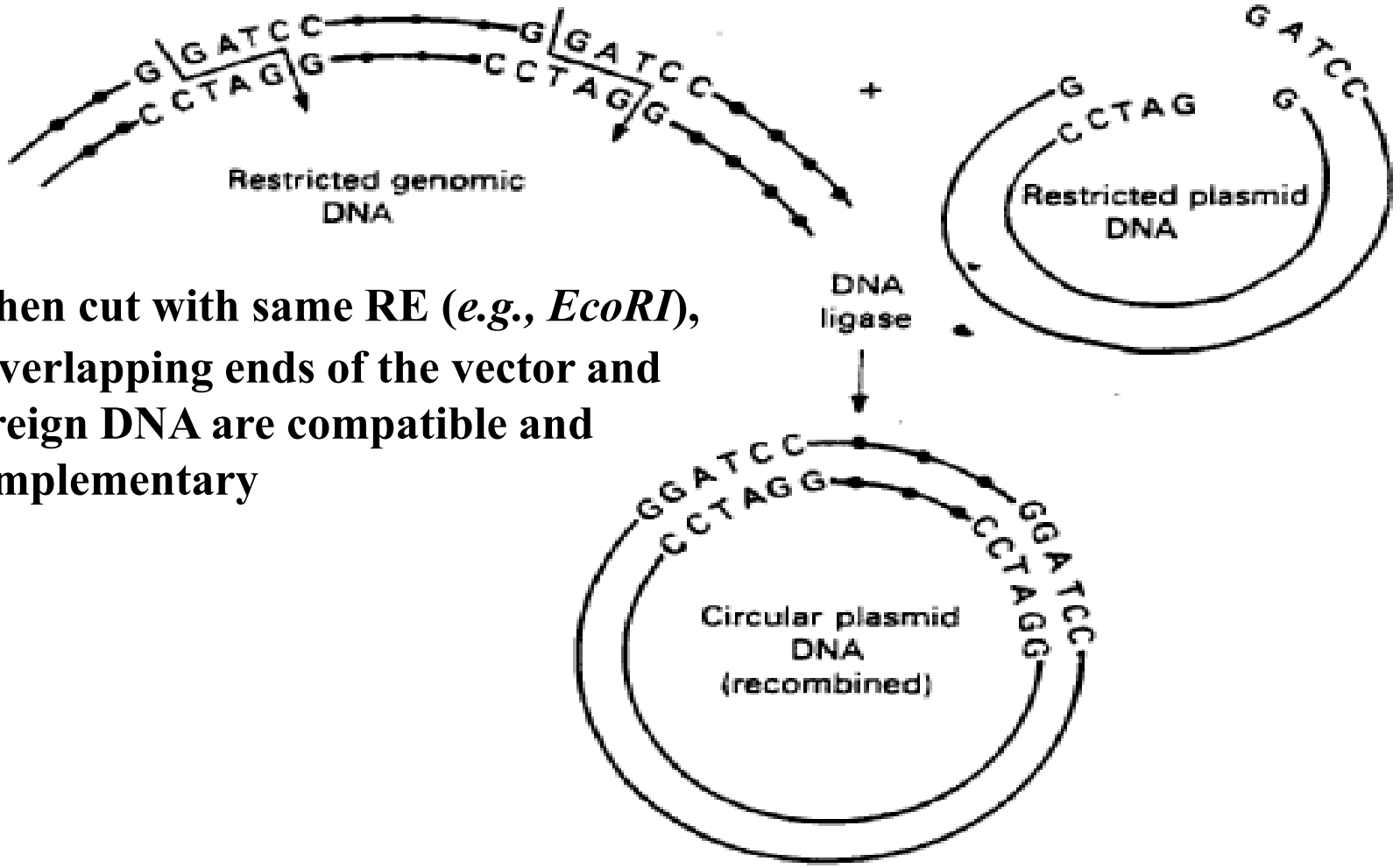
When cut with a restriction enzyme (RE), the ends of the cut DNA fragment can be cohesive or blunt-ended depending on the enzyme.

Enzyme	Recognition Sequence
EcoRI	G↓AATTC
HindIII	A↓AGCTT
BamHI	G↓GATCC
BglI	GCCNNNN↓NGGC
PvuI	CGATC↓G
HaeIII	GG↓CC
MboI	GAT↓C

Generation of Cohesive & Blunt-ended Fragments



Restriction enzyme digestion of genomic DNA and plasmid vector DNA



When cut with same RE (*e.g.*, *EcoRI*),

- overlapping ends of the vector and foreign DNA are compatible and complementary

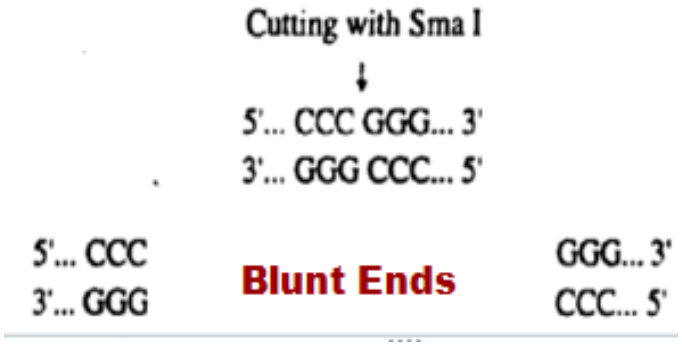
How does one cut a DNA if it doesn't contain desired RE sites?

Or

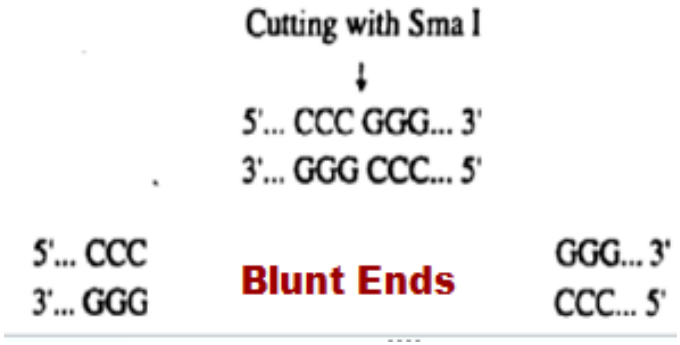
If the RE site is present within the DNA of interest?

Or

If the RE result in blunt-ended DNA fragments?



If the RE result in blunt-ended DNA fragments, then how do we clone it?

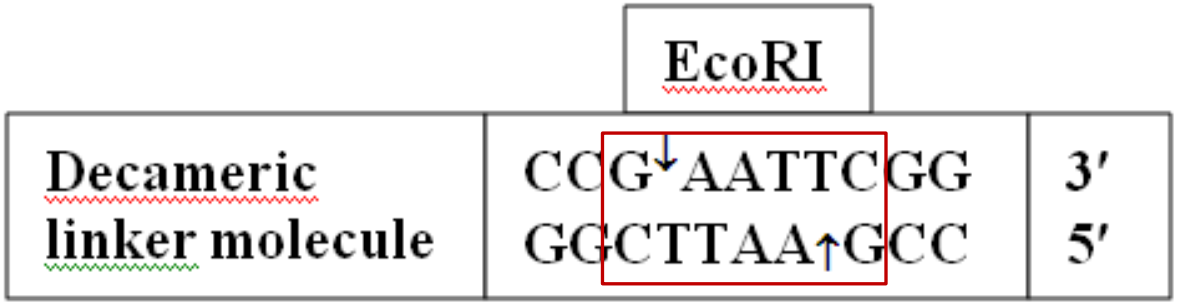


To insert this DNA in the vector,

- a linker molecule can be ligated on either side by DNA ligase, cut with the RE contained in the linker molecule to obtain cohesive ends.

Linkers & Adaptors

Linkers - short, double-stranded DNA molecules (~ 8-14bp) with one internal site for RE (~ 3-8bp)



Problem?

- **the sites for the enzyme used to generate cohesive ends may be present in the target DNA fragment, limiting its use for cloning.**
- **This problem can be solved using adaptors.**

Linkers & Adaptors

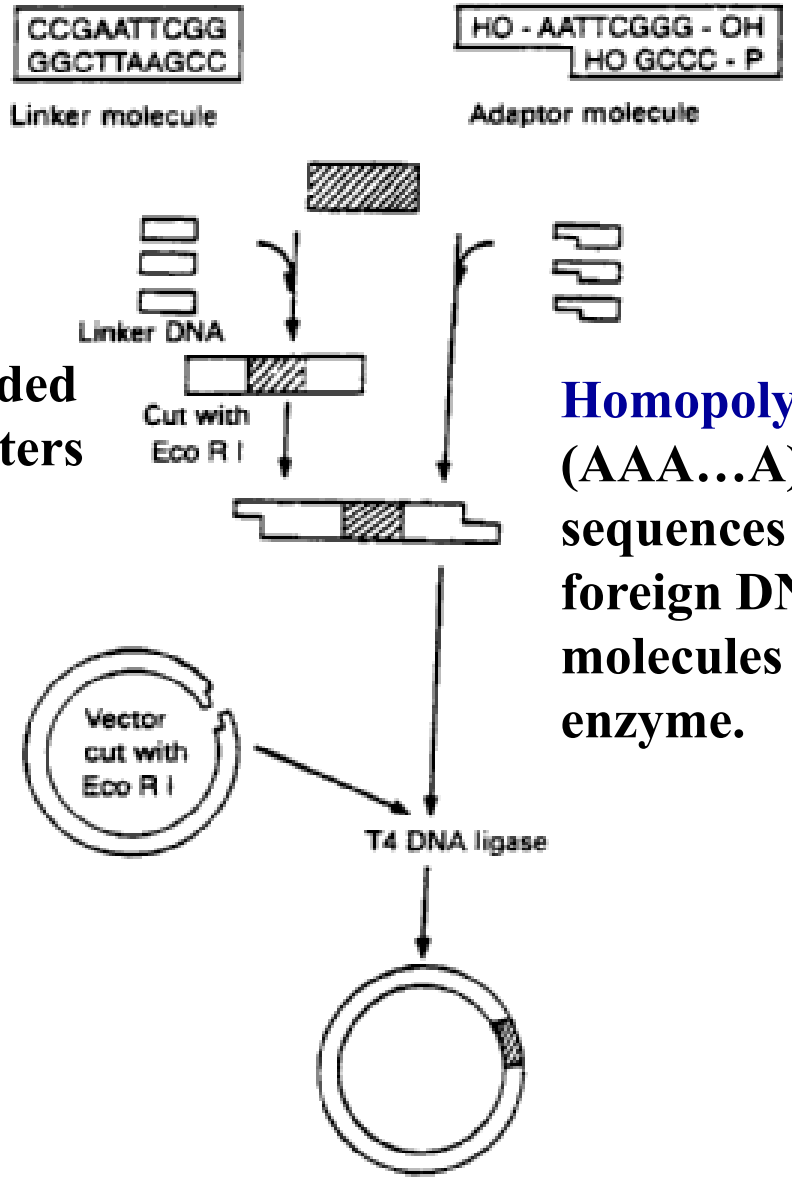
Adaptors - chemically synthesized DNA molecules with pre-formed cohesive ends

- it has one blunt end bearing a 5′ phosphate group and another cohesive end for a specific RE which is not phosphorylated to prevent self ligation.

Adaptor molecule

- reduces the need for restriction digestion following ligation

Use of Linker & Adaptor Molecules in the Formation of Recombinant Plasmids



This step avoided on using adapters

Homopolymer method - oligo A (AAA…A) or oligo T (TTTT…T) sequences added to the ends of foreign DNA & vector DNA molecules by terminal transferase enzyme.

Features of Restriction Enzymes

- **Length of recognition sequence dictates how frequently the enzyme will cut a DNA sequence**

  Which of the recognition sites - of length, 4, 6, or 8, will occur at higher frequency? At what distances will they occur?

- **Different REs can have the same recognition site and are called isoschizomers, e.g., *SacI* & *SstI* : GAGCTC**

- **Restriction recognitions sites can be unambiguous, e.g., *BamH* I recognizes the sequence GGATCC and no other, or ambiguous, e.g., *Hinf* I has a recognition site, GANTC.**

  Recognition sites for Hinf I will occur at what frequency?

Features of Restriction Enzymes

- **Recognition site for one enzyme may contain the restriction site for another, e.g., *BamH* I recognition site (GGATCC) contains the recognition site for *Sau3A* I (GATC).**

Sau3A I recognizes the sequence GATC and produces the same sticky ends as BamH I upon cutting

Will the two REs give the same results? If not, which one will give larger number of fragments?

- **Most recognition sequences are palindromes - they read the same forward and backward**

Can we use the property of palindrome sequence to identify restriction recognition sites?

Applications of Restriction Enzymes

Danna & Nathans showed that it was possible:

- **to prepare a physical map of the SV40 genome**
- **to localize the origin of replication**
- **to position early & late genes of SV40 onto this "restriction map"**
- **that any individual gene could be mapped by testing for biological activity during transformation experiments**
- **informative mutants could be made by deleting one or more of the specific fragments**

Applications of Restriction Enzymes

- **Variations in DNA sequences, *viz.*, mutations in recognition sites, as a result of VNTRs, insertions, deletions, inversions and translocations, can be identified by RE analysis**

  - The length variations is known as restriction fragment length polymorphisms (RFLPs).

