Gene Technology
&
DNA Sequencing

Lecture-3

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.

To sequence a gene, we need to

- Identify the region of interest
- Isolate it from the organism – DNA fragmentation
- move 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

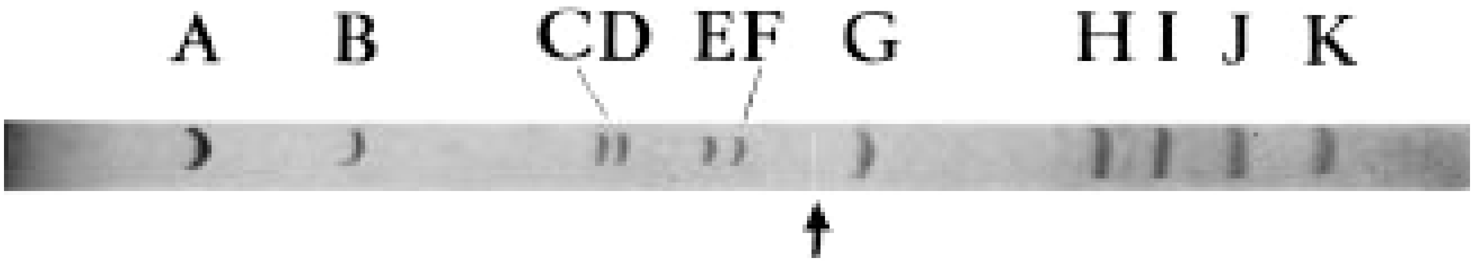
Ability to cut DNA into discrete fragments allows to understand

- **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

Radioautogram of ^{14}C-labeled SV40 DNA cleaved with endonuclease R



SV40 DNA (a tumor virus) - after cutting, or "digesting" it with *H. influenzae* restriction enzyme, analyzed the pieces using a polyacrylamide gel electrophoresis.

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

Background

How were these restriction endonucleases identified?

Bacteria are under constant attack by viruses, e.g., bacteriophages

To protect themselves, bacteria have developed a method to chop up any foreign DNA, by an enzyme, called endonuclease

- it circulates in the bacterial cytoplasm, waiting for any attacking virus.

- also called restriction enzymes because they restrict the infection of bacteriophages.

Why do the restriction enzymes 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 (attached to the cytosine in dinucleotide CG) 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

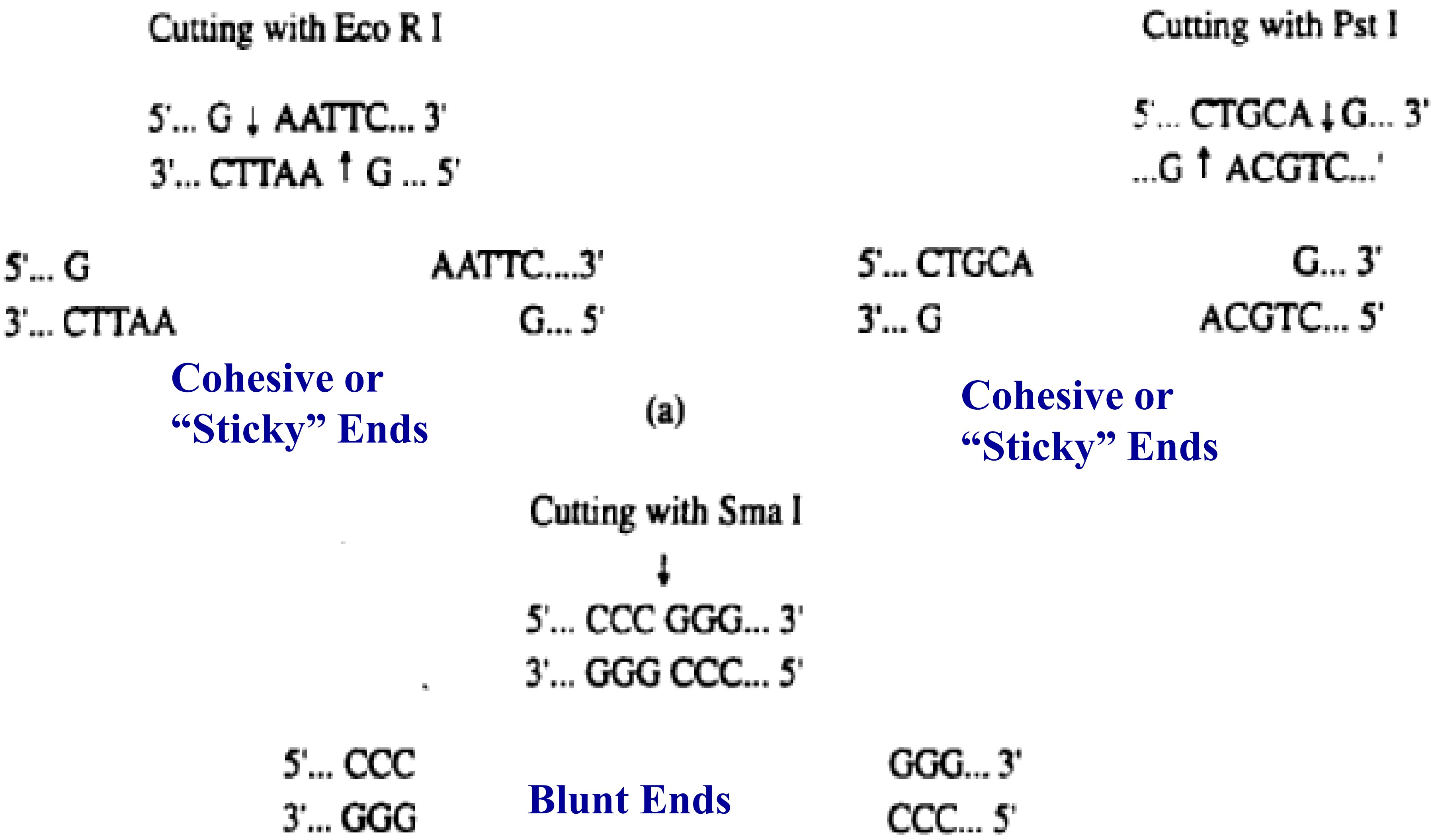
Different endonucleases present in different bacteria recognize different nucleotide sequences

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.

Generation of Cohesive & Blunt-ended Fragments



Cutting with Eco R I

5'... G ↓ AATTC... 3'
3'... CTTAA ↑ G ... 5'

5'... G AATTC....3'
3'... CTTAA G... 5'

Cohesive or "Sticky" Ends

(a)

Cutting with Pst I

5'... CTGCA ↓ G... 3'
...G ↑ ACGTC...'

5'... CTGCA G... 3'
3'... G ACGTC... 5'

Cohesive or "Sticky" Ends

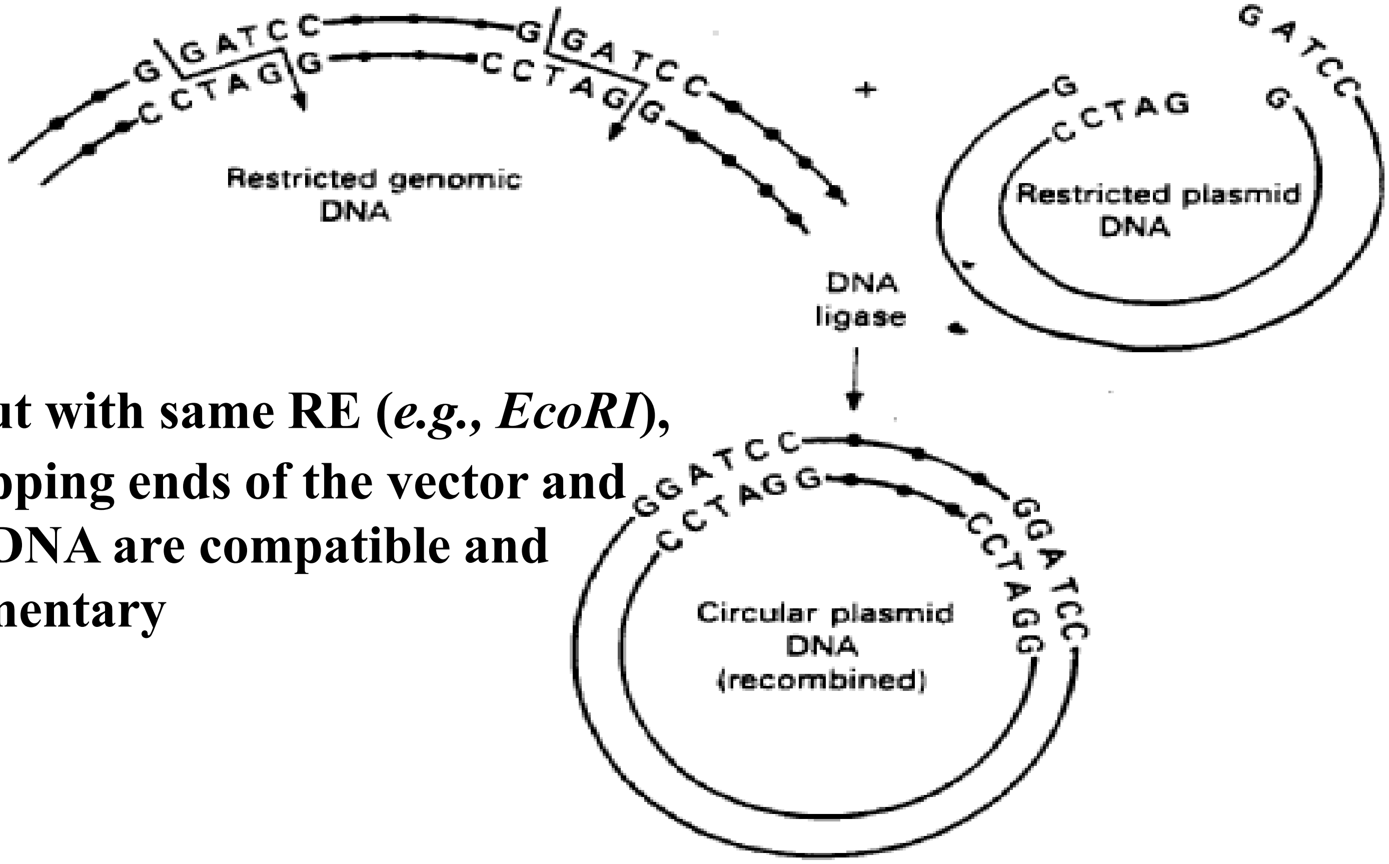
Cutting with Sma I

↓
5'... CCC GGG... 3'
3'... GGG CCC... 5'

5'... CCC GGG... 3'
3'... GGG CCC... 5'

Blunt Ends

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

A plasmid is a small, circular, double-stranded DNA molecule that is distinct from a cell's chromosomal DNA.

Features of Restriction Enzymes

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

  Frequency of recognition sites of length, 4, 6, or 8?

- 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., *Hinf* I recognition site: GANTC – it's frequency of occurrence?
- 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

- To prepare a physical map of the genome
- In genetic engineering - 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

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

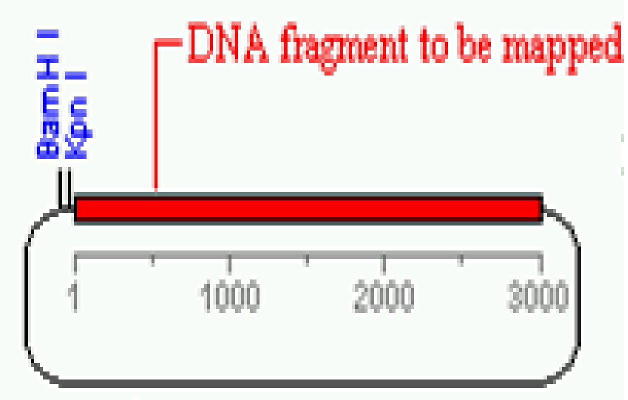
Multiple Complete Digest Mapping – creates a map by digesting DNA with multiple REs

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

DNA to be restriction mapped is usually contained within a well-characterized plasmid or bacteriophage vector for which the sequence is known, which facilitates making the map.

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:

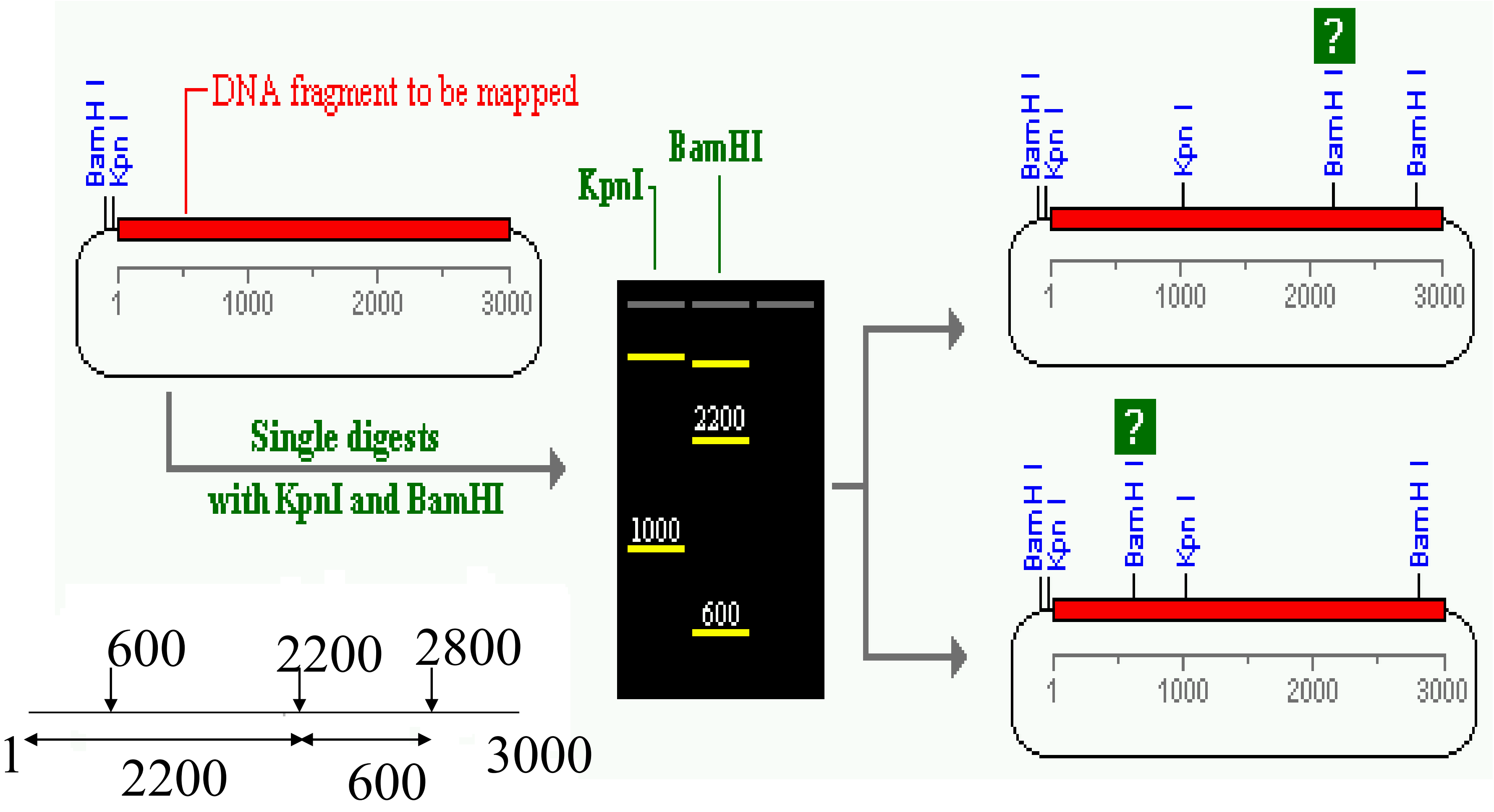
Kpn I yields 2 fragments: 1000bp & "big"

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

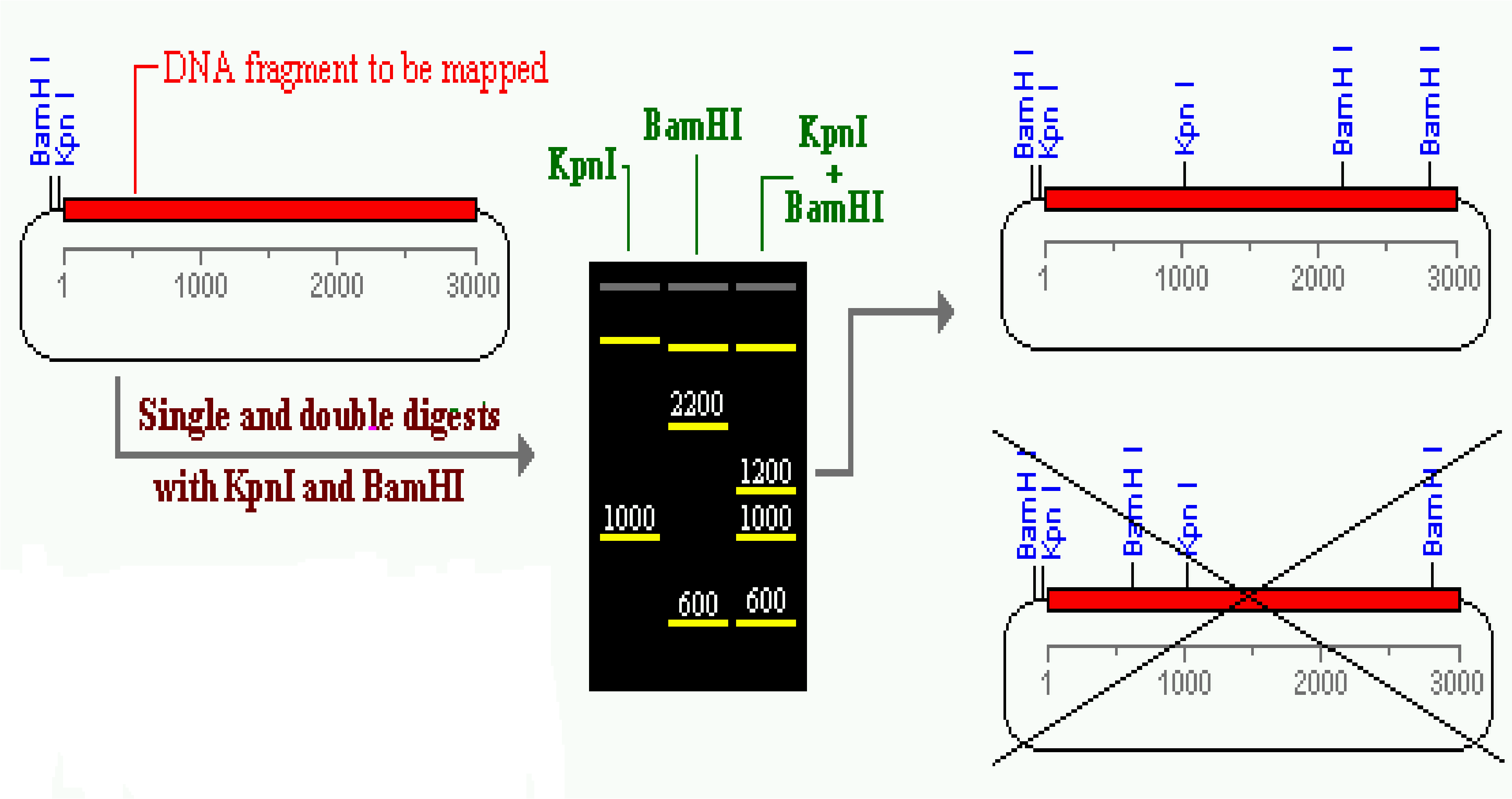
Restriction Mapping



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.

