# CAP 5510: Introduction to Bioinformatics
# CGS 5166: Bioinformatics Tools

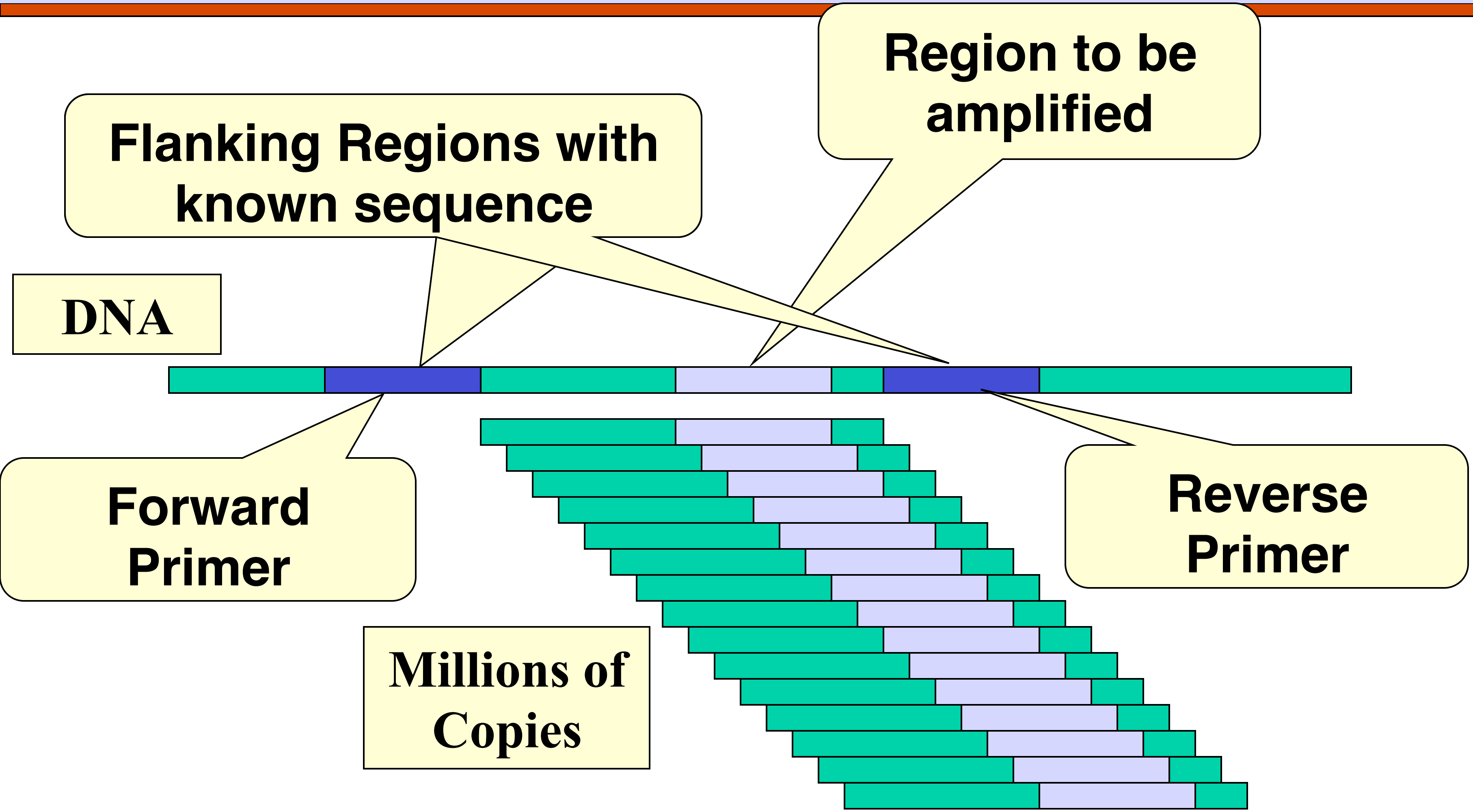
Giri Narasimhan
ECS 254; Phone: x3748
giri@cis.fiu.edu
www.cis.fiu.edu/~giri/teach/BioinfS15.html

# PCR and Sequencing

# Polymerase Chain Reaction (PCR)

- For testing, large amount of DNA is needed
  - Identifying individuals for forensic purposes
    - (0.1 microliter of saliva contains enough epithelial cells)
  - Identifying pathogens (viruses and/or bacteria)
- 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