- **In genetic engineering - using REs DNA may be cut at precise locations & using DNA ligase, reassembled in any desired order, allowing the researchers to assemble customized genomes; create designer bacteria that make insulin, or growth hormones, or add genes for disease resistance to agricultural plants, etc.**

- **in DNA sequencing – first step is to cut the DNA in manageable pieces**

Restriction Map

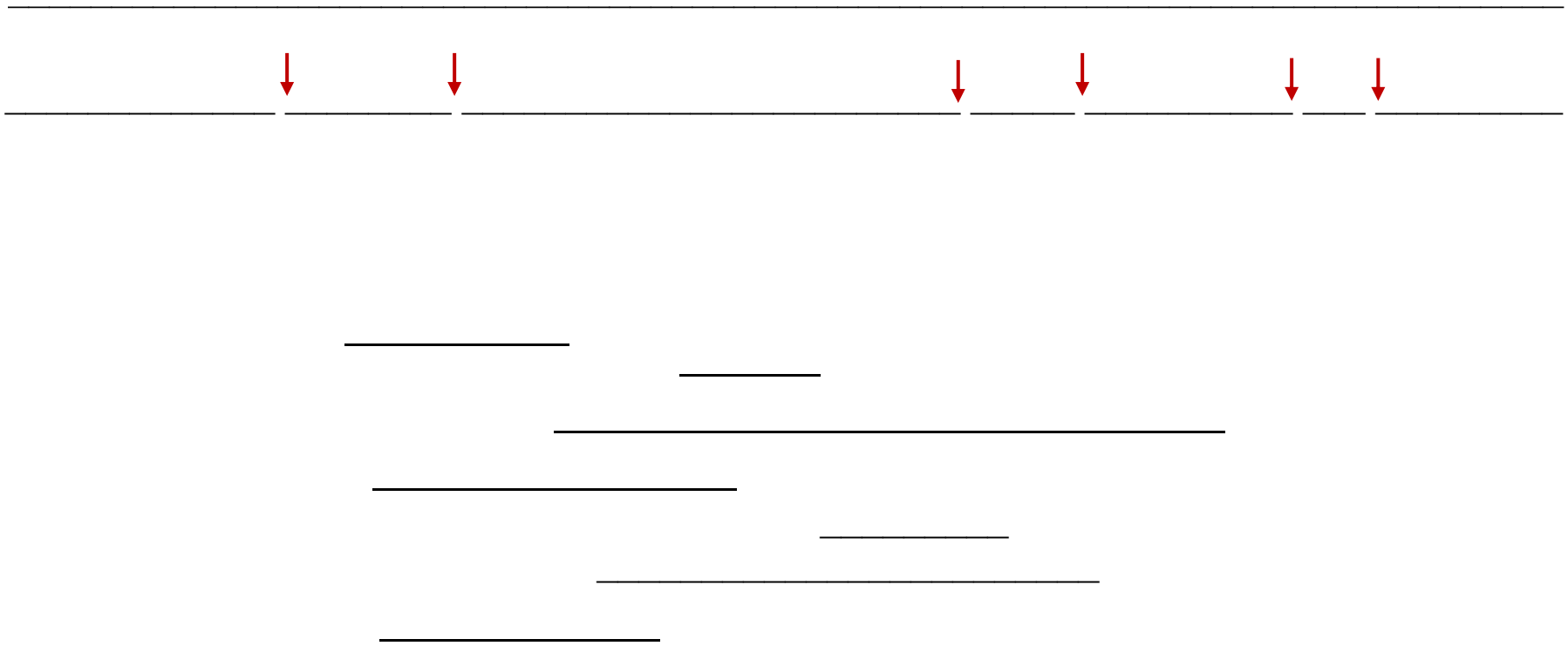
Restriction map is a description of restriction endonuclease cleavage sites within a piece of DNA

- **generating such a map is the first step in characterizing an unknown DNA**

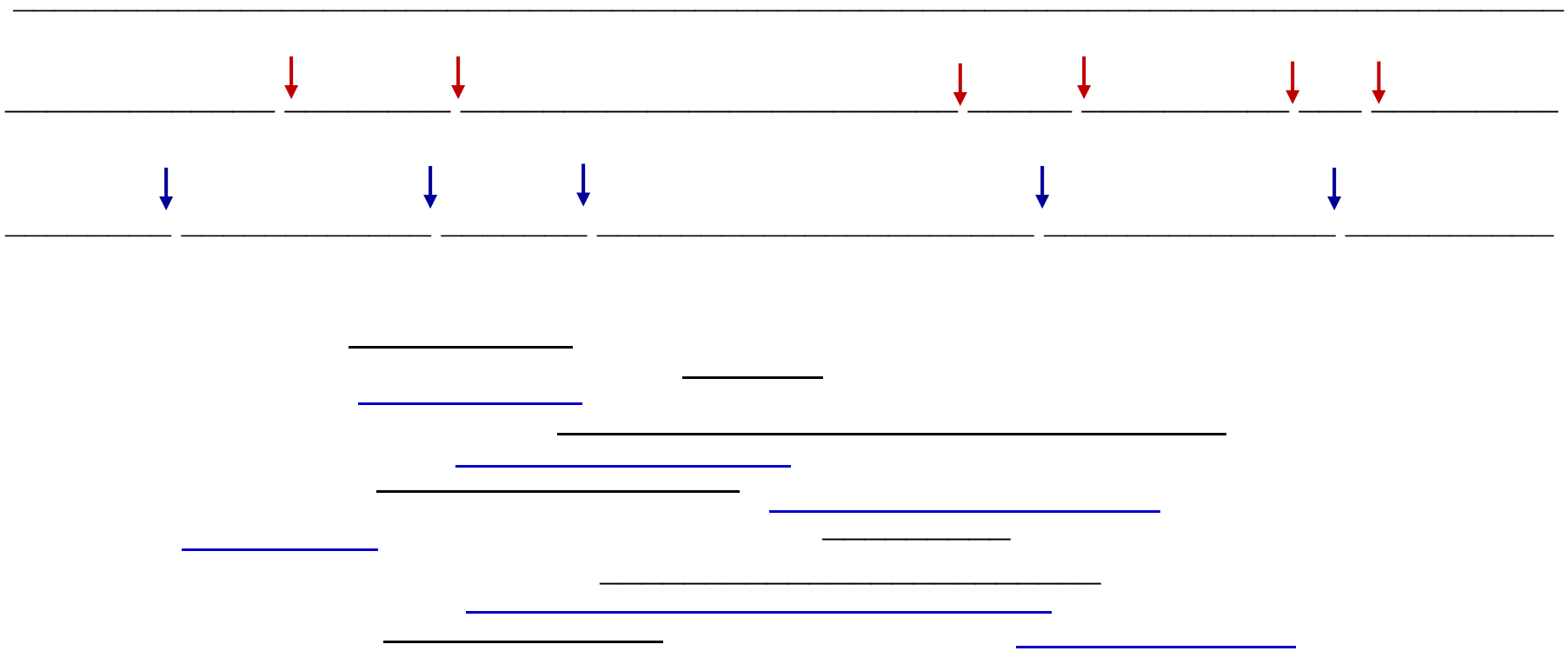
Multiple Complete Digest Mapping – creates a map by digesting DNA with multiple REs

- each recognizing a different specific short DNA sequence and producing a separate fingerprint for each clone

Because of the frequent occurrence of these sites, restriction mapping produces a relatively fine scale of physical map.



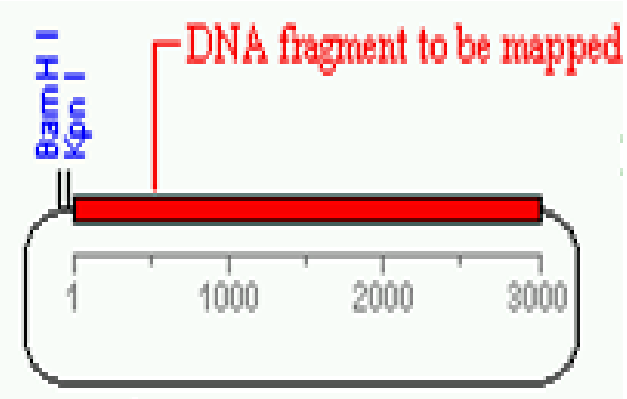
How do you order the fragments in the correct order?



The fragments can be arranged in the correct order by finding the overlapping fragments

Restriction Mapping

Ex: Consider a plasmid that contains a 3000 bp fragment of unknown DNA & unique recognition sites for enzymes Kpn I & BamH I.



Consider first separate digestions with Kpn I & BamH I:

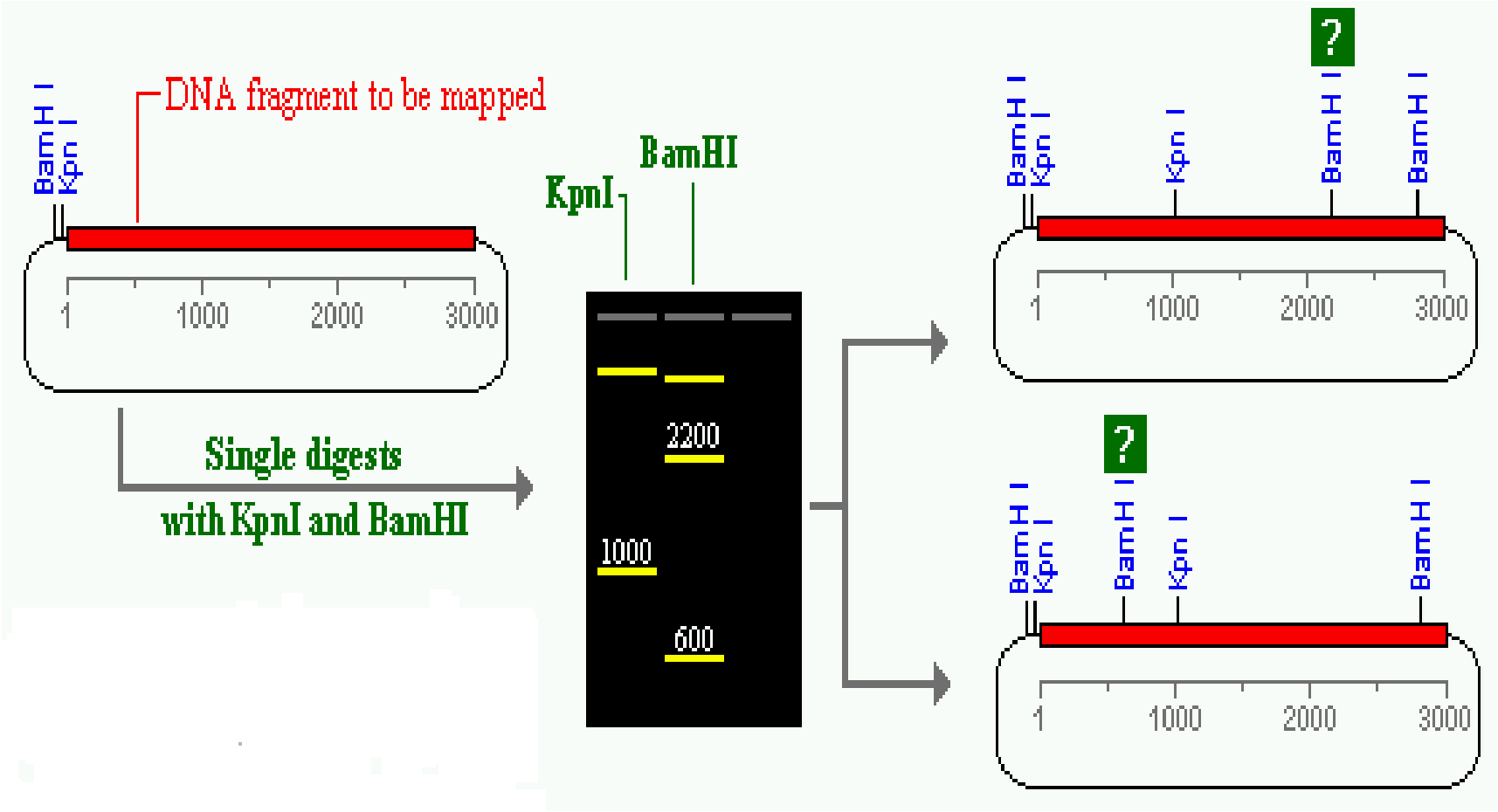
Kpn I yields 2 fragments: 1000bp & "big"

BamH I yields 3 fragments: 600, 2200 & "big"

big – part of unknown DNA sequence + vector

⇒ one Kpn I site & two BamH I sites are present in the unknown DNA sequence, given 1 each on the vector sequence

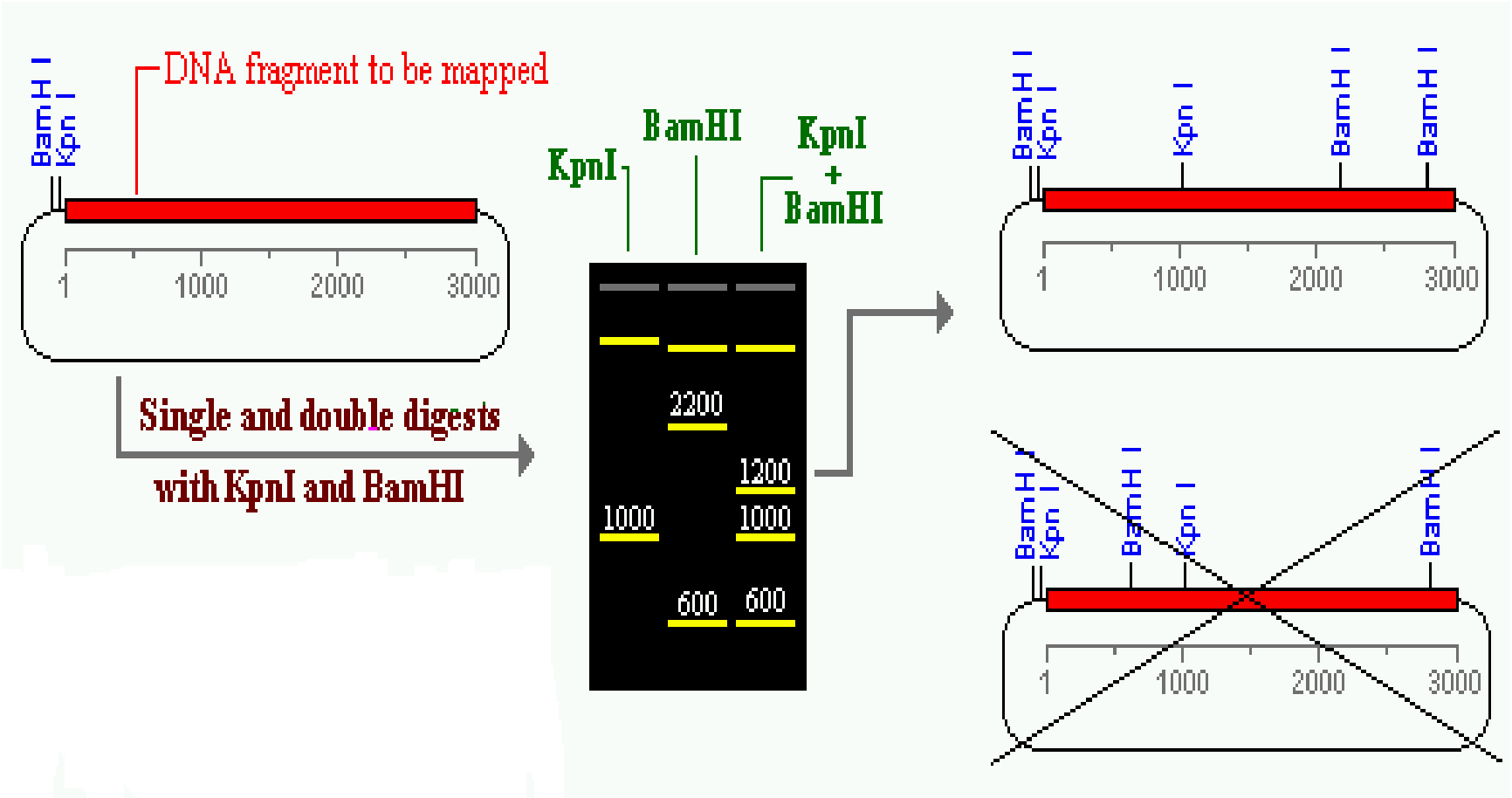
Restriction Mapping



One BamH I site is at 2800 bp. Trick to determine the location of 2nd BamH I site is to digest the plasmid with Kpn I & BamH I together

Restriction Mapping

Double digest yields fragments of 600, 1000 & 1200 bp (plus the "big" fragment).