Can we computationally generate a restriction map of DNA sequence?

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/

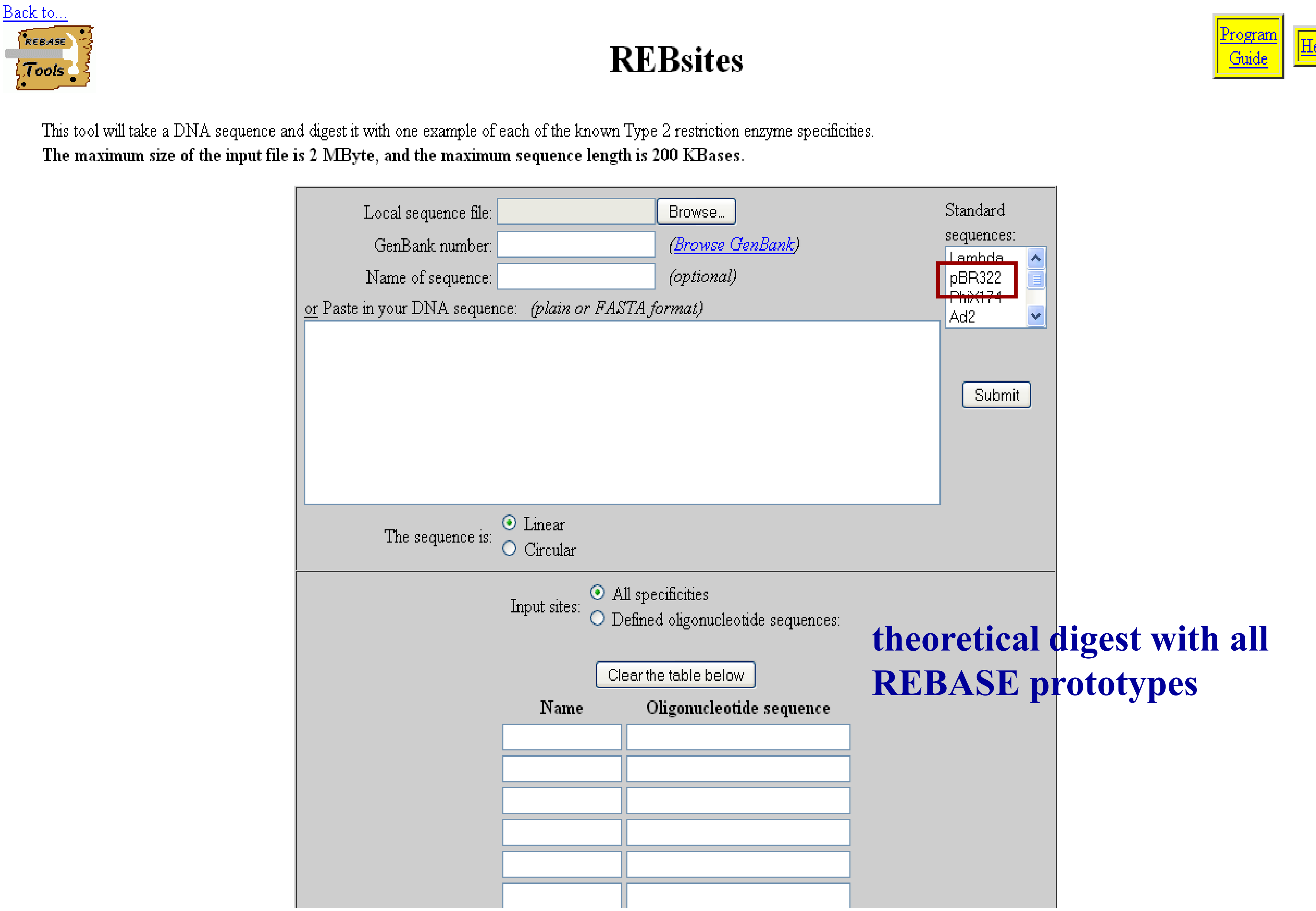
- **Webcutter**

  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



Back to...

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.

Local sequence file: Browse...

GenBank number: (*Browse GenBank*)

Name of sequence: (*optional*)

or Paste in your DNA sequence: (*plain or FASTA format*)

Standard sequences:
Lambda
pBR322
PhiX174
Ad2

Submit

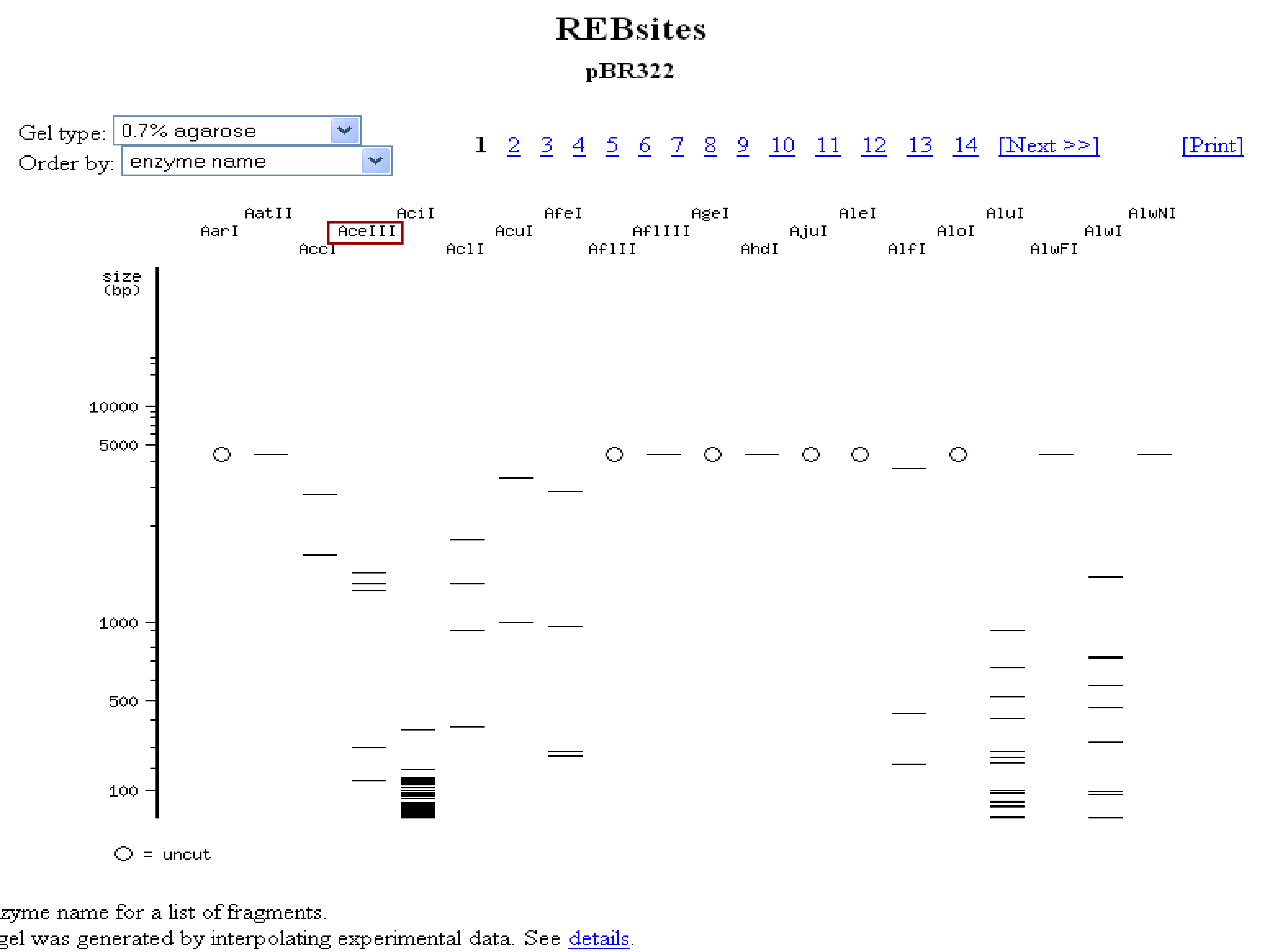
The sequence is: ● Linear ○ Circular

Input sites: ● All specificities ○ Defined oligonucleotide sequences:

Clear the table below

Name	Oligonucleotide sequence

theoretical digest with all REBASE prototypes



REBsites

pBR322

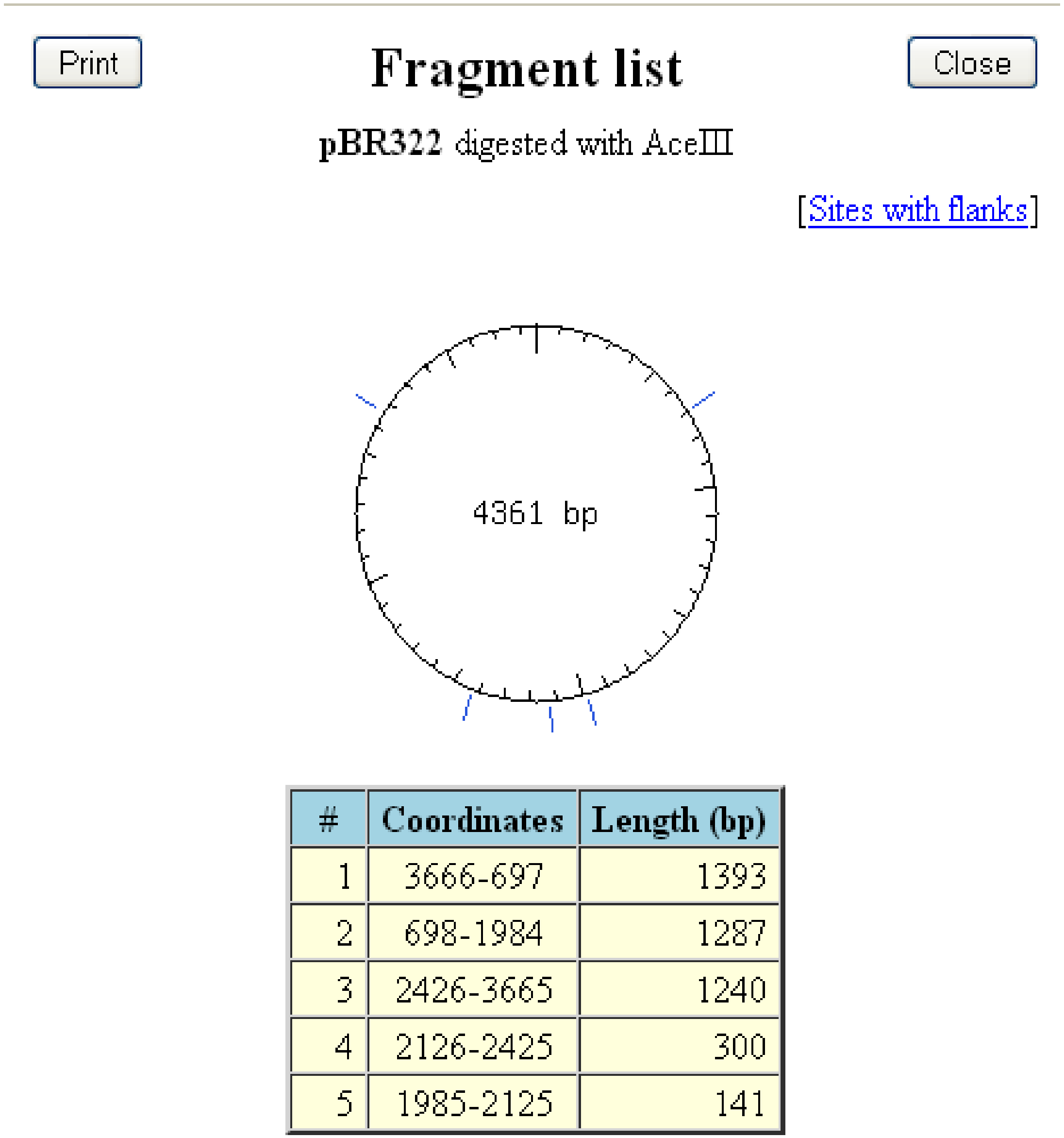
Gel type: 0.7% agarose

Order by: enzyme name

1 2 3 4 5 6 7 8 9 10 11 12 13 14 [Next >>] [Print]

zyme name for a list of fragments.

gel was generated by interpolating experimental data. See details.