# PCR



Flanking Regions with known sequence

Region to be amplified

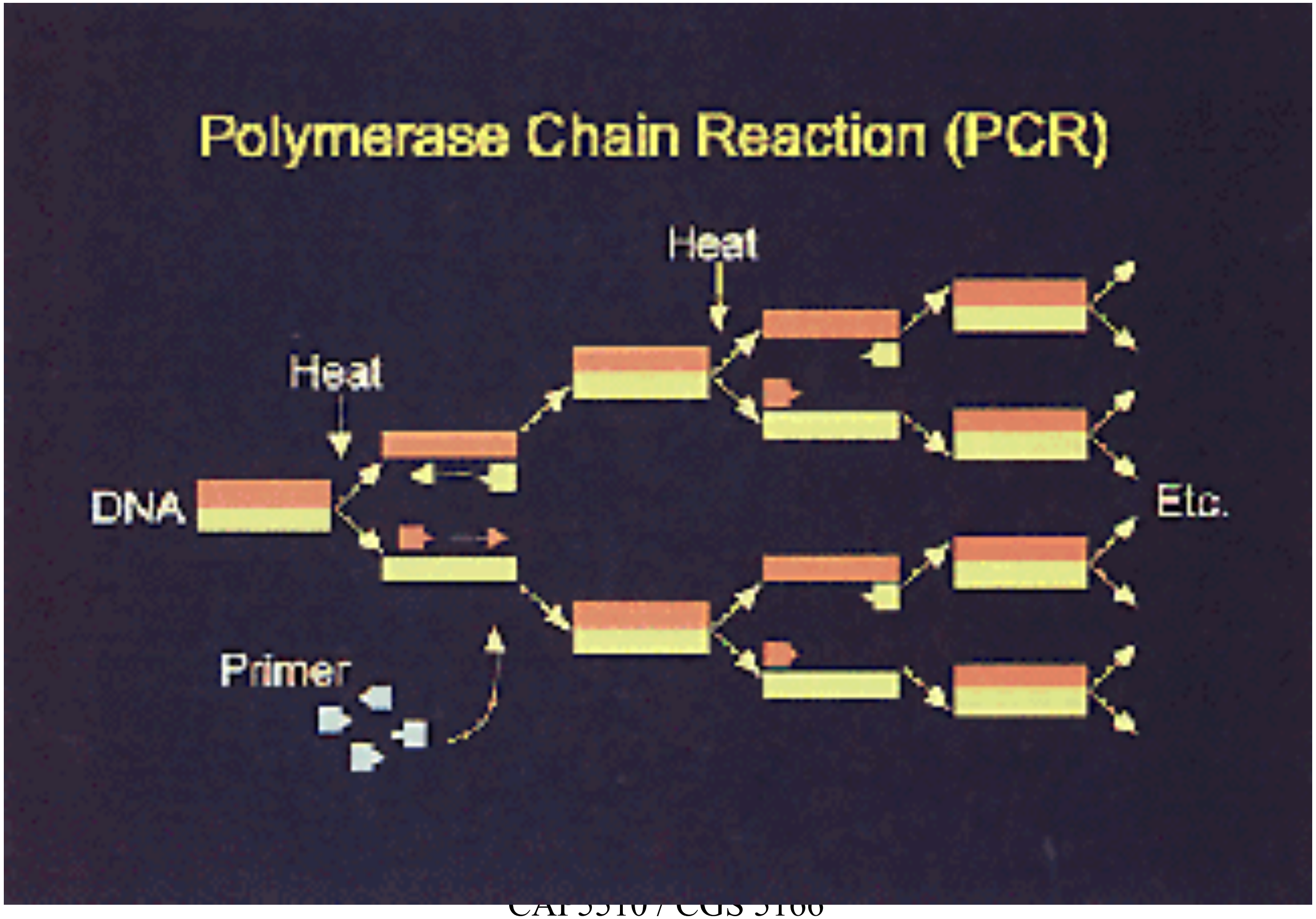
DNA

Forward Primer

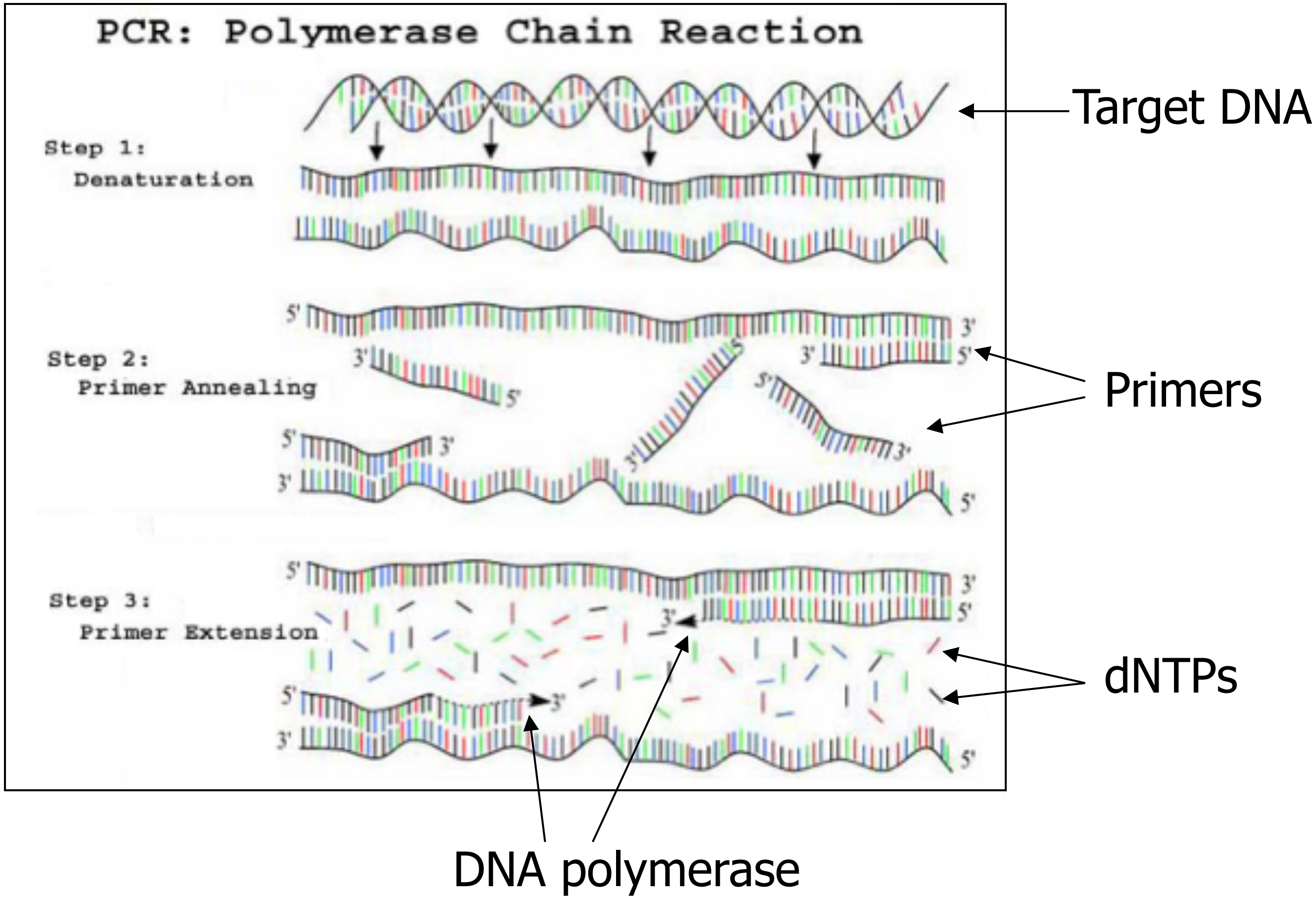
Reverse Primer

Millions of Copies

# PCR



# Schematic outline of a typical PCR cycle



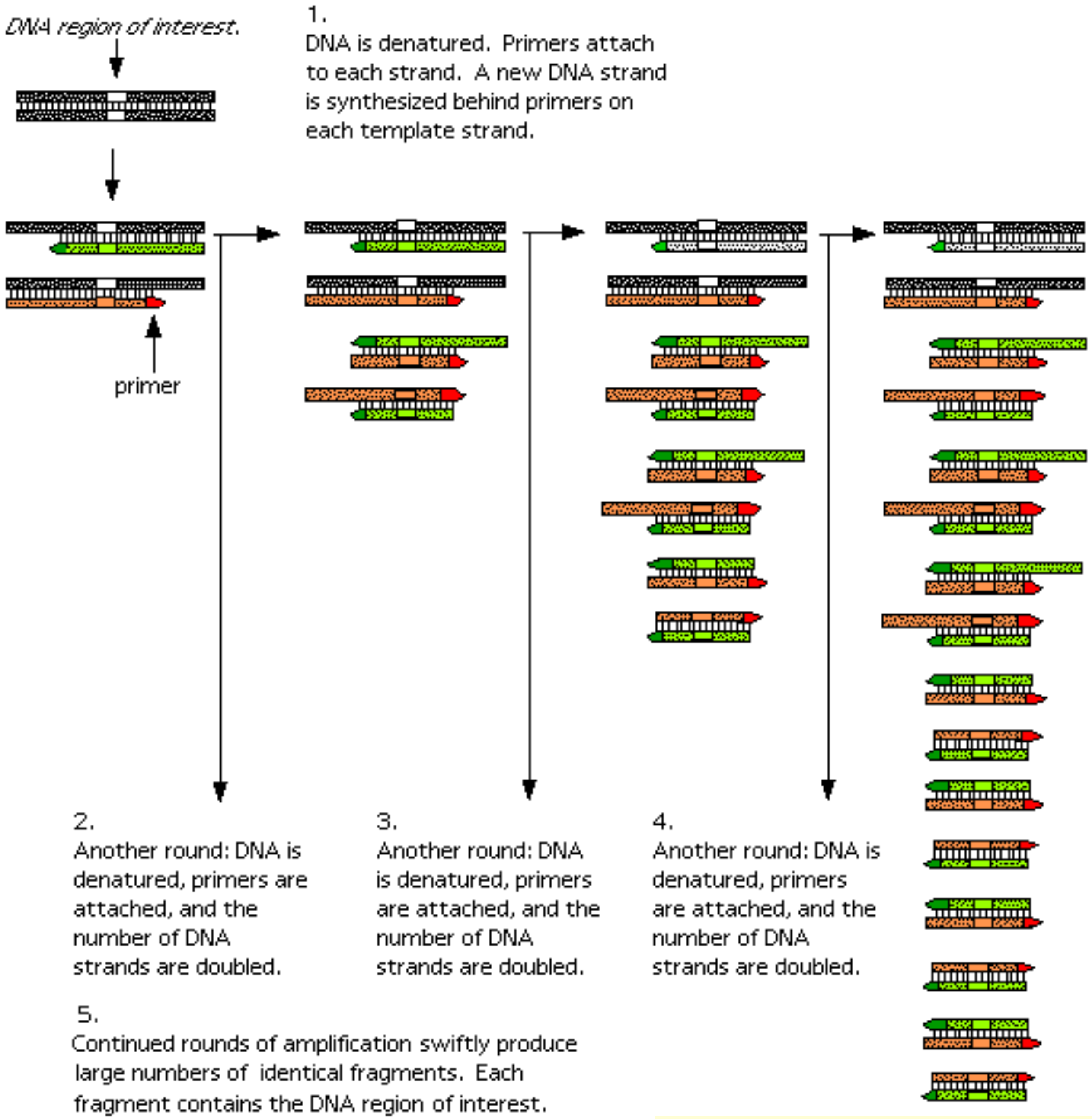
Target DNA

Primers

dNTPs

DNA polymerase

# POLYMERASE CHAIN REACTION



1\.
DNA is denatured. Primers attach to each strand. A new DNA strand is synthesized behind primers on each template strand.

2\.
Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

3\.
Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

4\.
Another round: DNA is denatured, primers are attached, and the number of DNA strands are doubled.