Restriction Mapping

If the above process is conducted with a larger set of enzymes, a much more complete map would result

single digests - are used to determine which fragments are in the unknown DNA, and

multiple digests - to order and orient the fragments correctly.

For any novel genome, e.g., SARS-CoV-2, can a physical map be constructed computationally?

Restriction Mapping

Using a Computer to Generate Restriction Maps

If the sequence is known, feed it to computer programs, which will search the sequence for various RE recognition sites and build a map.

- **Mapper - available as part of Molecular Toolkit**

  http://arbl.cvmbs.colostate.edu/molkit/mapper/

- **Webcutter**

  http://www.firstmarket.com/cutter/cut2.html

- **RebSite – as part of the REBASE Tools**

  http://tools.neb.com/REBsites/index.php3

REBASE
The Restriction Enzyme dataBASE

A comprehensive database containing information:

- **restriction enzymes, methylases & related proteins involved in restriction-modification processes**
- **recognition and cleavage sites, isoschizomers, neoschizomers, commercial availability, methylation sensitivity, crystal & sequence data.**

All newly sequenced genomes are analyzed for the presence of putative restriction systems and these data included in REBASE

It is updated daily (http://rebase.neb.com/)

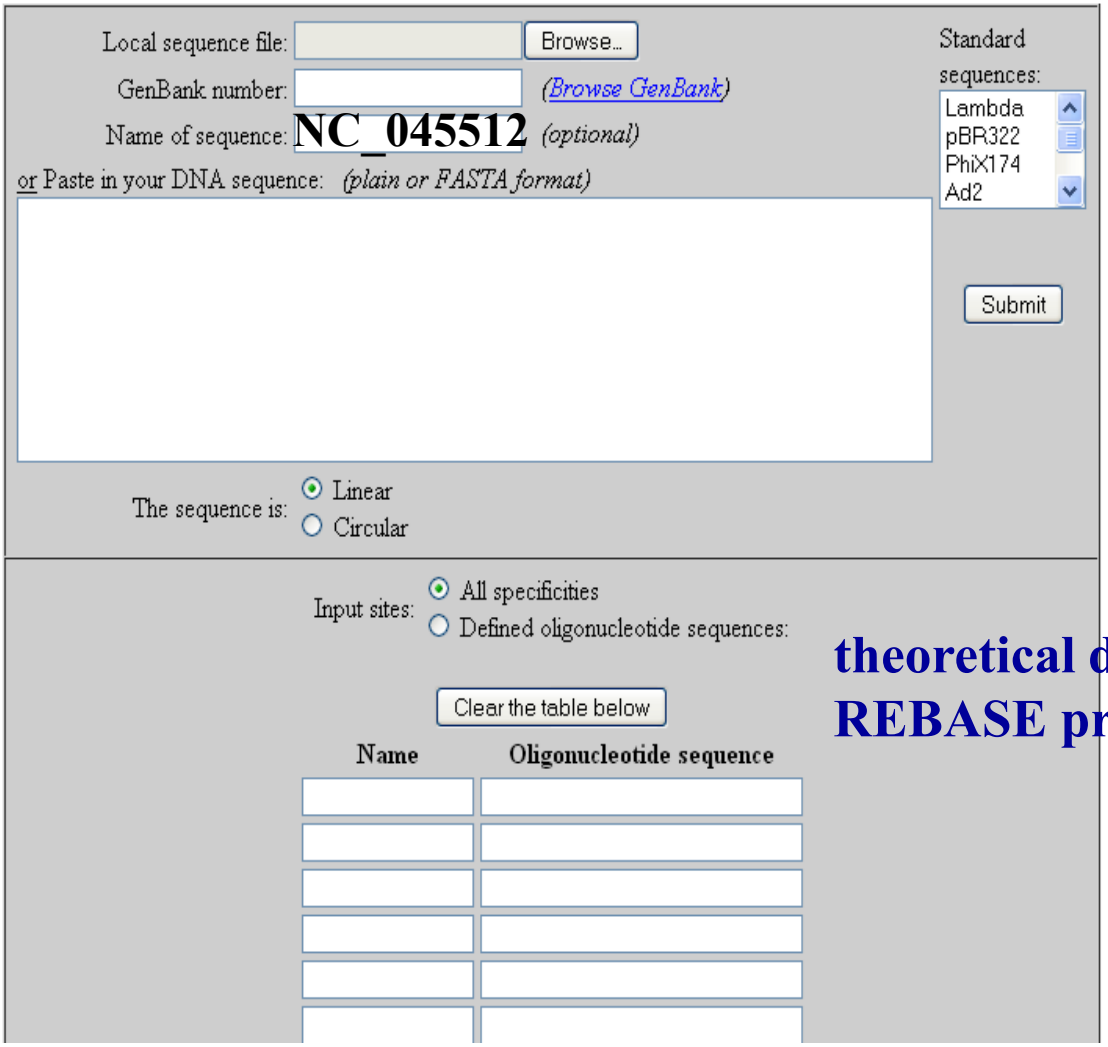
Ref: Robert et al, *Nucl. Acids Res.* 43: D298-D299 (2015)

Back to...

Program Guide

REBsites

This tool will take a DNA sequence and digest it with one example of each of the known Type 2 restriction enzyme specificities.
The maximum size of the input file is 2 MByte, and the maximum sequence length is 200 KBases.



theoretical digest with all REBASE prototypes

[New DNA]

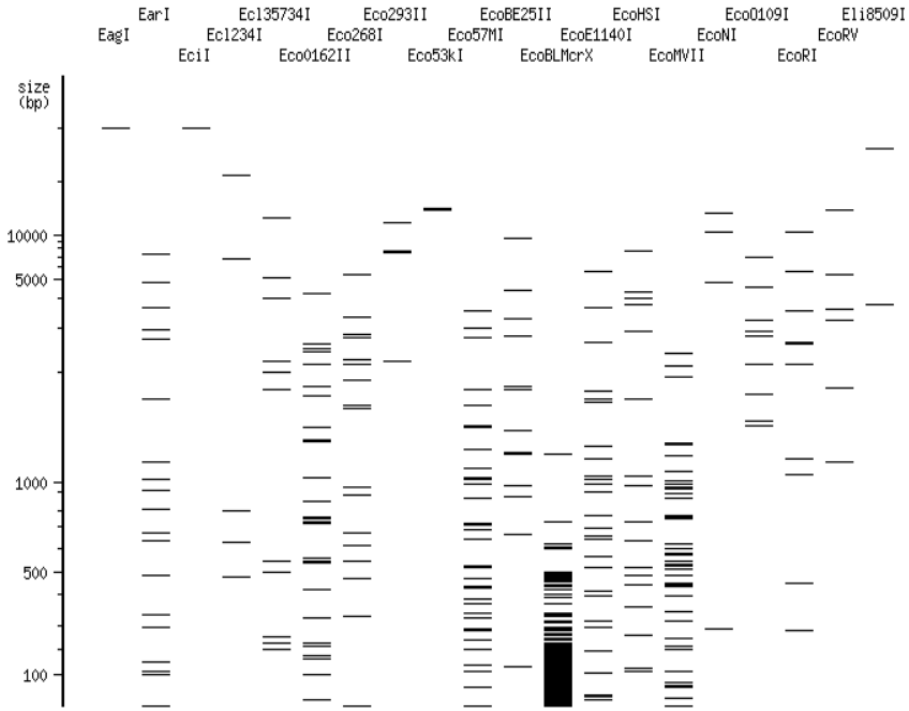
REBsites

NC 045512

Gel: Agarose 0.7%
Order by: enzyme name

[<< Prev] 1 2 3 4 5 6 7 8 9 10 11 12 **13** 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 [Next >>]

[Print]



Click on an enzyme name for a list of fragments/sites.

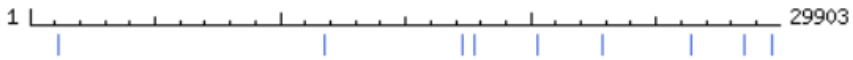
Print

Fragment list

Close

NC 045512 digested with EcoRI

[Sites with flanks]



#	Location	Size [bp]
1	1162-11734	10573
2	11735-17280	5546
3	22871-26439	3569
4	20279-22870	2592
5	17729-20278	2550
6	26440-28551	2112
7	1-1161	1161
8	28552-29620	1069
9	17281-17728	448
10	29621-29903	283

Assignment

- **Write a program to generate a restriction map for Wuhan isolate-1 genome (Acc. Id.: NC_045512) using EcoRI as RE compare your results with REBsites.**
- **Write a program to identify restriction recognition sites in a given DNA sequence.**

Cloning

What is cloning?

The process of cloning involves the production of multiple copies of a DNA fragment of interest by amplification *in vivo*

- depends upon the ability of vectors to continue their life cycles in bacterial or yeast cells in spite of having foreign DNA inserted into them.

Cloning vector - a DNA molecule that carries foreign DNA into a host cell, replicates inside a bacterial (or yeast) cell and produces many copies of itself and the foreign DNA

Gene Cloning involves two major steps:

- **Gene of interest is isolated & incorporated into a small, self-replicating DNA molecule (called a vector) which can be an *E. coli* plasmid or a virus.**

  - Vector with an incorporated gene is called a recombinant vector.

- **Recombinant vector is then introduced into a suitable host cell. As the vector replicates, many identical copies of the cloned gene are produced.**

Major requirement of all vectors - an origin of replication for a given host cell in order that they may replicate autonomously (i.e., independently of the host's chromosome)

Features of Cloning Vectors:

- **sequences that permit the propagation of itself in bacteria (or yeast)**
- **a cloning site to insert foreign DNA; the most versatile vectors contain a site that can be cut by many REs**
- **a method of selecting for bacteria (or yeast) containing a vector with foreign DNA; usually accomplished by selectable markers for drug resistance**

Types of Vectors

Vector	Insert size (kb)
Plasmids	<10 kb
Bacteriophage	9 - 20 kb
Cosmids	33 - 47 kb
Bacterial artificial chromosomes (BACs)	75 - 125 kb
Yeast artificial chromosomes (YACS)	100-1000 kb

Types of Vectors

Plasmids - an extra-chromosomal double-stranded circular DNA molecules that replicates autonomously inside the bacterial cell

Plasmids are important as one can:

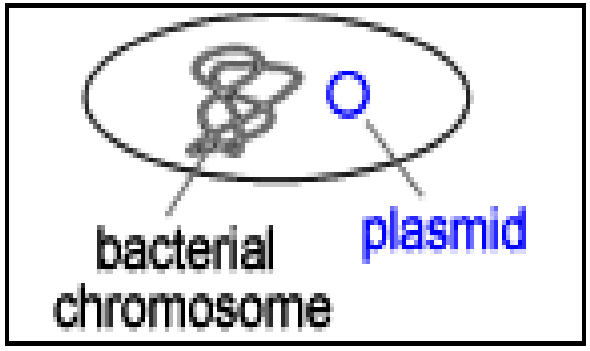
(i) isolate them in large quantities,

(ii) cut & splice them, add DNA of choice,

(iii) put them back into bacteria, where they replicate along with the bacteria's own DNA,

(iv) isolate them again to get billions of copies of inserted DNA

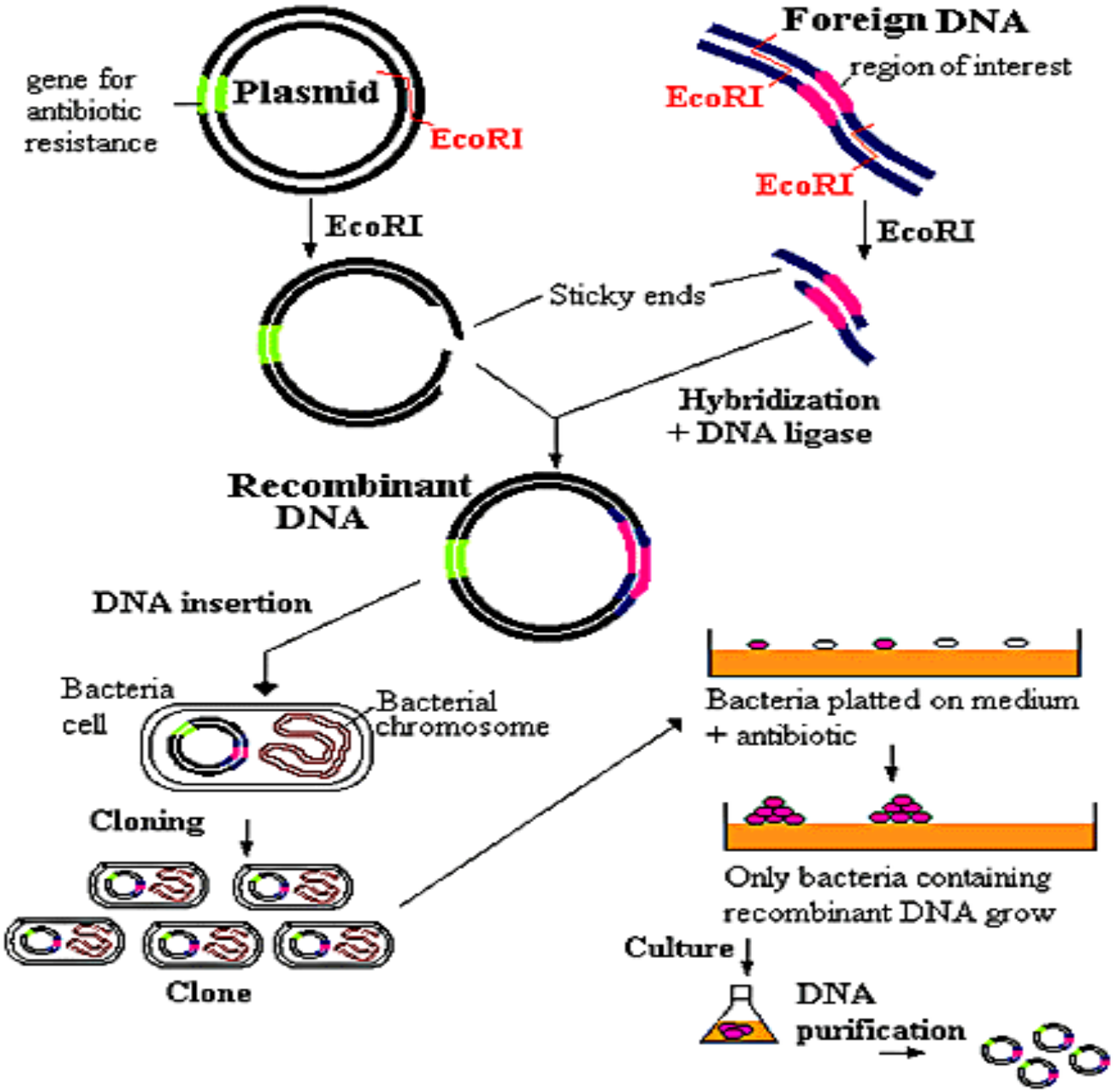
Limitation: size of DNA that can be introduced into the cell by transformation (~ 2 - 10kb)



Plasmid vectors are derived from naturally occurring plasmids of *E. coli* such as ColE1 or from related plasmid pMB1

pBR322 – most widely used cloning vectors of *E. coli*, is a hybrid between ColE1 & genes coding for resistance to antibiotics tetracycline & ampicillin

What's the advantage of inserting genes coding for resistance to antibiotics into a vector?