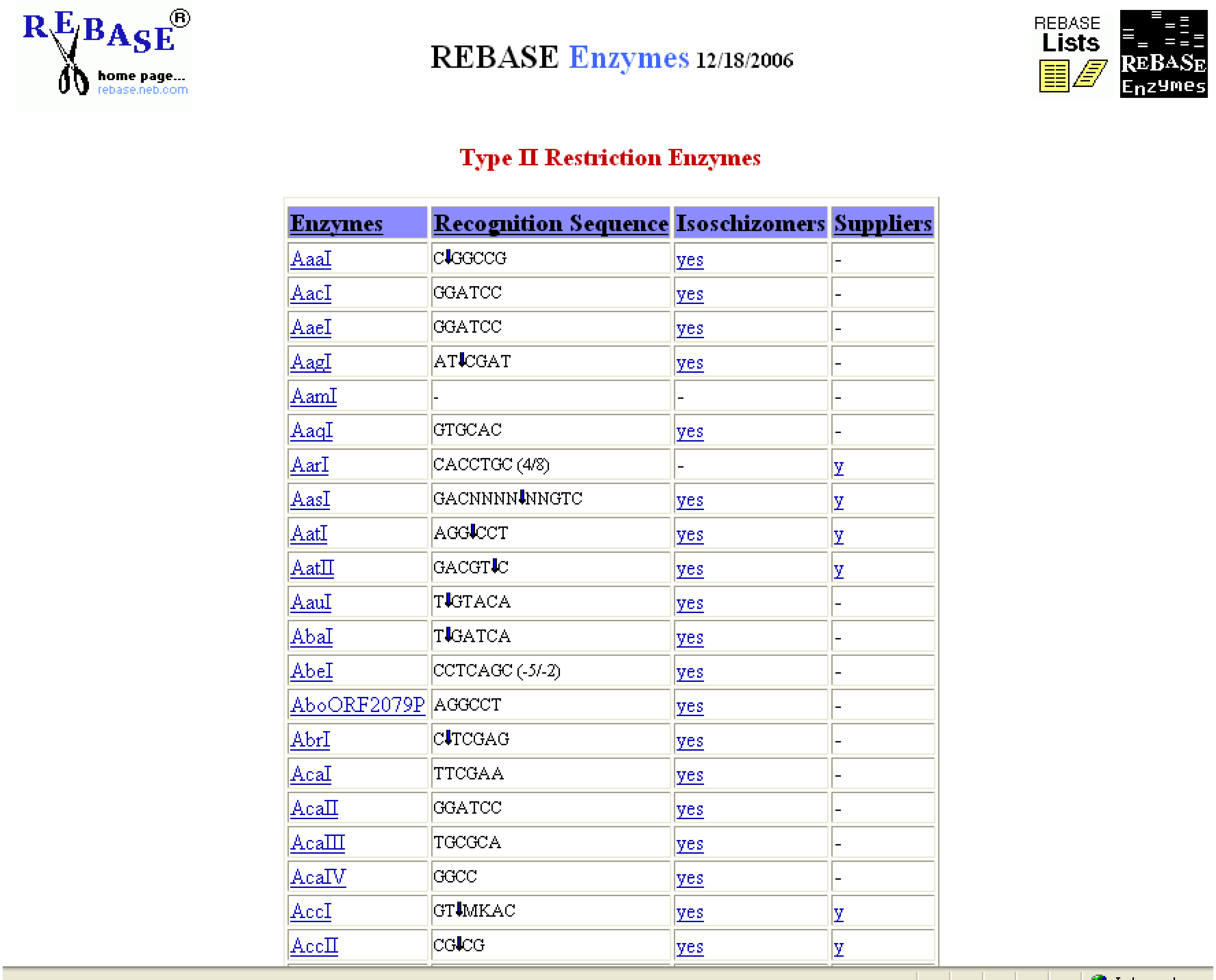


Fragment list

pBR322 digested with AceIII

[Sites with flanks]

#	Coordinates	Length (bp)
1	3666-697	1393
2	698-1984	1287
3	2426-3665	1240
4	2126-2425	300
5	1985-2125	141



REBASE Enzymes 12/18/2006

Type II Restriction Enzymes

Enzymes	Recognition Sequence	Isoschizomers	Suppliers
AaaI	C↓GGCCG	yes	-
AacI	GGATCC	yes	-
AaeI	GGATCC	yes	-
AagI	AT↓CGAT	yes	-
AamI	-	-	-
AaqI	GTGCAC	yes	-
AarI	CACCTGC (4/8)	-	y
AasI	GACNNNN↓NNGTC	yes	y
AatI	AGG↓CCT	yes	y
AatII	GACGT↓C	yes	y
AauI	T↓GTACA	yes	-
AbaI	T↓GATCA	yes	-
AbeI	CCTCAGC (-5/-2)	yes	-
AboORF2079P	AGGCCT	yes	-
AbrI	C↓TCGAG	yes	-
AcaI	TTCGAA	yes	-
AcaII	GGATCC	yes	-
AcaIII	TGCGCA	yes	-
AcaIV	GGCC	yes	-
AccI	GT↓MKAC	yes	y
AccII	CG↓CG	yes	y

REBASE® home page... rebase.neb.com

AaaI

Type II restriction enzyme
subtype: P

Different enzyme:

Go Cle

Recognition Sequence:

C^GGCCG

```
        ↓
5' ..  C  G  G  C  C  G  3' ..
3' ..  G  C  C  G  G  C  5' ..
                        ↑
```

REBASE Enz Num 1 entered Jan 1 1987 ... modified May 24 2004

Prototype: XmaIII
Org #: 1
Organism: *Acetobacter aceti ss aceti*
Organism source: M. Fukaya
Growth Temperature: 26 °

# sites on	
Adeno2:	19
Lambda:	2
pBR322:	1
PhiX174:	0
SV40:	0

Site frequency in sequenced genomes...

Related References...
sorted by date in new win
sorted by authors in new w

NOT commercially avail

Similar enzymes...

REBASE Enzymes

REBASE Lists

12/18/2006
Dr. Richard J. Roberts
and Dana Macelis

HELP REBASE ?

REBASE Tools

Assignment

- **Write a program to generate a restriction map for a specific RE and compare your results with Mapper.**
- **Write a program to identify restriction recognition sites in a given DNA sequence.**
- **Write to program to obtain the reverse strand in the forward strand of a DNA sequence in given**

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

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:

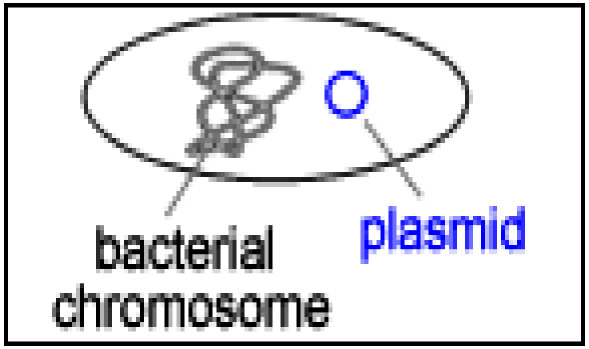
(i) isolate them in large quantities,

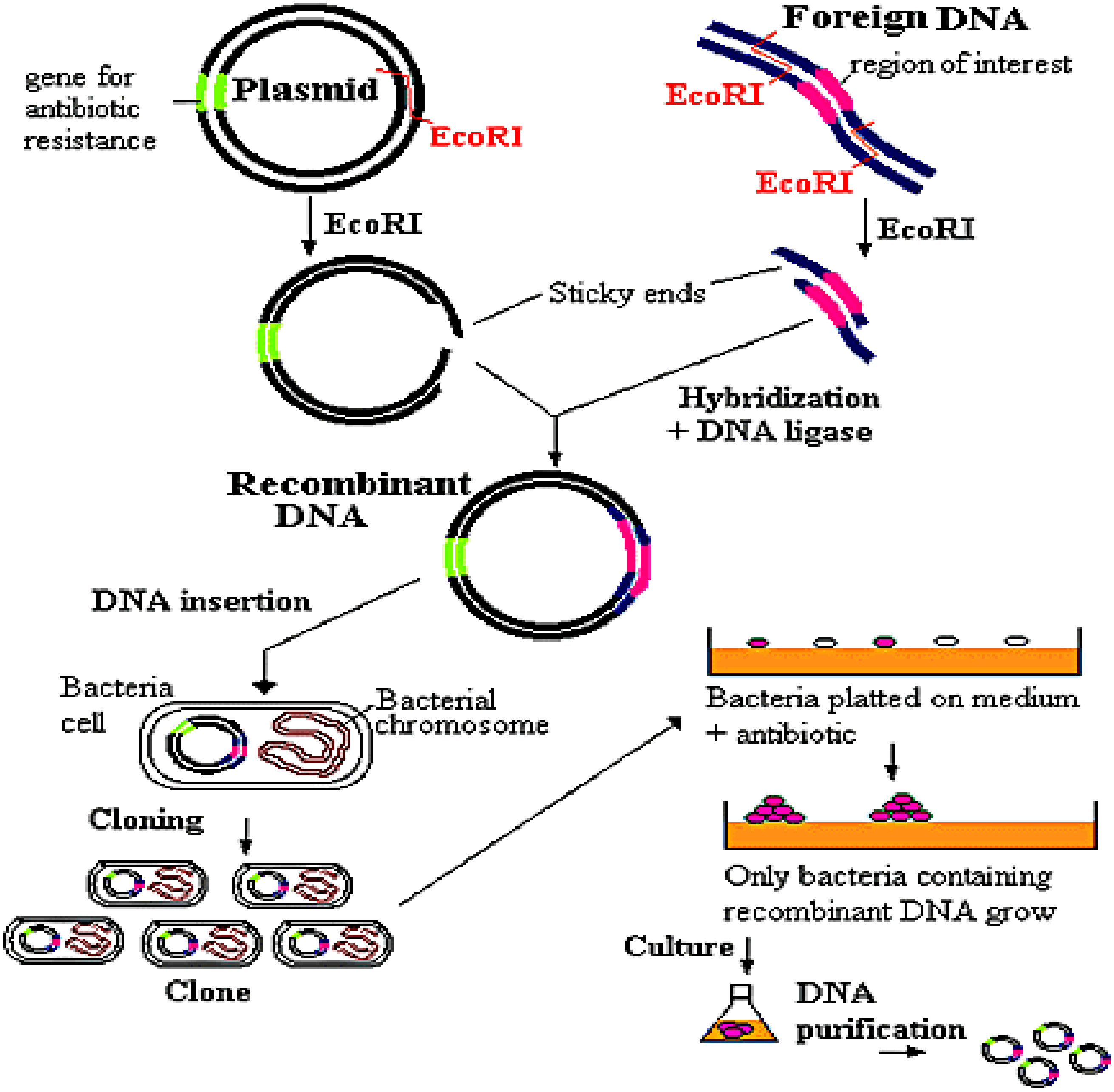
(ii) cut & splice them, add DNA of choice,

(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)





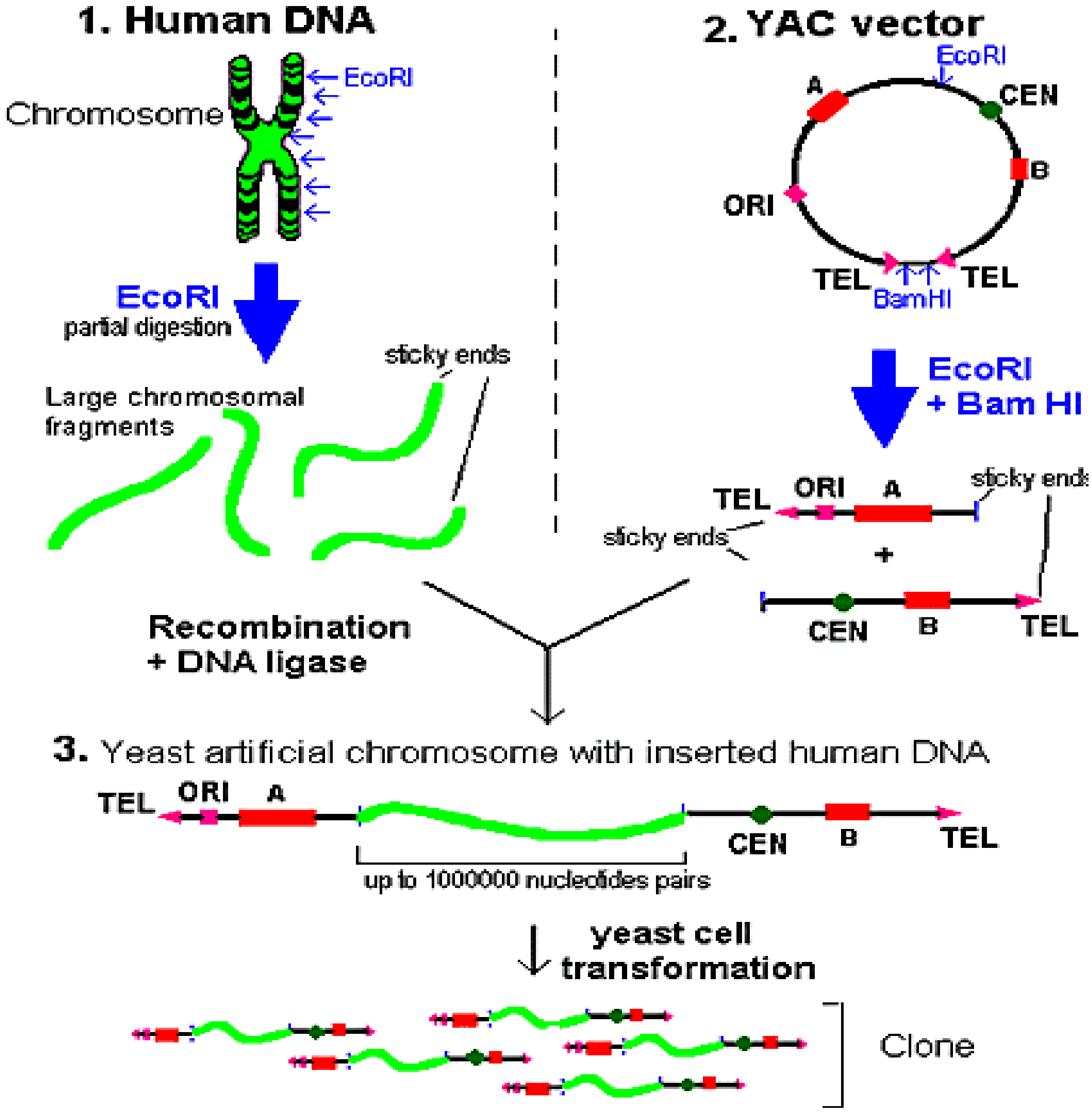
Cloning into a plasmid

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)

DNA Sequencing

DNA Sequencing - determine the precise sequence of nucleotides in a sample of DNA – the order of A, T, G, C

Various types of sequencing:

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

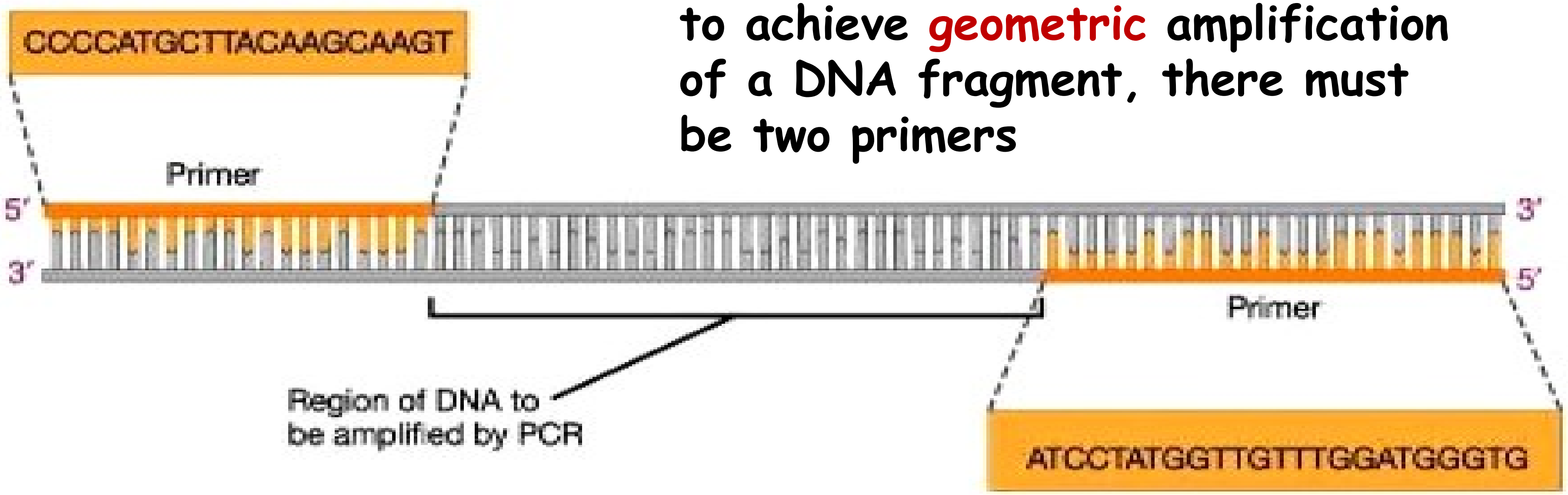
Essential requirements for a PCR:

- 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.