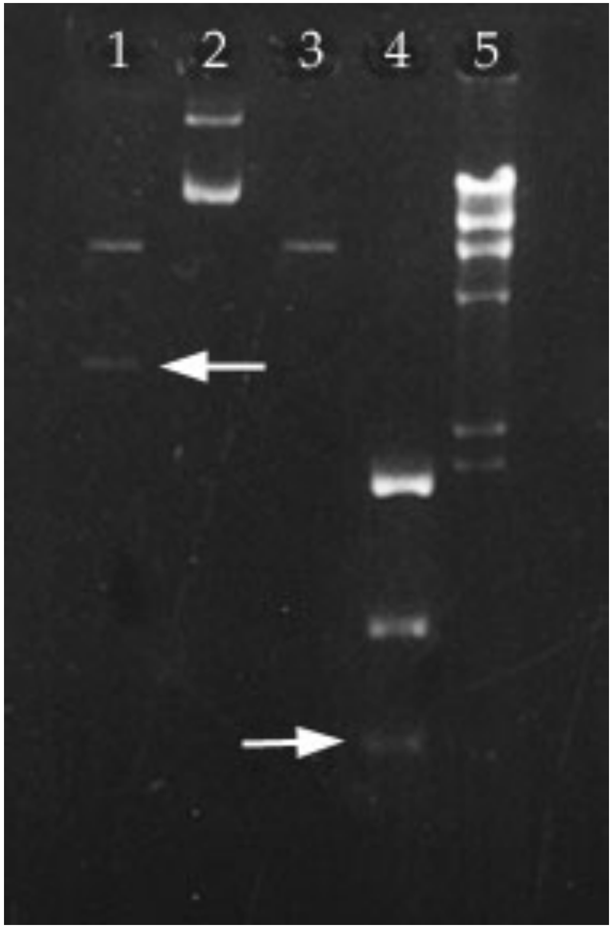
5\.
Continued rounds of amplification swiftly produce large numbers of identical fragments. Each fragment contains the DNA region of interest.

Picture Copyright: AccessExcellence @ the National Museum of Health

# Gel Electrophoresis

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

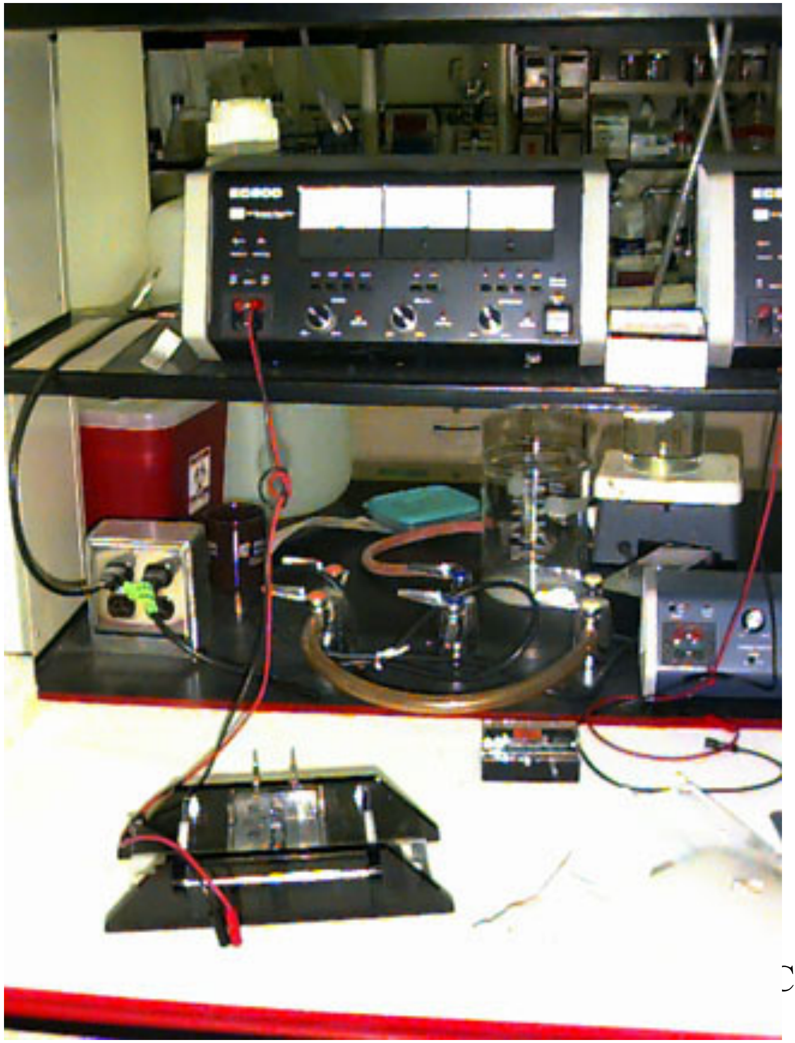
# Gel Pictures

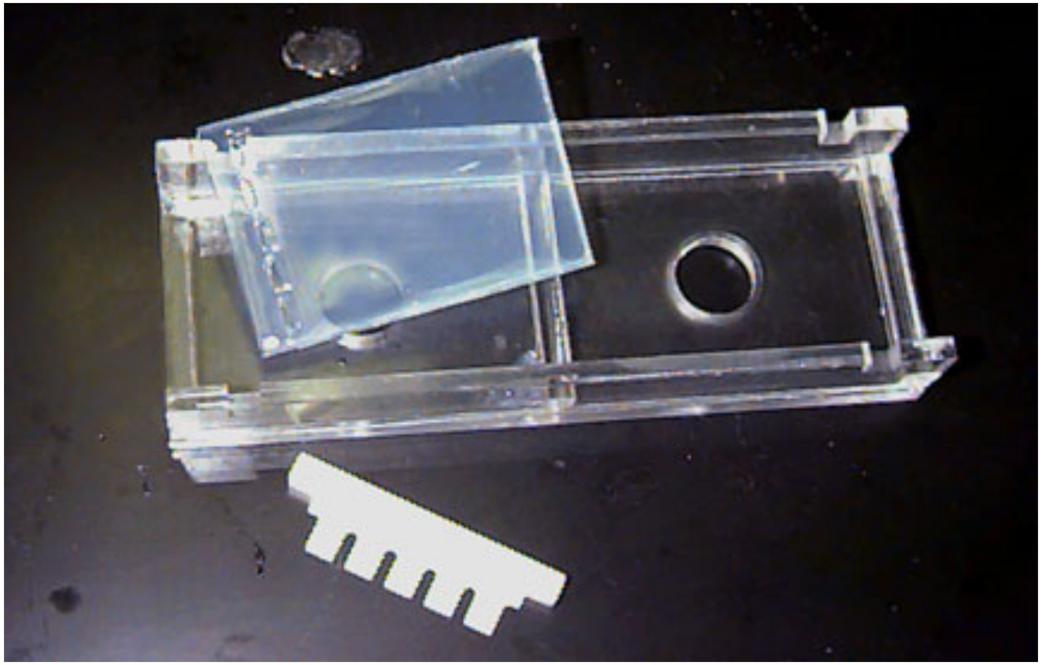


# Gel Electrophoresis: Measure sizes of fragments

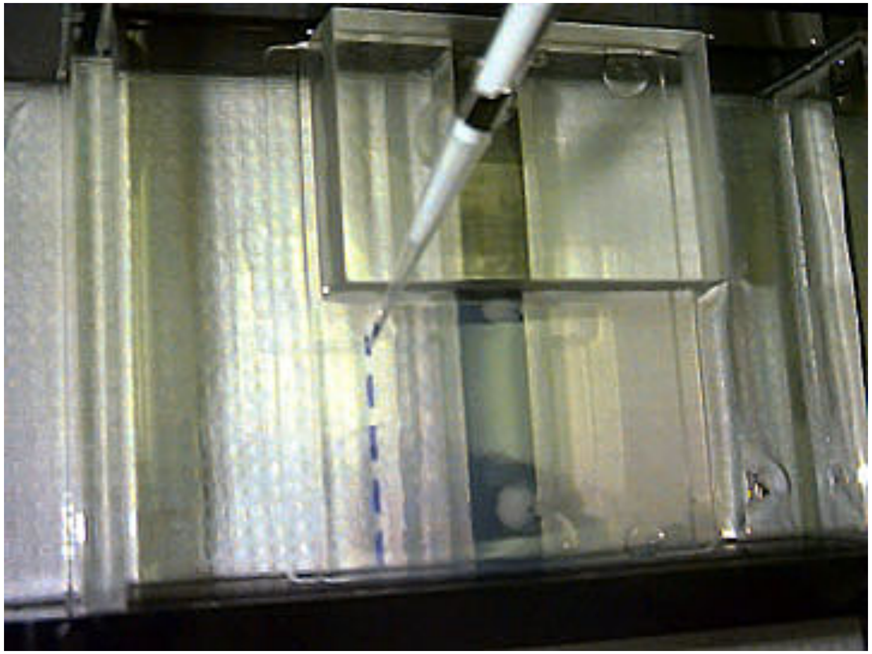
- ❑ The phosphate backbone makes DNA a highly negatively charged molecule.
- ❑ DNA can be separated according to its size.
- ❑ Gel: allow hot 1% solution of purifed agarose to cool and solidify/polymerize.
- ❑ DNA sample added to wells at the top of a gel and voltage is applied. Larger fragments migrate through the pores slower.
- ❑ Varying concentration of agarose makes different pore sizes & results.
- ❑ Proteins can be separated in much the same way, only acrylamide is used as the crosslinking agent.