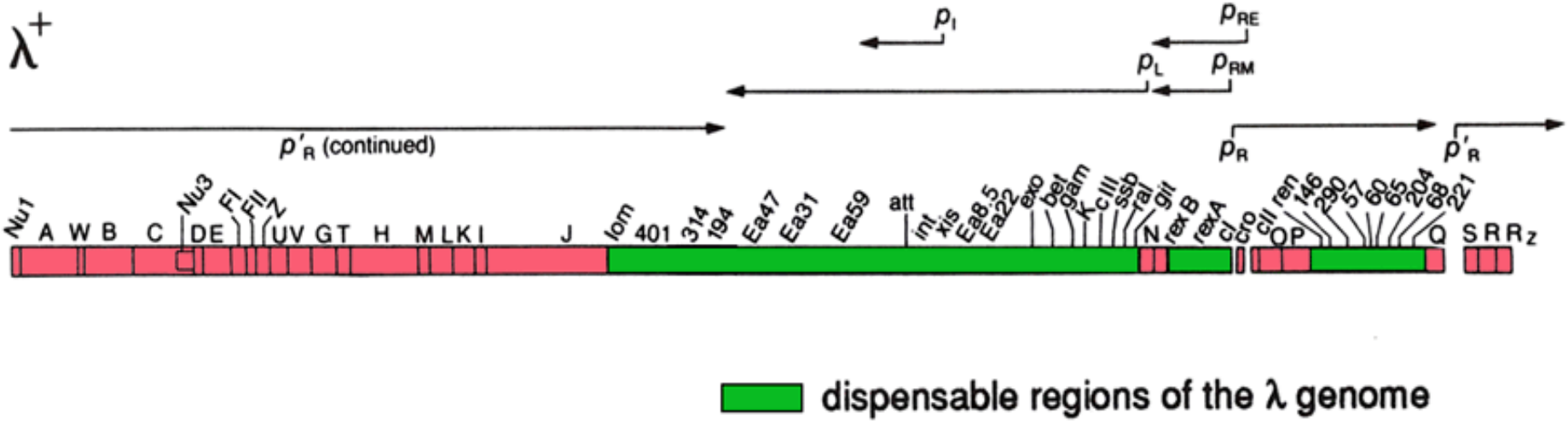
Cloning into a plasmid

Types of Vectors

Bacteriophage Vectors

– a double-stranded linear molecule of size 49.5Kbp

Cloning limit: 9 - 20 kb



Enterobacteria phage λ is a bacterial virus, or bacteriophage, that infects the bacterial species *E. coli*.

Artificially Constructed Vectors

Cosmids - an extra-chromosomal circular DNA molecule that combines features of plasmids and cos gene of phage lambda

Cloning limit: 35 - 50 kb

BAC - Bacterial Artificial Chromosome

- based on naturally occurring F-factor plasmid found in the bacterium E. coli.

Cloning limit: 100-300 kb

YAC - Yeast Artificial Chromosomes

- it is a vector constructed from yeast DNA, used to clone large DNA fragments

Cloning limit: 100-1000 kb

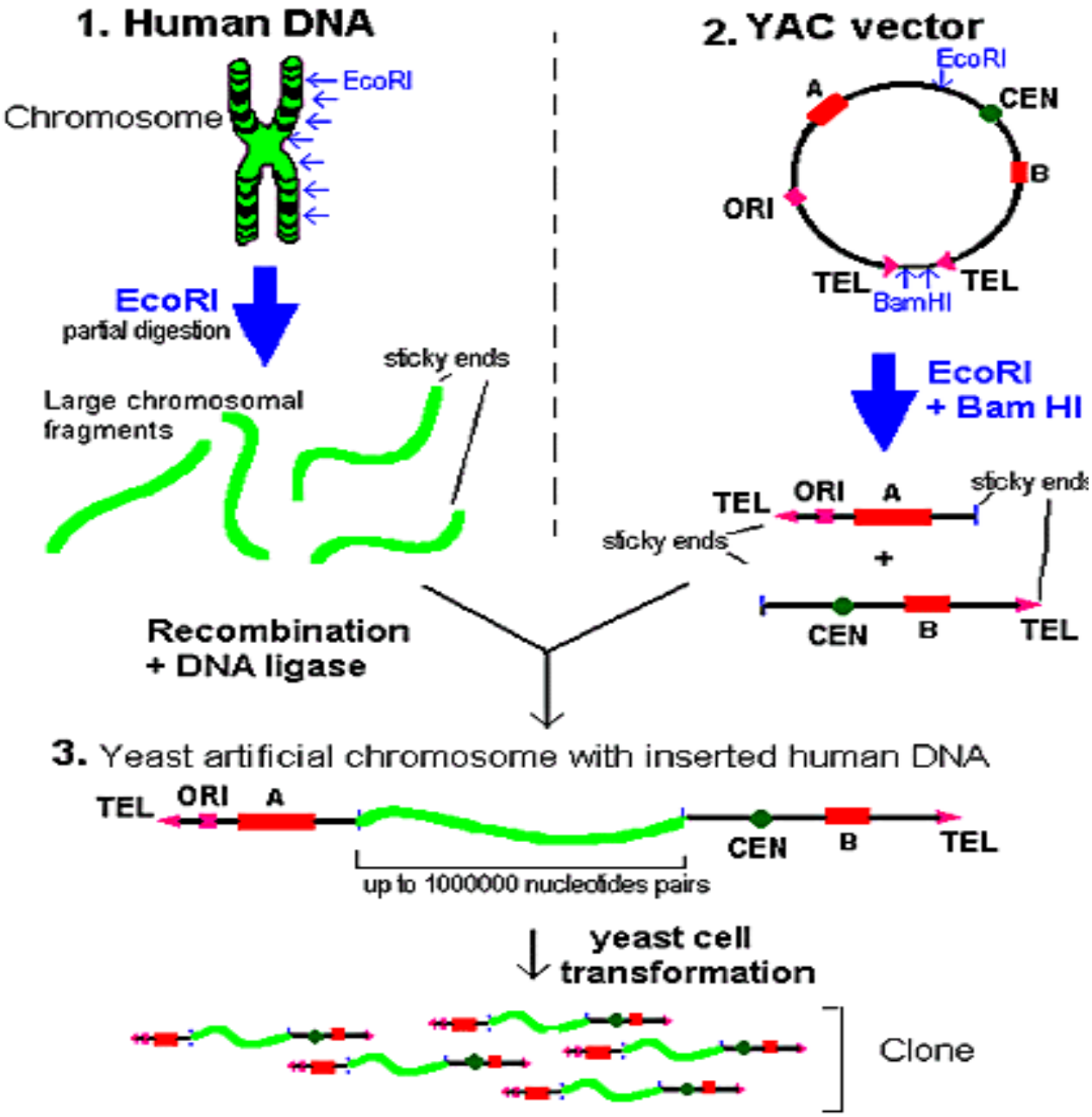
Useful for cloning long segments of eukaryotic DNA

YAC - a functional self-replicating artificial chromosome. It includes three specific DNA sequences that enable it to propagate from one cell to its offspring:

- **TEL: The telomere which is located at each chromosome end, protects the linear DNA from degradation by nucleases**
- **CEN: The centromere which is the attachment site for mitotic spindle fibers, "pulls" one copy of each duplicated chromosome into each new daughter cell.**
- **ORI: Replication origin sequences, specific DNA sequences that allow the DNA replication machinery to assemble on the DNA and move at the replication forks**

It also contains few other specific sequences like:

- **A and B: selectable markers that allow easy isolation of yeast cells that have taken up the artificial chromosome.**
- **Recognition site for two REs: EcoRI & BamHI**



Cloning into a Yeast Artificial Chromosome (YAC

Why is it important to be able to clone large sequences?

To map the entire human genome ($3x10^9$ bps) would require more than 1000,000 plasmid clones (~10Kb limit).

In principle, the human genome could be represented in about 10,000 YAC clones (~1Mb limit)

What determines the choice vector?

- **insert size**
- **vector size**
- **restriction sites**
- **copy number**
- **cloning efficiency**
- **ability to screen for inserts**

DNA Sequencing

DNA Sequencing - determine the precise sequence of nucleotides in a sample of DNA

Various types of sequencing:

- **Sequencing a region of interest, e.g., gene.**
- **Whole Genome/Exome Sequencing**
- **cDNA Sequencing – sequencing cDNA libraries of expressed genes**
- **High-throughput sequencing – next-generation, 3rd & 4th generation sequencing**
- **Metagenome sequencing - sequencing of environmental samples**

- depending on the nature of analysis, type of sample, or type of sequencer used

Sequencing a Region of Interest

First requirement in sequencing a region of DNA is

- to have enough starting template for sequencing.

This is achieved by PCR - Polymerase Chain Reaction

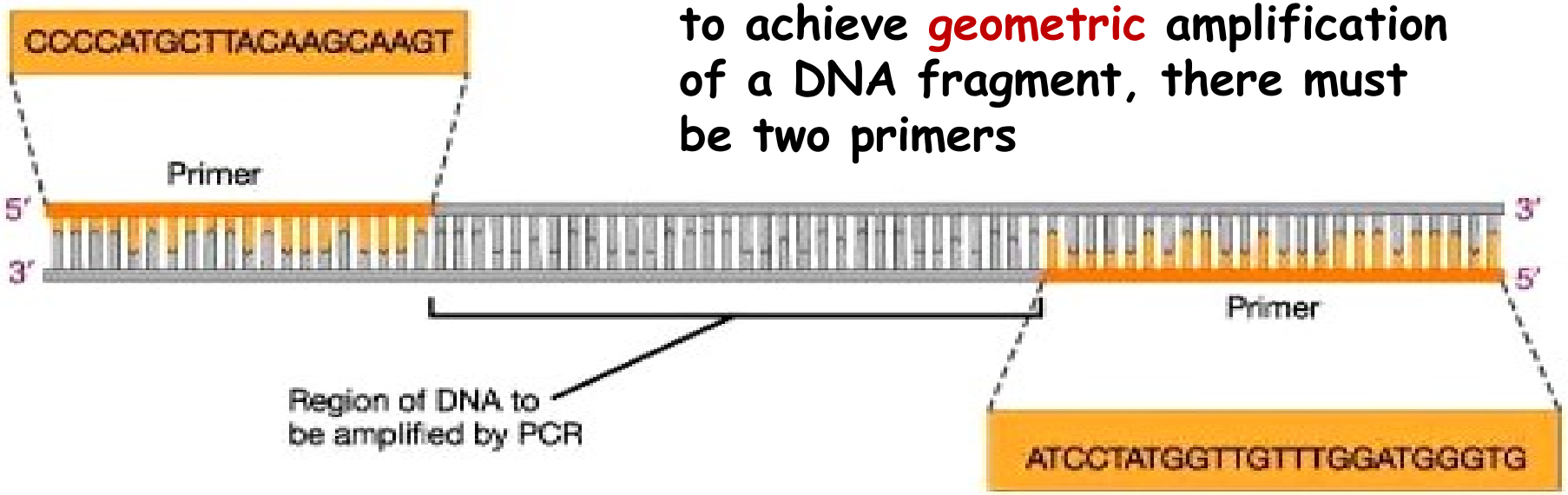
- **carried out in an automated cycler for 30 - 40 cycles.**

Essential requirements for a PCR:

- **a mixture of 4 deoxy-nucleotides in ample quantities**
- **dATP, dGTP, dCTP, dTTP**
- **Taq DNA polymerase**
- **primers**
- **Genomic DNA of interest**

What is the advantage of using PCR over traditional gene cloning?

Region of DNA to be amplified by PCR



to achieve geometric amplification of a DNA fragment, there must be two primers

Primers - short single-stranded oligonucleotides which anneal to the DNA template and serve as a starting point for DNA synthesis

Why are primers required?

The Cycling Reactions

Step-1: Denaturation at 94°C

- opens up double stranded DNA, all enzymatic reactions stop.

Step-2: Annealing at 54°C

- Primers jiggling around because of Brownian motion, binds to single stranded template once an exact match is found; the polymerase then attaches and start copying the template.