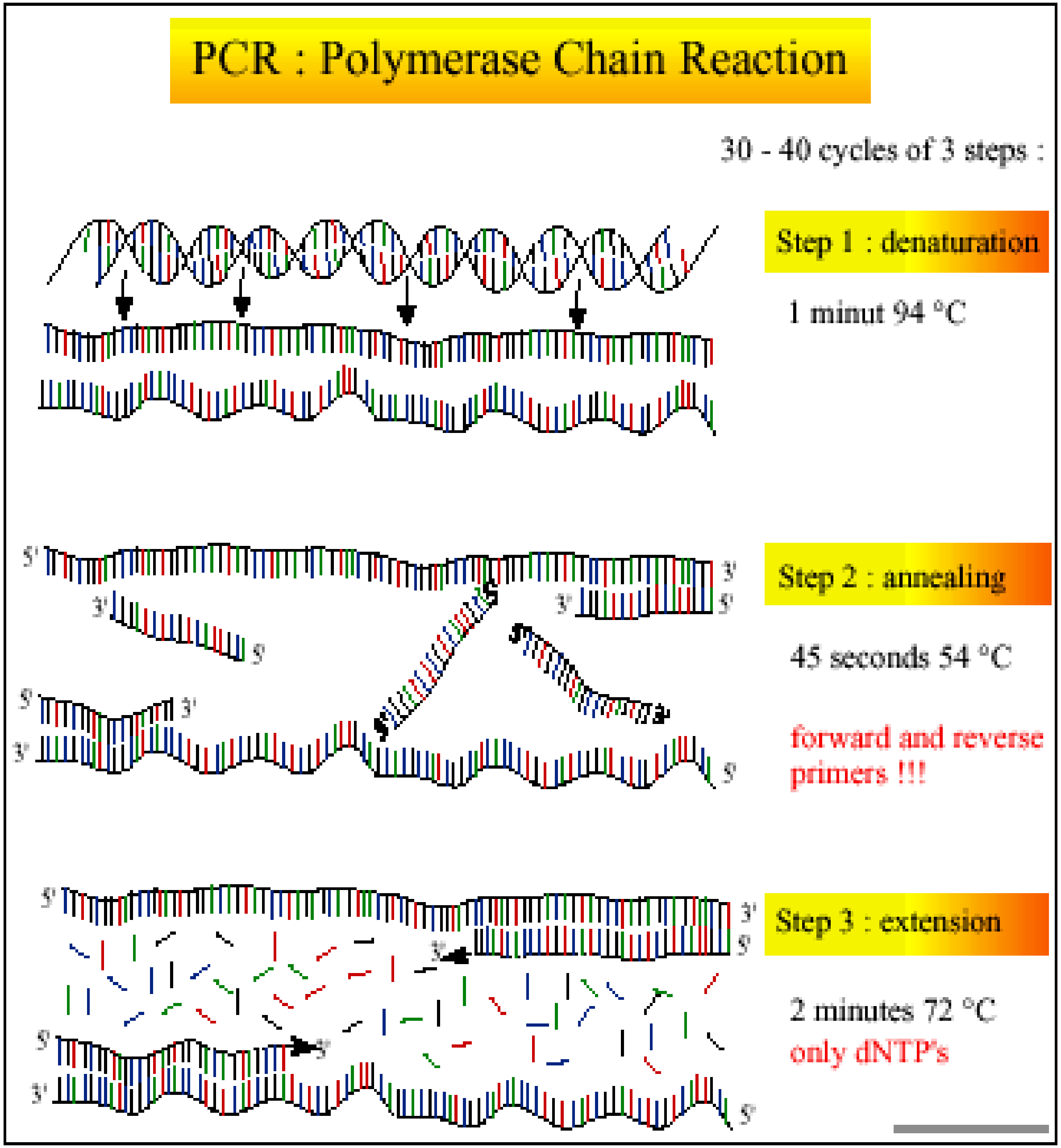
Step-2: Annealing at 54°C

- 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

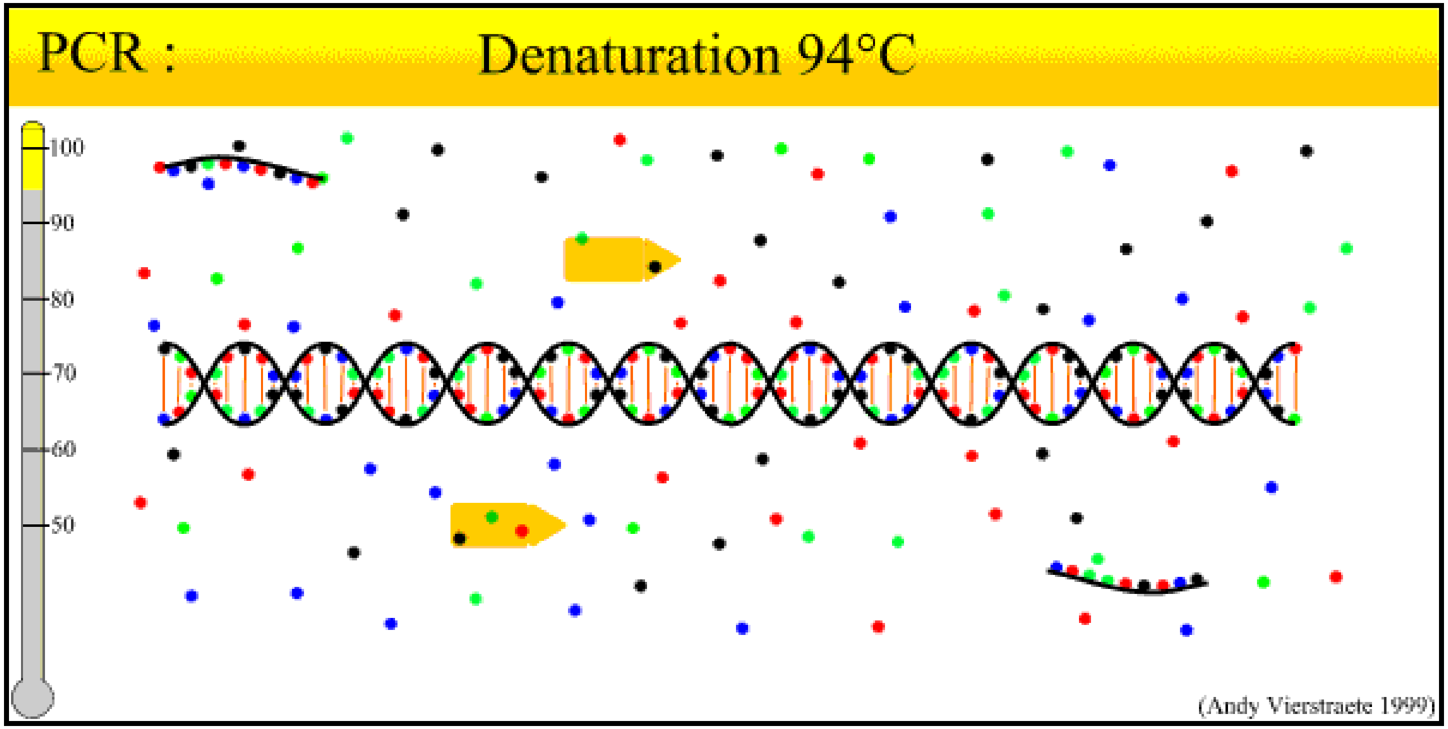
forward and reverse primers !!!

Step 3 : extension

2 minutes 72 °C

only dNTP's

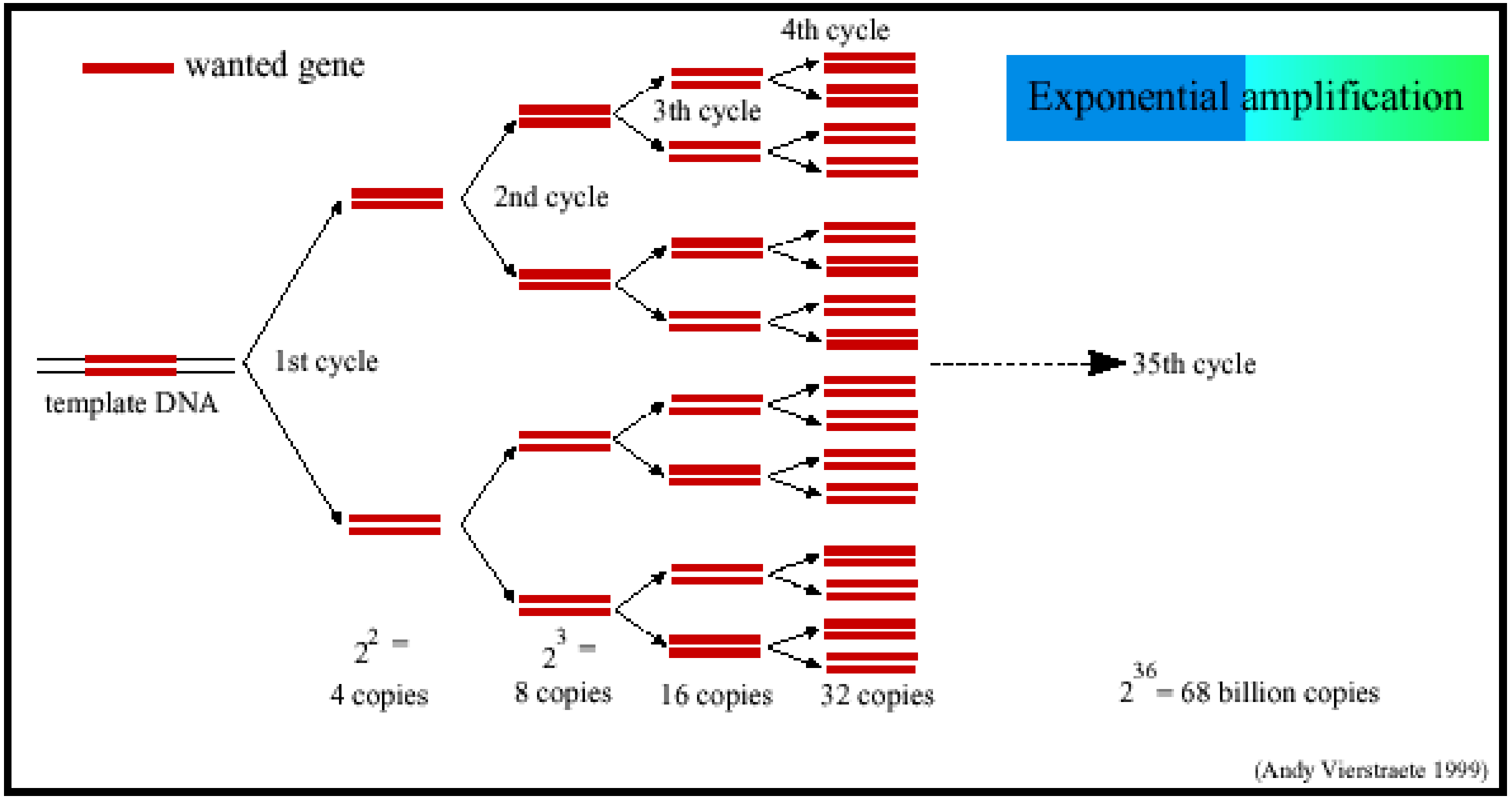
Different Steps in PCR



Exponential amplification of region of interest

Both strands are copied during PCR

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



(Andy Vierstraete 1999)

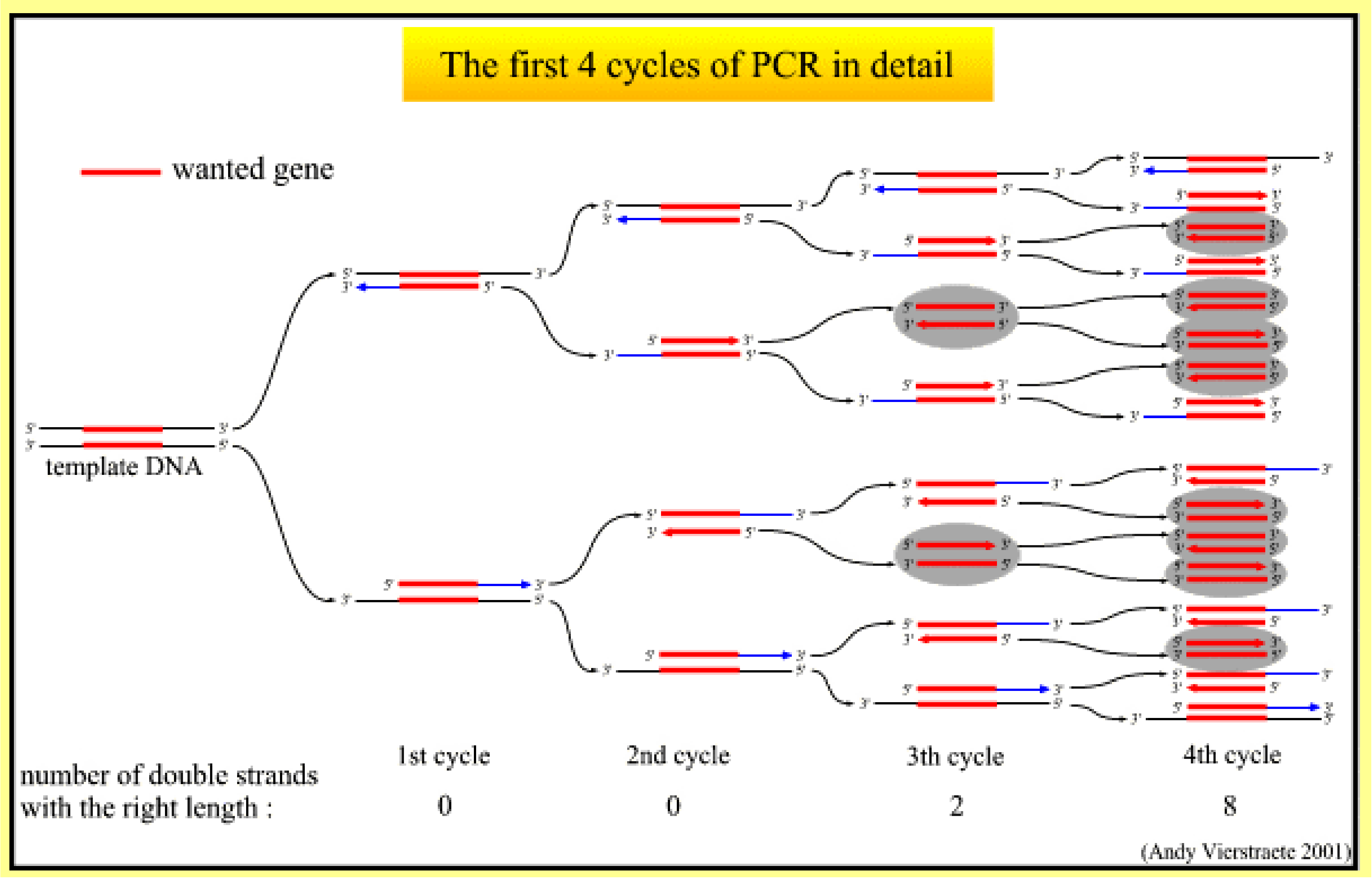
Verification of PCR Product

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**

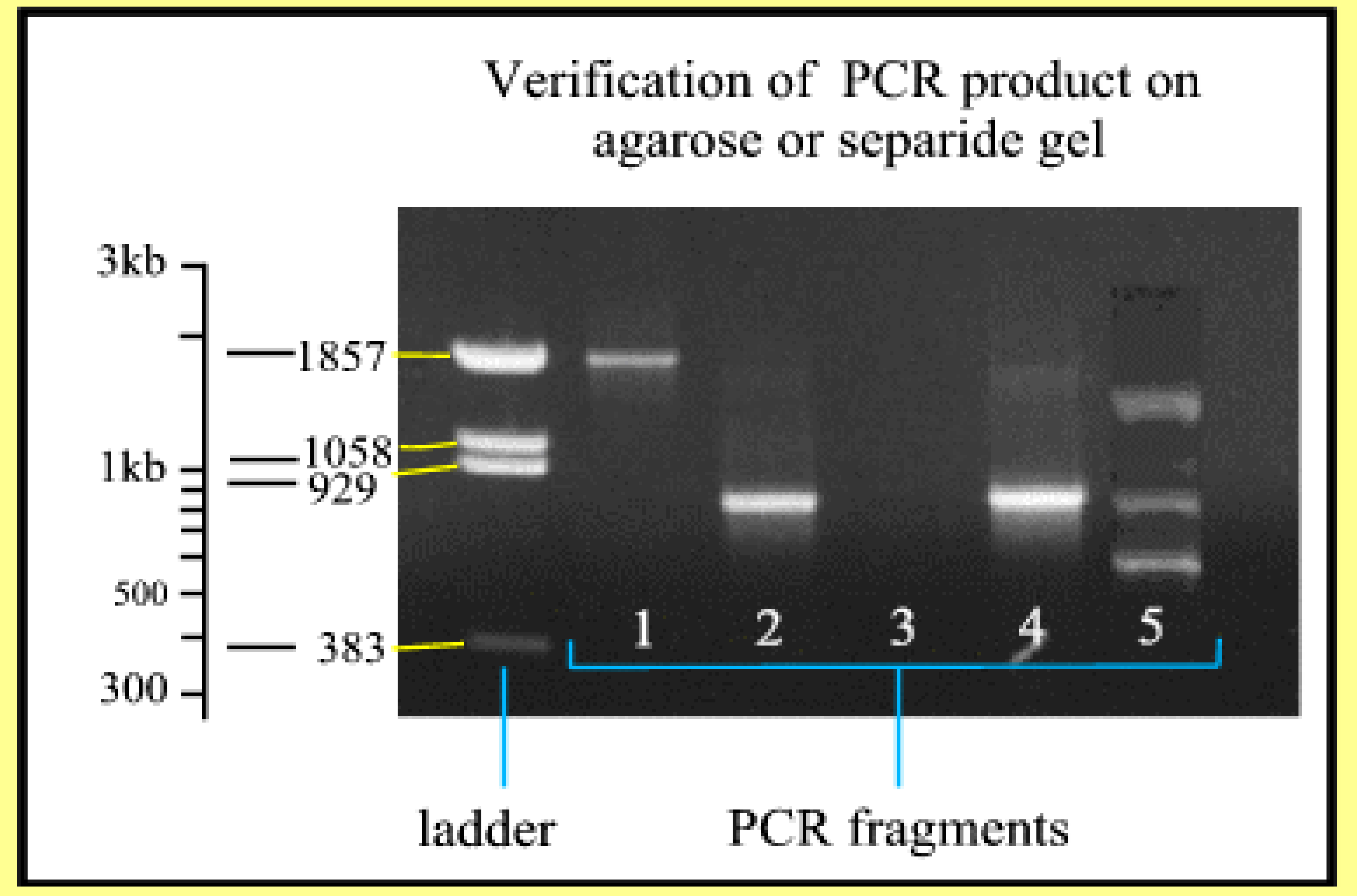
First 4 cycles of a PCR reaction



The first 4 cycles of PCR in detail

(Andy Vierstraete 2001)