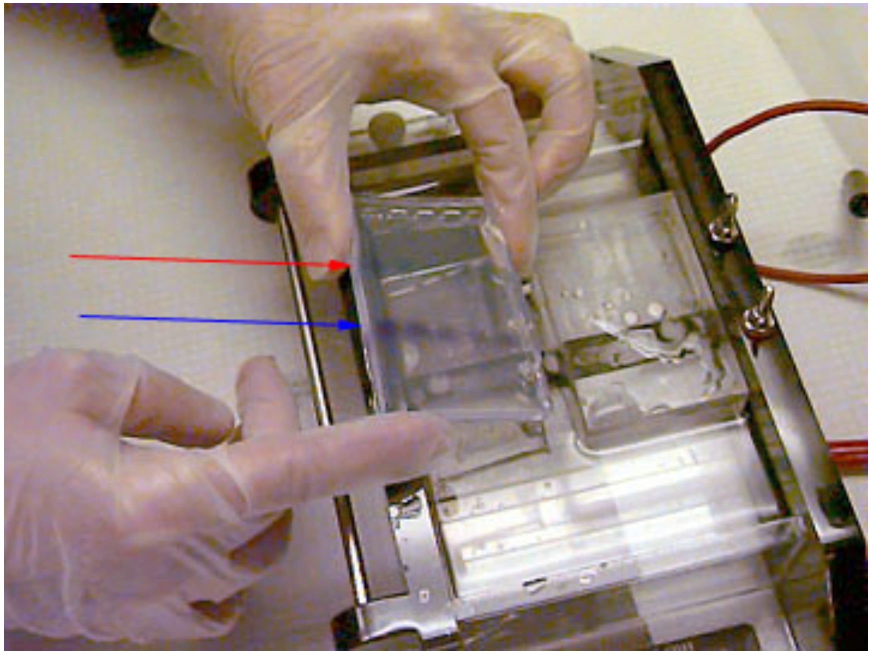
# Gel Electrophoresis

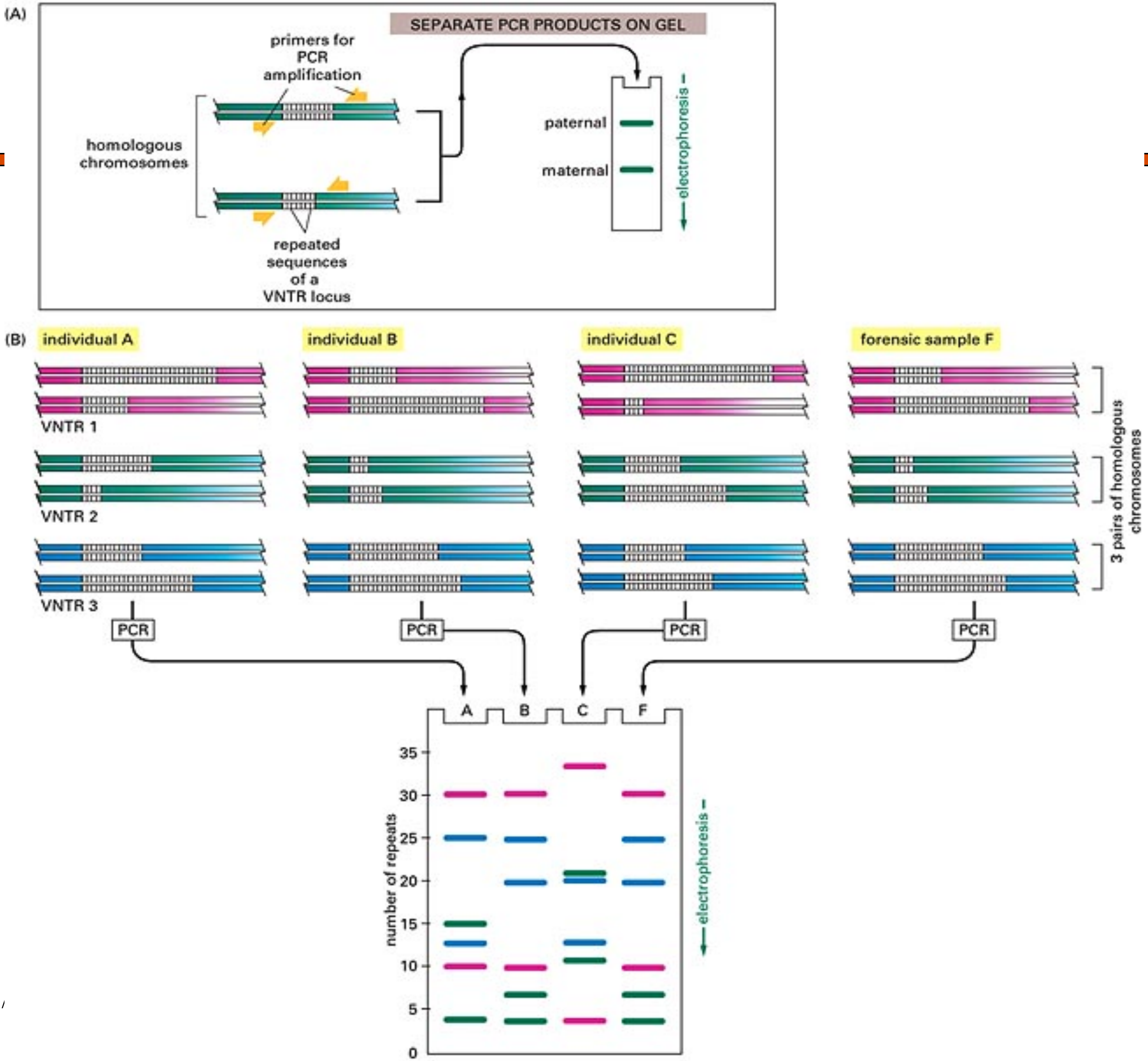




# Gel Electrophoresis






(A)
SEPARATE PCR PRODUCTS ON GEL
primers for PCR amplification
homologous chromosomes
repeated sequences of a VNTR locus
paternal
maternal
electrophoresis
(B)
individual A
individual B
individual C
forensic sample F
VNTR 1
VNTR 2
VNTR 3
3 pairs of homologous chromosomes
PCR
PCR
PCR
PCR
A
B
C
F
number of repeats
35
30
25
20
15
10
5
0
electrophoresis

# Sequencing

# Why sequencing?

- Useful for further study:
  - Locate gene sequences, regulatory elements
  - Compare sequences to find similarities
  - Identify mutations
  - Use it as a basis for further experiments