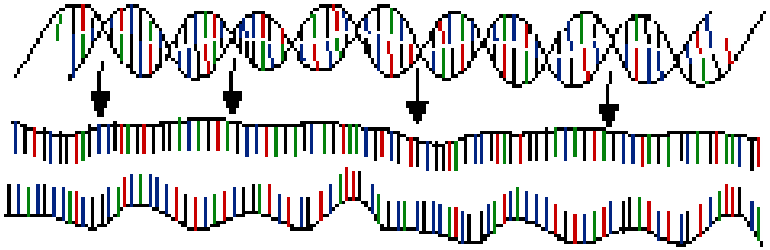
Step-3: Extension at 72°C

- ideal working temperature for the polymerase. Bases complementary to the template are coupled to the primer on 3' side (reading the template from 3' to 5' side)

Different Steps in PCR

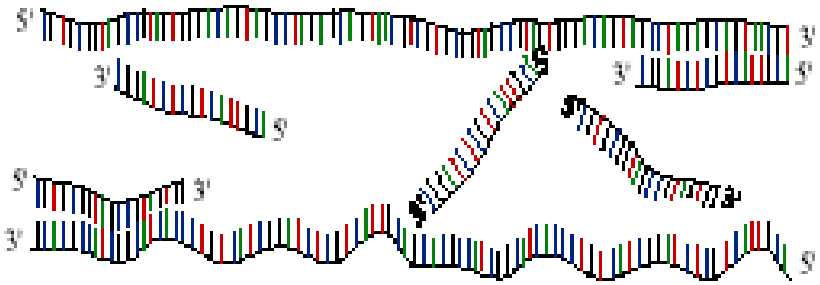
PCR : Polymerase Chain Reaction

30 - 40 cycles of 3 steps :



Step 1 : denaturation

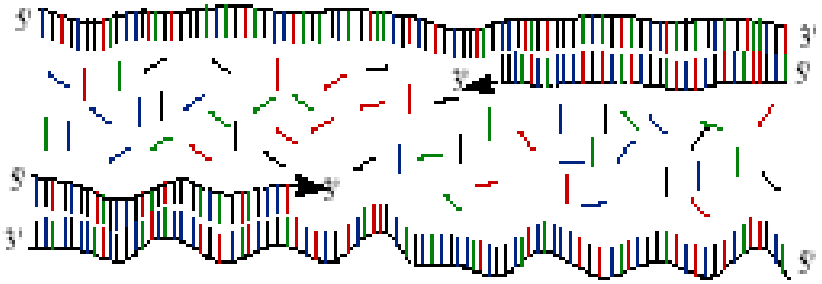
1 minut 94 °C



Step 2 : annealing

45 seconds 54 °C

forward and reverse primers !!!

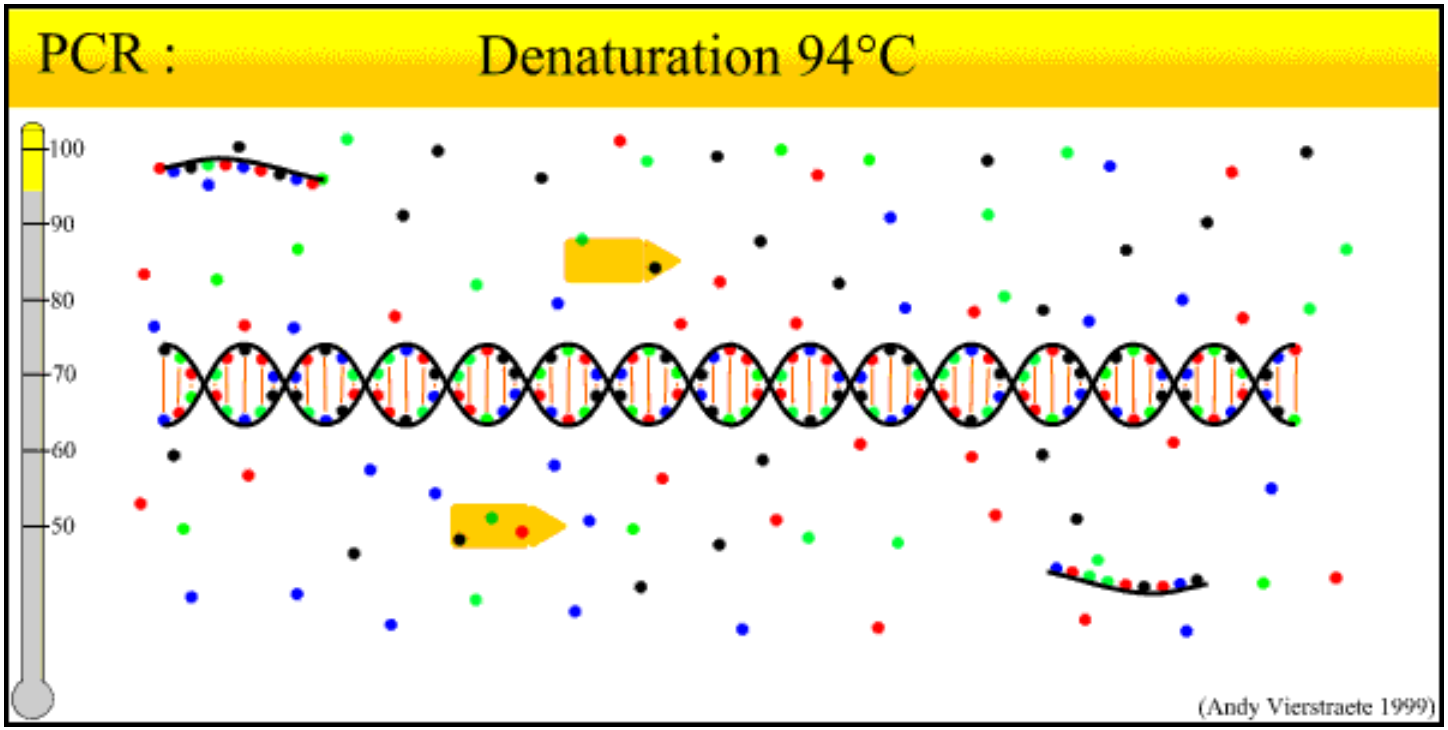


Step 3 : extension

2 minutes 72 °C

only dNTP's

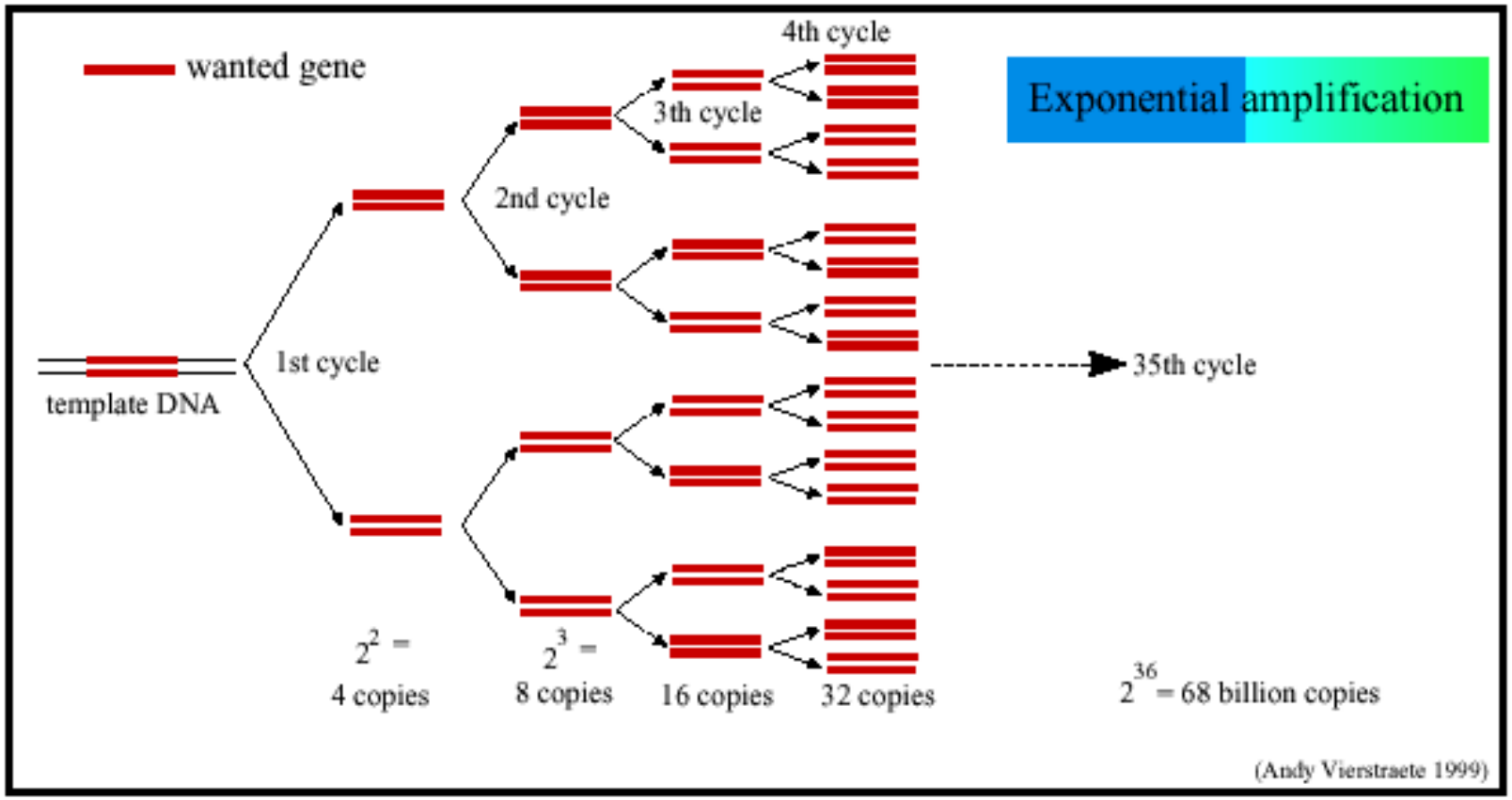
Different Steps in PCR



Exponential amplification of region of interest

<u>Both strands</u> are copied during PCR

- leading to an exponential increase of the number of copies of the region of interest.



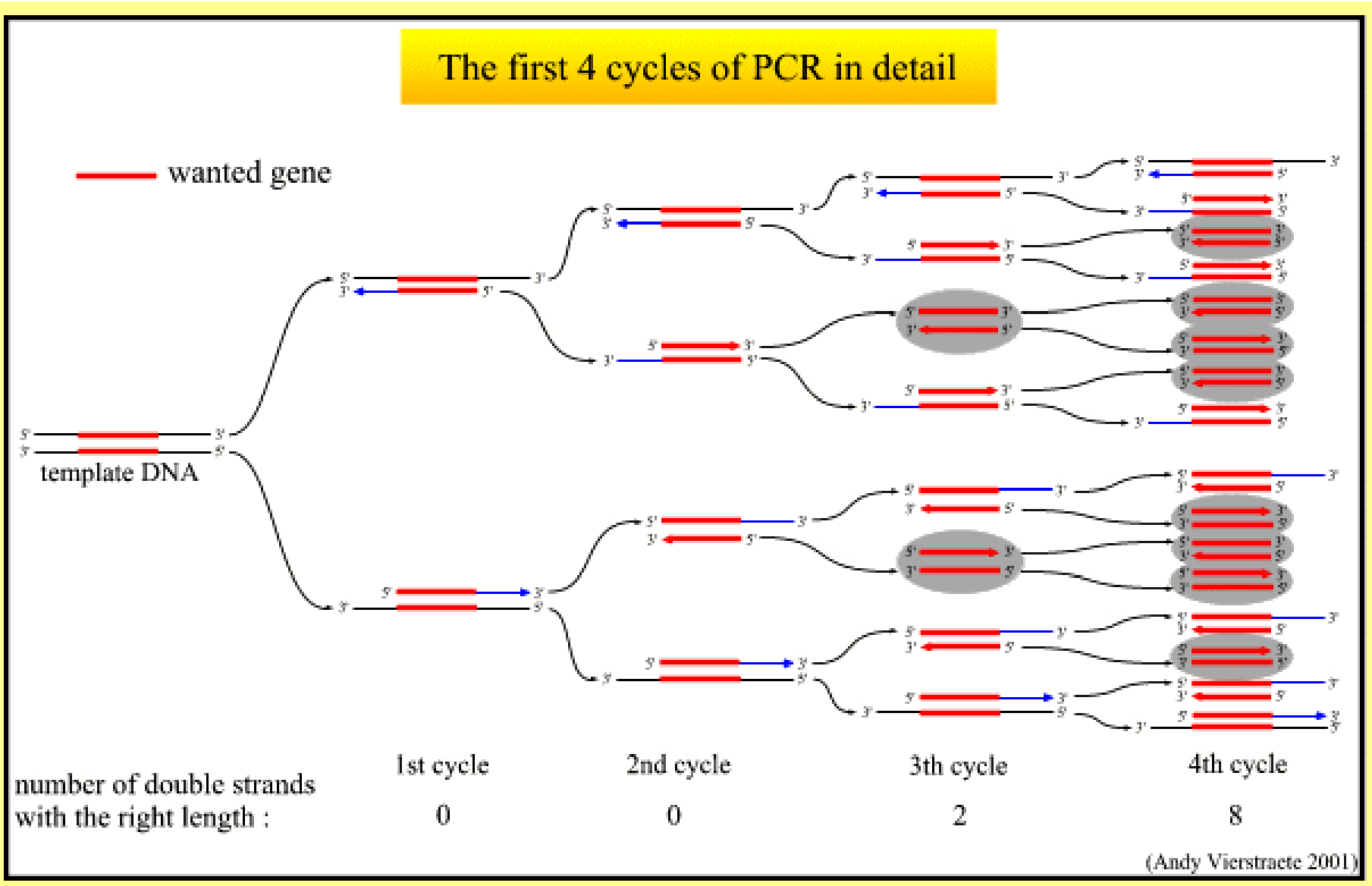
Verification of PCR Product

Is the template copied during PCR and is it the right size?