Verification of the PCR product



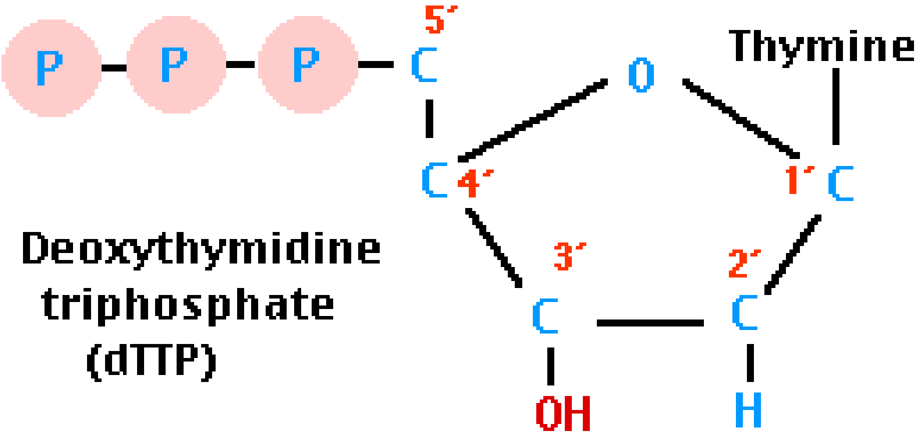
Verification of PCR product on agarose or separide gel

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 normal (deoxy) nucleotides in ample quantities
  - dATP
  - dGTP
  - dCTP
  - dTTP

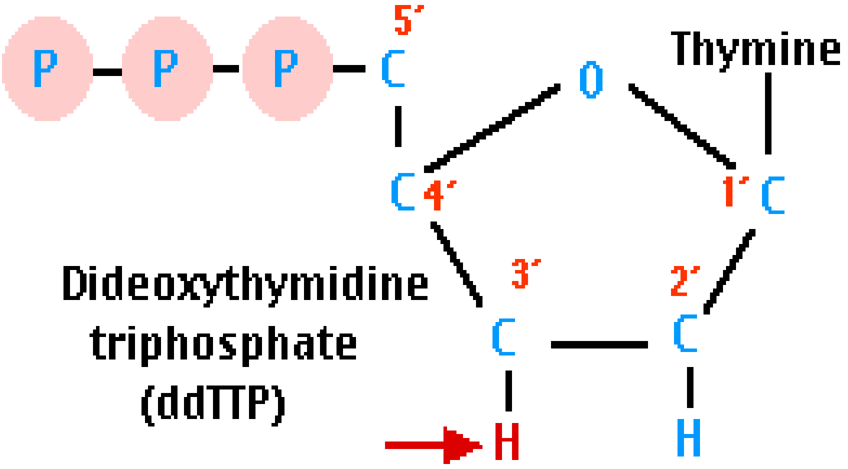


Deoxythymidine triphosphate (dTTP)

- *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



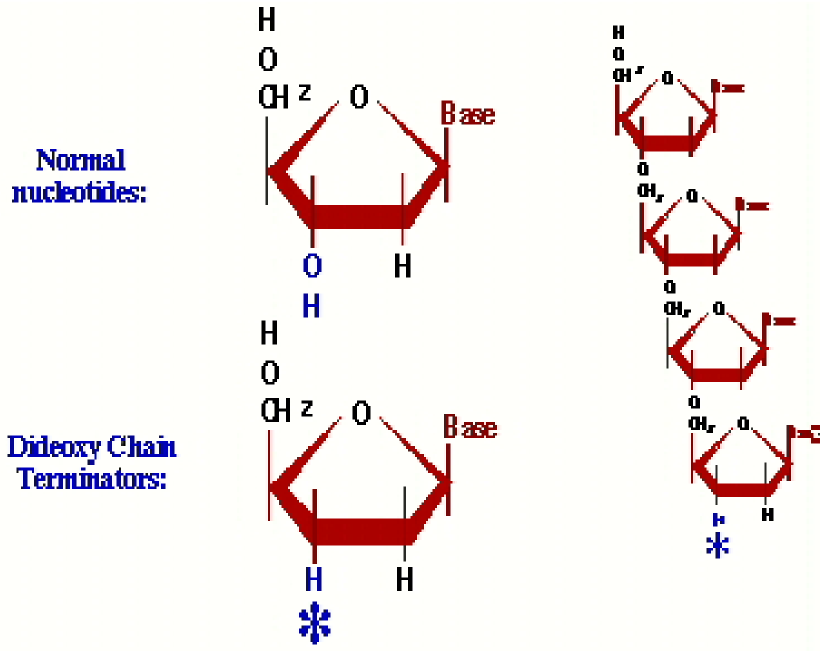
This method of DNA sequencing is called dideoxy method, or chain termination method, or Sanger's method.

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

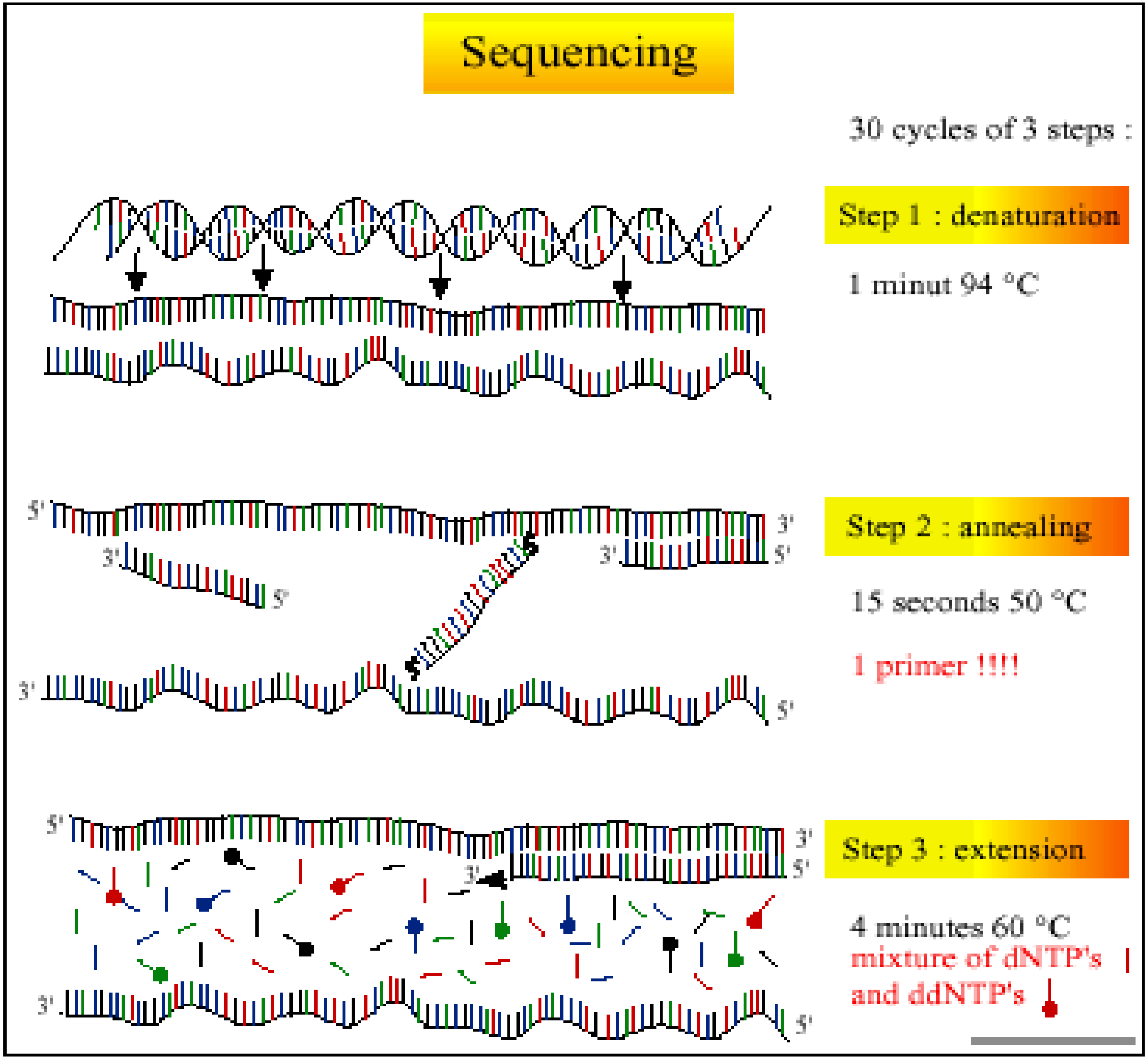
- Denaturation at 94°C
- Annealing at 50°C
- Extension at 60°C ← instead of 72°C

II Separation of the fragments

III Detection on an automated sequencer

IV Assembling the sequenced parts

Different steps in Sequencing



Sequencing

30 cycles of 3 steps :

Step 1 : denaturation

1 minut 94 °C

Step 2 : annealing

15 seconds 50 °C

1 primer !!!!

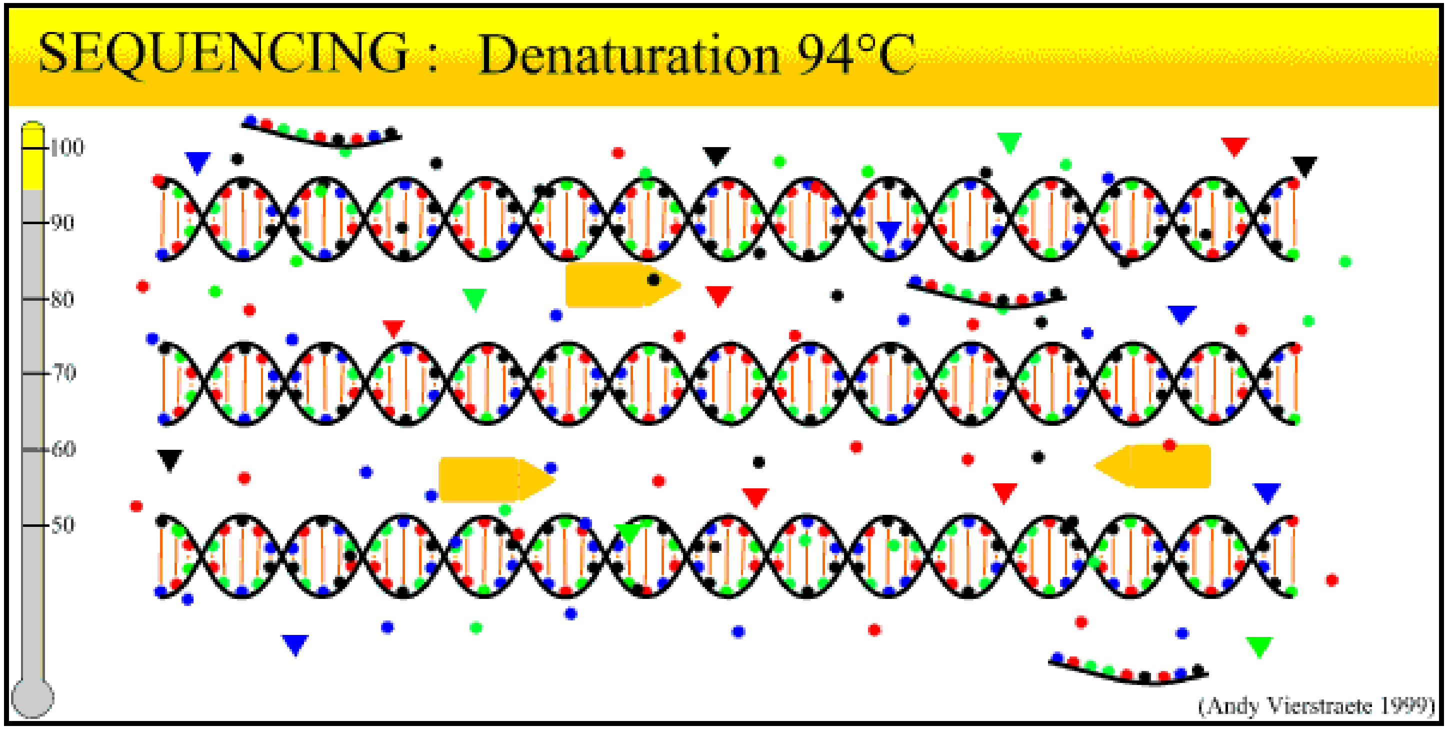
Step 3 : extension

4 minutes 60 °C

mixture of dNTP's

and ddNTP's

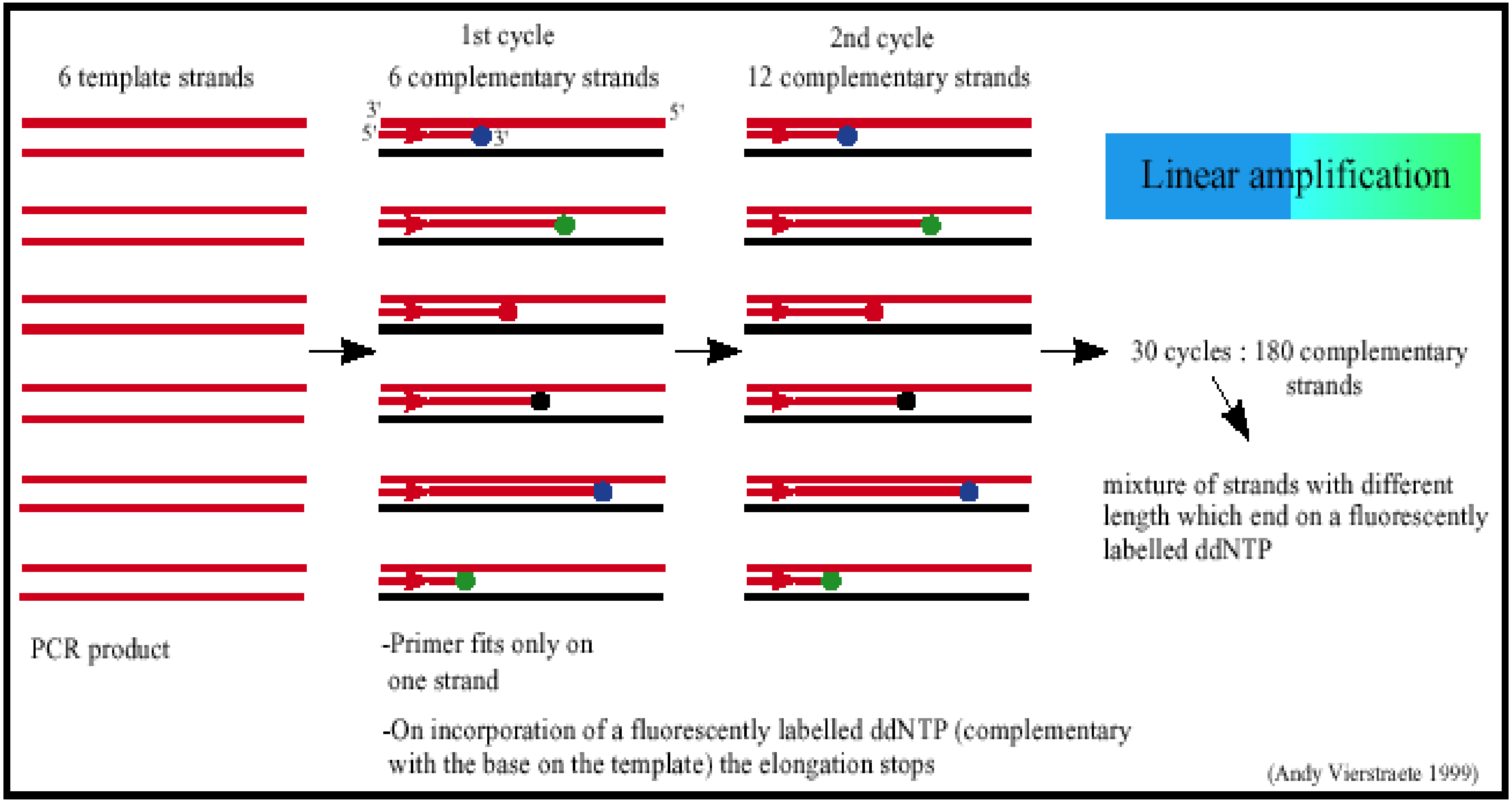
Different steps in Sequencing



PCR Sequencing

Since only one primer is used, only one strand is copied during sequencing – results in linear increase of the No. of copies.

⇒ large amount of DNA in the starting mixture is required.



1st cycle
2nd cycle
6 template strands
6 complementary strands
12 complementary strands
3'
5'
5'
3'
Linear amplification
30 cycles : 180 complementary strands
mixture of strands with different length which end on a fluorescently labelled ddNTP
PCR product
-Primer fits only on one strand
-On incorporation of a fluorescently labelled ddNTP (complementary with the base on the template) the elongation stops

(Andy Vierstraete 1999)

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 are loaded on an acrylamide gel for separation

- gel electrophoresis.

During electrophoresis, a voltage is created across the gel making one end positive and the other negative.

