BSC 4934: Q'BIC Capstone Workshop

Dr. Giri Narasimhan

ECS 254A; Phone: x3748

giri@cis.fiu.edu

http://www.cis.fiu.edu/~giri/teach/BSC4934_Su09.html

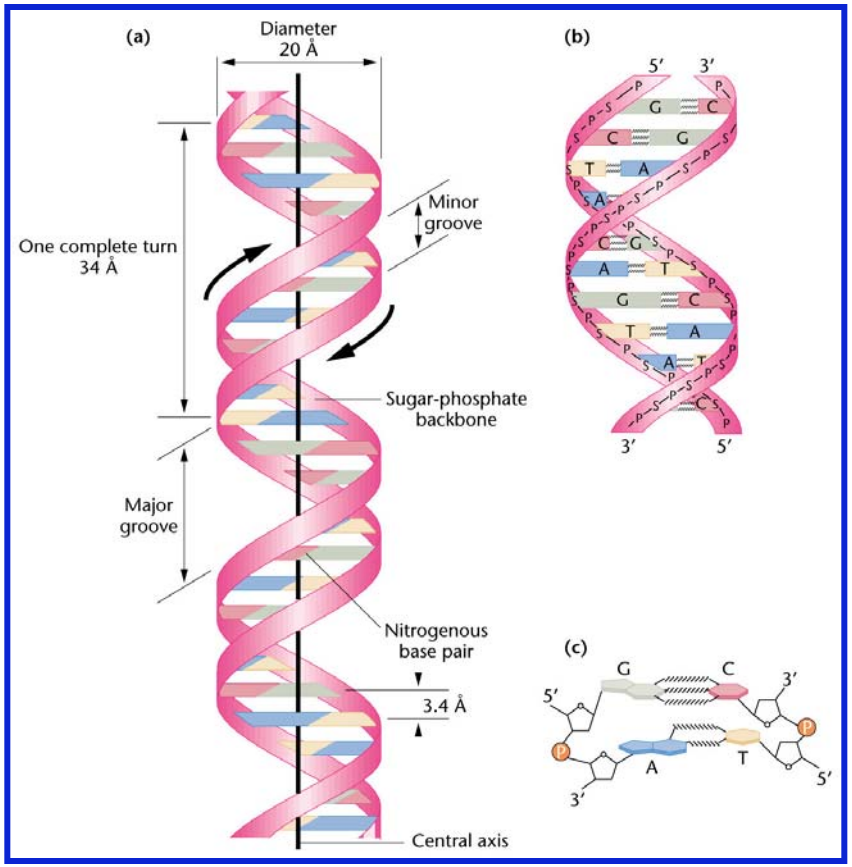
24 June through 7 July, 2009

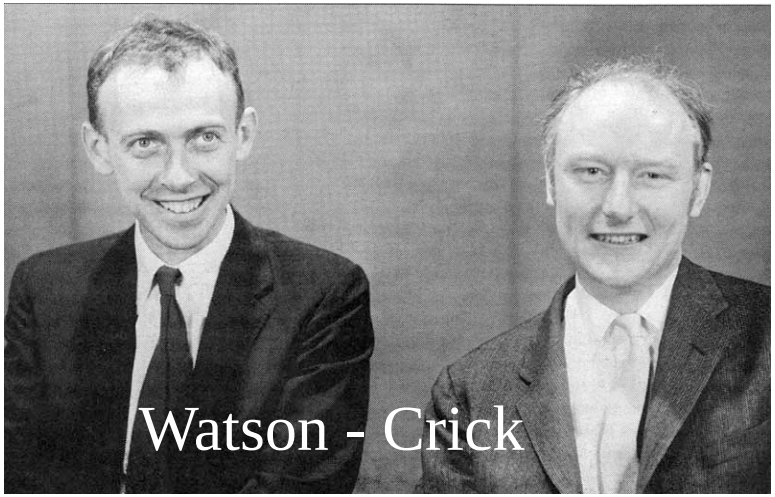
Dr. Kalai Mathee

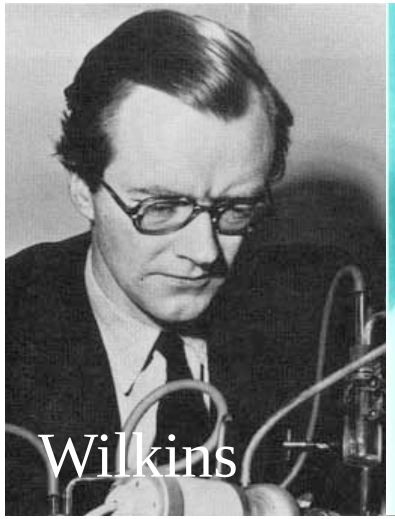
Department of Molecular Microbiology & Infectious Diseases

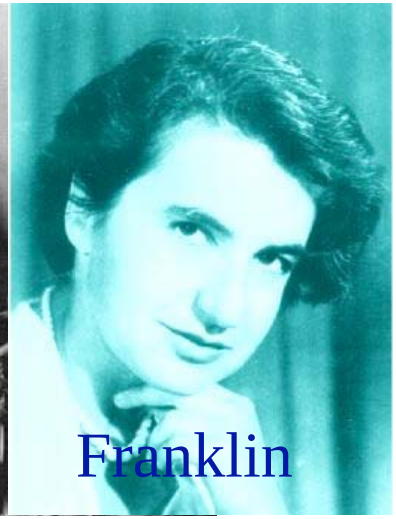
www.fiu.edu/~matheek

DNA Structure - 1953









DNA Controversy

1. Double Helix by Jim Watson - Personal Account (1968)
2. Rosalind Franklin by Ann Sayre (1975)
3. The Path to the Double Helix by Robert Olby (1974)
4. Rerelease of Double Helix by Jim Watson with Franklin's paper
5. Rosalind Franklin: The Dark Lady of DNA by Brenda Maddox (2003)
6. Secret of Photo 51 - 2003 NOVA Series

What are the next big Qs?

1. What is order of DNA sequence in a chromosome?
2. How does the DNA replicate?
3. How does the mRNA transcribe?
4. How is the protein gets translated?

Etc

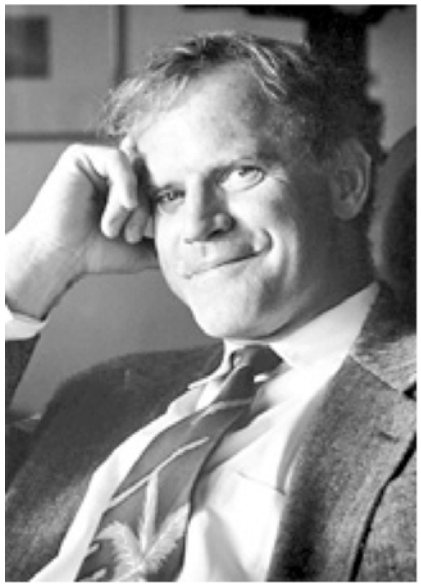
One of the tool that made a difference

Polymerase Chain Reaction

Polymerase Chain Reaction

1983 - technique was developed by Kary Mullis & others (1944-)

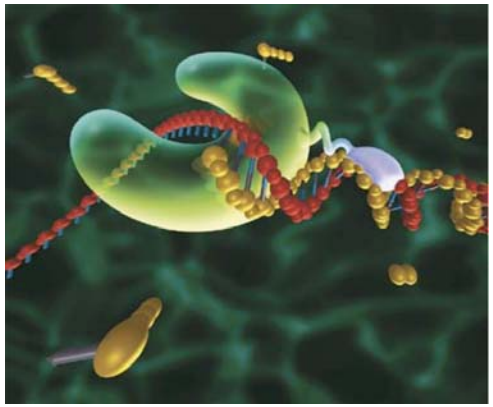
1993 Nobel prize for Chemistry



Controversy: Kjell Kleppe, a Norwegian scientist in 1971, published paper describing the principles of PCR

Stuart Linn, professor at University of California, Berkeley, used Kleppe's papers in his own classes, in which Kary Mullis was a student at the time