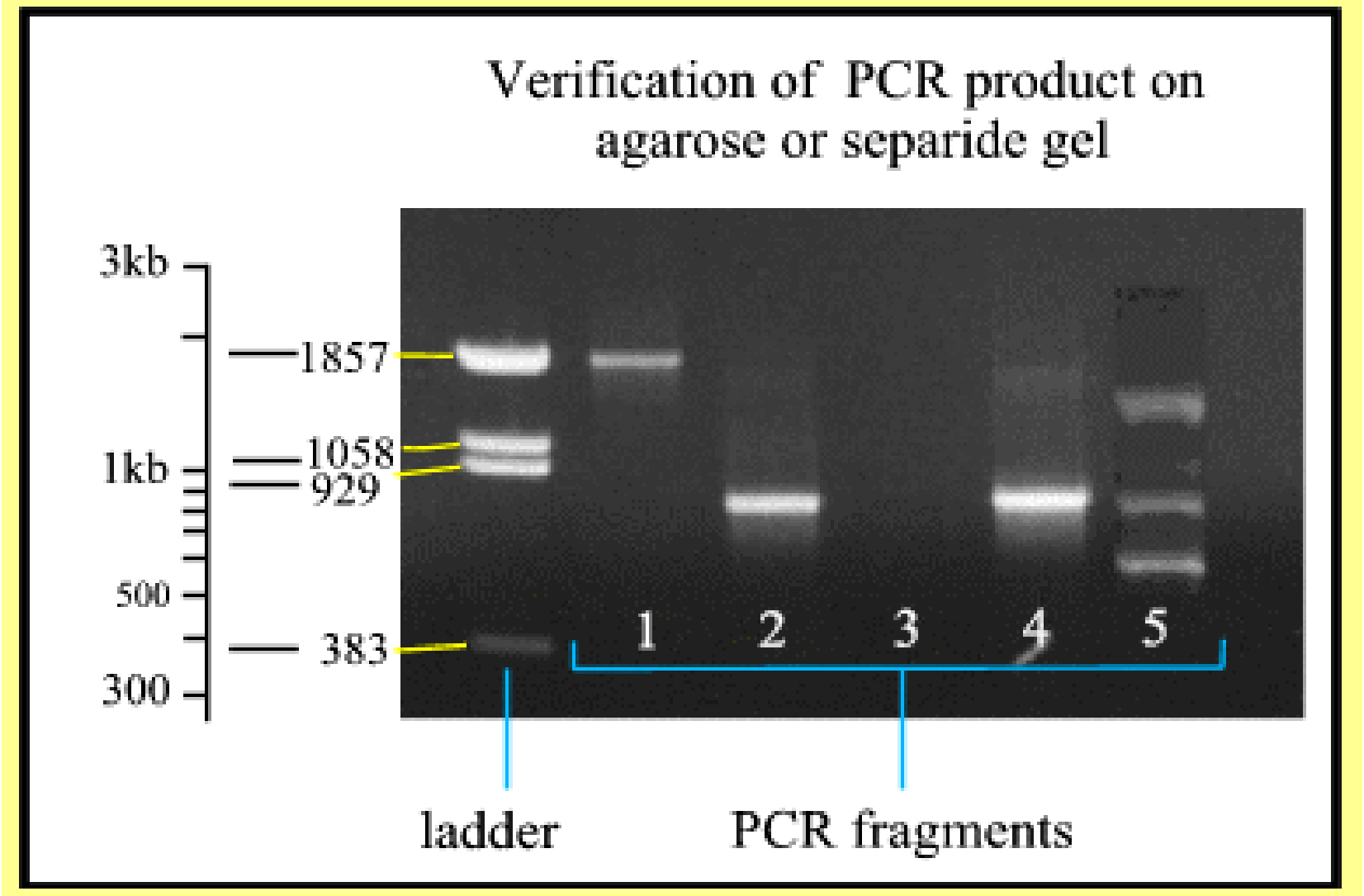
Before the PCR product is used in further applications, it has to be checked if:

1. **A product is formed**
2. **The product is of the right size**
3. **Only one band is formed**

First 4 cycles of a PCR reaction



Verification of the PCR product

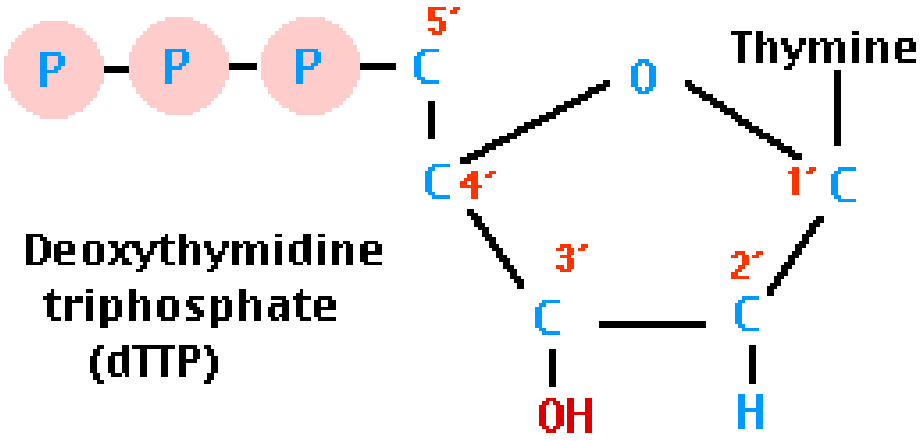


PCR Sequencing

For sequencing, we don't start from gDNA (like in PCR) but mostly from PCR fragments or cloned genes.

Amplified PCR product is supplied with

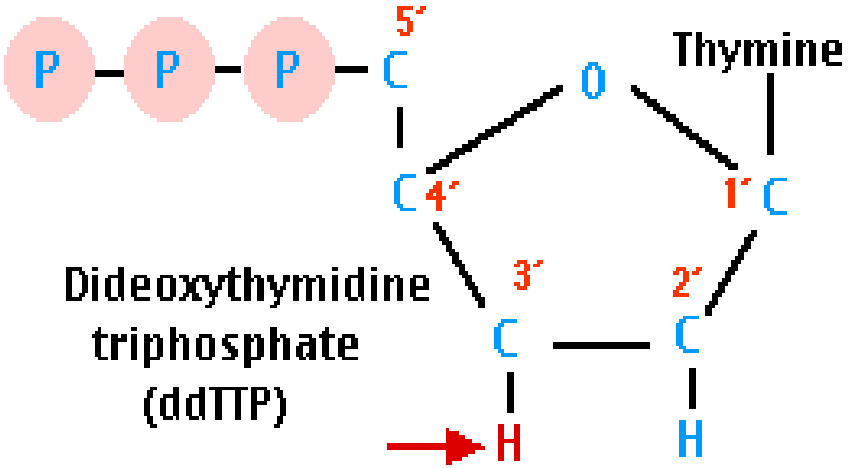
- **a mixture of all four <u>normal</u> (deoxy) nucleotides in ample quantities**
  - **dATP**
  - **dGTP**
  - **dCTP**
  - **dTTP**



- ***Taq* DNA polymerase**

PCR Sequencing

- a mixture of all four dideoxynucleotides, each present in limiting quantities and each labeled with a "tag" that fluoresces a different color:
  - ddATP
  - ddGTP
  - ddCTP
  - ddTTP



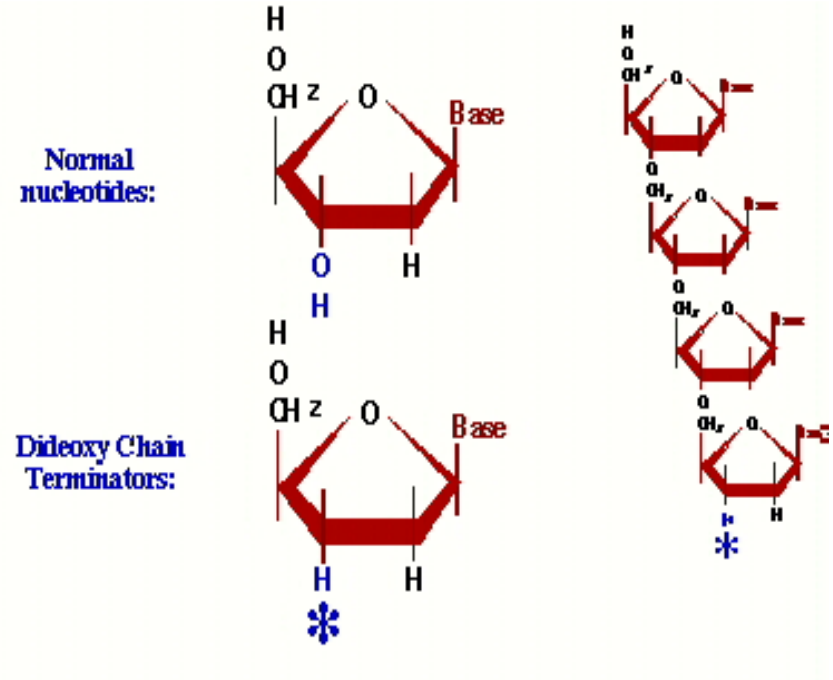
This method of DNA sequencing is called dideoxy method, or chain termination method, or Sanger's method.

PCR Sequencing

Dideoxy method: DNA is synthesized from four deoxynucleotide triphosphates.

Each new nucleotide is added to 3′ -OH group of the last nucleotide added.

When a dideoxynucleotide, ddNTP is added to the growing DNA strand, chain elongation stops because there is no 3′-OH for the next nucleotide to be attached to.



Steps in PCR Sequencing

I The sequencing reaction

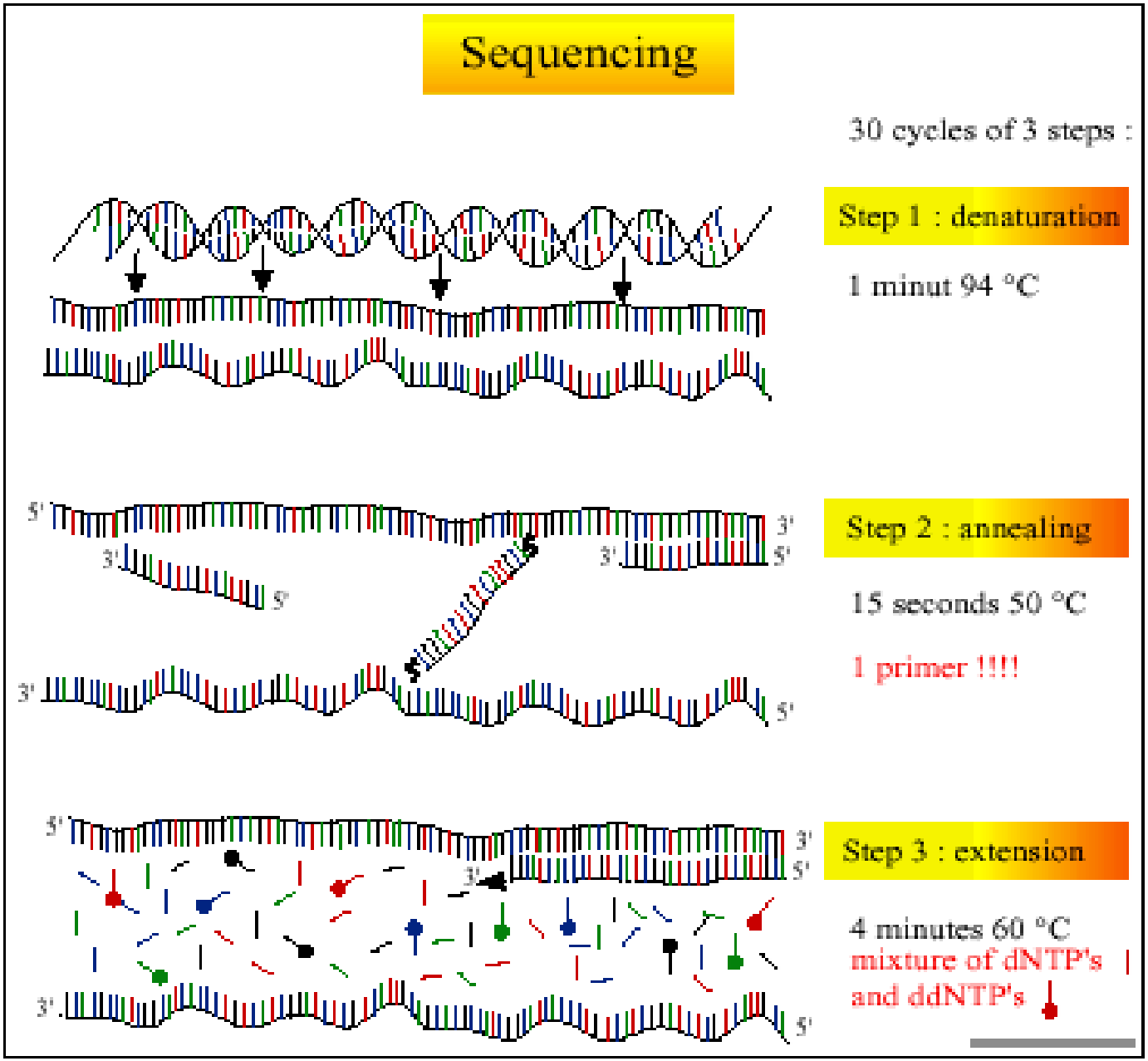
- **Denaturation at 94°C**
- **Annealing at 50°C**
- **Extension at 60°C** ← **instead of 72°C**