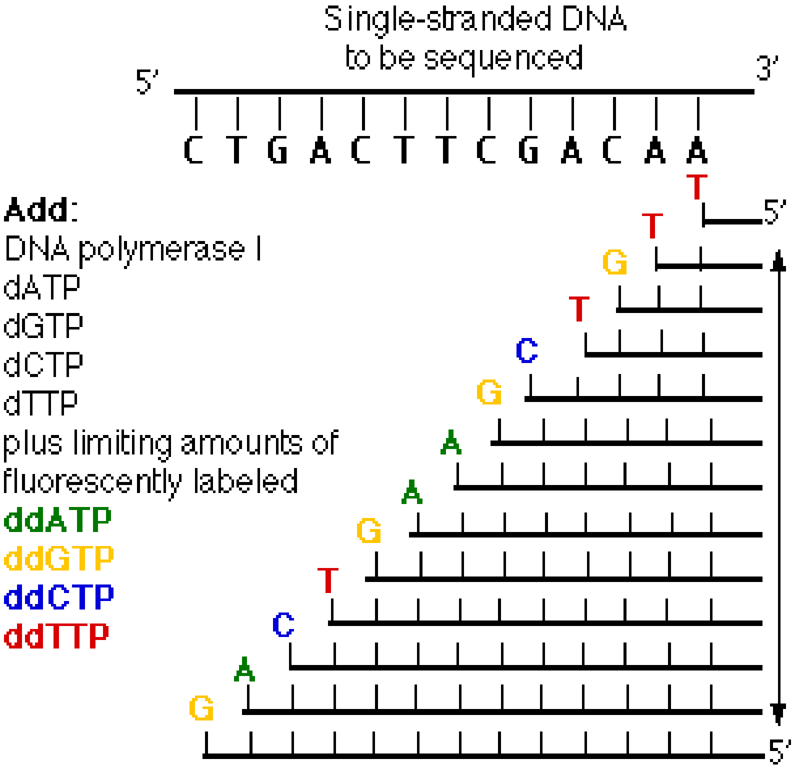
DNA being –vely charged, migrates to the positive side;

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

Separation of molecules with electrophoresis

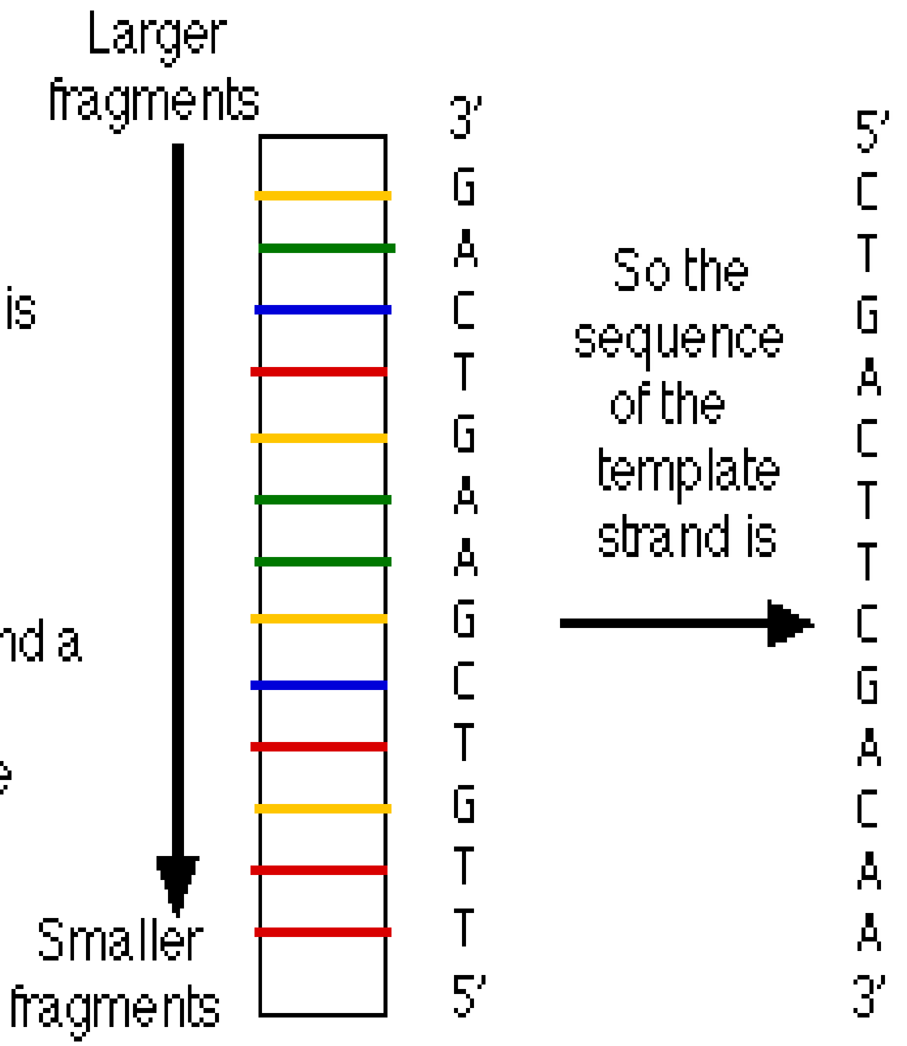
Very good resolution - a difference of even one nucleotide is enough to separate a strand from the next shorter or longer strand.

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


Single-stranded DNA to be sequenced
5'
3'
C T G A C T T C G A C A A
Add:
DNA polymerase I
dATP
dGTP
dCTP
dTTP
plus limiting amounts of fluorescently labeled
ddATP
ddGTP
ddCTP
ddTTP
T
T
G
T
C
G
A
A
G
T
C
A
G
5'
5'

Separation of Molecules with Electrophoresis

Electrophoresis using laser to activate the fluorescent dideoxy nucleotides and a detector to distinguish the colors



Larger fragments

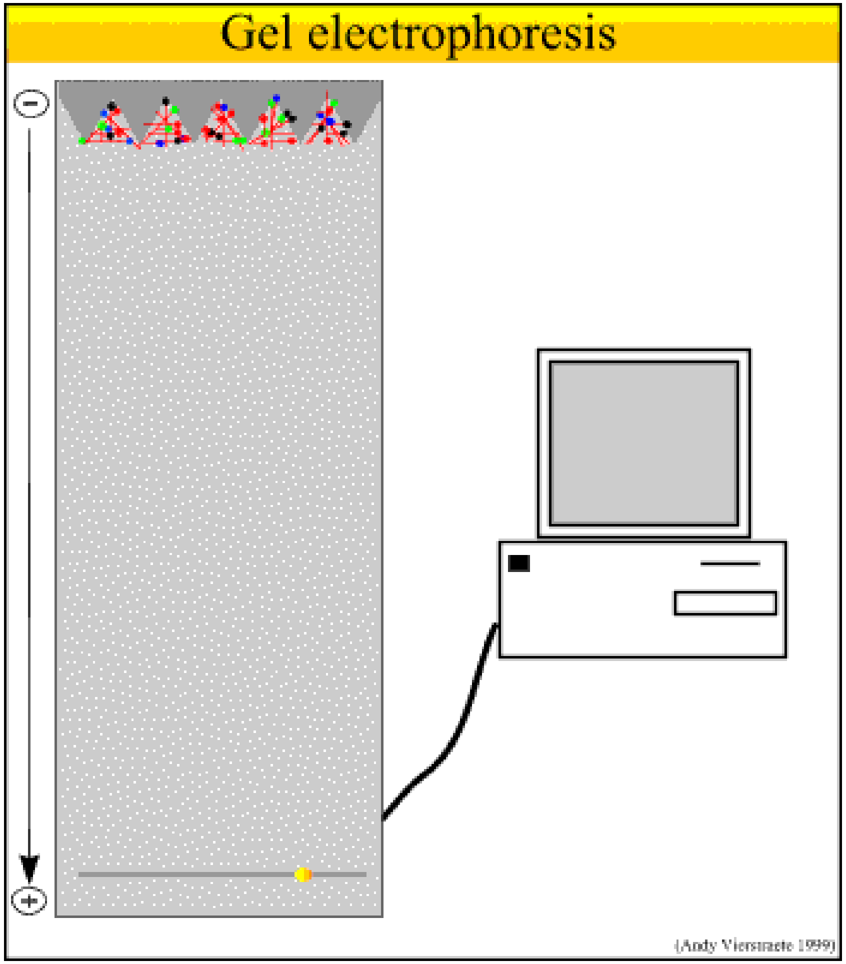
Smaller fragments

3′ G A C T G A A G C T G T T 5′

So the sequence of the template strand is

5′ C T G A C T T C G A C A A 3′

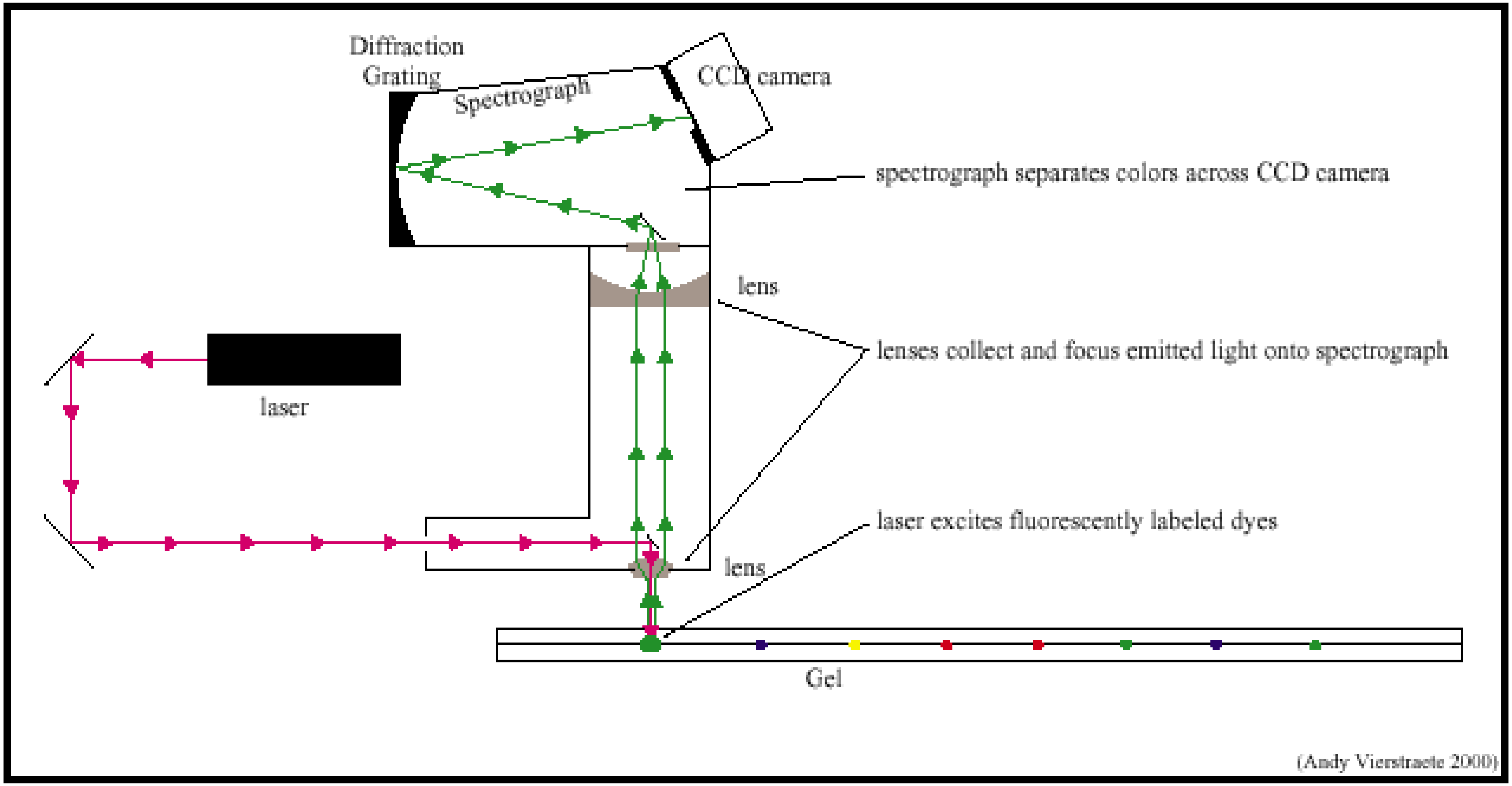
Separation of the Molecules with Electrophoresis



PCR Sequencing

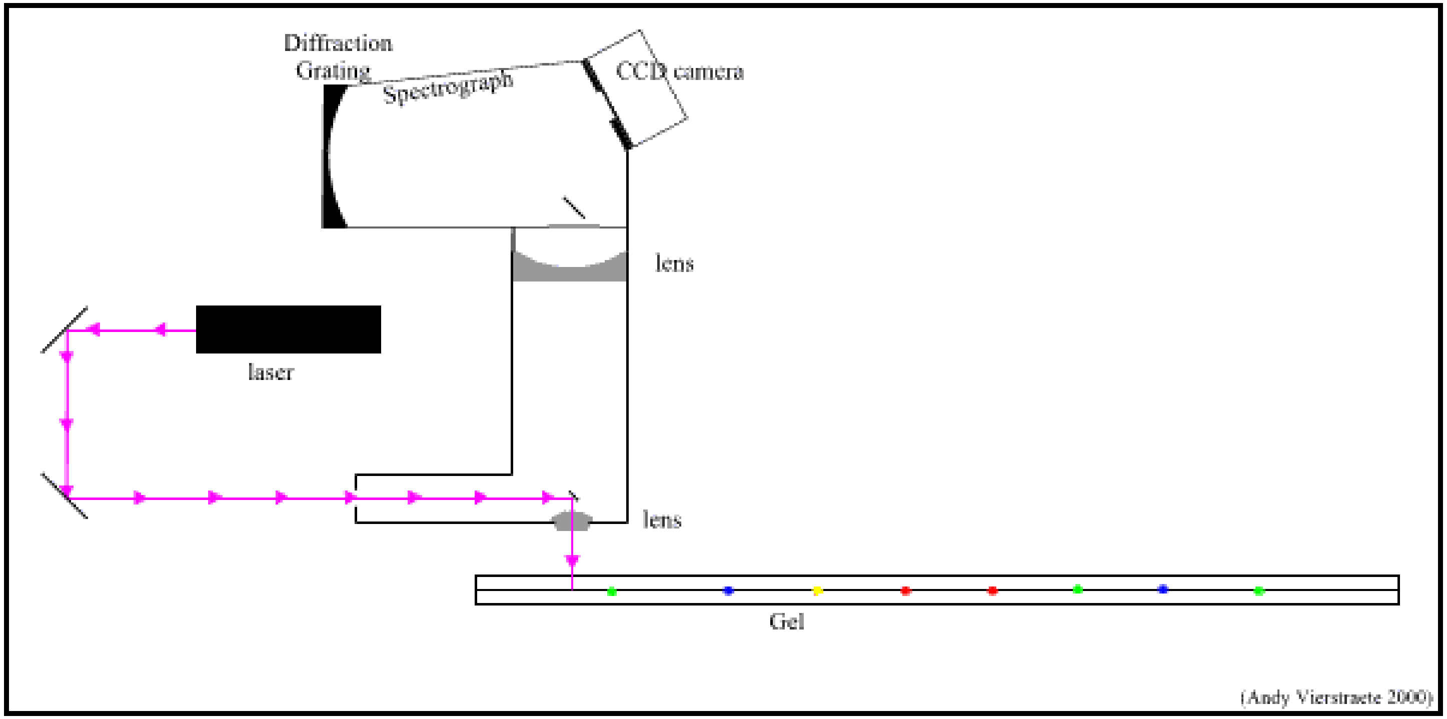
III Detection on an automated sequencer:

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


Diffraction Grating
Spectrograph
CCD camera
spectrograph separates colors across CCD camera
lens
lenses collect and focus emitted light onto spectrograph
laser
laser excites fluorescently labeled dyes
lens
Gel

(Andy Vierstraete 2000)

Scanning & Detection System on a Sequencer



(Andy Vierstraete 2000)

PCR Sequencing

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

The picture can't be displayed.

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

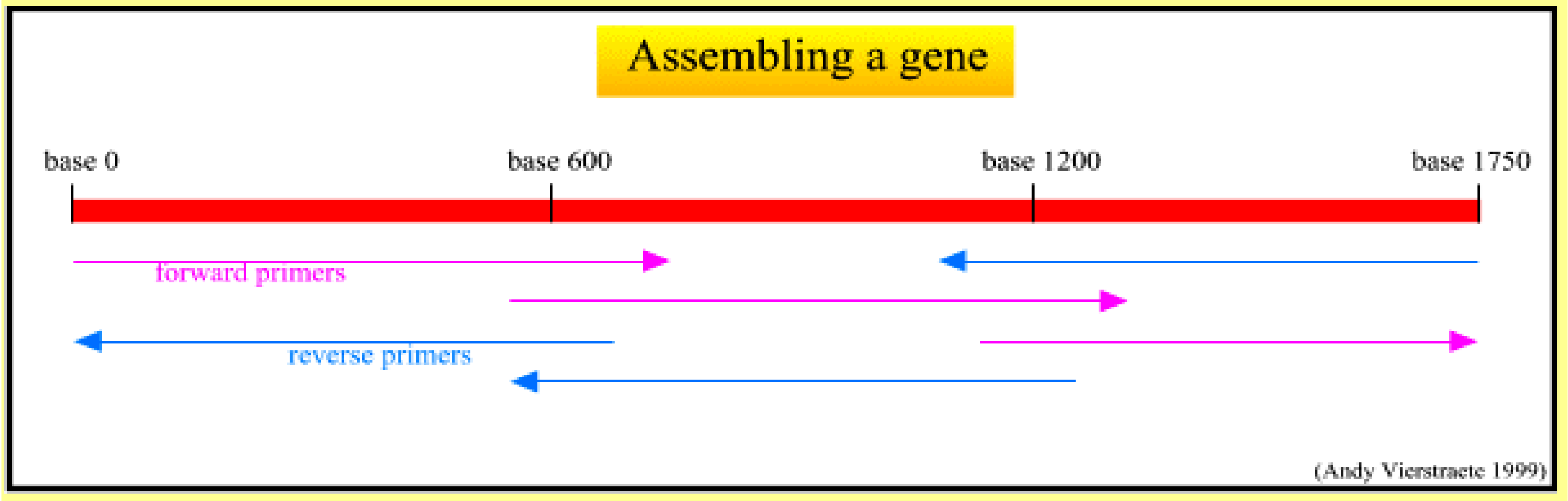
PCR Sequencing

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, is sequenced with internal primers.