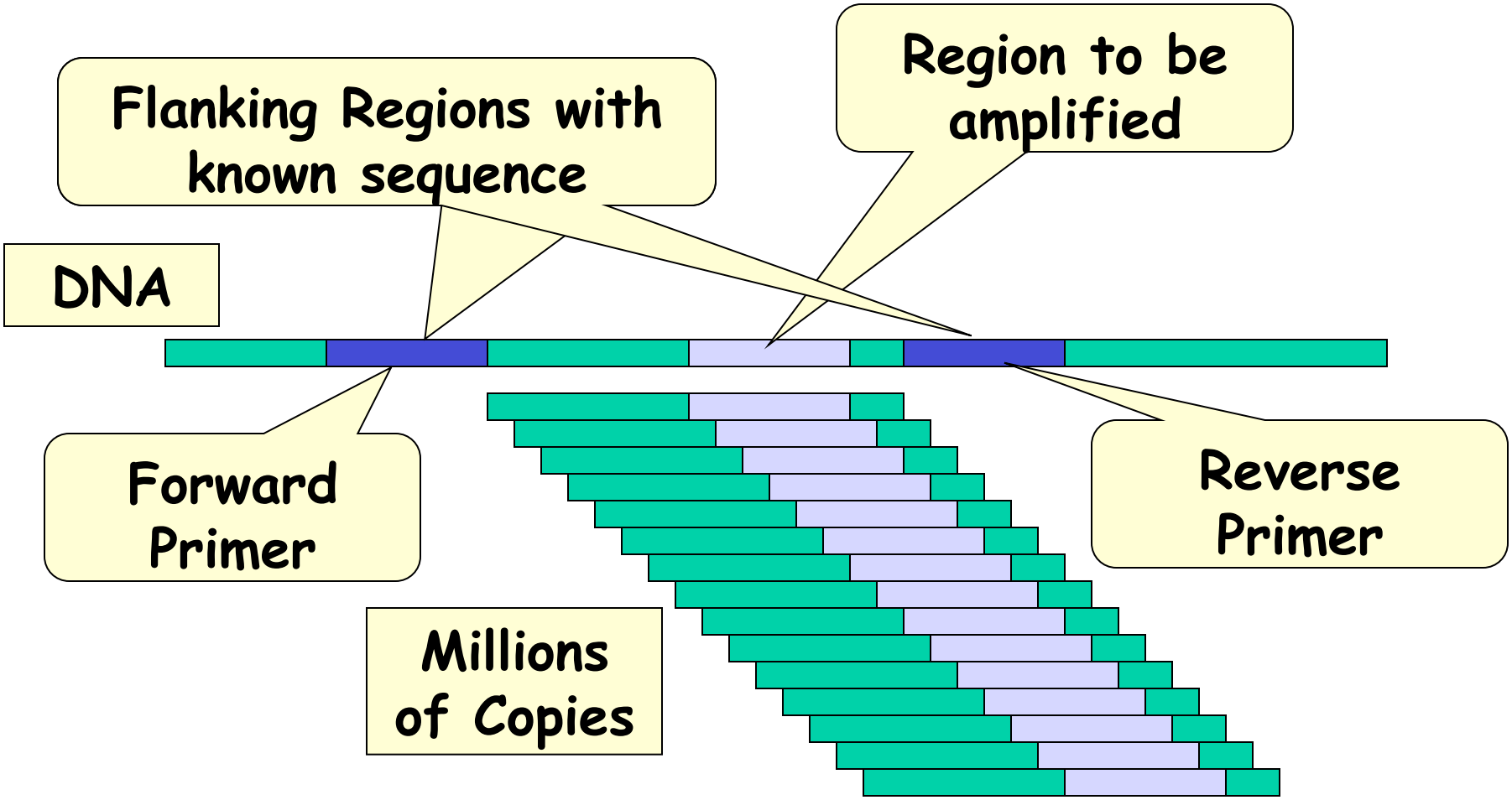
DNA Replication & Polymerase



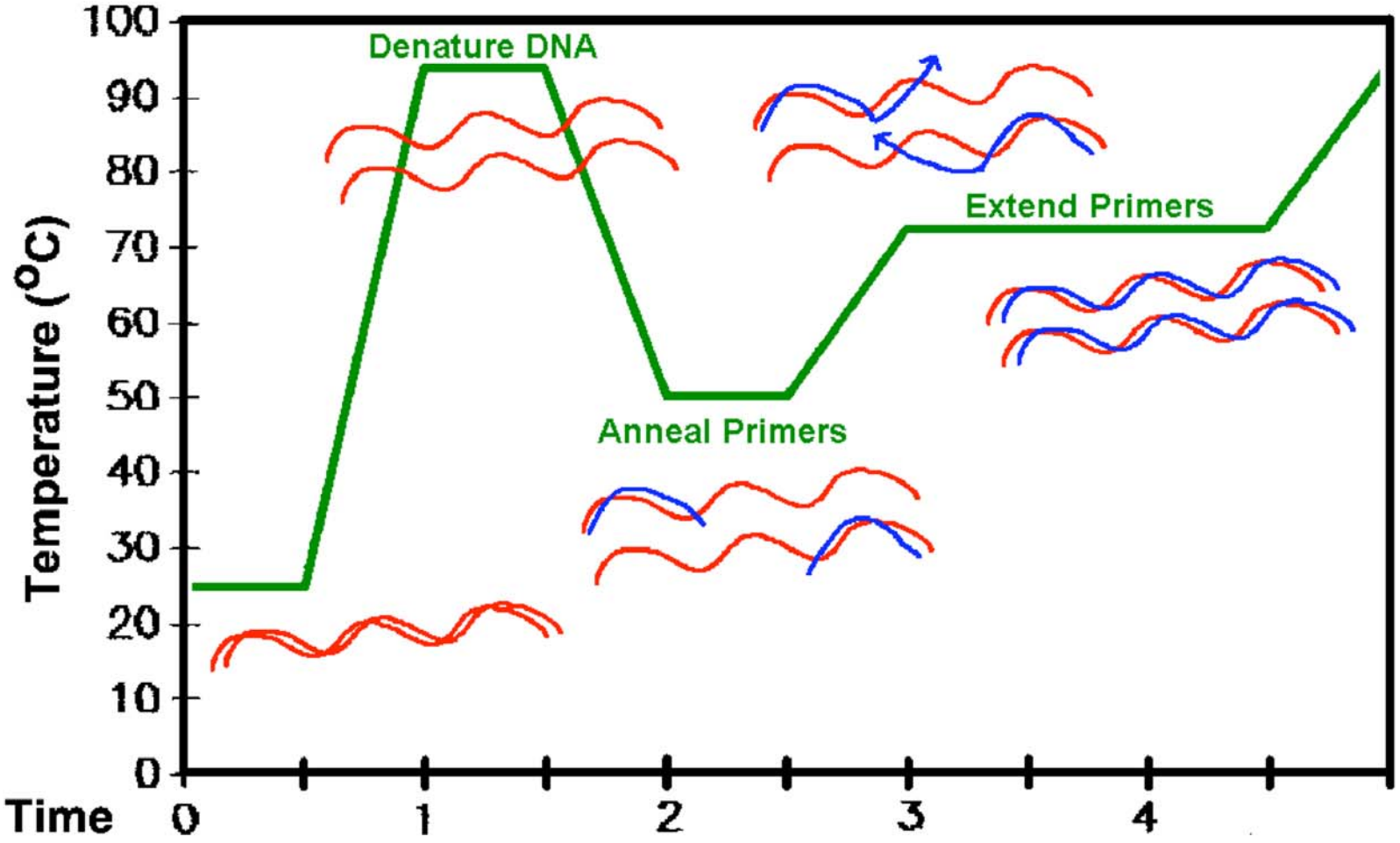
Polymerase Chain Reaction (PCR)

- PCR is a technique to amplify the number of copies of a specific region of DNA.
- Useful when exact DNA sequence is unknown
- Need to know "flanking" sequences
- Primers designed from "flanking" sequences
- If no info known, one can add adapters (short known sequence) then use a primer that recognizes the adaptor

PCR

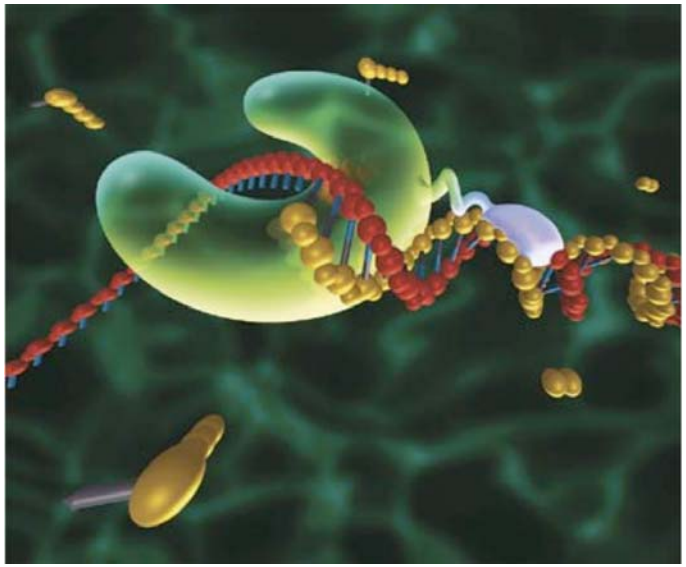

Flanking Regions with known sequence
Region to be amplified
DNA
Forward Primer
Reverse Primer
Millions of Copies

PCR

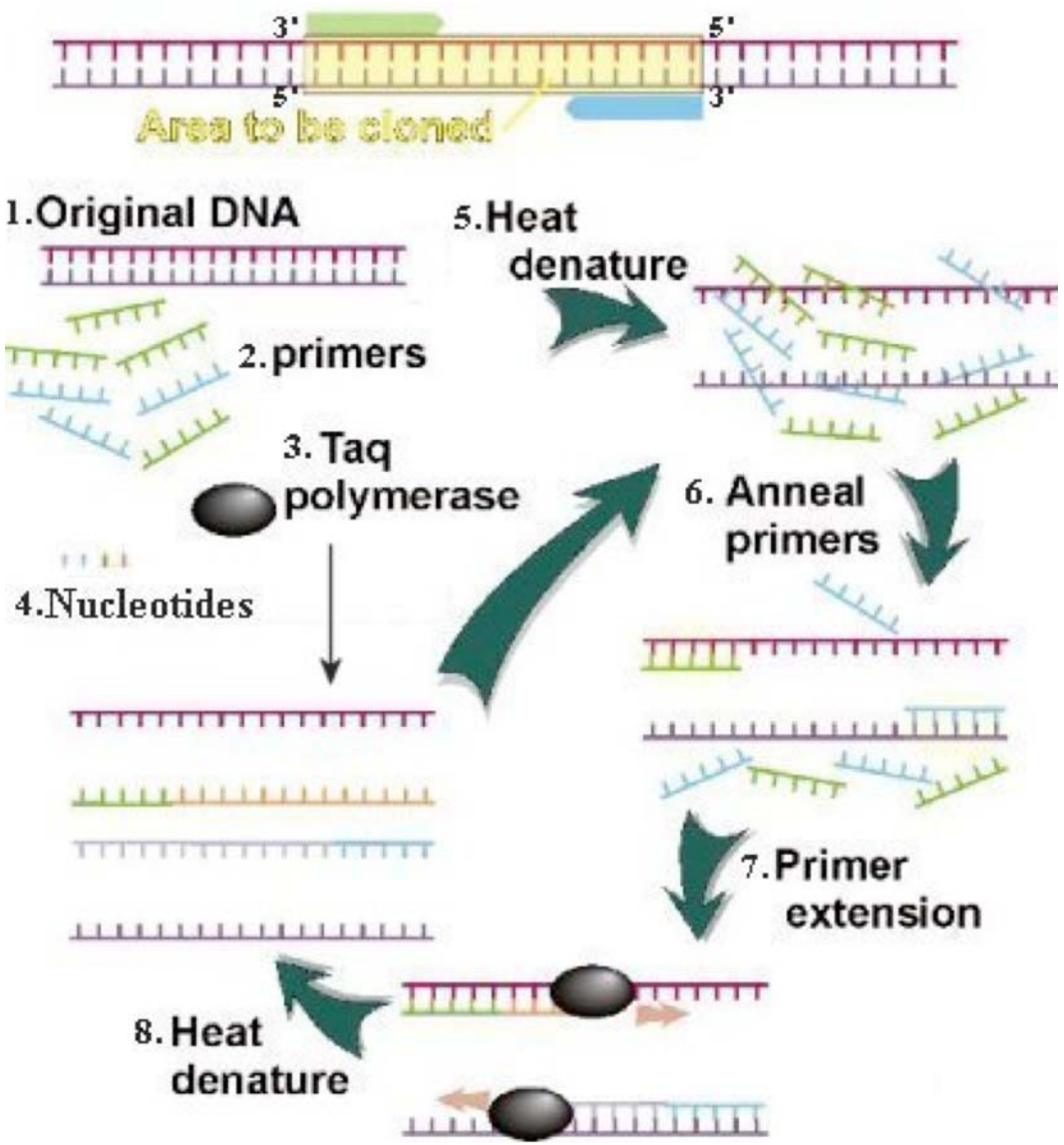

Denature DNA
Extend Primers
Anneal Primers
Temperature (°C)
100
90
80
70
60
50
40
30
20
10
0
Time
0
1
2
3
4

Taq polymerase

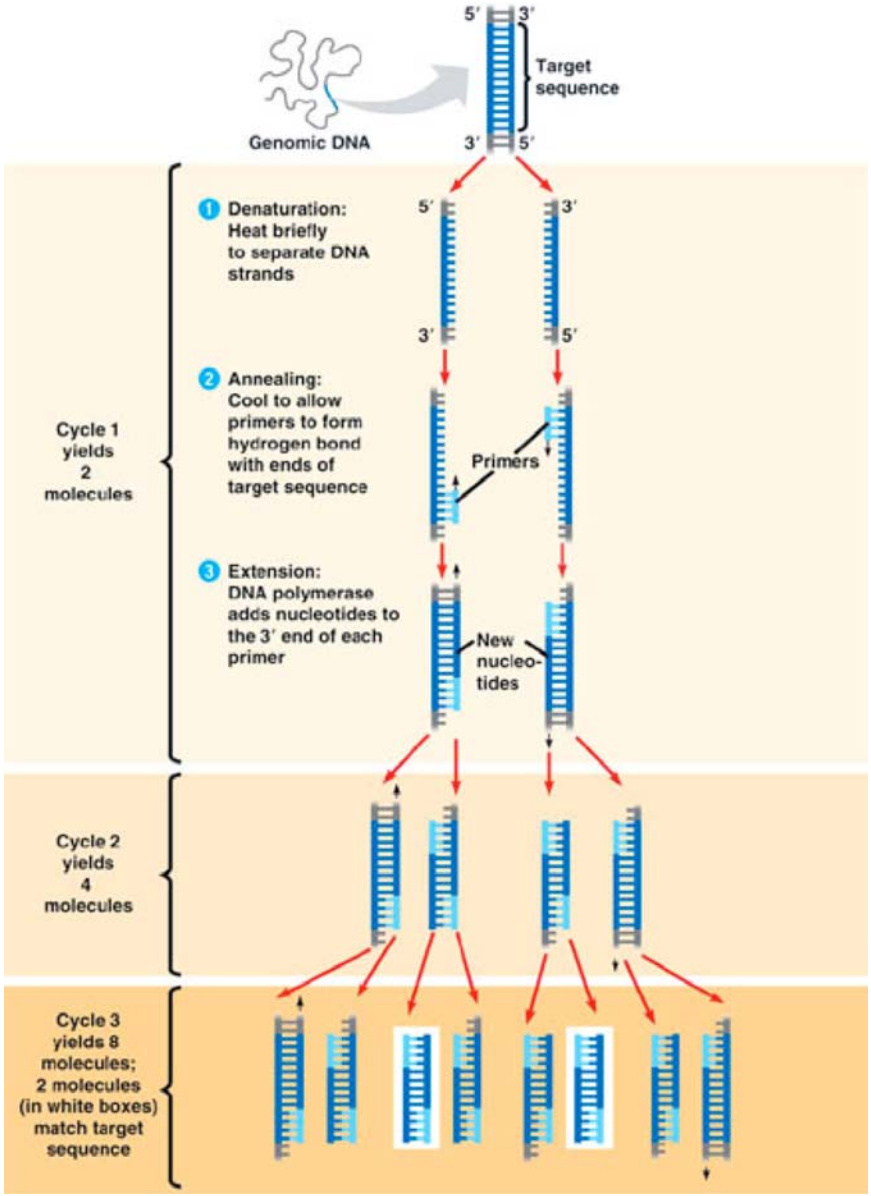
- Thermostable DNA polymerase named after the thermophilic bacterium *Thermus aquaticus*
- Originally isolated by Thomas D. Brock in 1965
- Molecule of the 80s
- Many versions of these polymerases are available
- Modified for increased fidelity



Schematic outline of a typical PCR cycle

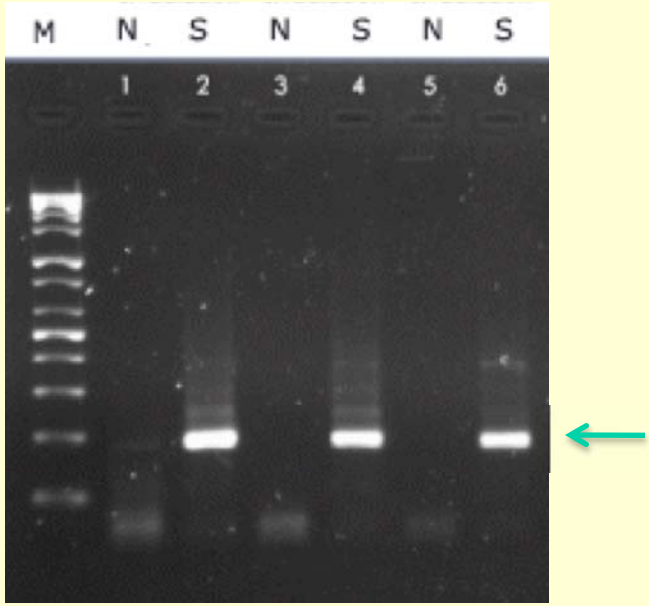


PCR


5′
3′
Target sequence
Genomic DNA
3′
5′
1 Denaturation: Heat briefly to separate DNA strands
5′
3′
3′
5′
2 Annealing: Cool to allow primers to form hydrogen bond with ends of target sequence
Primers
3 Extension: DNA polymerase adds nucleotides to the 3′ end of each primer
New nucleo-tides
Cycle 1 yields 2 molecules
Cycle 2 yields 4 molecules
Cycle 3 yields 8 molecules; 2 molecules (in white boxes) match target sequence