II Separation of the fragments

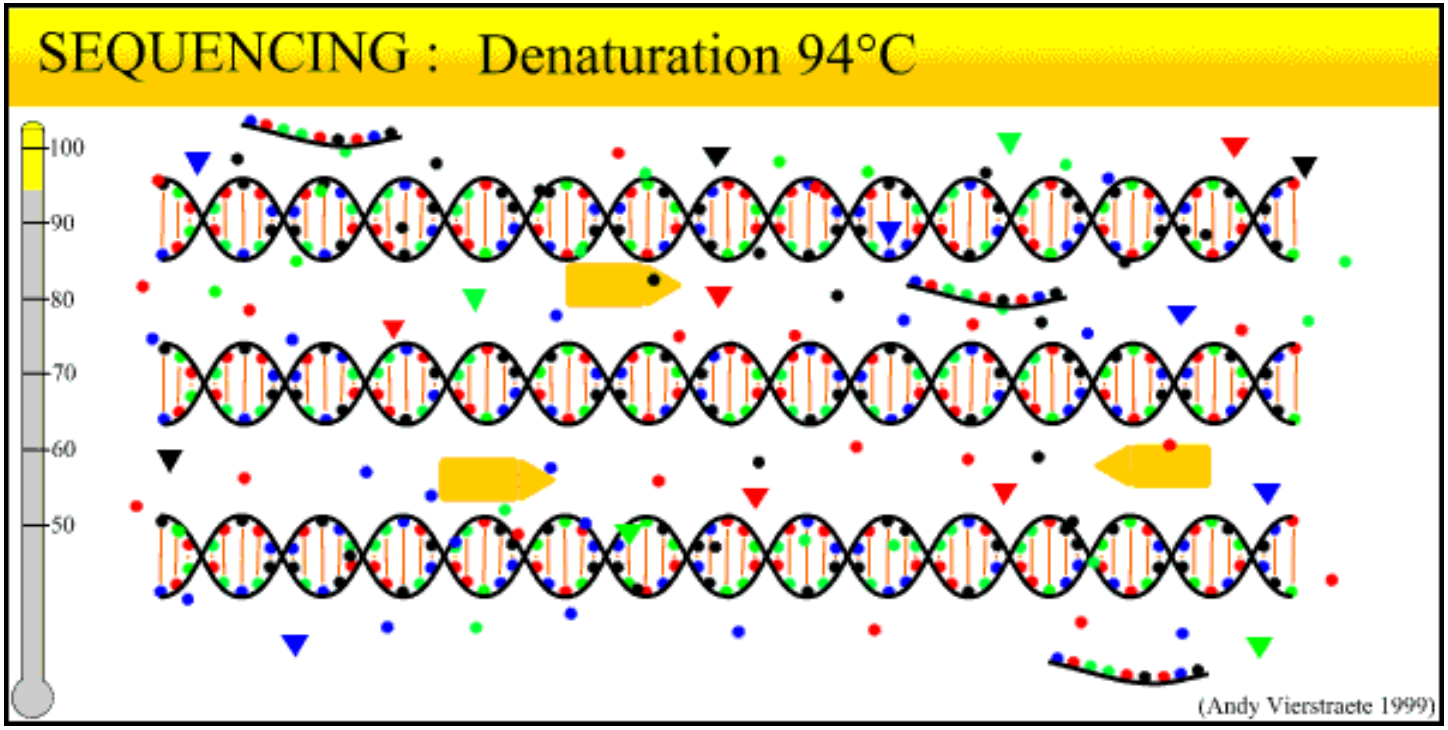
III Detection on an automated sequencer

IV Assembling the sequenced parts

Different steps in Sequencing

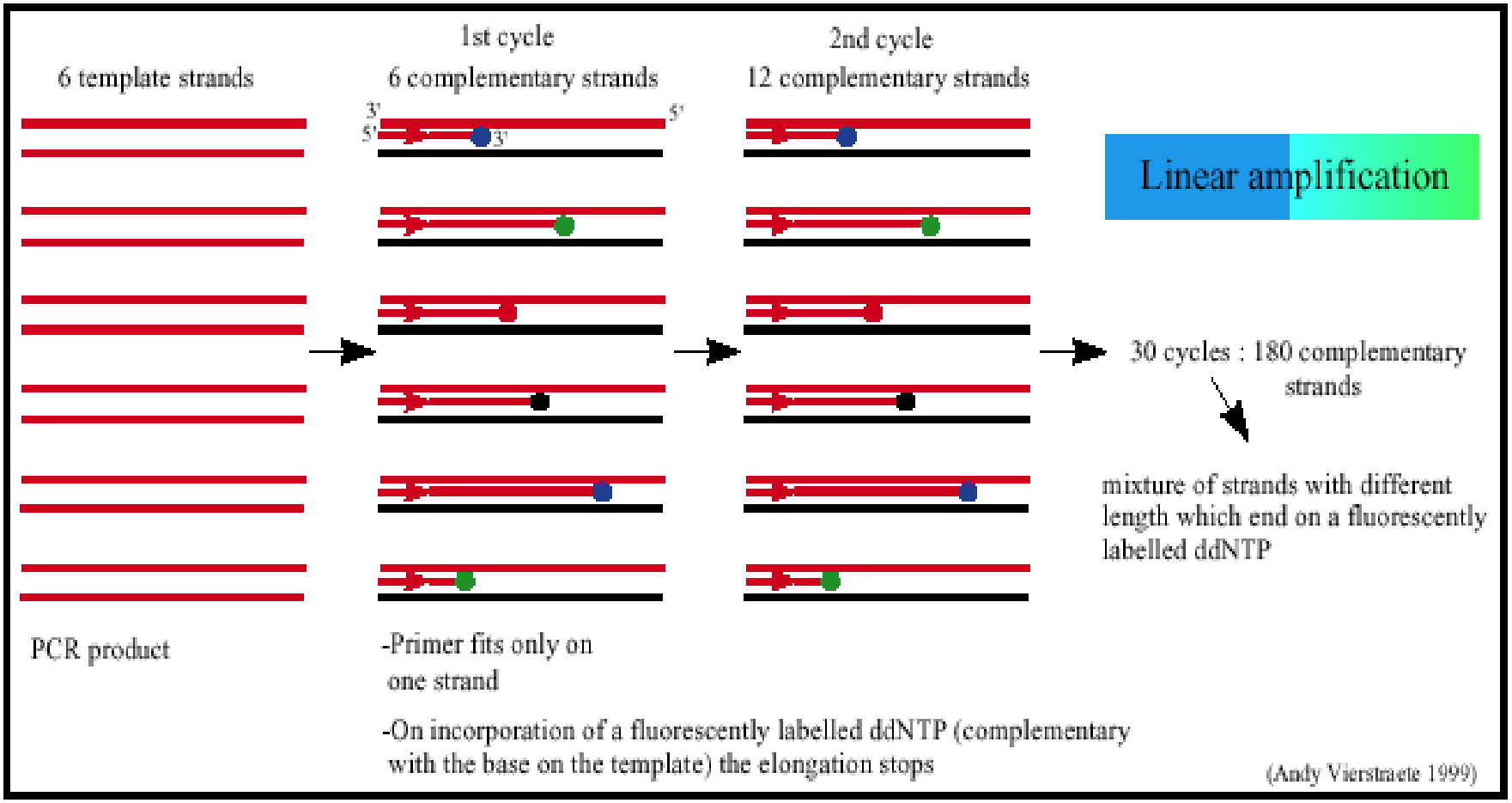


Different steps in Sequencing



PCR Sequencing

Since only one primer is used, only one strand is copied during sequencing – resulting in a linear increase of the number of copies of one strand of the gene. Hence, a large amount of DNA in the starting mixture for sequencing is required.



PCR Sequencing

II Separation of the molecules:

After the sequencing reactions, the mixture of strands of different lengths, all ending on a fluorescently labeled ddNTP, need to be separated

- done by loading the mix on an acrylamide gel - gel electrophoresis.

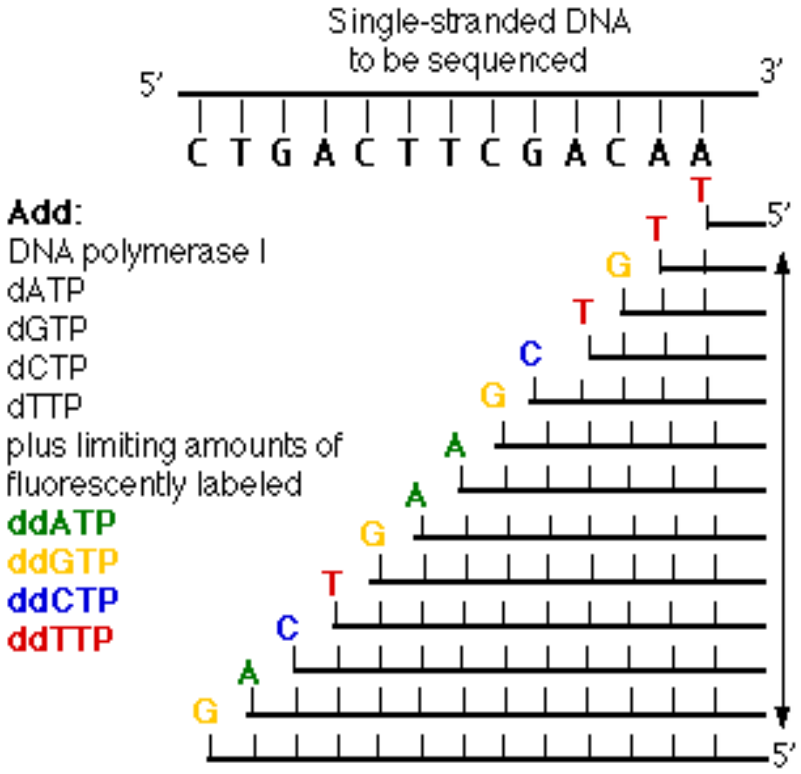
During electrophoresis, a voltage is created across the gel making one end positive and the other negative. DNA being –vely charged, migrates to the positive side.

DNA strands of different length migrate at different rates and thus can be separated based on their size - the smallest strand travels the fastest.

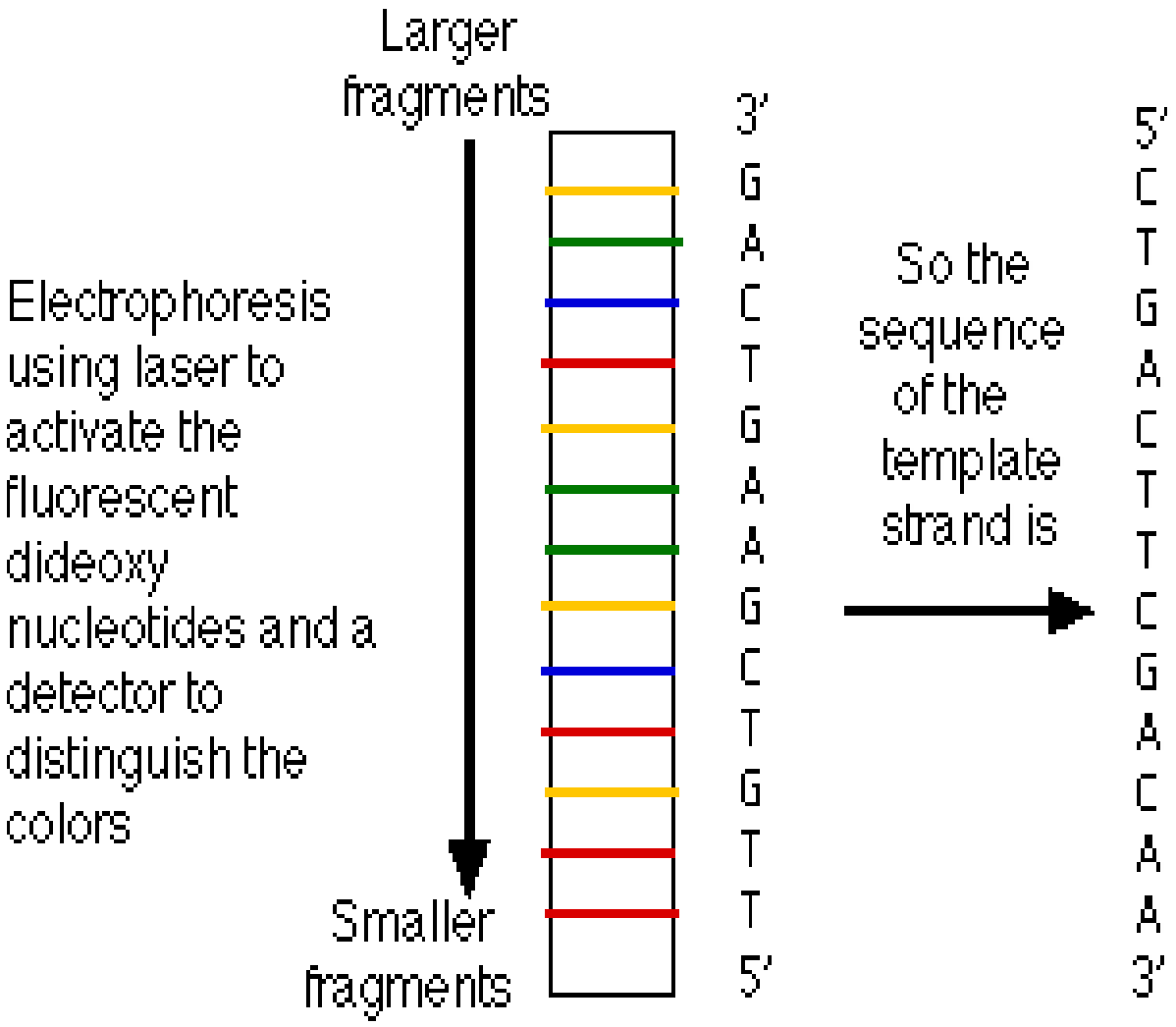
Separation of molecules with electrophoresis

Very good resolution - a difference of even one nucleotide is enough to separate a strand from the next shorter or longer strand.

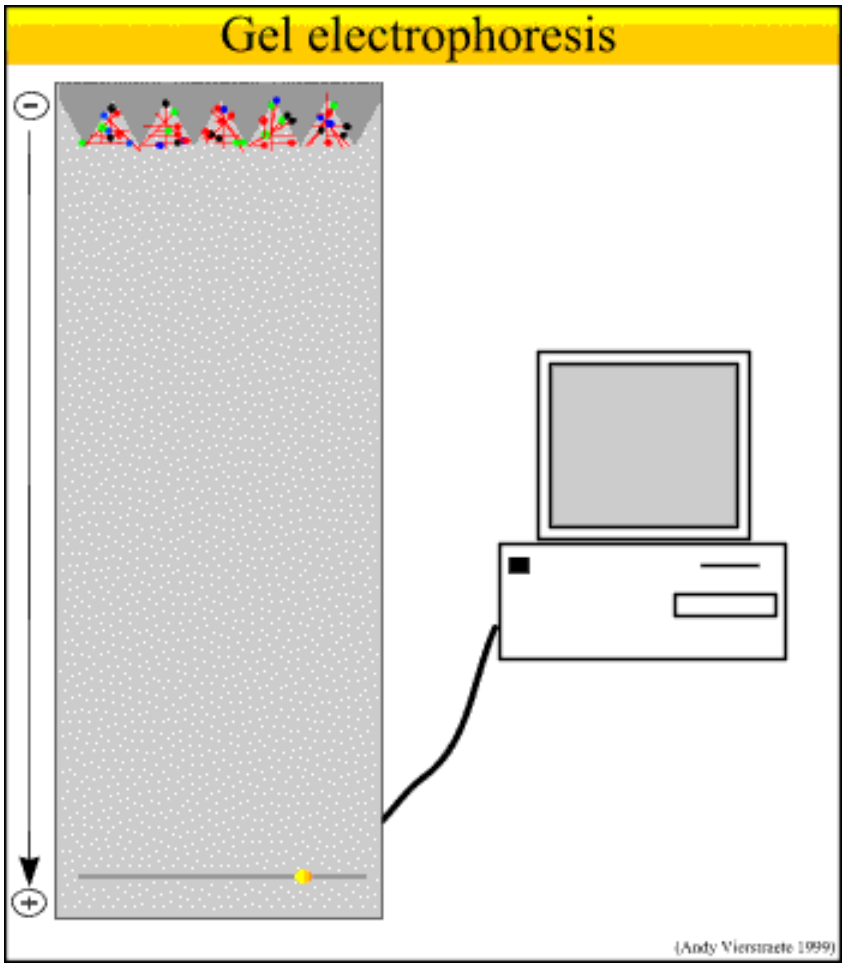
Four dideoxynucleotides fluoresces a different color when illuminated by a laser beam and an automatic scanner provides a printout of the sequence.