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



(Andy Vierstraete 1999)

Genome Sequencing

Genome Sequencing

By Sanger's method, we can sequence a fragment of DNA ~ 1000bp long.

But what about longer pieces?

Human genome is 3 billion bases long, arranged on 23 pairs of chromosomes.

Sequencing machine reads just a drop in the ocean!

Genome Sequencing

Solution: Break the entire genome into manageable pieces and sequence them.

Two approaches used for sequencing Human genome:

- Publicly funded Human Genome Project (HGP) – clone-by-clone or hierarchical shotgun sequencing method
- Privately Funded Sequencing Project - Celera Genomics – whole genome shotgun sequencing method

Genome Sequencing

Hierarchical shotgun sequencing approach:

- **genomic DNA is cut into pieces of about 150 Mb**
- **inserted into BAC cloning vectors,**
- **transformed into *E. coli* where they are replicated and stored.**

BAC inserts are isolated & mapped to determine the order of each cloned 150 Mb fragment

- referred to as the Golden Tiling Path

Begun formally in 1990, Human Genome Project was a 13-yr effort coordinated by the U.S. DAE and NIH.
- completed in 2003

Genome Sequencing

Each BAC fragment in the Golden Path is

- fragmented randomly into smaller pieces,
- each piece is cloned into a plasmid and sequenced on both strands.

These sequences are aligned so that identical regions overlap.

Contiguous pieces are then assembled into finished sequence once each strand has been sequenced about 5 times to produce 10× coverage of high-quality data.

Genome Sequencing

Whole genome shotgun sequencing (WGS)

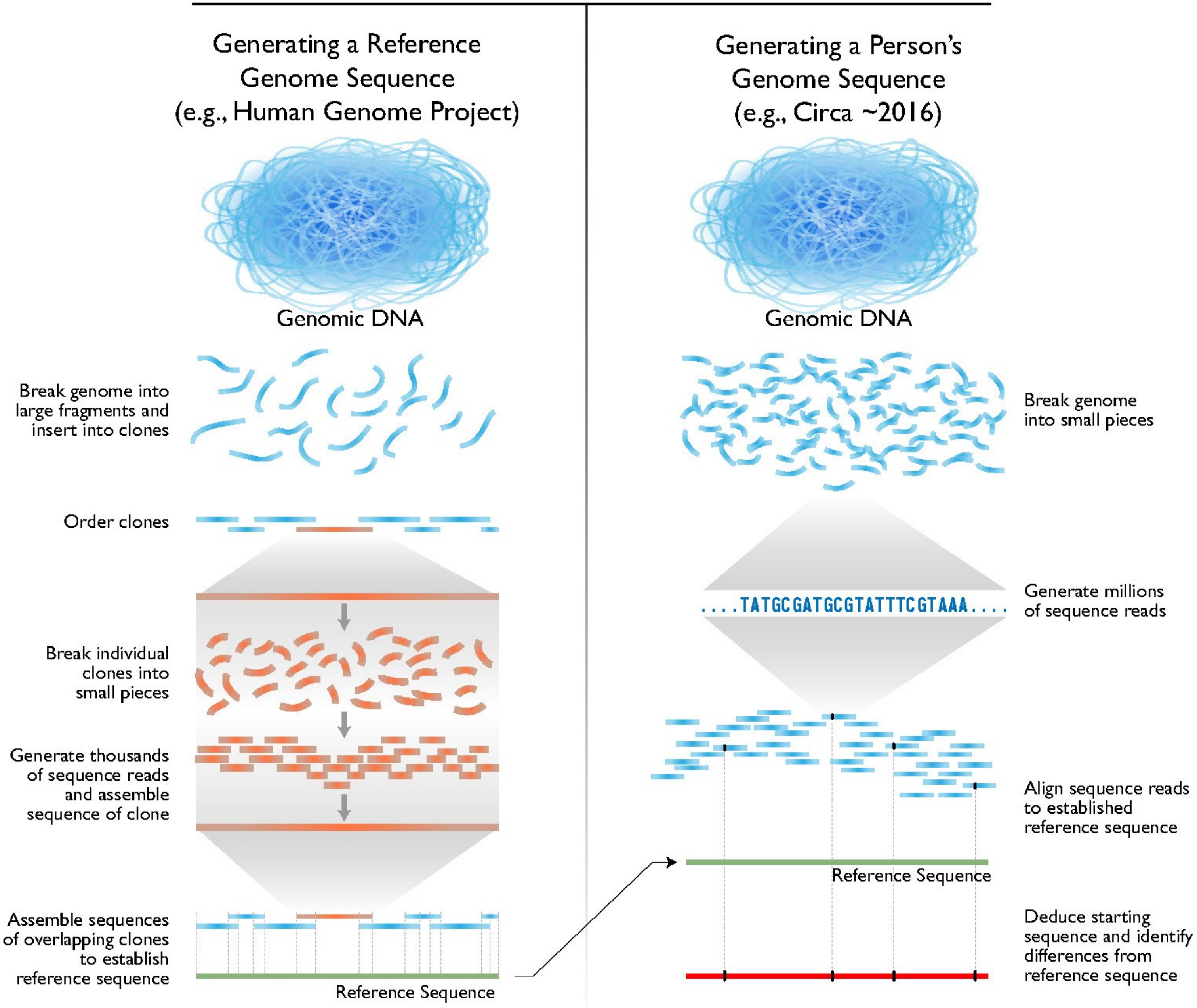
- method developed and preferred by Celera Genomics

- skips the entire step of making libraries of BAC clones

Blast apart the entire human genome into fragments of 2 - 10 kb and sequence them.

Challenge is then to assemble these fragments into the whole genome sequence.

Human Genome Sequencing



Generating a Reference Genome Sequence (e.g., Human Genome Project)

Genomic DNA

Break genome into large fragments and insert into clones

Order clones

Break individual clones into small pieces

Generate thousands of sequence reads and assemble sequence of clone

Assemble sequences of overlapping clones to establish reference sequence

Reference Sequence

Generating a Person's Genome Sequence (e.g., Circa ~2016)

Genomic DNA

Break genome into small pieces

Generate millions of sequence reads

....TATGCGATGCGTATTTCGTAAA....

Align sequence reads to established reference sequence

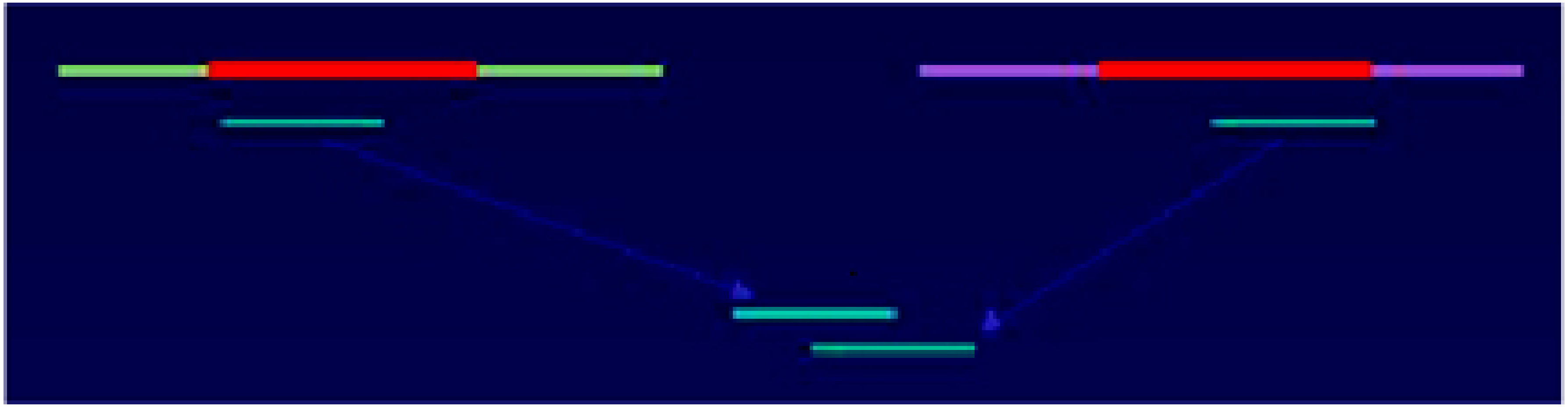
Reference Sequence

Deduce starting sequence and identify differences from reference sequence

Whole Genome Shotgun Method

What makes the task of assembling the genome fragments especially challenging

- repeats in the genome (~ 50% in human genome).



Because of the various ways a fragment could align with a repeat, and the different areas adjacent to the repeats in the original genome, assemblers need to be designed so as not to incorrectly join fragments

Whole Genome Shotgun Method

Adding to the challenge is the sheer computational complexity of the task.

Size of H. genome = 3 x 10^9 bp. Given length of read ~500 bps, for desired coverage of 10x, No. of reads required is:

RequiredReads = GenomeLength * DesiredCoverage / ReadLength

= 6 * 10^7

With 60M reads to assemble, we need algorithms that run in near linear time (*O*(*nlogn*))

Whole Genome Shotgun Method

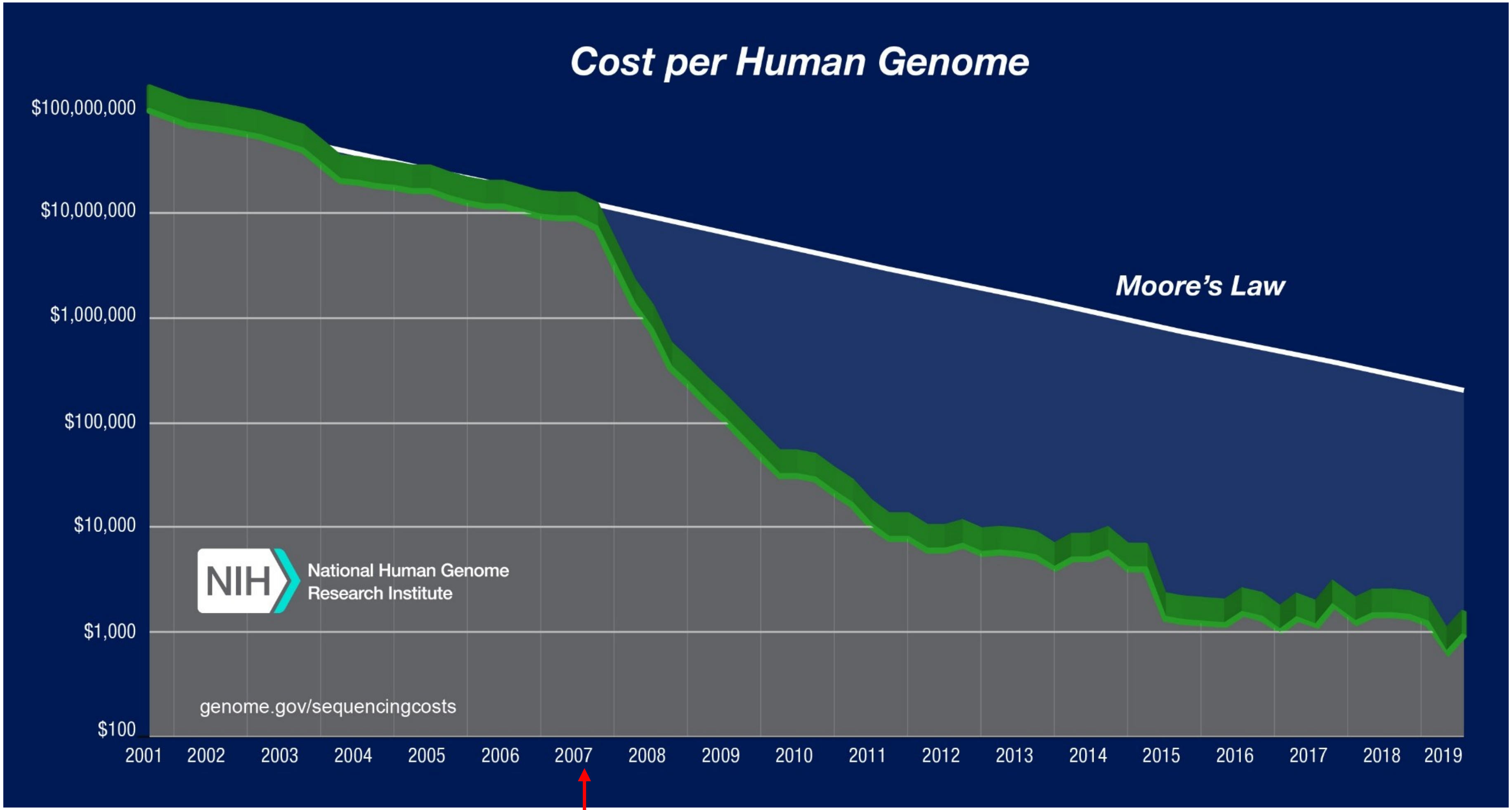
Which method is better?

Depends on the size and complexity of the genome

Note: Celera had access to the HGP data but the HGP did not have access to Celera data.

Which method is preferable for sequencing the genome of a novel coronavirus – SAR-CoV-2? Why?

High-throughput / Next-Generation Sequencing



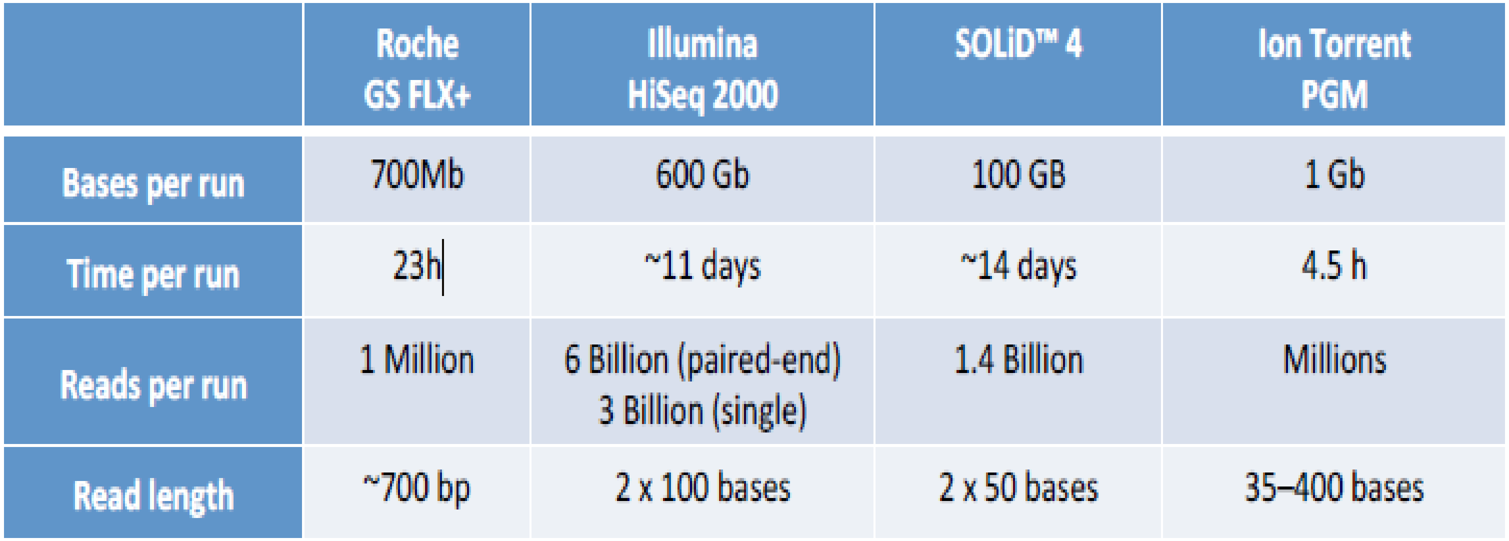
13yrs, $3 billion **S.S.** **8days, $10,000** **15min, <$1,000**

DNA sequencing beating Moore's law

HTS/NGS Sequencing

High-throughput sequencing (HTS) technologies have revolutionized the way biologists acquire and analyze genomic data.