Gel Electrophoresis

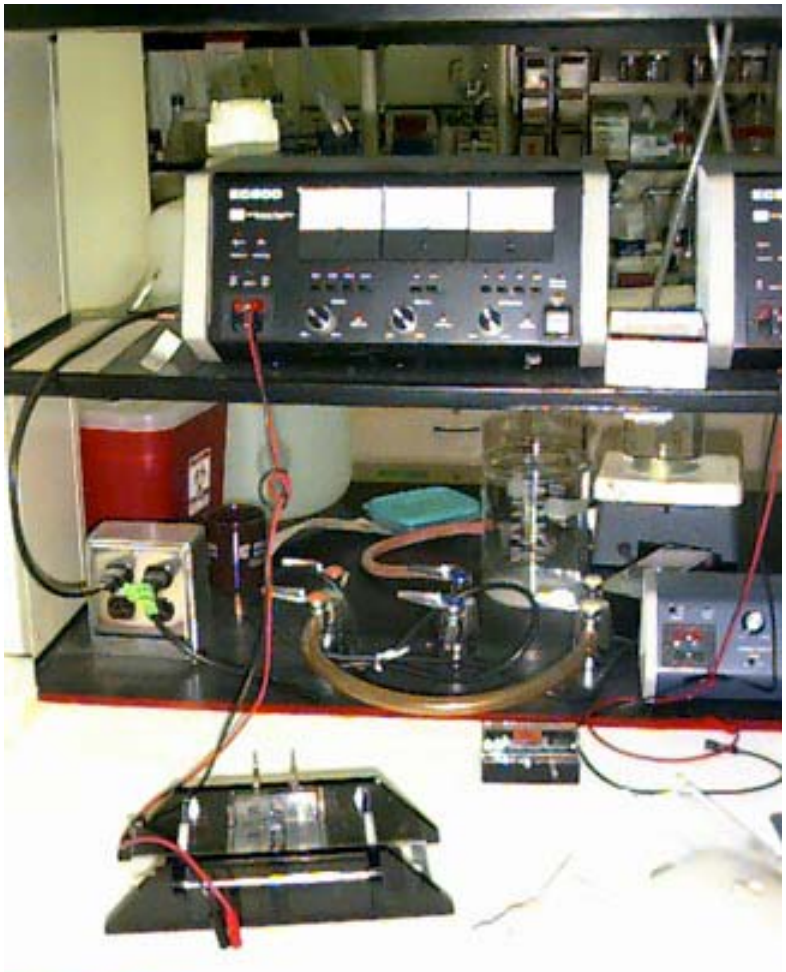
- Used to measure the size of DNA fragments.
- When voltage is applied to DNA, different size fragments migrate to different distances (smaller ones travel farther).

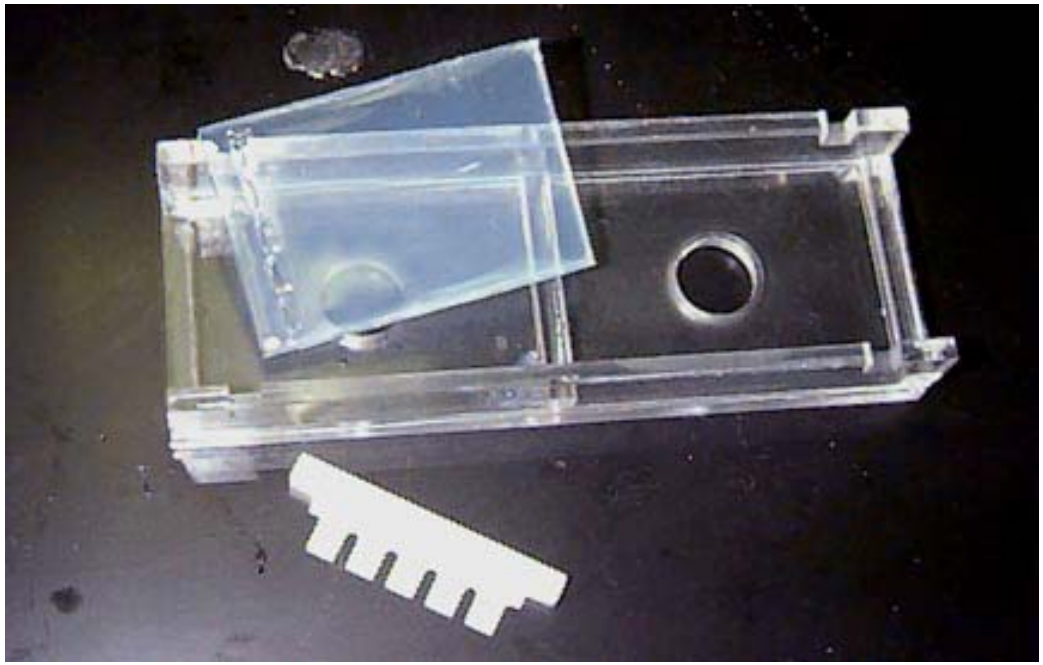


Gel Electrophoresis for DNA

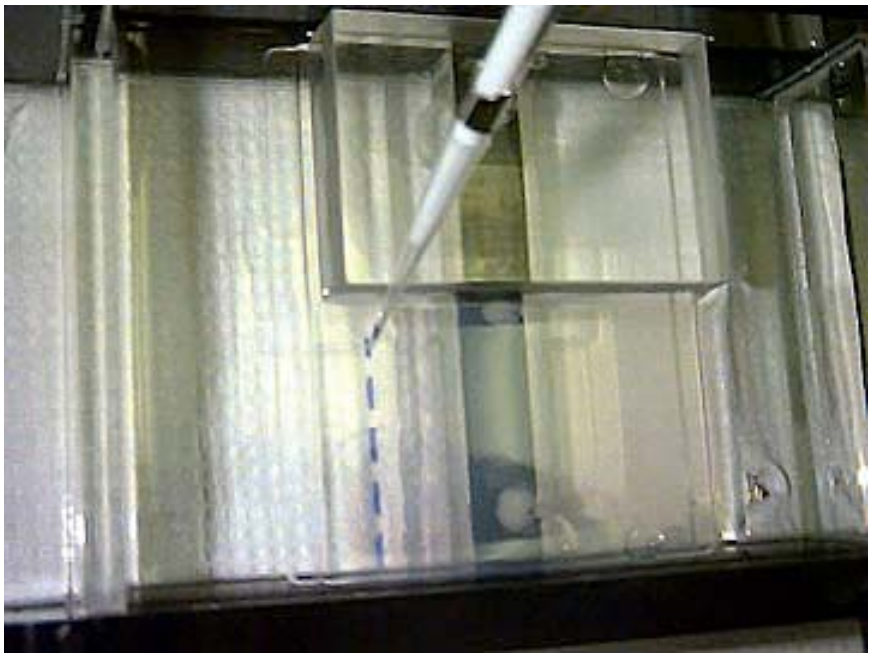
- DNA is negatively charged – WHY?
- DNA can be separated according to its size
- Use a molecular sieve – Gel
- Varying concentration of agarose makes different pore sizes & results
- Boil agarose to cool and solidify/polymerize
- Add DNA sample to wells at the top of a gel
- Add DNA loading dye (color to assess the speed and make it denser than running buffer)
- Apply voltage
- Larger fragments migrate through the pores slower
- Stain the DNA – EtBr, SyberSafe, etc

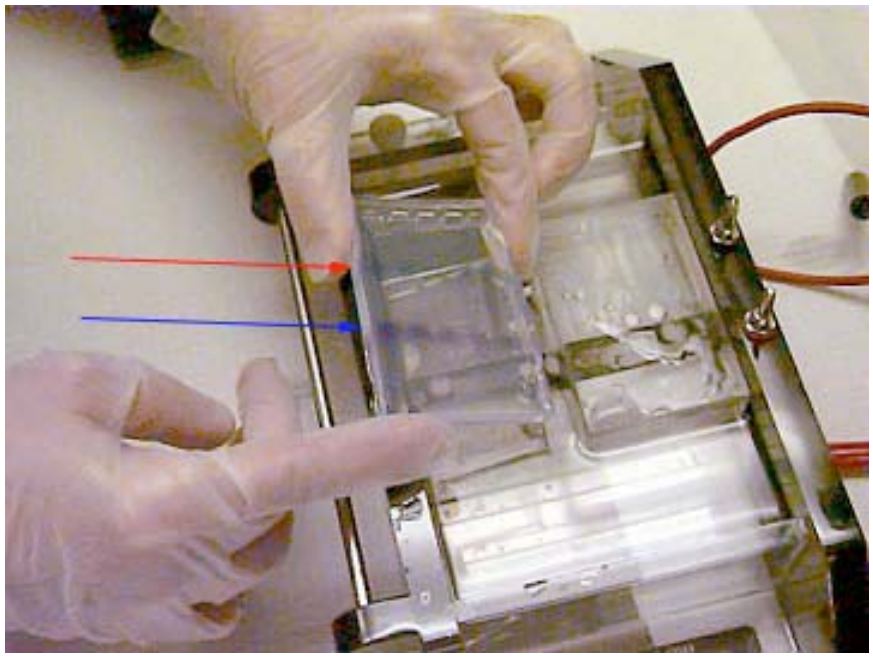
Gel Electrophoresis





Gel Electrophoresis





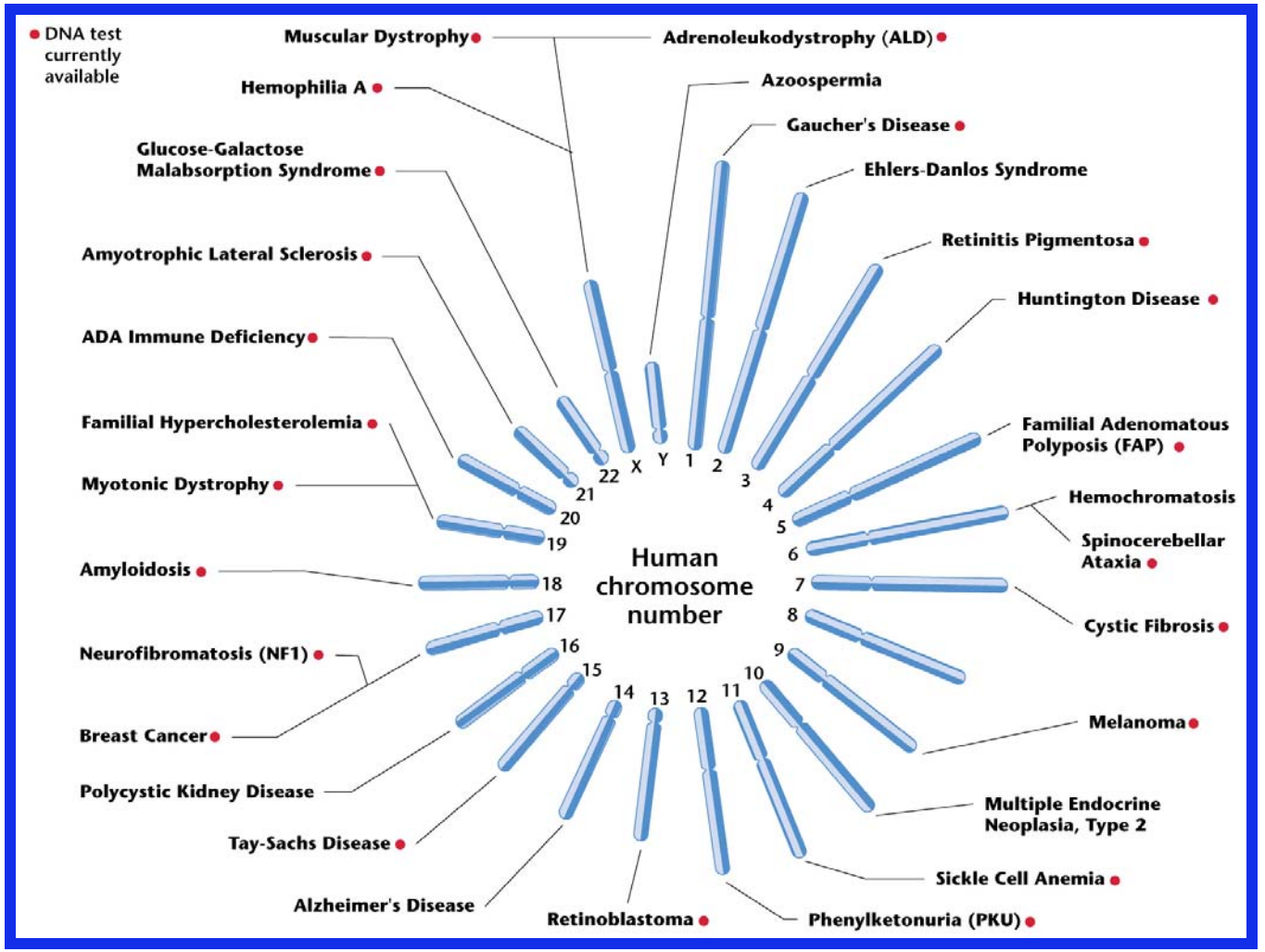
Sequencing

Why sequencing?