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

# History

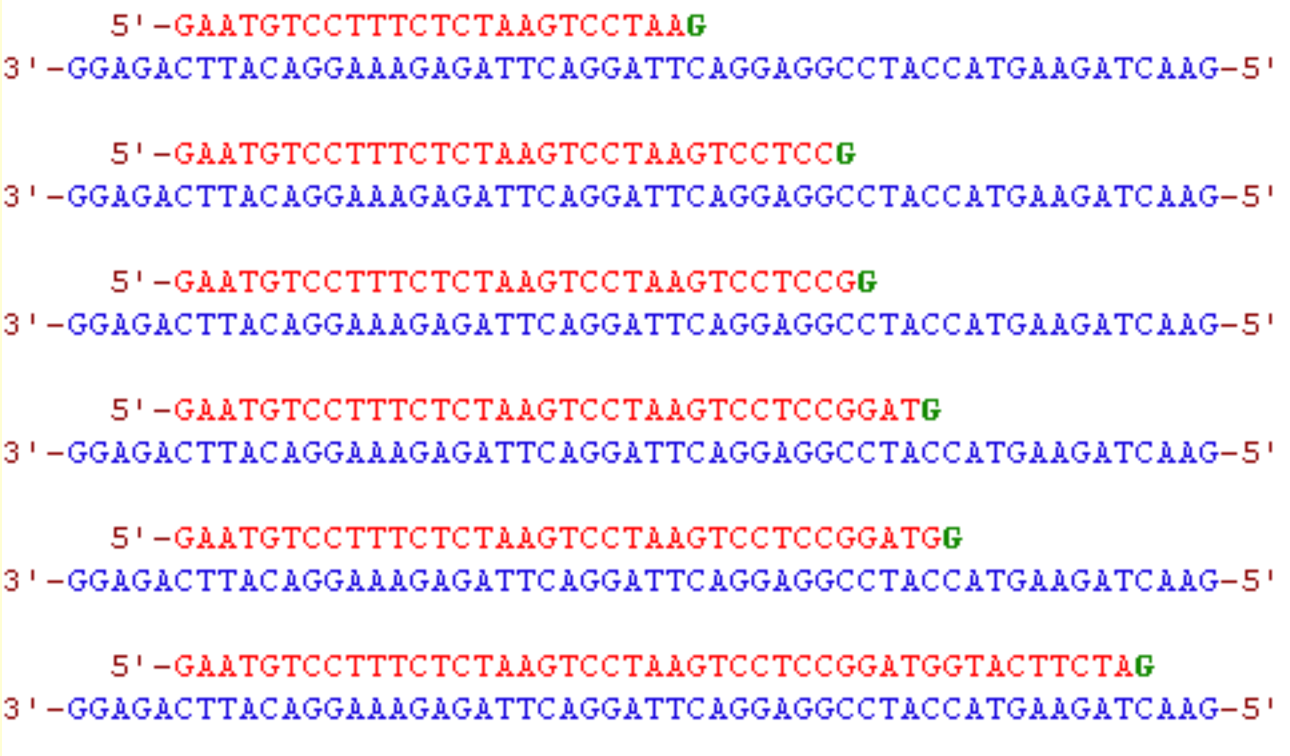
- Two methods independently developed in 1974
  - Maxam & Gilbert method
  - Sanger method: became the standard
- Nobel Prize in 1980

# 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.
- http://www.wellcome.ac.uk/Education-resources/Teaching-and-education/Animations/DNA/WTDV026689.htm

# Sanger Method

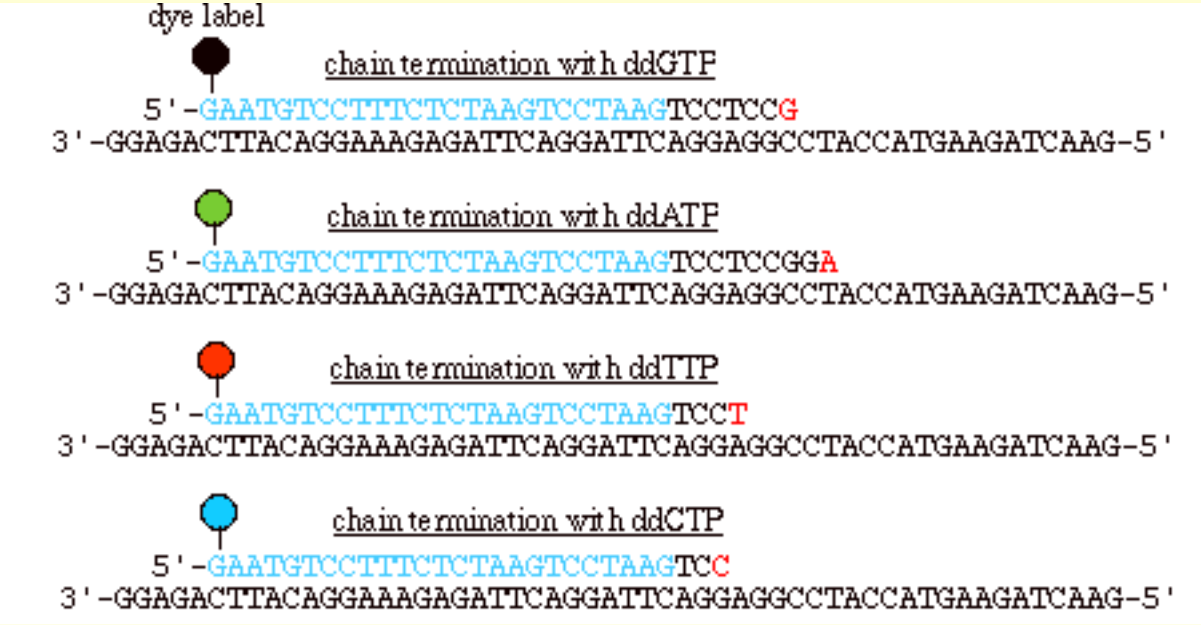
- ❑ Example of sequences seen in gel from "G" tube:



```
     5'-GAATGTCCTTTCTCTAAGTCCTAAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

     5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGGTACTTCTAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
```

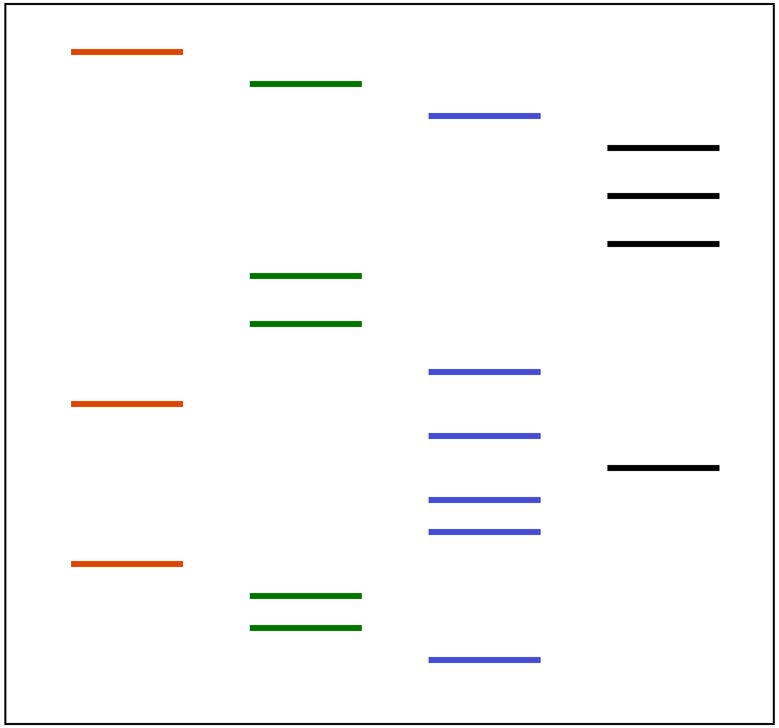
# Modified Sanger

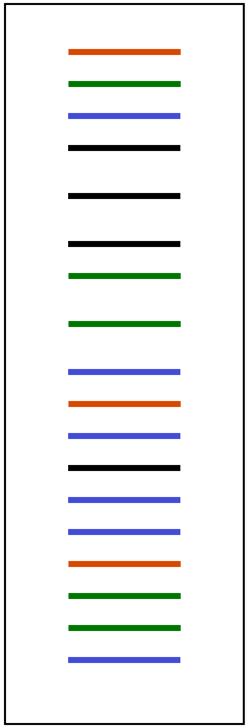
- Reactions performed in a single tube containing all four ddNTP's, each labeled with a different color dye


dye label
chain termination with ddGTP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
chain termination with ddATP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGA
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
chain termination with ddTTP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCT
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'
chain termination with ddCTP
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCC
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