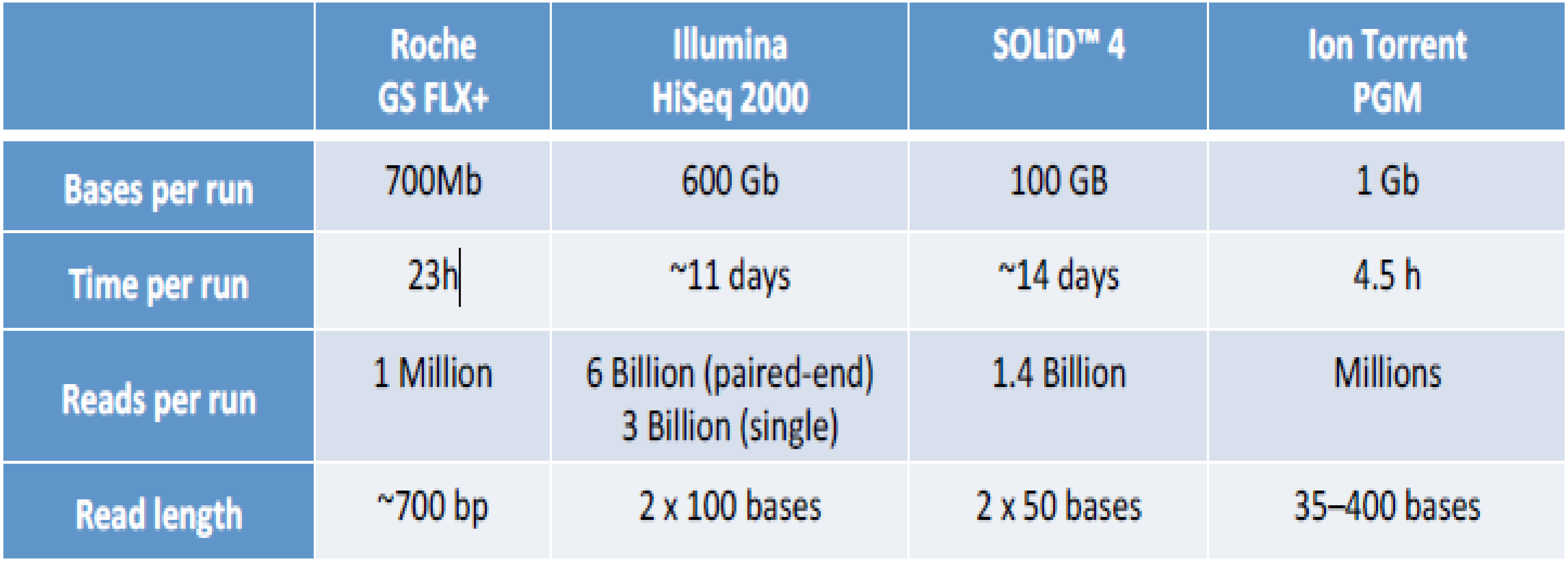
- massively parallel sequencing



	Roche GS FLX+	Illumina HiSeq 2000	SOLiD™ 4	Ion Torrent PGM
Bases per run	700Mb	600 Gb	100 GB	1 Gb
Time per run	23h	~11 days	~14 days	4.5 h
Reads per run	1 Million	6 Billion (paired-end) 3 Billion (single)	1.4 Billion	Millions
Read length	~700 bp	2 x 100 bases	2 x 50 bases	35–400 bases

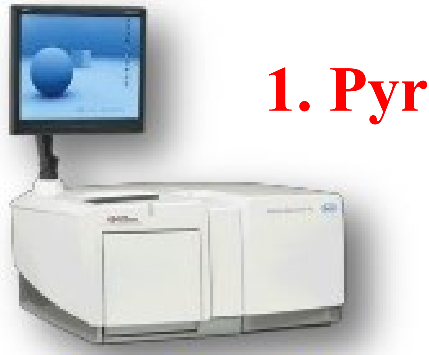
- can generate tens of gigabases per week, at a cost 200-fold less than previous methods.

Sequencing Machines: Overview



	Roche GS FLX+	Illumina HiSeq 2000	SOLiD™ 4	Ion Torrent PGM
Bases per run	700Mb	600 Gb	100 GB	1 Gb
Time per run	23h	~11 days	~14 days	4.5 h
Reads per run	1 Million	6 Billion (paired-end) 3 Billion (single)	1.4 Billion	Millions
Read length	~700 bp	2 x 100 bases	2 x 50 bases	35–400 bases

Sequencing Machines: Overview



Roche GS-FLX

1. Pyrosequencing

Life Technologies SOLiD

3. Sequence by ligation

Illumina HiSeq

2. Sequence by Synthesis

Life Technologies Ion Torrent

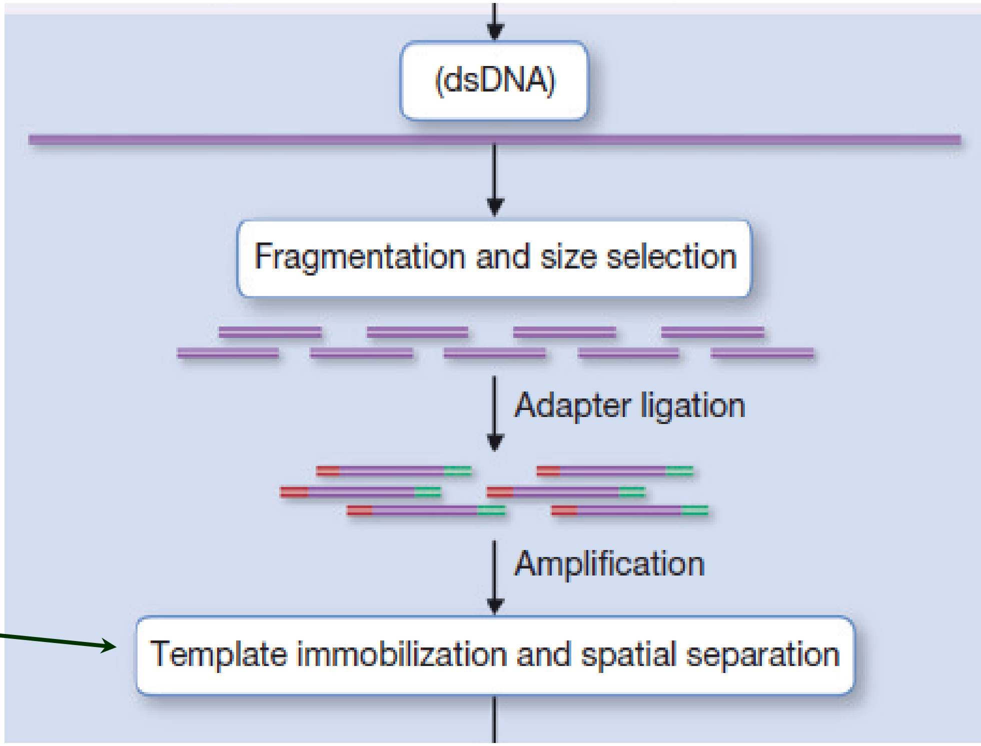
4. Proton Detection

Basic workflow: Template Generation

Sequence library – convert starting material into a library of sequencing reaction templates.

Require common steps:

- **Fragmentation**
- **Size selection**
- **Adapter ligation**


(dsDNA)
Fragmentation and size selection
Adapter ligation
Amplification
Template immobilization and spatial separation

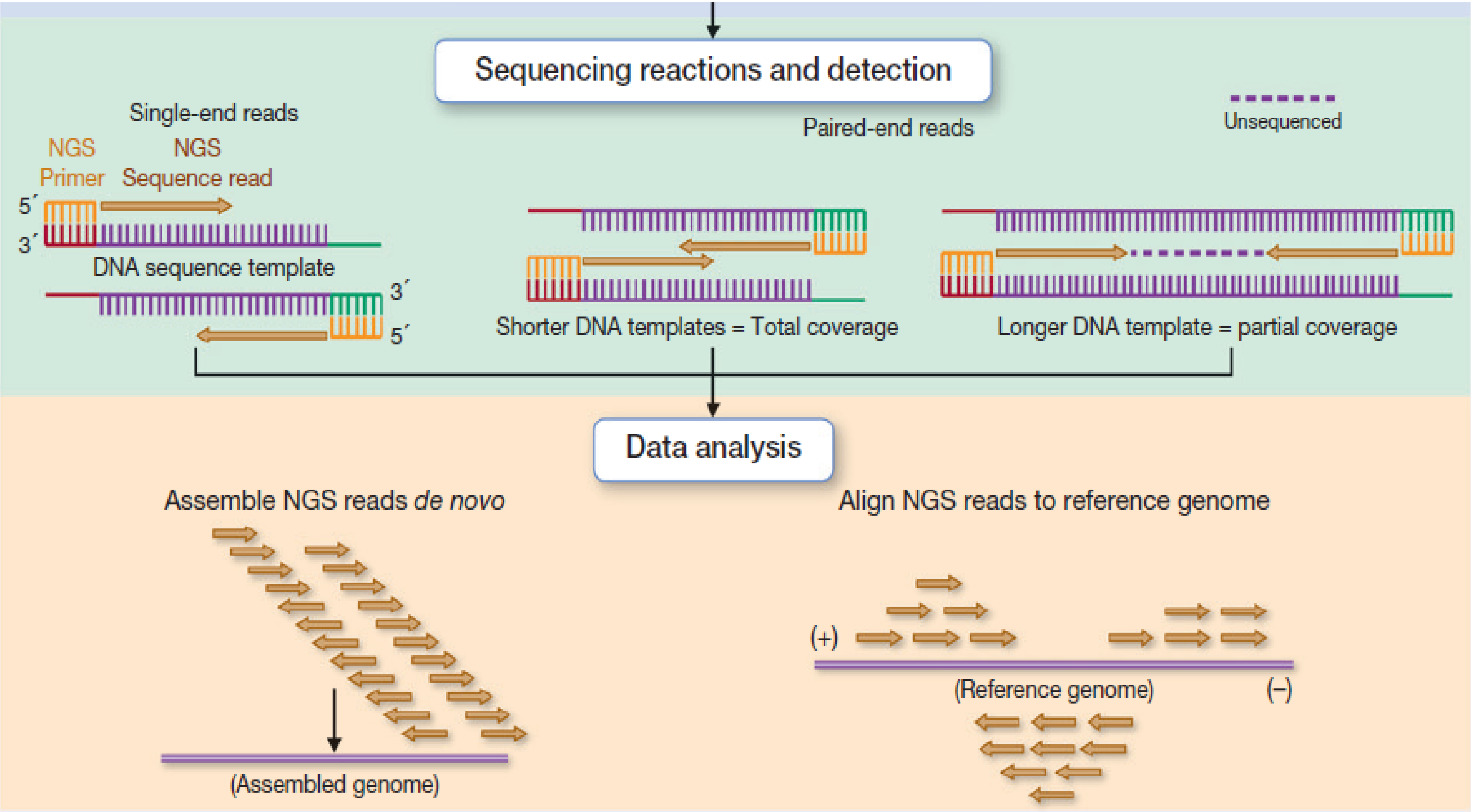
by attachment to solid surfaces or beads

Amplification-based - "second-generation" sequencing technology

Single-molecule - "third-generation" sequencing technology

A library is either sequenced directly - Single-molecule templates or, amplified then sequenced - Clonally amplified templates

Basic workflow: Detection & Data Analysis


Sequencing reactions and detection
Single-end reads
NGS
Primer
NGS
Sequence read
5´
3´
DNA sequence template
3´
5´
Paired-end reads
Unsequenced
Shorter DNA templates = Total coverage
Longer DNA template = partial coverage
Data analysis
Assemble NGS reads de novo
(Assembled genome)
Align NGS reads to reference genome
(+)
(Reference genome)
(–)

Data Analysis

The scale and nature of data produced by all NGS platforms place substantial demands on IT at all stages of sequencing, including data tracking, storage, and quality control:

- **base calling - by proprietary software**
- **Quality check and filtering or reads**

- aligning sequencing data to Reference genome. if available, or a *de novo* assembly is conducted.

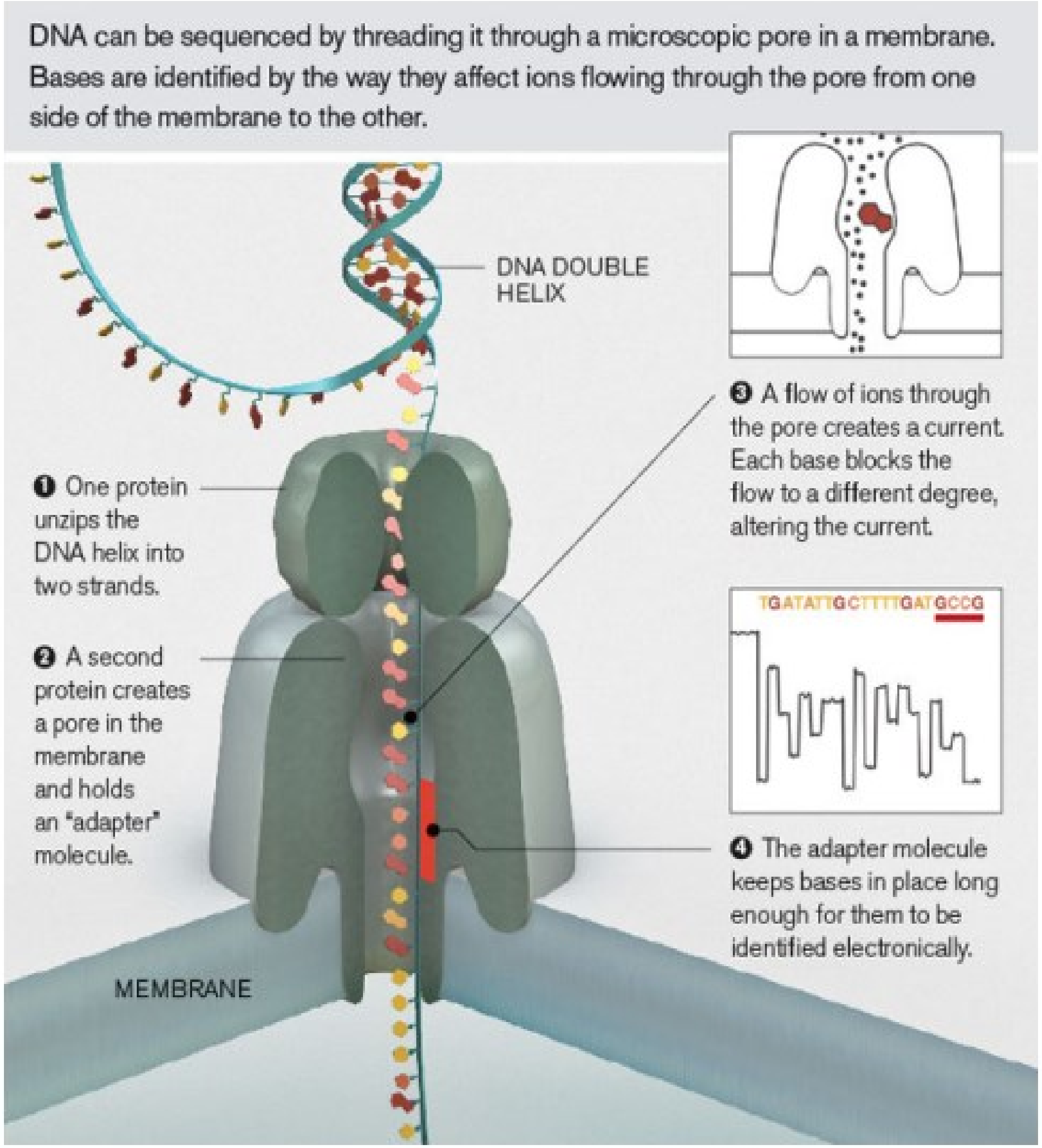
Once the sequence is aligned to a reference genome, the data needs to be analyzed in an experiment-specific fashion.

Sequence alignment & assembly is an active area of computational research

Third Generation Sequencing (TGS)

- **‘Long read sequencing’ – read length: ~ 10 – 60Kb**
- **Single molecule sequencing**
- **No PCR step involved**
- **Faster and portable**
- **Under active development**
- **e.g., PacBio Single molecule real time sequencing (SMRT) and Oxford Nanopore**

Oxford Nanopore - MinION


DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

HTS Applications

One of the most prominent applications of NGS is re-sequencing:

- **whole genome resequencing**
- **target-region resequencing**
- **exome resequencing**

Any human individual's genome available in NCBI?

- **genome-wide analysis of single nucleotide variations and other structural variations, multiple individuals, or strains, cancer sequencing, population-based sampling of a species, migration patterns of a virus, e.g., SARS-CoV-2, etc.**

PCR Sequencing

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

What technique is used for diagnostic testing of COVID-19?

While sequencing a novel genome for the first time, how are primers identified?

PCR Sequencing

Real time RT-PCR used for diagnostic testing of COVID-19

It is a laboratory technique combining reverse transcription of RNA into DNA (called complementary DNA or cDNA) and amplification of specific DNA targets using polymerase chain reaction (PCR).

It is primarily used to measure the amount of a specific RNA.

This is achieved by monitoring the amplification reaction using fluorescence, a technique called real-time PCR or quantitative PCR (qPCR).

- routinely used for analysis of gene expression and quantification of viral RNA in research and clinical settings.

Can we now answer these Qs:

- **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 is the specific probe regions extracted and amplified for detection?**
- **Is it possible to store the DNA sample for re-testing? How?**

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)