- Useful for further study:
  - Locate gene sequences, regulatory elements
  - Compare sequences to find similarities
  - Identify mutations – genetic disorders
  - Use it as a basis for further experiments
  - Better understand the organism
  - Forensics

Next 4 slides contains material prepared by Dr. Stan Metzenberg. Also see:
http://stat-www.berkeley.edu/users/terry/Classes/s260.1998/Week8b/week8b/node9.html

Human Hereditary Diseases

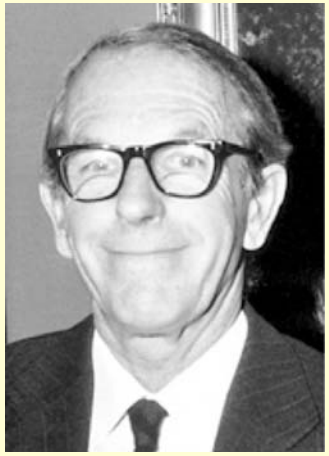


Those inherited conditions that can be diagnosed using DNA analysis are indicated by a (•)

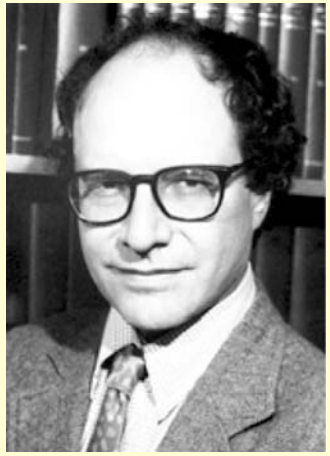
History

- Two methods independently developed in 1974
  - Maxam & Gilbert method
  - Sanger method: became the standard
- Nobel Prize in 1980

Insulin; Sanger, 1958



Sanger

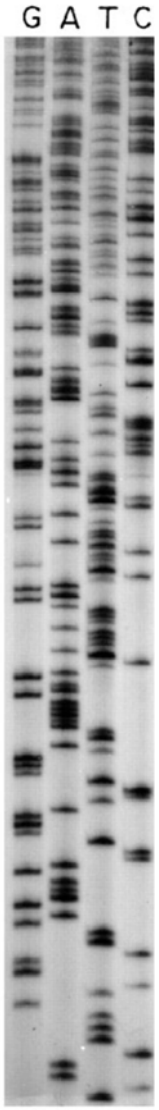


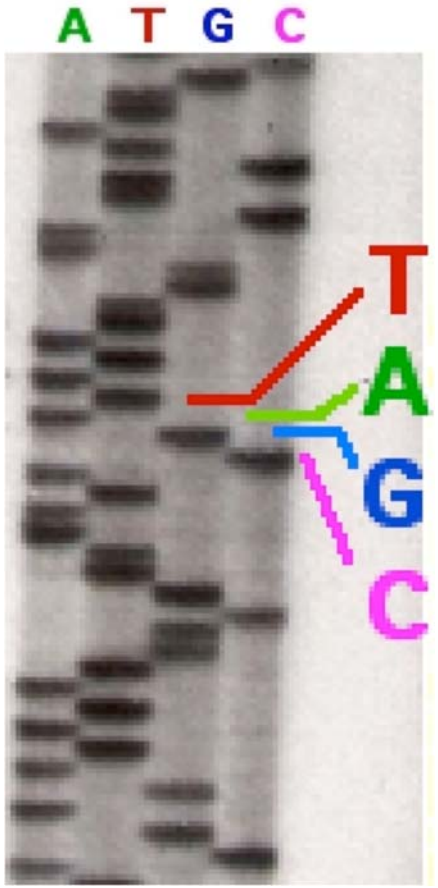
Gilbert

Original Sanger Method

- (Labeled) Primer is annealed to template strand of denatured DNA. This primer is specifically constructed so that its 3' end is located next to the DNA sequence of interest. Once the primer is attached to the DNA, the solution is divided into four tubes labeled "G", "A", "T" and "C". Then reagents are added to these samples as follows:
  - "G" tube: ddGTP, DNA polymerase, and all 4 dNTPs
  - "A" tube: ddATP, DNA polymerase, and all 4 dNTPs
  - "T" tube: ddTTP, DNA polymerase, and all 4 dNTPs
  - "C" tube: ddCTP, DNA polymerase, and all 4 dNTPs
- DNA is synthesized, & nucleotides are added to growing chain by the DNA polymerase. Occasionally, a ddNTP is incorporated in place of a dNTP, and the chain is terminated. Then run a gel.
- All sequences in a tube have same prefix and same last nucleotide.

Sequencing Gel



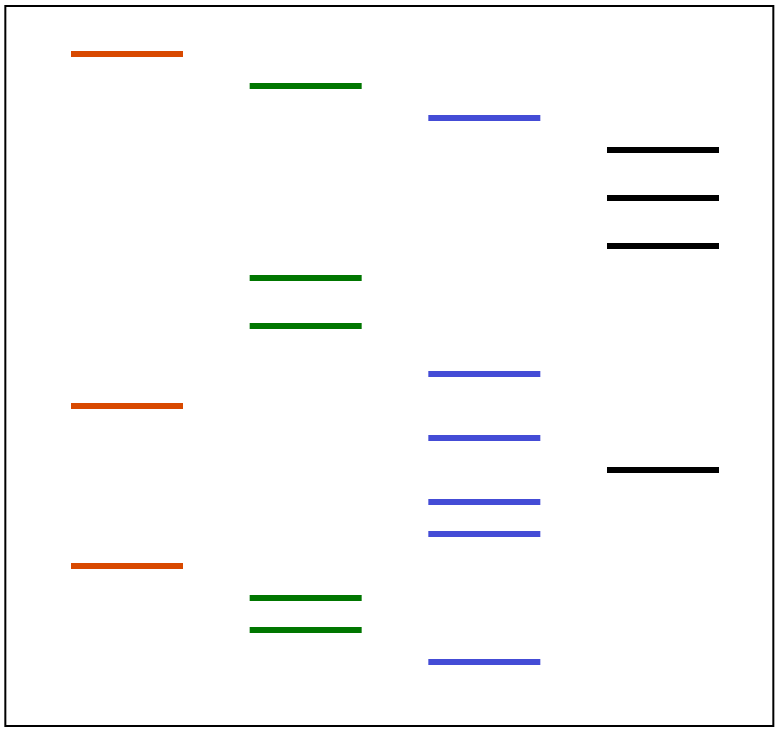


Modified Sanger

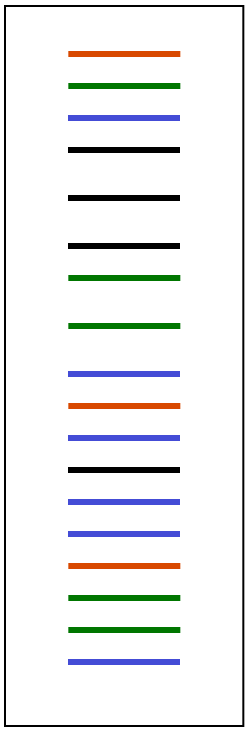
- Reactions performed in a single tube containing all four ddNTP's, each labeled with a different color fluorescent dye

Sequencing Gels: Separate vs Single Lanes

GCCAGGTGAGCCTTTGCA



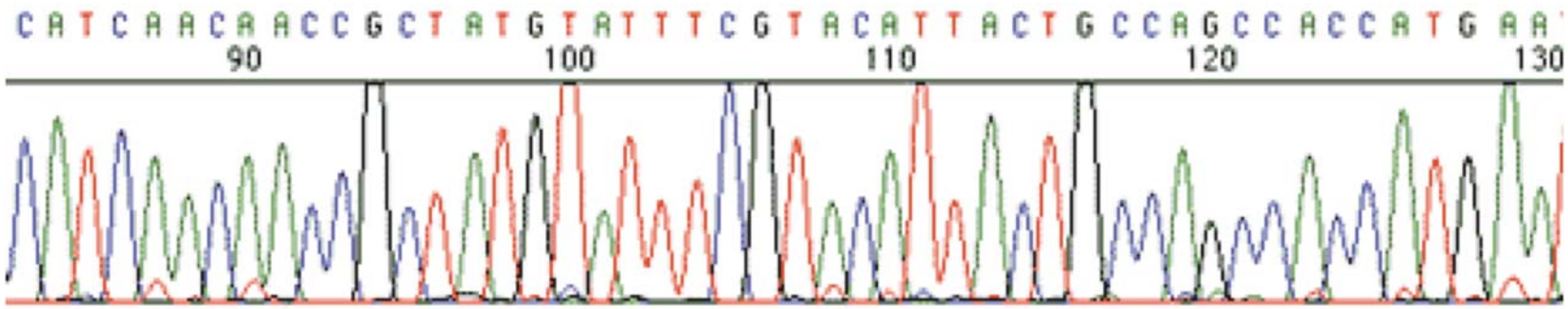
A C G T



Automated Sequencing Instruments

Sequencing

- Flourescence sequencer
- Computer detects specific dye
- Peak is formed
- Base is detected
- Computerized



Maxam-Gilbert Sequencing

- Not popular
- Involves putting copies of the nucleic acid into separate test tubes
- Each of which contains a chemical that will cleave the molecule at a different base (either adenine, guanine, cytosine, or thymine)
- Each of the test tubes contains fragments of the nucleic acid that all end at the same base, but at different points on the molecule where the base occurs.
- The contents of the test tubes are then separated by size with gel electrophoresis (one gel well per test tube, four total wells), the smallest fragments will travel the farthest and the largest will travel the least far from the well.
- The sequence can then be determined from the picture of the finished gel by noting the sequence of the marks on the gel and from which well they came from.

Human Genome Project

Play the Sequencing Video:

- **Download Windows file from http://www.cs.fiu.edu/~giri/teach/6936/Papers/Sequence.exe**
- **Then run it on your PC.**

Human Genome Project

1980 The sequencing methods were sufficiently developed

International collaboration was formed: International Human Genome Consortium of 20 groups - a Public Effort (James Watson as the chair!)

Estimated expense: $3 billion dollars and 15 years

Part of this project is to sequence: *E. coli, Sacchromyces cerevisiae, Drosophila melanogaster, Arabidopsis thaliana, Caenorhabdidtis elegans*

- Allow development of the sequencing methods

Got underway in October 1990

Automated sequencing and computerized analysis

Public effort: 150,000 bp fragments into artificial chromosomes (unstable - but progressed)

In three years large scale physical maps were available

Venter vs Collins

National Human Genome Research Institute

Venter's lab in NIH (joined NIH in 1984) is the first test site for ABI automated sequences; he developed strategies (Expressed Sequence Tags - ESTs)

1992 - decided to patent the genes expressed in brain - "Outcry"

Resistance to his idea

Watson publicly made the comment that Venter's technique during senate hearing - "wasn't science - it could be run by monkeys"

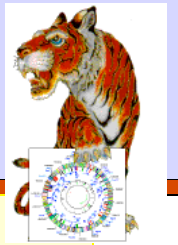
In April 1992 Watson resigned from the HGP

Craig Venter and his wife Claire Fraser left the NIH to set up two companies

- the not-for-profit TIGR The Institute for Genomic Research, Rockville, Md
- A sister company FOR-profit with William Hazeltine - HGSI - Human Genome Sciences Inc., which would commercialize the work of TIGR
- Financed by Smith-Kline Beecham ($125 million) and venture capitalist Wallace Steinberg.

Francis Collins of the University of Michigan replaced Watson as head of NHGRI.

Venter vs Collins



HGSI promised to fund TIGR with $70 million over ten years in exchange for marketing rights TIGR's discoveries

PE developed the automated sequencer & Venter - Whole-genome short-gun approach

"While the NIH is not very good at funding new ideas, once an idea is established they are extremely good," Venter

In May 1998, Venter, in collaboration with Michael Hunkapiller at PE Biosystems (aka Perkin Elmer / Applied Biosystems / Applera), formed Celera Genomics

Goal: sequence the entire human genome by December 31, 2001 - 2 years before the completion by the HGP, and for a mere $300 million

April 6, 2000 - Celera announces the completion "Cracks the human code"

Agrees to wait for HGP

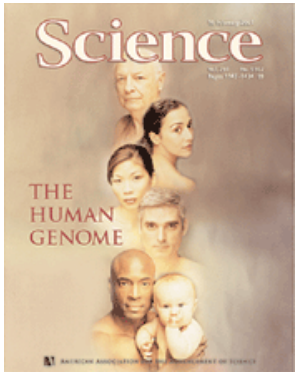
Summer 2000 - both groups announced the rough draft is ready

Human Genome Sequence

6 months later it was published - 5 years ahead of schedule with $ 3 billion dollars

50 years after the discovery of DNA structure

Human Genome Project was completed - 3.1 billion basepairs



Pros: No guessing of where the genes are

Study individual genes and their contribution

Understand molecular evolution

Risk prediction and diagnosis

Con: Future Health Diary --> physical and mental

Who should be entrusted? Future Partners, Agencies, Government

Right to "Genetic Privacy"

Modern Sequencing methods

- 454 Sequencing (60Mbp/run) [Rosch]
- Solexa Sequencing (600Mbp/run) [Illumina]

Compare to

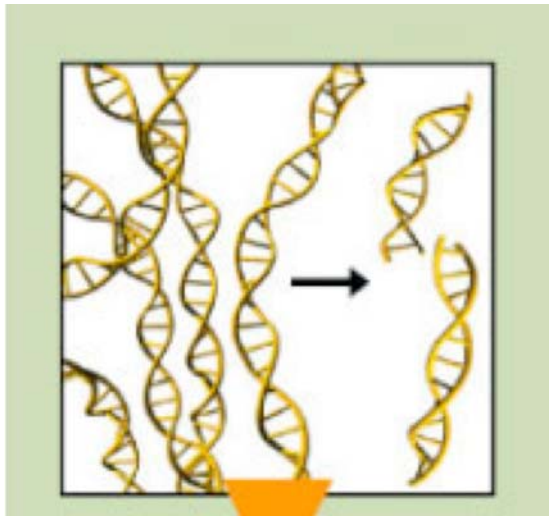
- Sanger Method (70Kbp/run)
- Short Gun Sequencing (??)