# Sequencing Gels: Separate vs Single Lanes

GCCAGGTGAGCCTTTGCA





# Sequencing

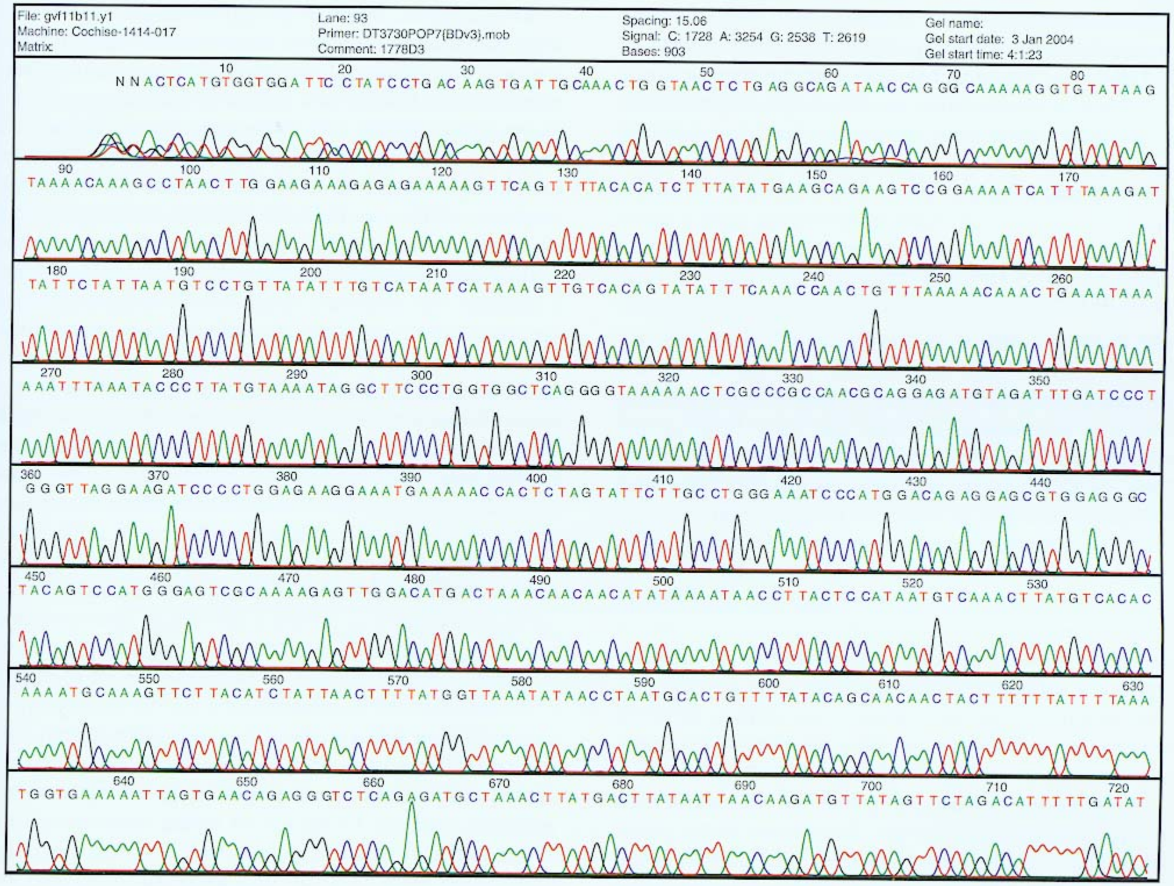
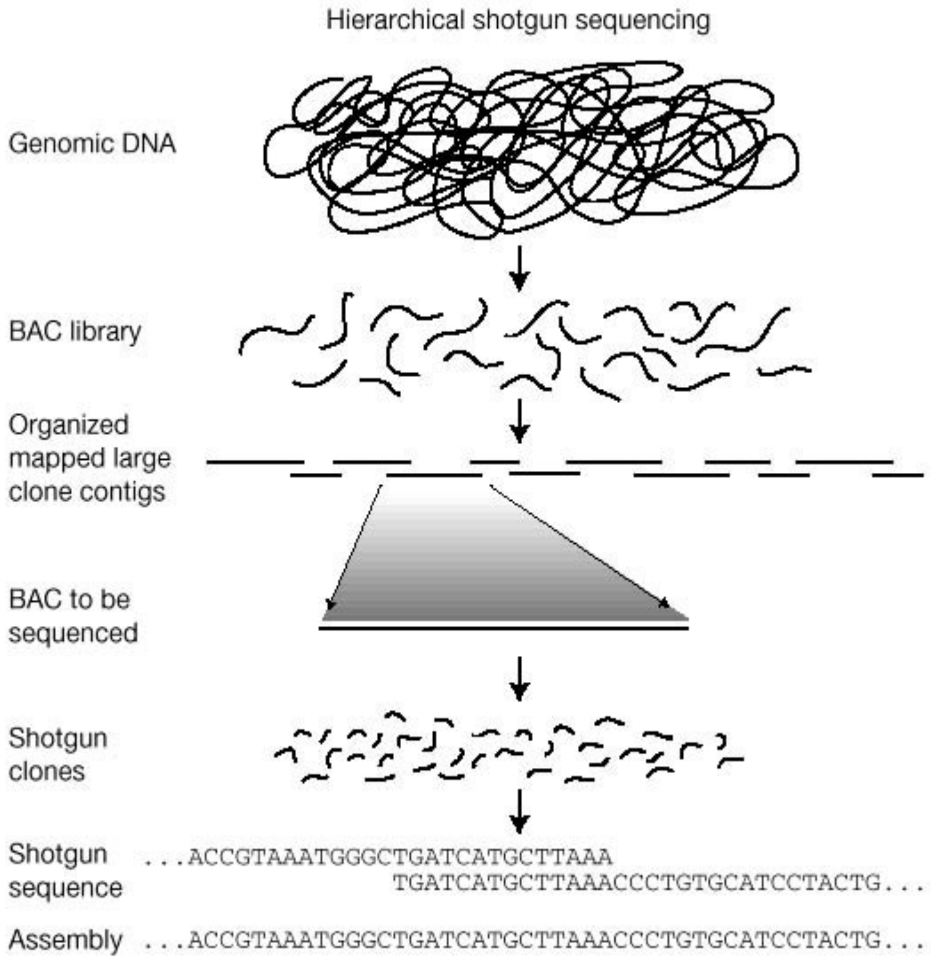


**FIGURE 13.3** A sample chromatogram, as viewed with the vtrace program (Ewing, 2002). Signal intensities corresponding to fragments ending with A (green), C (blue), G (black), and T (red) are shown out to approximately 722 bases.

# Shotgun Sequencing


Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
. . .ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG. . .
Assembly
. . .ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG. . .