Separation of Molecules with Electrophoresis



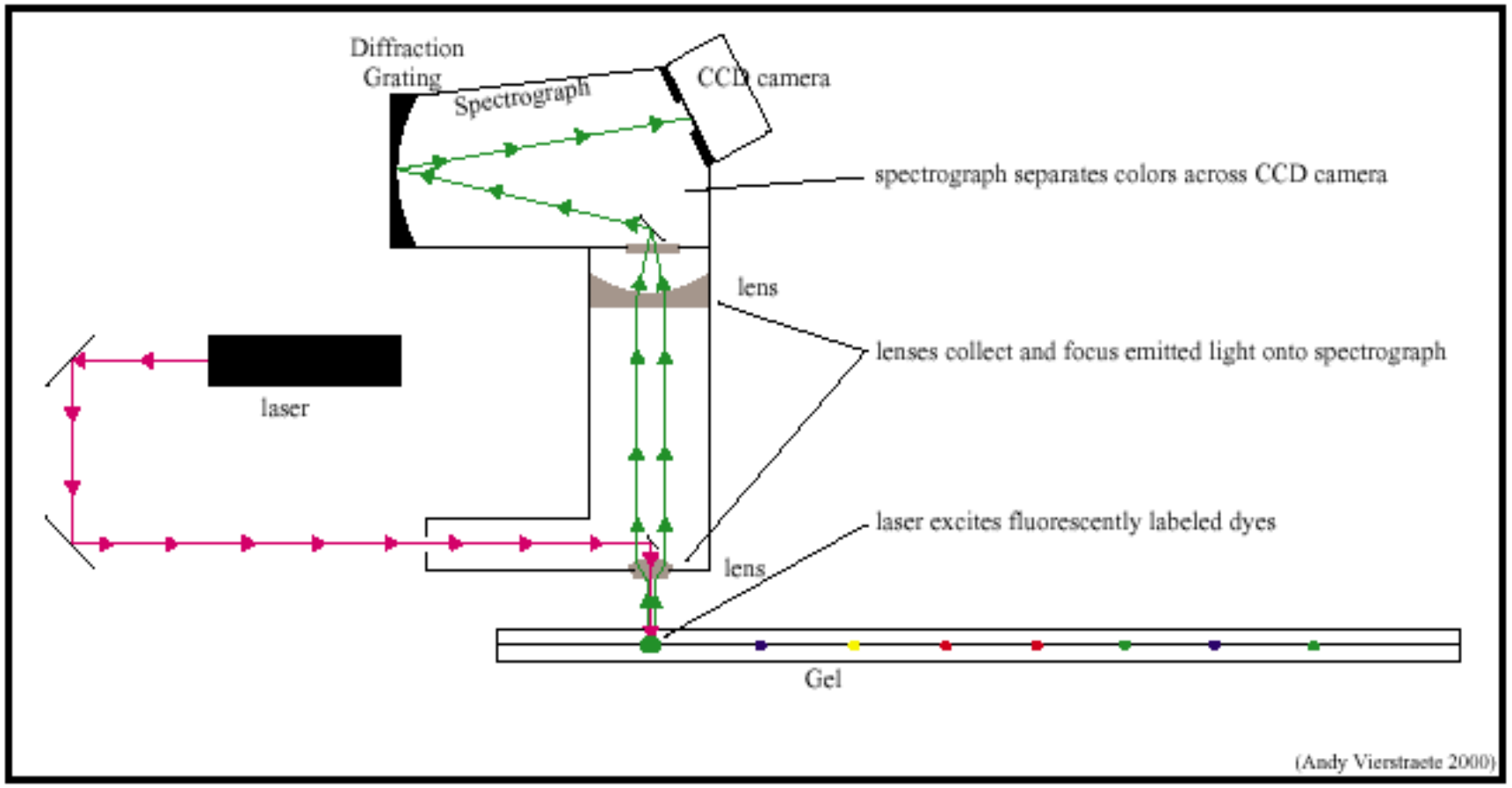
Separation of the Molecules with Electrophoresis



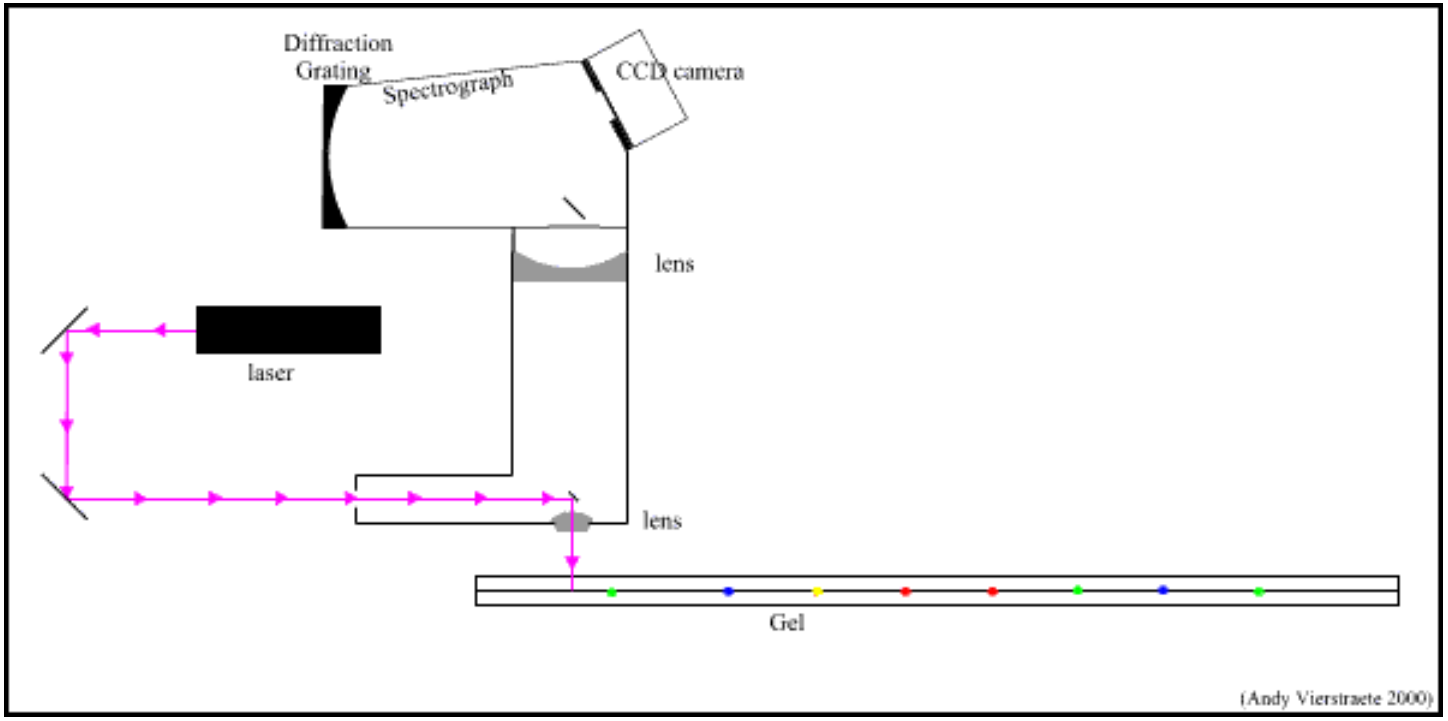
PCR Sequencing

III Detection on an automated sequencer:

Fluorescently labeled fragments that migrate through the gel pass a laser beam at the bottom of the gel.

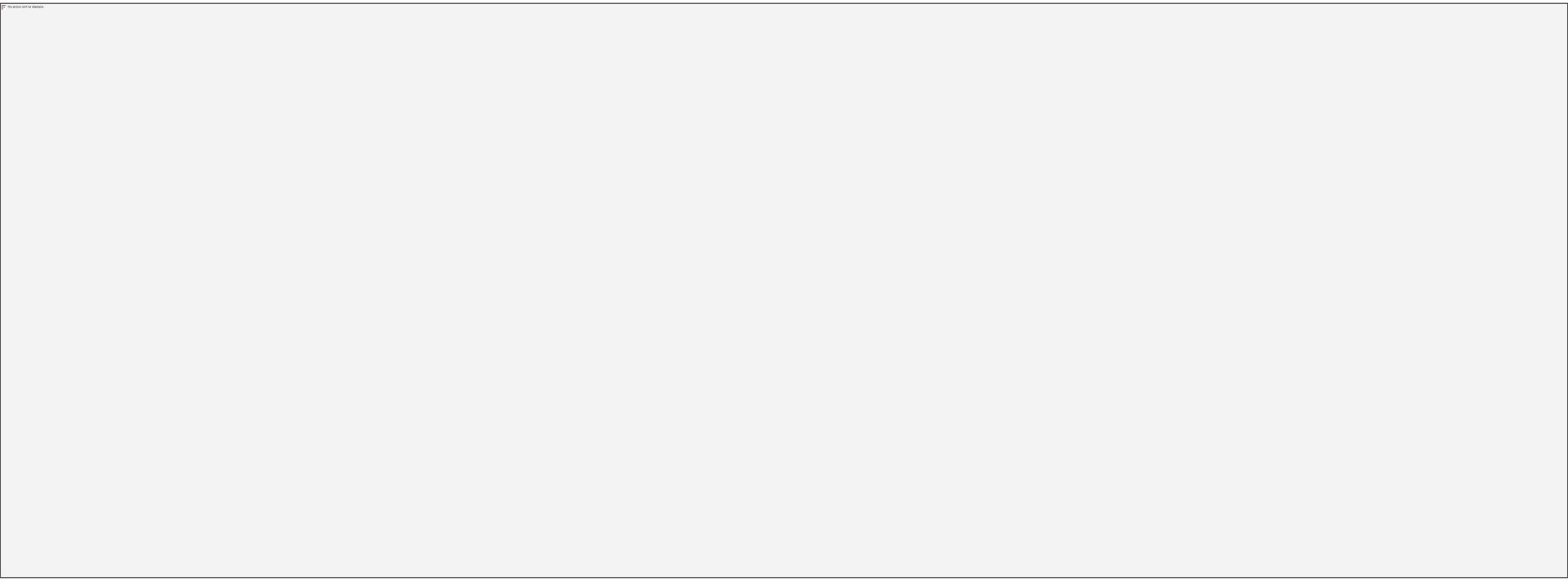


Scanning & Detection System on a Sequencer



PCR Sequencing

Plot of the colors detected in a 'lane' of the gel (one sample), scanned from smallest fragments to largest.



The computer interprets the colors by printing the nucleotide sequence across the top of the plot.

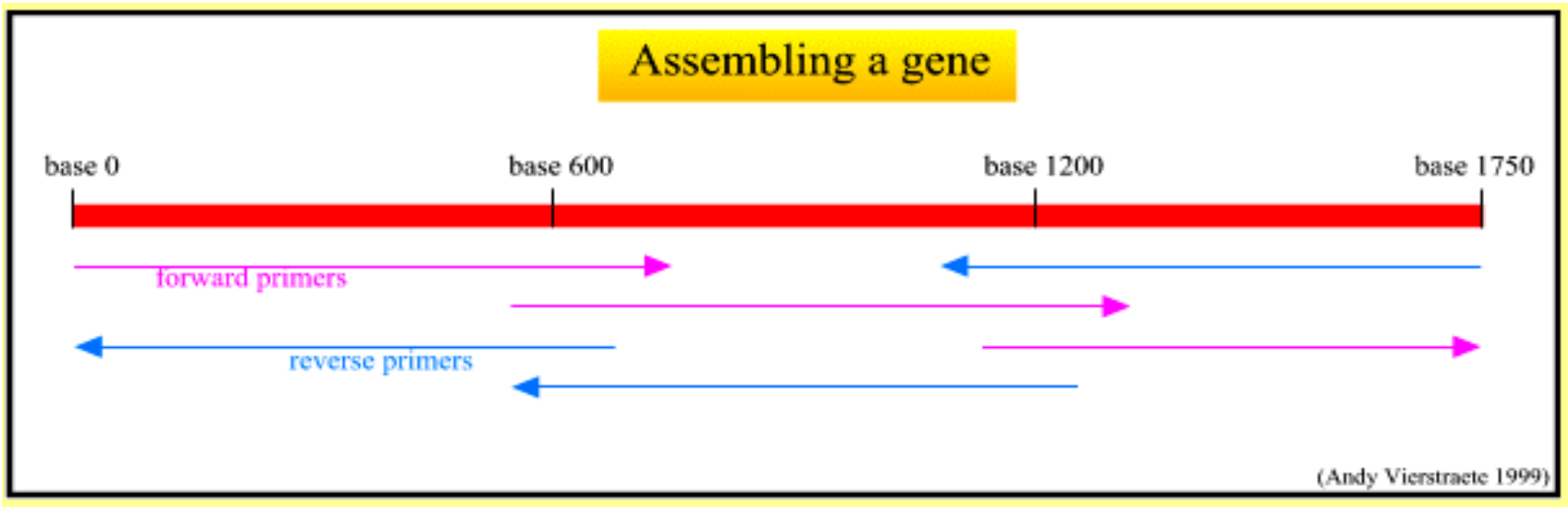
PCR Sequencing

IV Assembling the sequenced parts of a gene:

For publication, a gene sequence has to be confirmed in both directions using forward & reverse primers

Since it is only possible to sequence ~ 700-800 bases in one run, a gene of, say, 1800 bases, has to be sequenced with internal primers.

- the sequenced fragments are assembled using a computer program to obtain complete gene sequence.



PCR Sequencing

How would you go about sequencing SARS-CoV-2 genome, 29903 bases long?

- **How is the SARS-CoV-2 genome sequenced?**

- By PCR seq'g, Next-generation seq'g

- **How does one identify the coordinates of N gene on it? i.e., how to construct a physical map of a genome?**

- By constructing physical map of the genome using restriction recognition sites

- **How does one select which regions in this gene would give specificity for the presence of SARS-CoV-2?***

- By sequence comparison – we will discuss this later

- **How is the specific probe regions extracted and amplified for detection?**

- By making site-specific cuts in the region of interest

- **Is it possible to store the DNA sample for re-testing? How?**

- By cloning

References:

1. **Concepts in Biotechnology, ed. D. Balasubramanyam**
2. **Restriction Endonucleases and DNA Modifying Enzymes http://arbl.cvmbs.colostate.edu/hbooks/genetics/biotech/enzymes/index.html**
3. **REBASE: restriction enzymes and methyltransferases, Nucleic Acids Research, Vol. 31 (1), 418–420 (2003)**