454 Sequencing: New Sequencing Technology

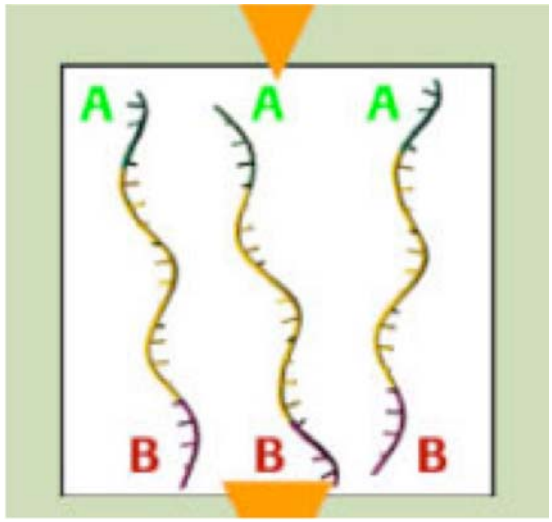
- 454 Life Sciences, Roche
- Sequencing by synthesis - pyrosequencing
- Parallel pyrosequenicng
- Fast (20 million bases per 4.5 hour run)
- Low cost (lower than Sanger sequencing)
- Simple (entire bacterial genome in on day with one person -- without cloning and colony picking)
- Convenient (complete solution from sample prep to assembly)
- PicoTiterPlate Device
  - Fiber optic plate to transmit the signal from the sequencing reaction
- Process:
  - Library preparation: Generate library for hundreds of sequencing runs
  - Amplify: PCR single DNA fragment immobilized on bead
  - Sequencing: "Sequential" nucleotide incorporation converted to chemilluminscent signal to be detected by CCD camera.

454 Sequening

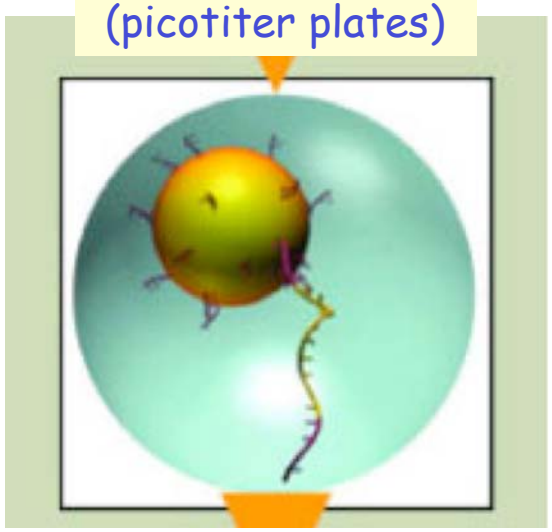
Fragment



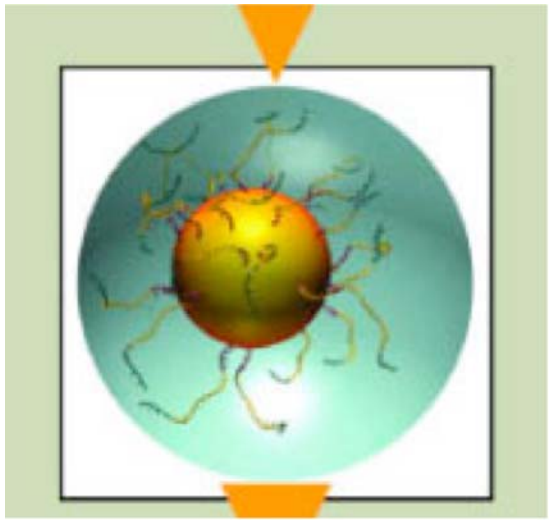
Add Adaptors



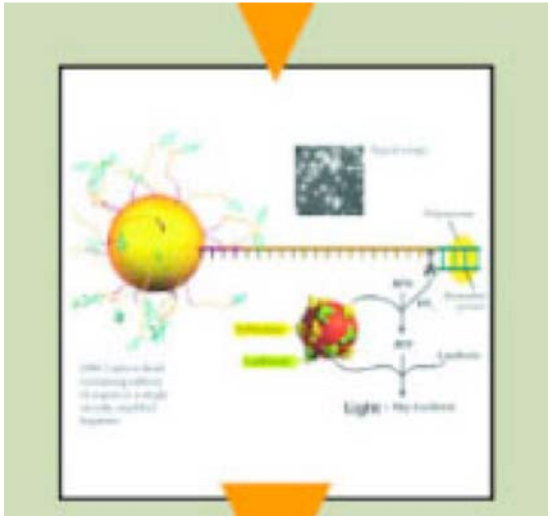
1 fragment-1 bead
(picotiter plates)



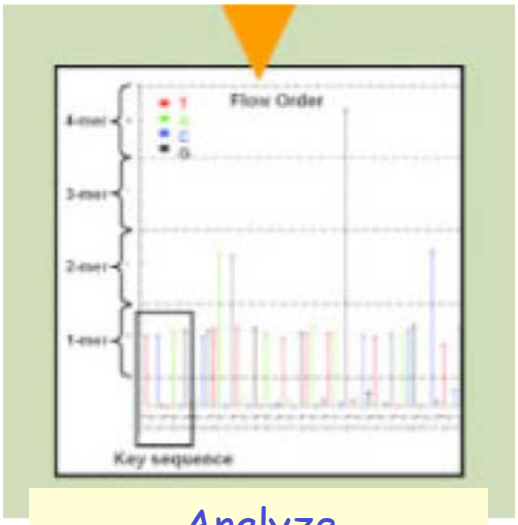
emPCR on bead



Sequence

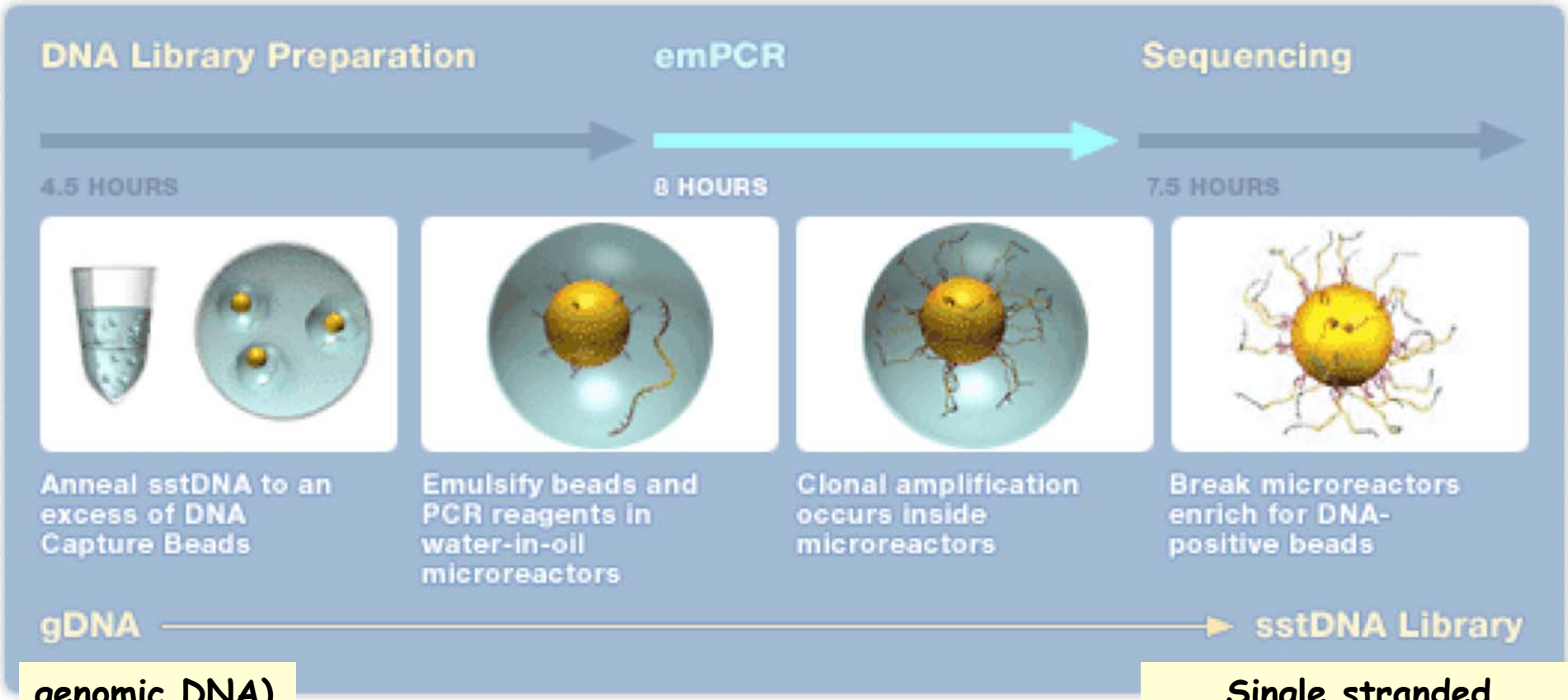


Analyze
one bead - one read)



emPCR

FIGURE 8

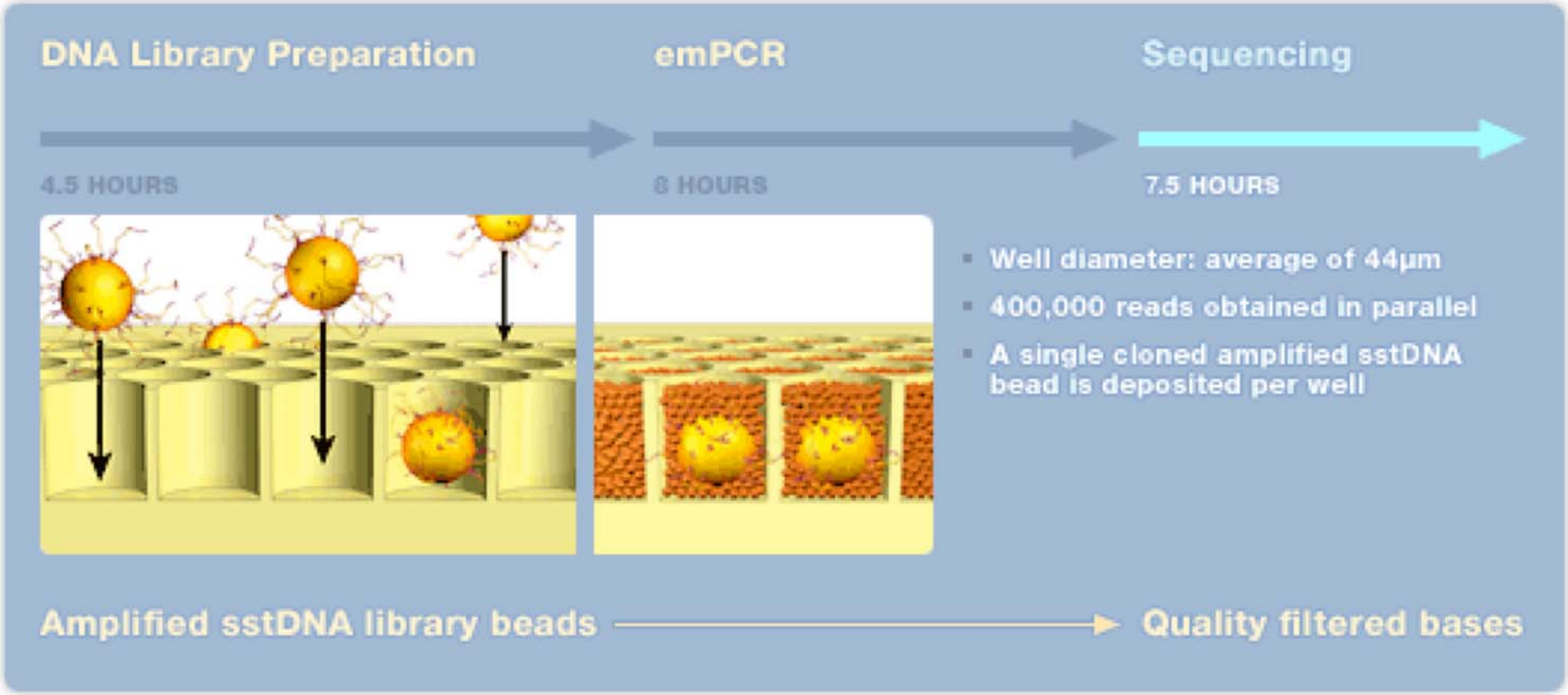


genomic DNA)

Single stranded template DNA library

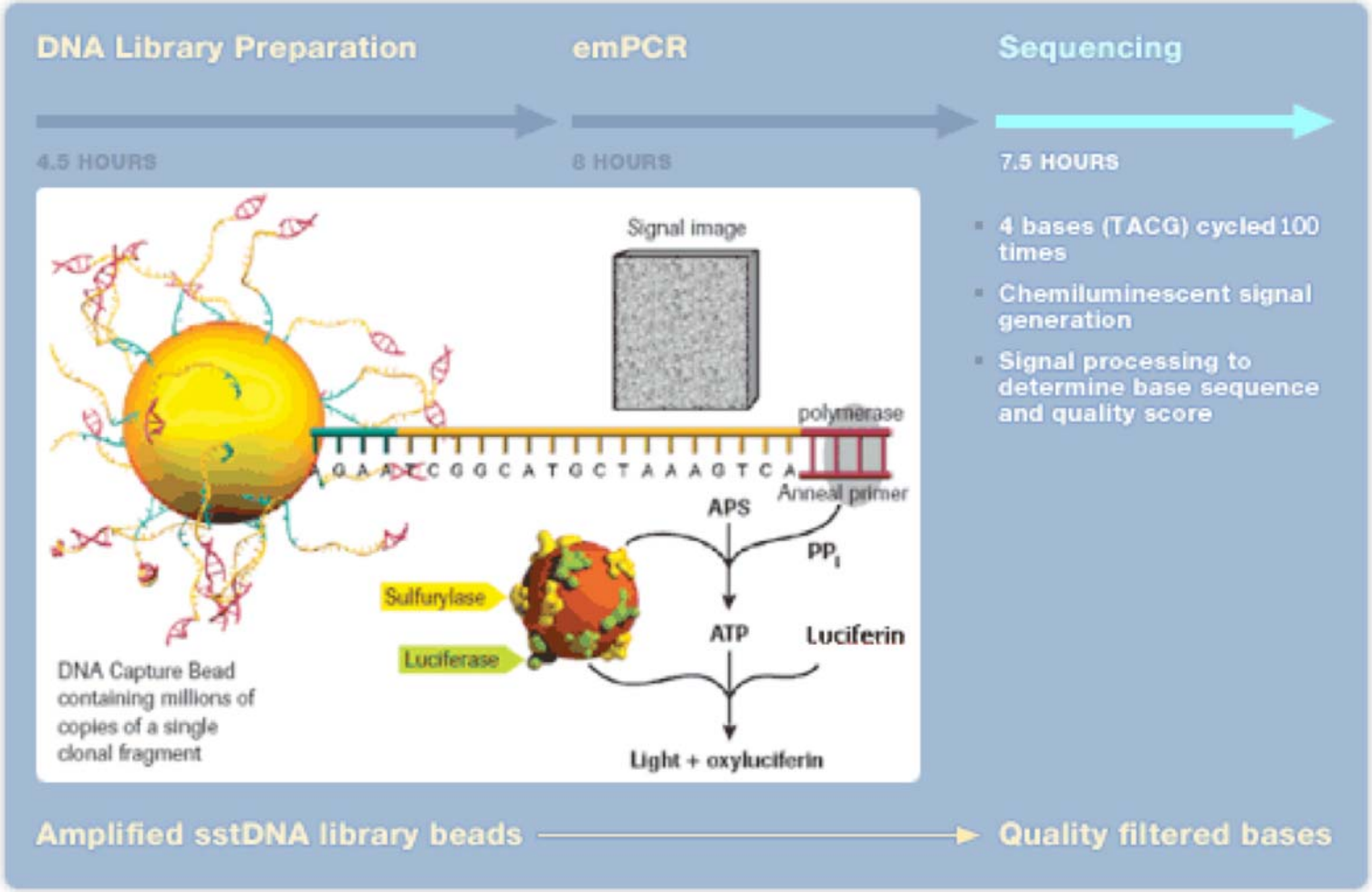
Sequencing

FIGURE 9

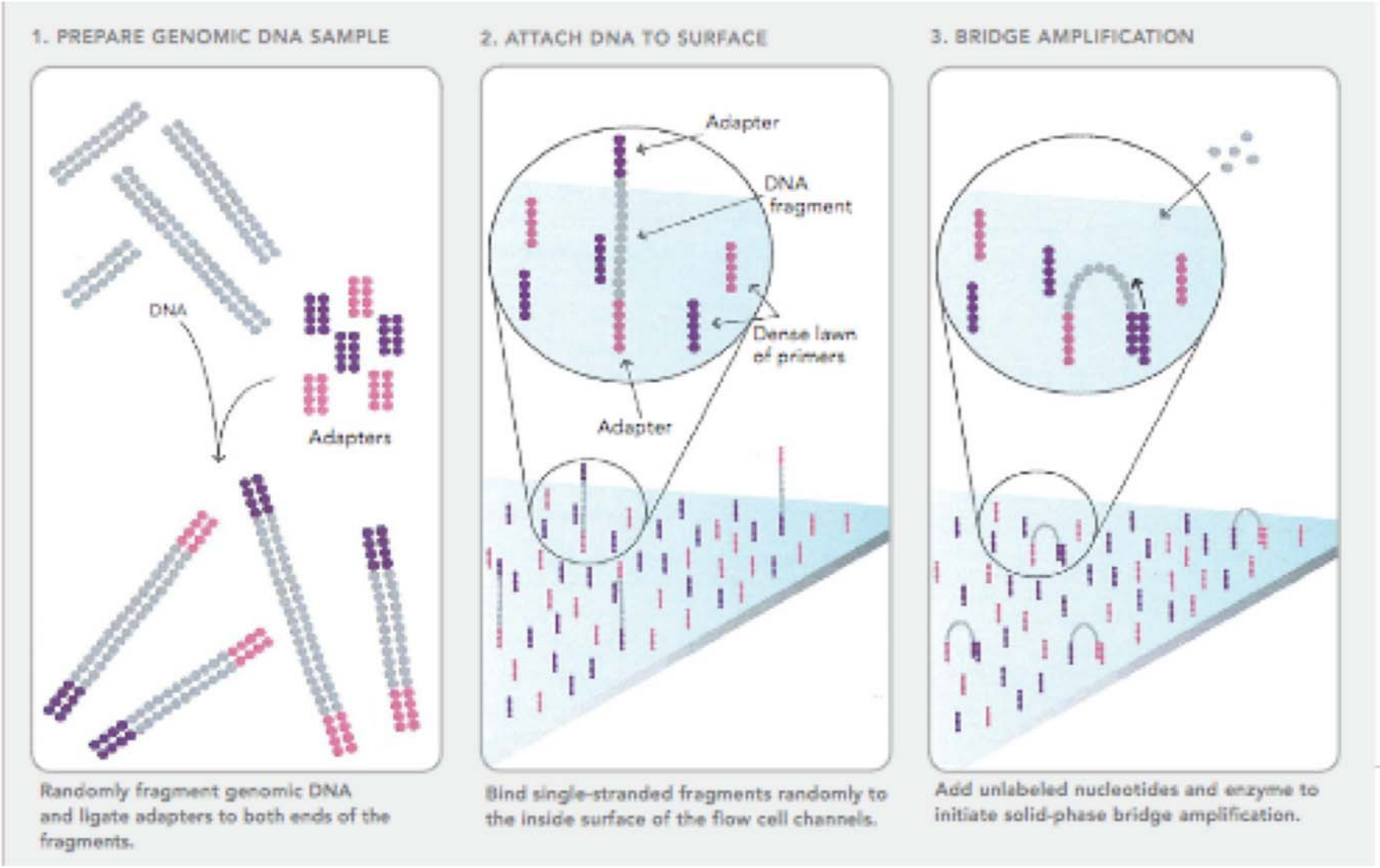


Sequencing

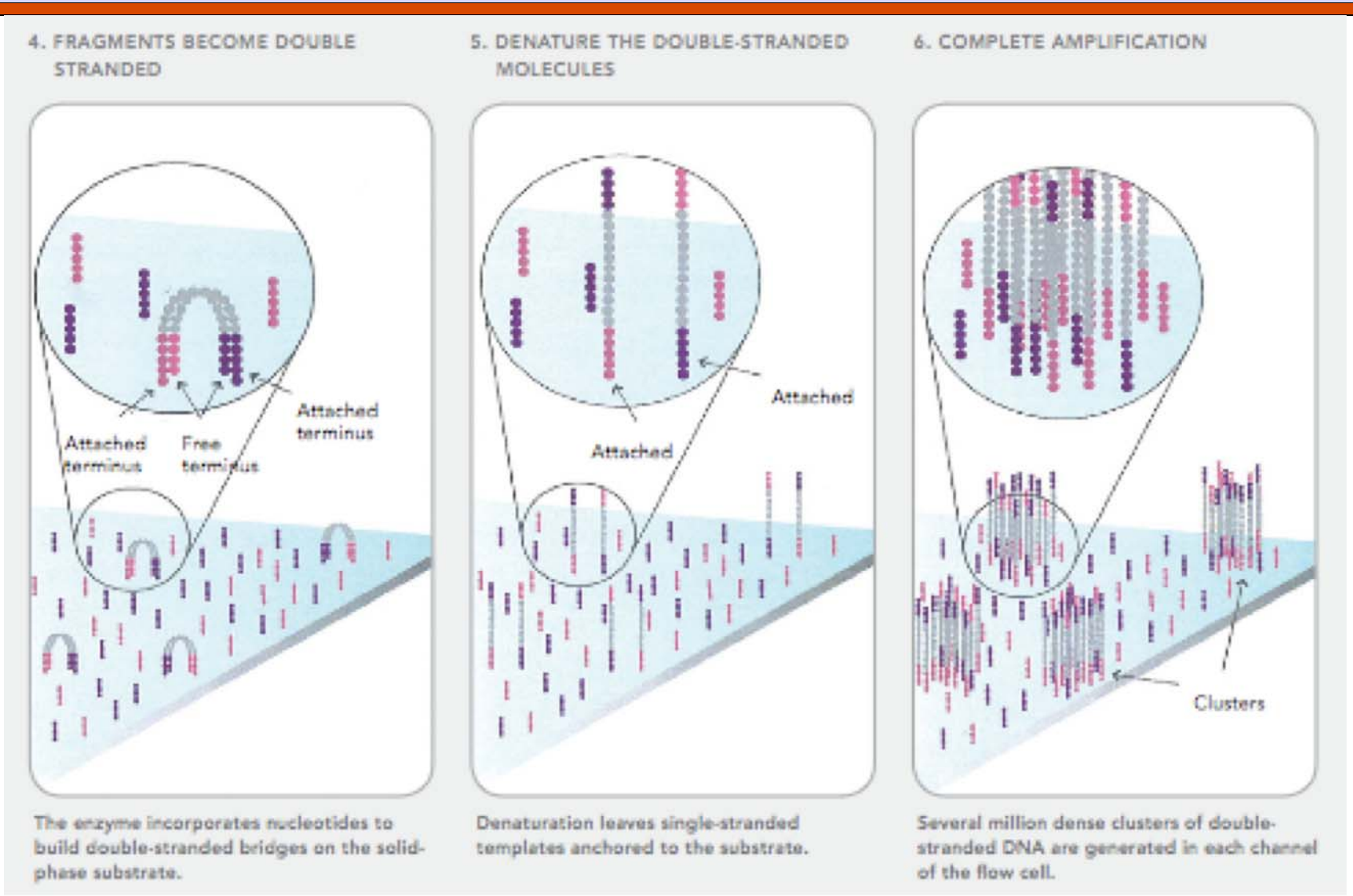
FIGURE 10



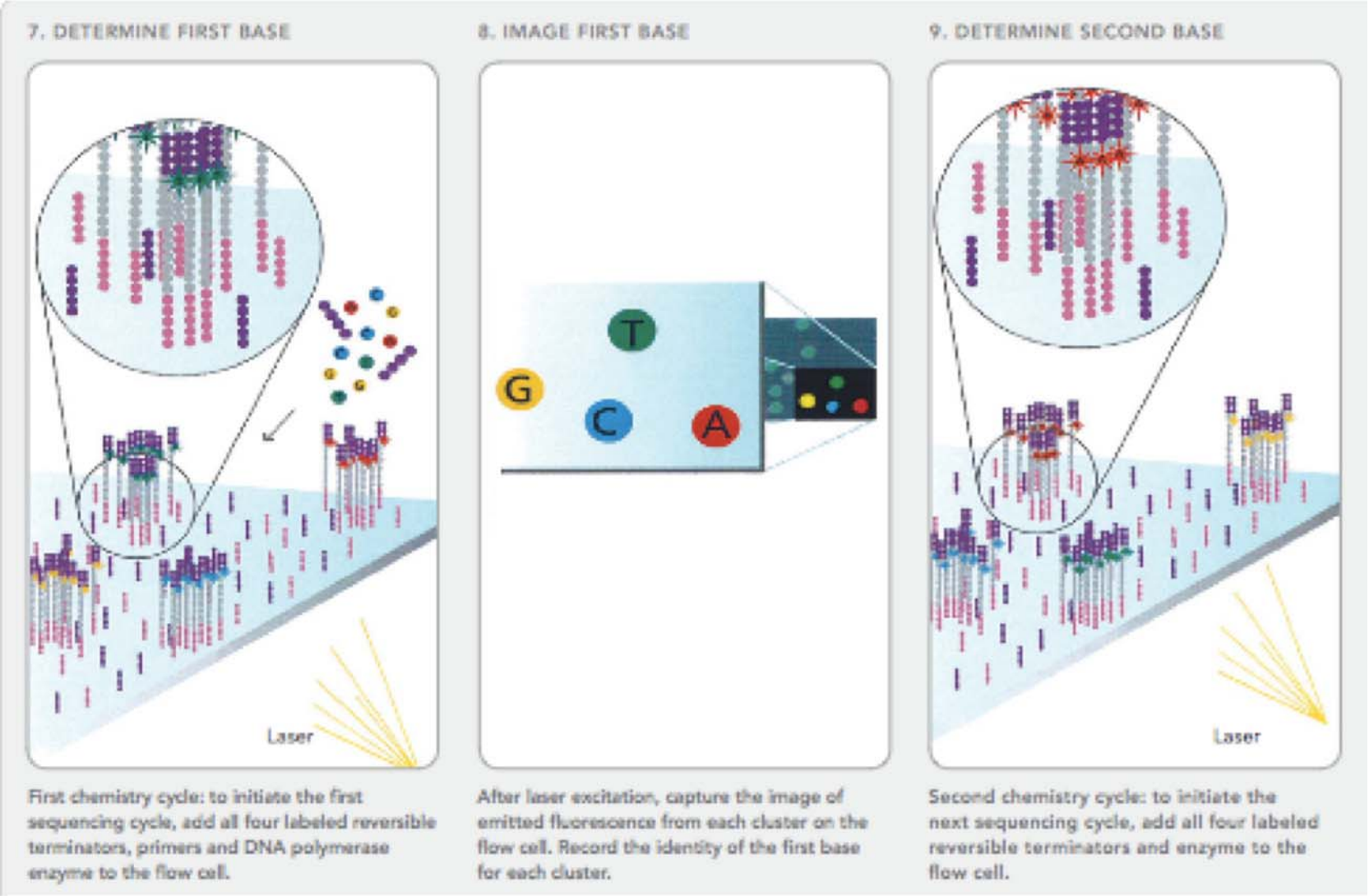
Solexa Sequencing



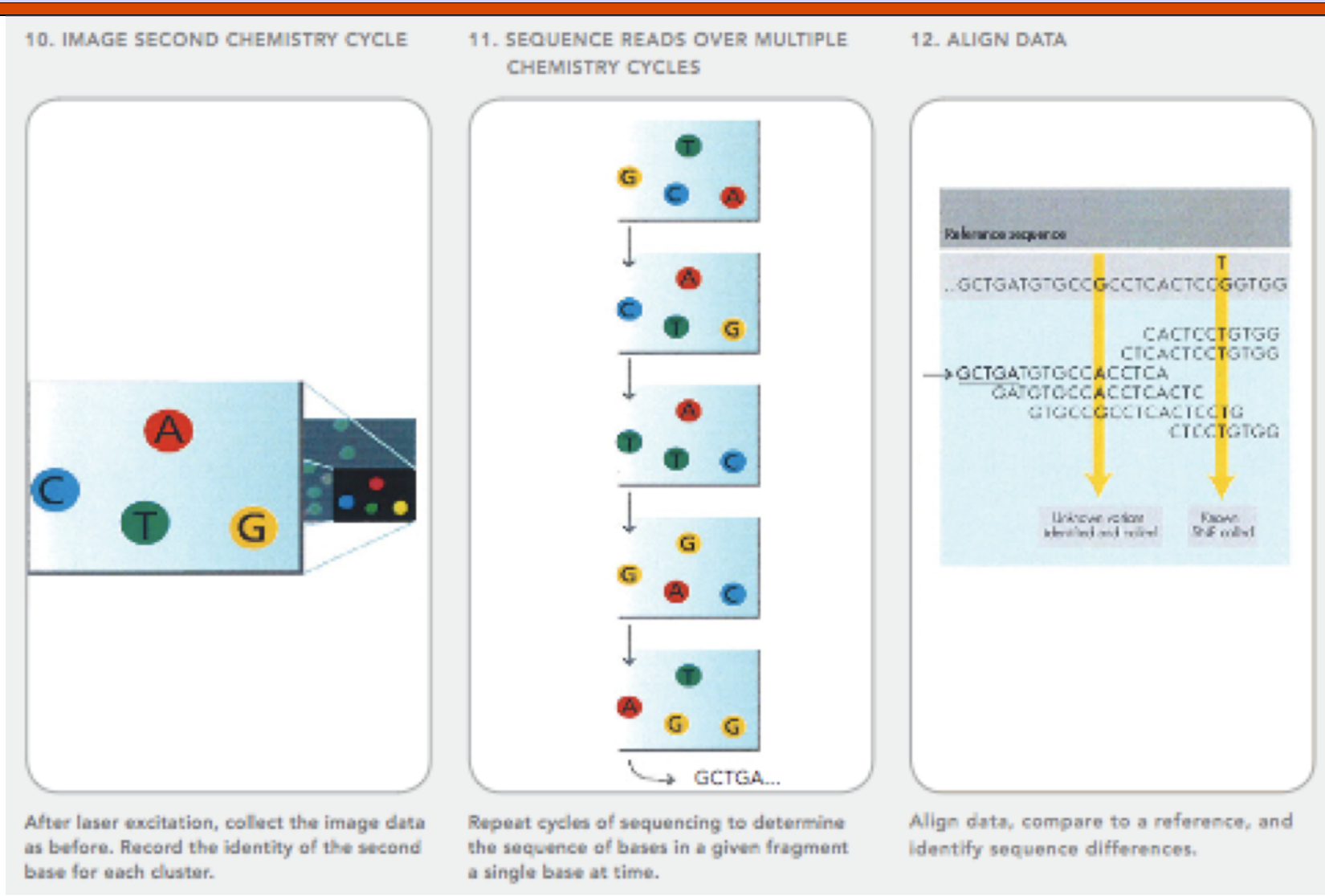
Solexa Sequencing



Solexa Sequencing

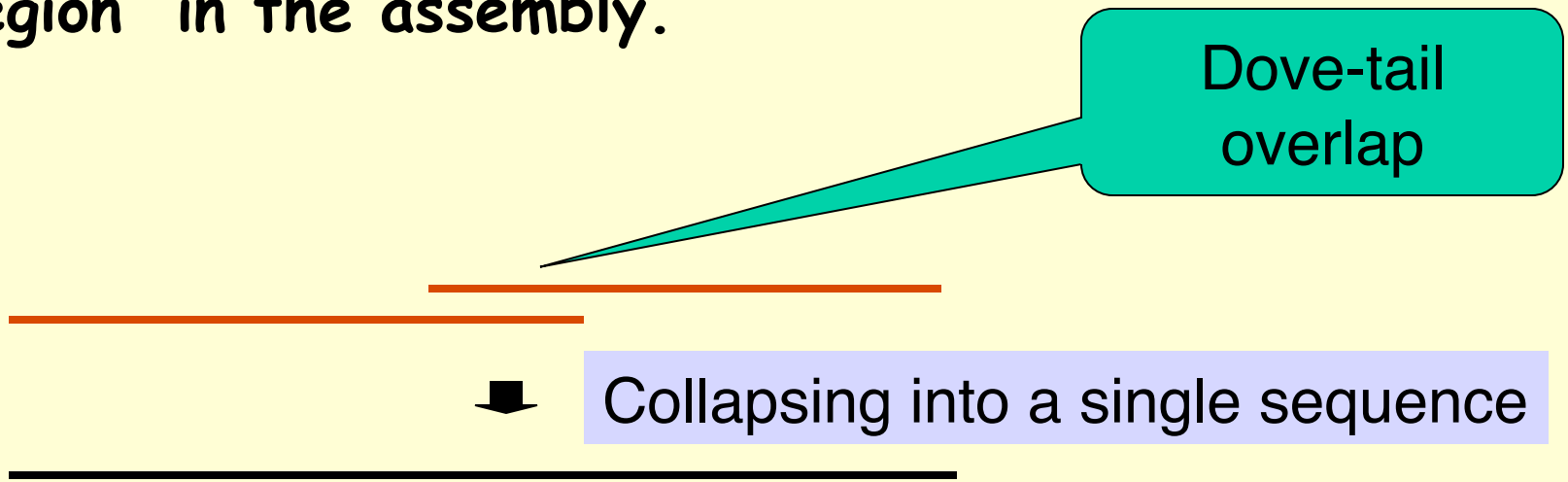


Solexa Sequencing



Sequencing: Generate Contigs

- Short for "contiguous sequence". A continuously covered region in the assembly.



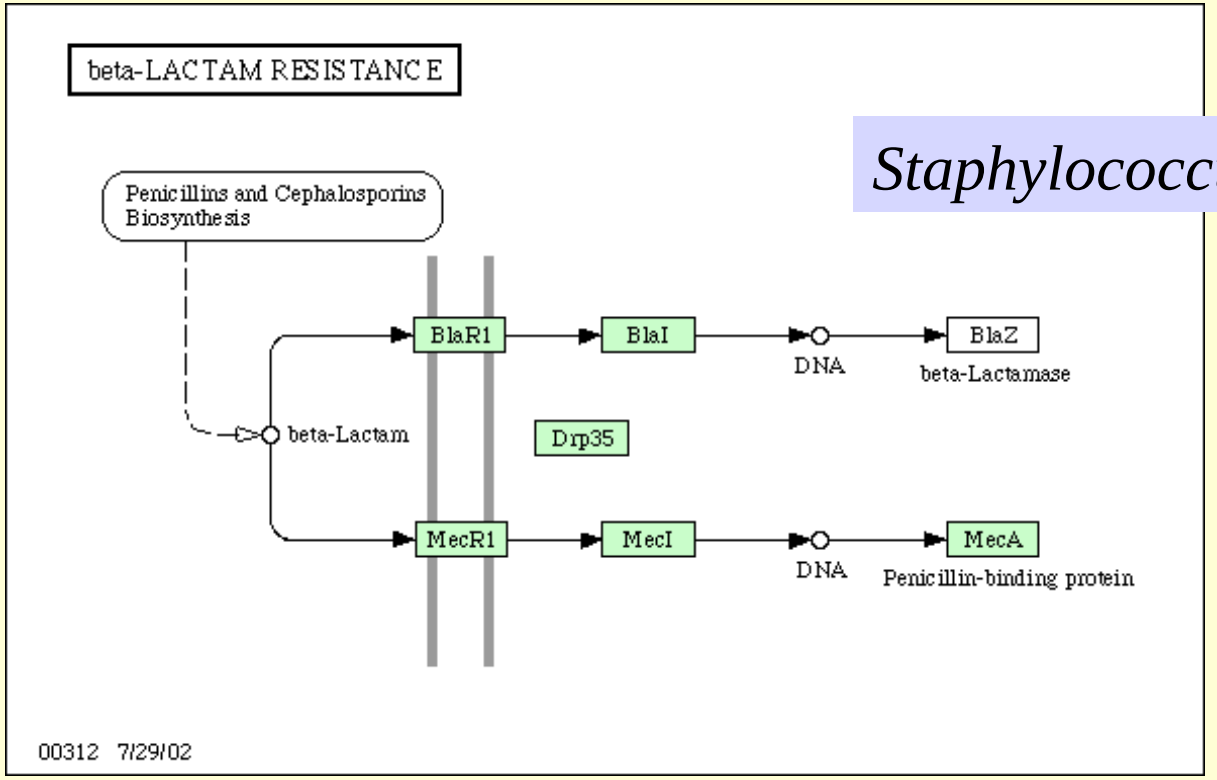
- Jang W et al (1999) Making effective use of human genomic sequence data. Trends Genet. 15(7): 284-6.
  Kent WJ and Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. Genome Res 11(9): 1541-8.

Assembly: Complications

- Errors in input sequence fragments (~3%)
  - Indels or substitutions
- Contamination by host DNA
- Chimeric fragments (joining of non-contiguous fragments)
- Unknown orientation
- Repeats (long repeats)
  - Fragment contained in a repeat
  - Repeat copies not exact copies
  - Inherently ambiguous assemblies possible
  - Inverted repeats
- Inadequate Coverage

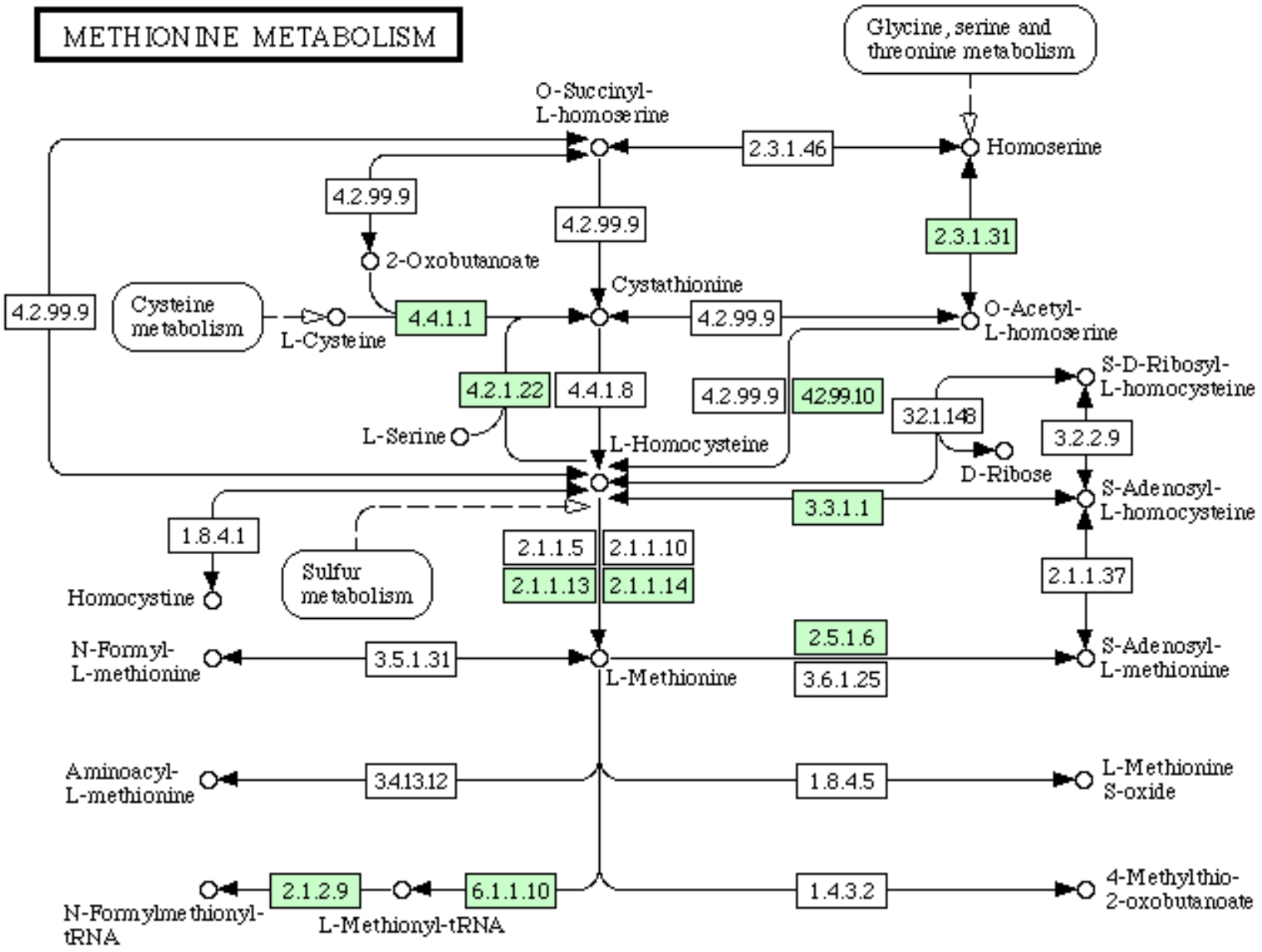
Gene Networks & Pathways

- **Genes & Proteins act in concert and therefore form a complex network of dependencies.**



Staphylococcus aureus

Pseudomonas aeruginosa



Omics

- Genomics: Study of all genes in a genome, or comparison of whole genomes.
  - Whole genome sequencing
- Metagenomics
  - Study of total DNA from a community (sample without separation or cultivation)
- Proteomics: Study of all proteins expressed by a genome
  - What is expressed at a particular time
  - 2D gel electrophoresis & Mass spectrometry
- Transcriptomics
  - Gene expression – mRNA (Microarray)
  - RNA sequencing
- Glycomics
  - Study of carbohydrates/sugars

Applications of NGS

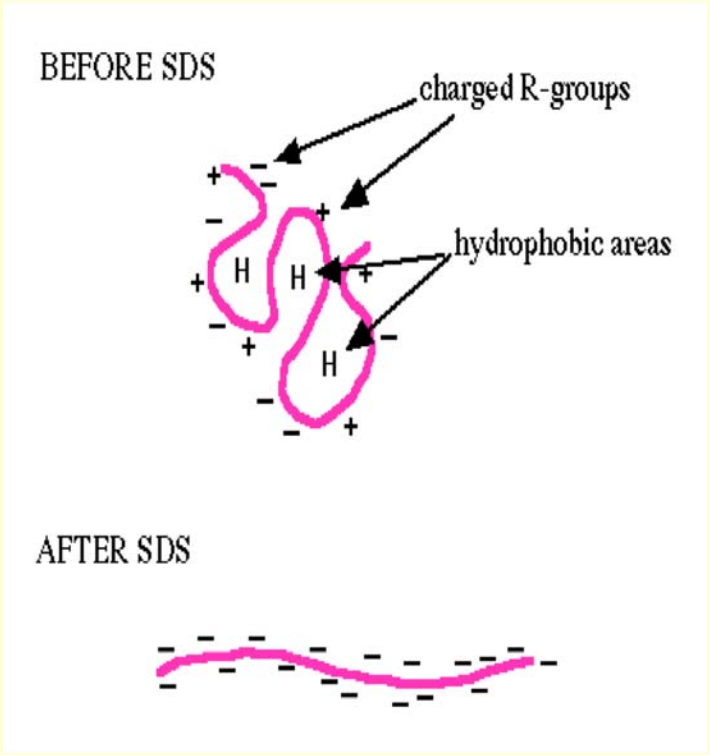
- Sequencing: Study new genomes
- RNA-Seq: Study transcriptomes and gene expression by sequencing RNA mixture
- ChIP-Seq: Analyze protein-binding sites by sequencing DNA precipitated with TF
- Metagenomics: Sequencinng metagenoms
- SNP Analysis: Study SNPs by deep sequencing of regions with SNPs
- Resequencing: Study variations, close gaps, etc.

Protein Sequence

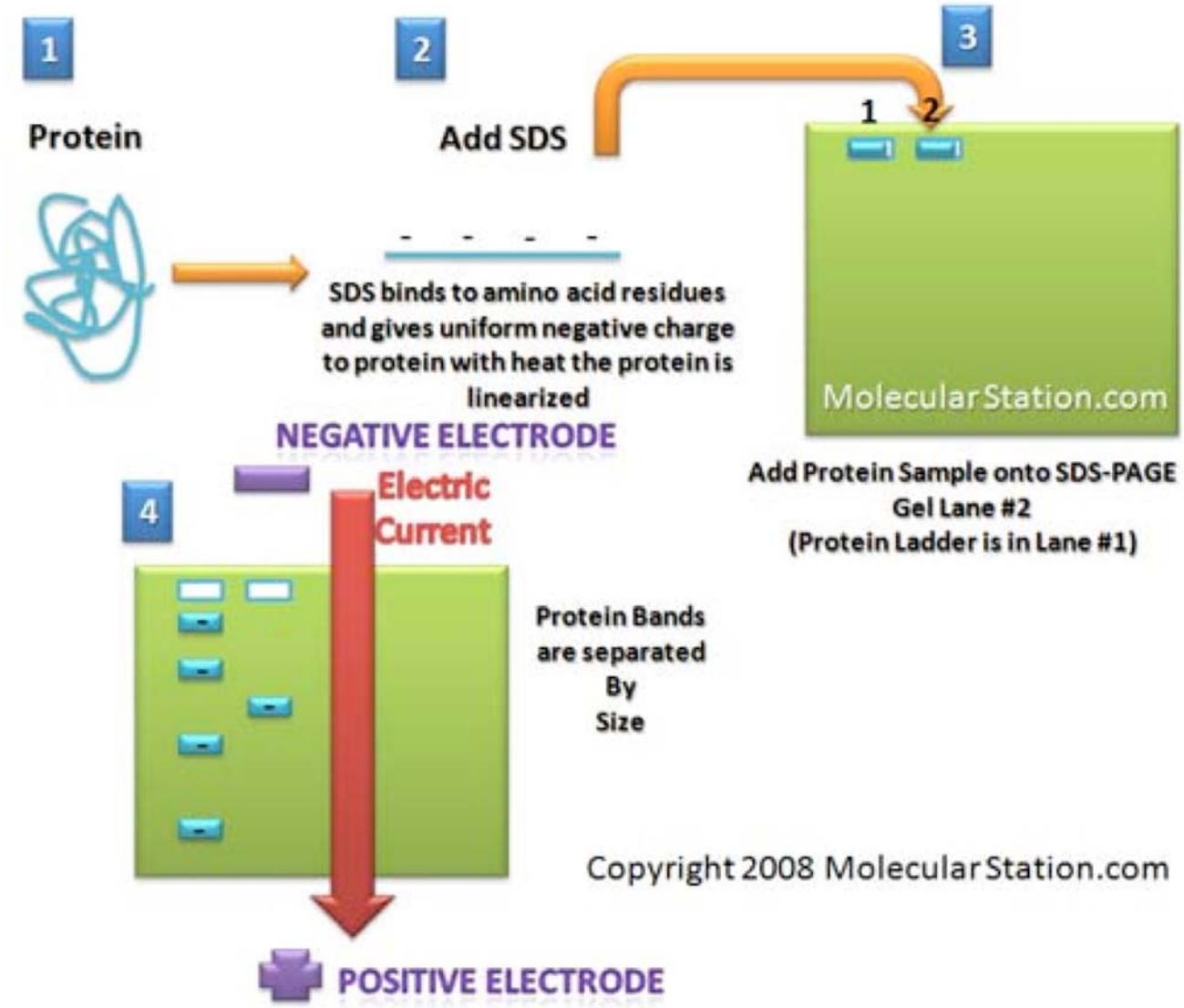
- 20 amino acids
- How is it ordered?
- Basis: Edman Degradation (Pehr Edman)
  - Limited ~30 residues
  - React with Phenylisothiocyanate
  - Cleave and chromatography
- First separate the proteins – Use 2D gels
- Then digest to get pieces
- Then sequence the smaller pieces
- Tedious
- Mass spectrometry

Gel Electrophoresis for Protein

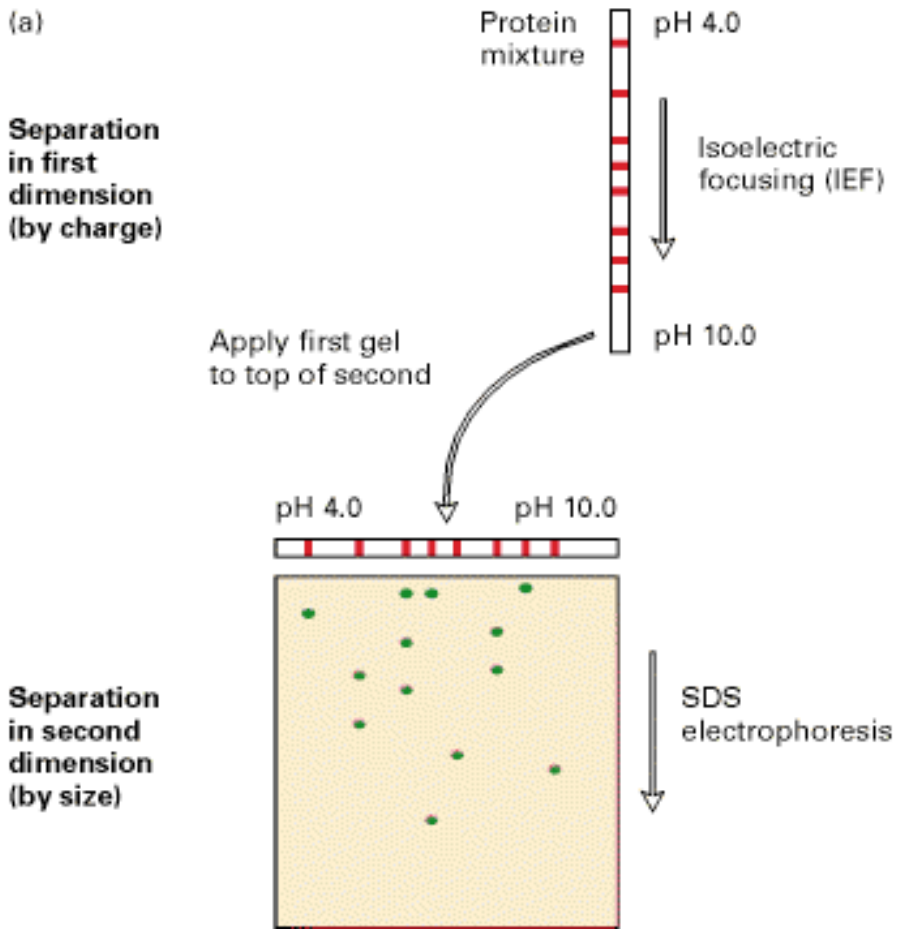
- Protein is also charged
- Has to be denatured - WHY
- Gel: SDS-Polyacrylamide gels
- Add sample to well
- Apply voltage
- Size determines speed
- Add dye to assess the speed
- Stain to see the protein bands



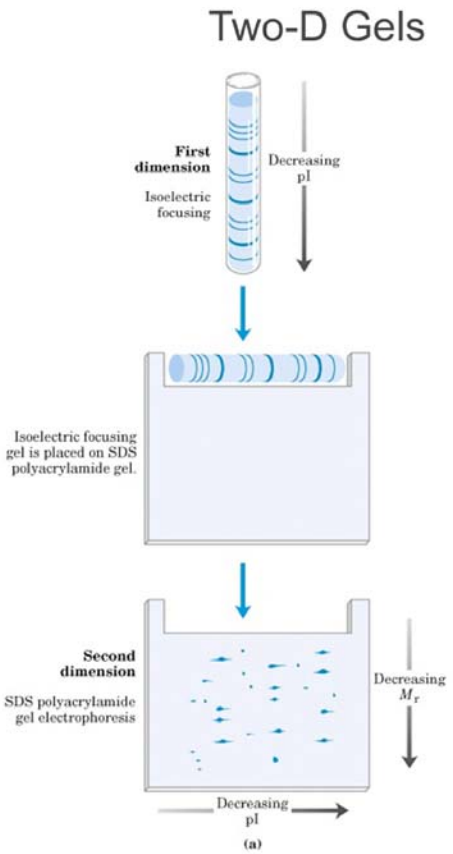
Protein Gel


1
Protein
2
Add SDS
3
1
2
SDS binds to amino acid residues and gives uniform negative charge to protein with heat the protein is linearized
MolecularStation.com
Add Protein Sample onto SDS-PAGE Gel Lane #2 (Protein Ladder is in Lane #1)
NEGATIVE ELECTRODE
4
Electric Current
Protein Bands are separated By Size
Copyright 2008 MolecularStation.com
POSITIVE ELECTRODE

2D-Gels


(a)
Protein mixture
pH 4.0
Separation in first dimension (by charge)
Isoelectric focusing (IEF)
pH 10.0
Apply first gel to top of second
pH 4.0
pH 10.0
Separation in second dimension (by size)
SDS electrophoresis

2D Gel Electrophoresis



(b)

Mass Spectrometry

- **Mass measurements By Time-of-Flight**
  Pulses of light from laser ionizes protein that is absorbed on metal target. Electric field accelerates molecules in sample towards detector. The time to the detector is inversely proportional to the mass of the molecule. Simple conversion to mass gives the molecular weights of proteins and peptides.

- **Using Peptide Masses to Identify Proteins:**
  One powerful use of mass spectrometers is to identify a protein from its peptide mass fingerprint. A peptide mass fingerprint is a compilation of the molecular weights of peptides generated by a specific protease. The molecular weights of the parent protein prior to protease treatment and the subsequent proteolytic fragments are used to search genome databases for any similarly sized protein with identical or similar peptide mass maps. The increasing availability of genome sequences combined with this approach has almost eliminated the need to chemically sequence a protein to determine its amino acid sequence.

Mass Spectrometry