From http://www.tulane.edu/~biochem/lecture/723/humgen.html

# Sequencing

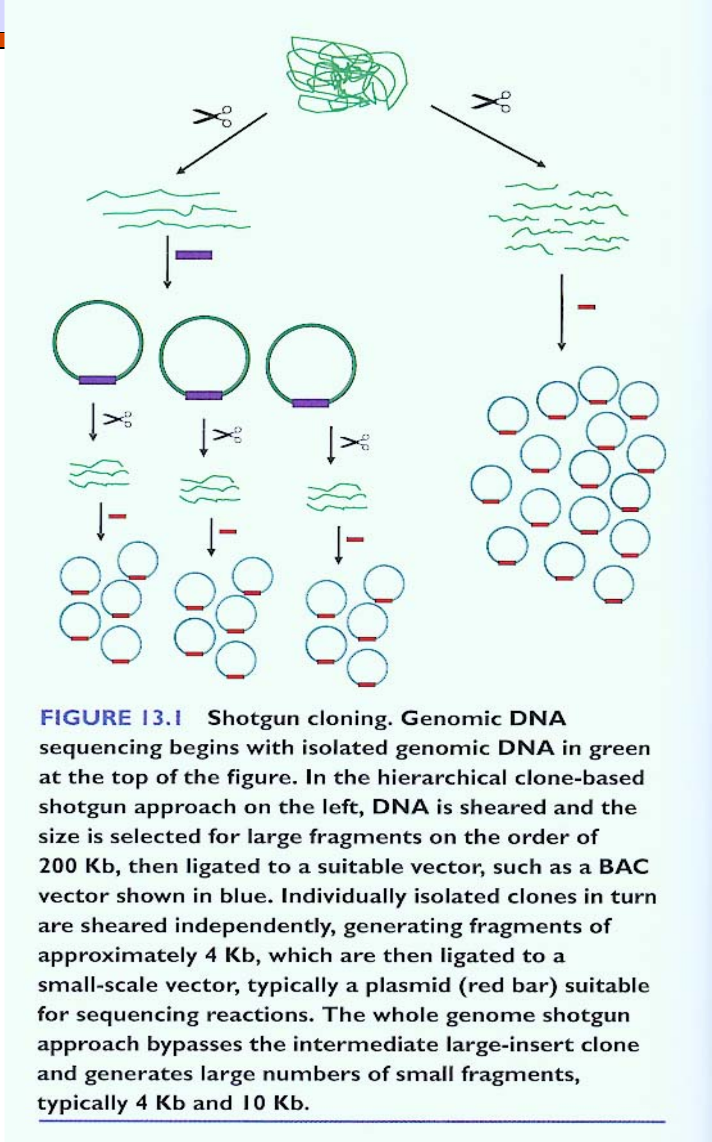
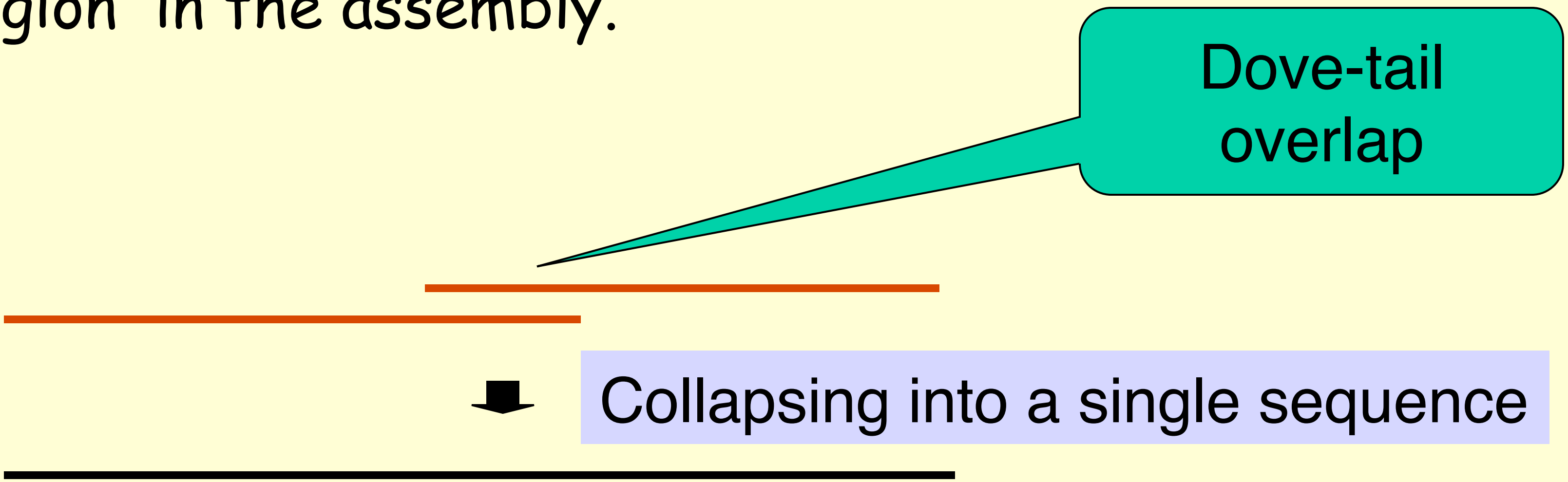


FIGURE 13.1 Shotgun cloning. Genomic DNA sequencing begins with isolated genomic DNA in green at the top of the figure. In the hierarchical clone-based shotgun approach on the left, DNA is sheared and the size is selected for large fragments on the order of 200 Kb, then ligated to a suitable vector, such as a BAC vector shown in blue. Individually isolated clones in turn are sheared independently, generating fragments of approximately 4 Kb, which are then ligated to a small-scale vector, typically a plasmid (red bar) suitable for sequencing reactions. The whole genome shotgun approach bypasses the intermediate large-insert clone and generates large numbers of small fragments, typically 4 Kb and 10 Kb.

# Sequencing: Generate Contigs

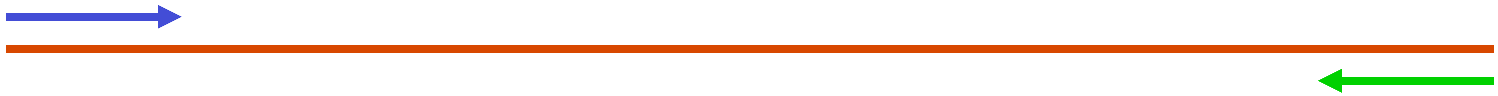
- Short for "contiguous sequence". A continuously covered region in the assembly.



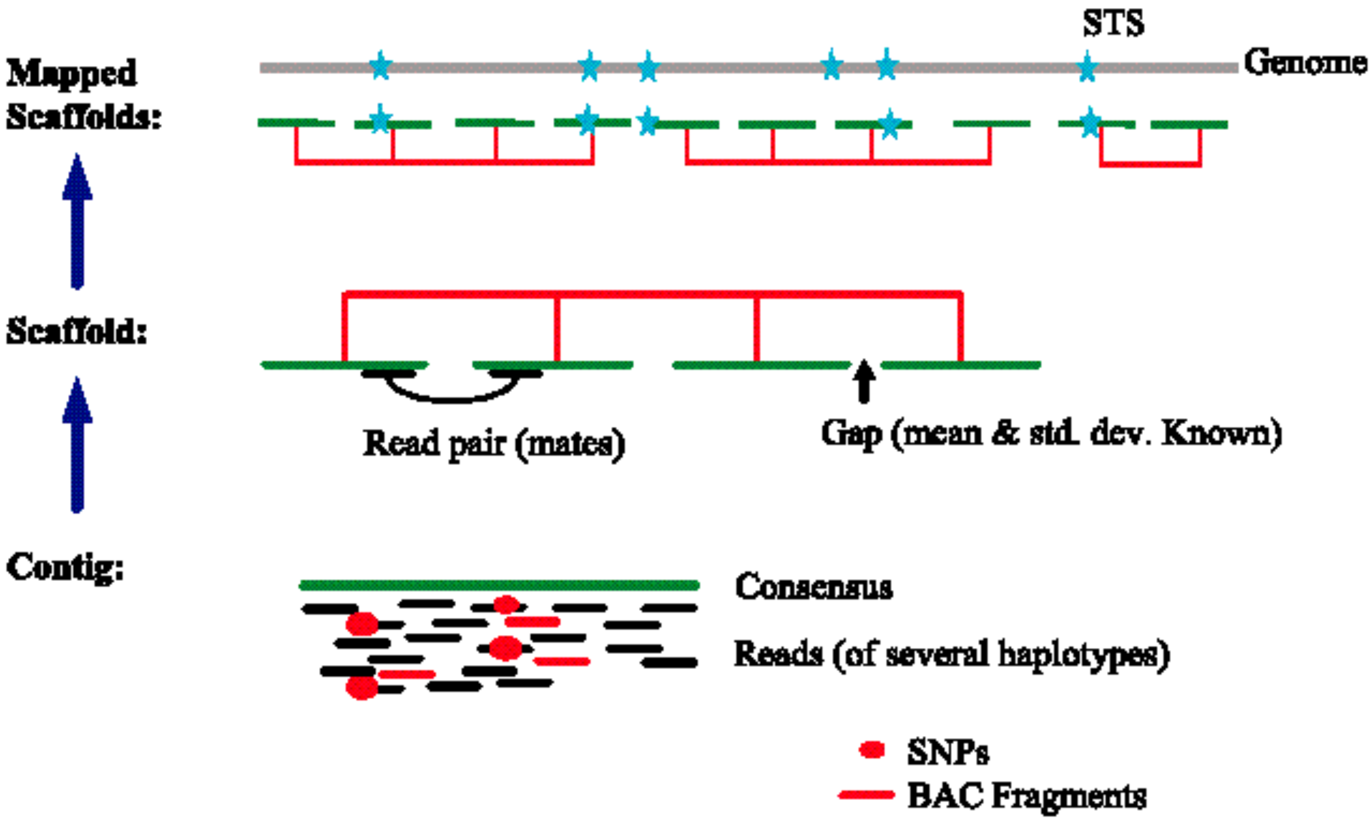
- Jang W et al (1999) Making effective use of human genomic sequence data. Trends Genet. 15(7): 284-6.
  Kent WJ and Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. Genome Res 11(9): 1541-8.

# Paired Reads

- Scaffold (supercontig): formed when two contigs with no sequence overlap can be linked
  - Data from paired end reads help create scaffolds with known gaps
    - If two reads end up in two different contigs, then we can link contigs to form scaffold.



# Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

# Human Genome Project

- ❑ Many videos available on youtube.com, dnatube.com, and elsewhere.
- ❑ Find some and watch them.

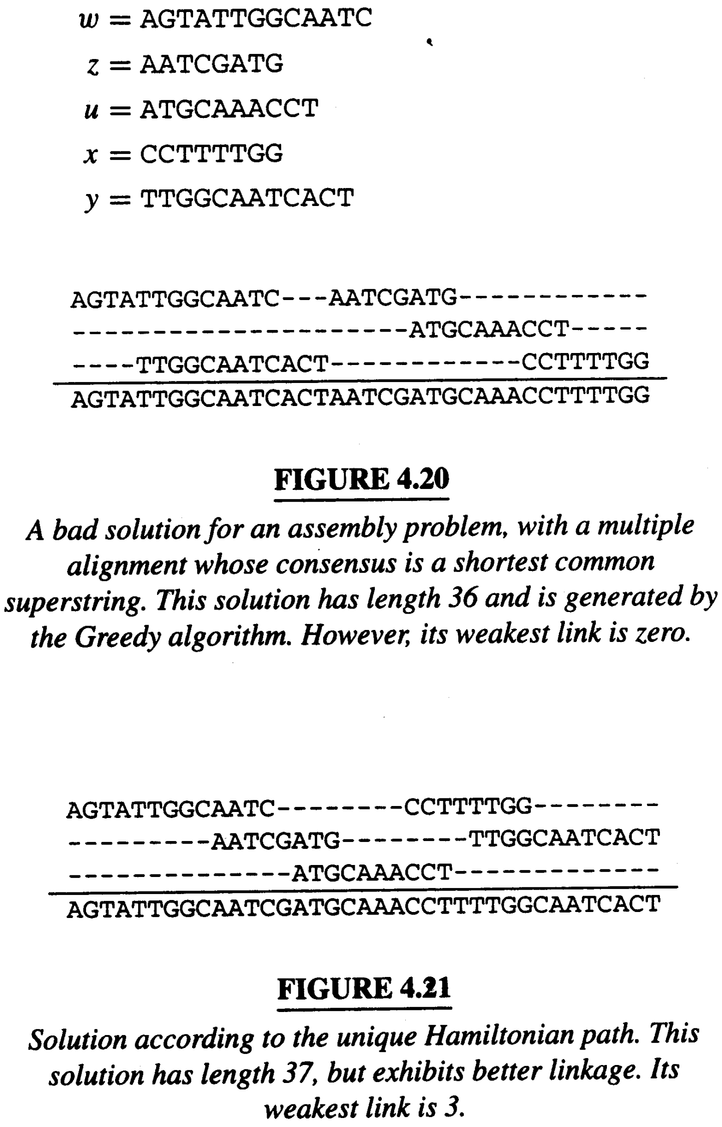
# Assembly: Simple Example

- ACCGT, CGTGC, TTAC, TACCGT
- Total length = ~10
- 
    - --**ACCGT**--
    - ----**CGTGC**
    - **TTAC**-----
    - -**TACCGT**–
    - **TTACCGTGC**

# 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

# Assembly: Complications



$w =$ AGTATTGGCAATC

$z =$ AATCGATG

$u =$ ATGCAAACCT

$x =$ CCTTTTGG

$y =$ TTGGCAATCACT

```
AGTATTGGCAATC---AATCGATG------------
---------------------ATGCAAACCT-----
----TTGGCAATCACT------------CCTTTTGG
AGTATTGGCAATCACTAATCGATGCAAACCTTTTGG
```

**FIGURE 4.20**

*A bad solution for an assembly problem, with a multiple alignment whose consensus is a shortest common superstring. This solution has length 36 and is generated by the Greedy algorithm. However, its weakest link is zero.*

```
AGTATTGGCAATC--------CCTTTTGG--------
---------AATCGATG--------TTGGCAATCACT
--------------ATGCAAACCT-------------
AGTATTGGCAATCGATGCAAACCTTTTGGCAATCACT
```

**FIGURE 4.21**

*Solution according to the unique Hamiltonian path. This solution has length 37, but exhibits better linkage. Its weakest link is 3.*

# Assembly: Complications

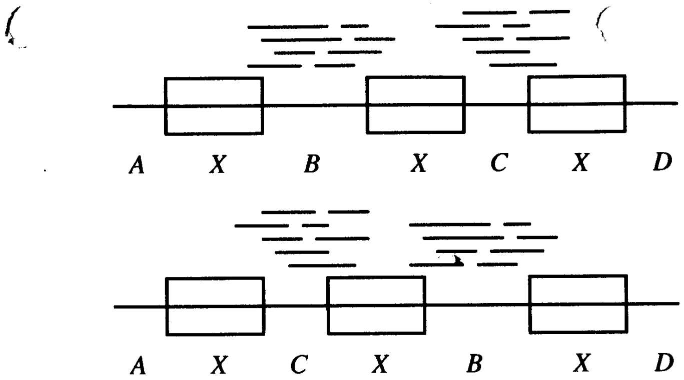


**FIGURE 4.8**

*Target sequence leading to ambiguous assembly because of repeats of the form XXX.*

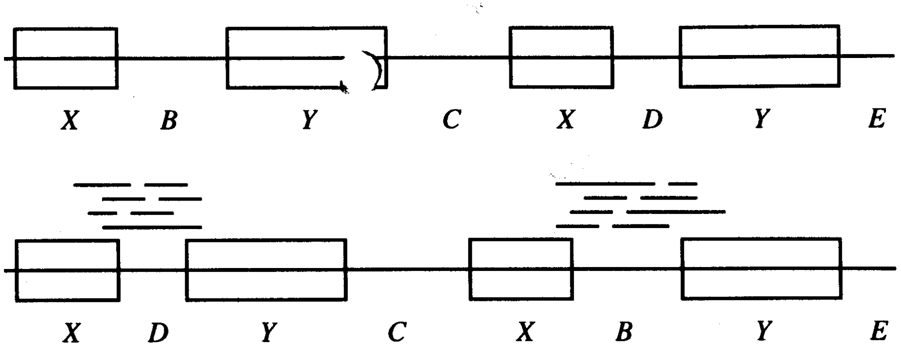


**FIGURE 4.9**

*Target sequence leading to ambiguous assembly because of repeats of the form XYXY.*

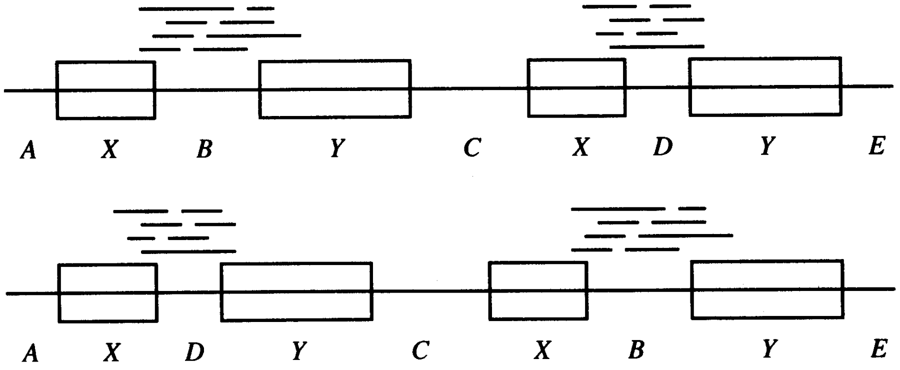


**FIGURE 4.9**

*Target sequence leading to ambiguous assembly because of repeats of the form XYXY.*

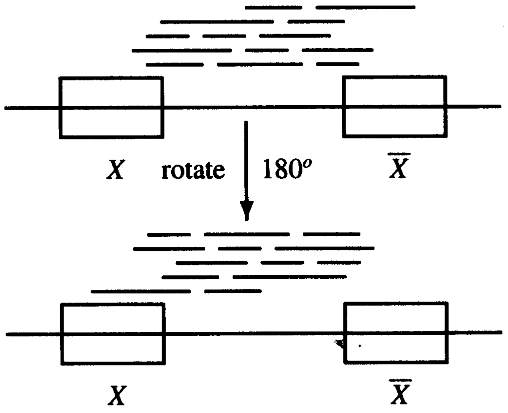
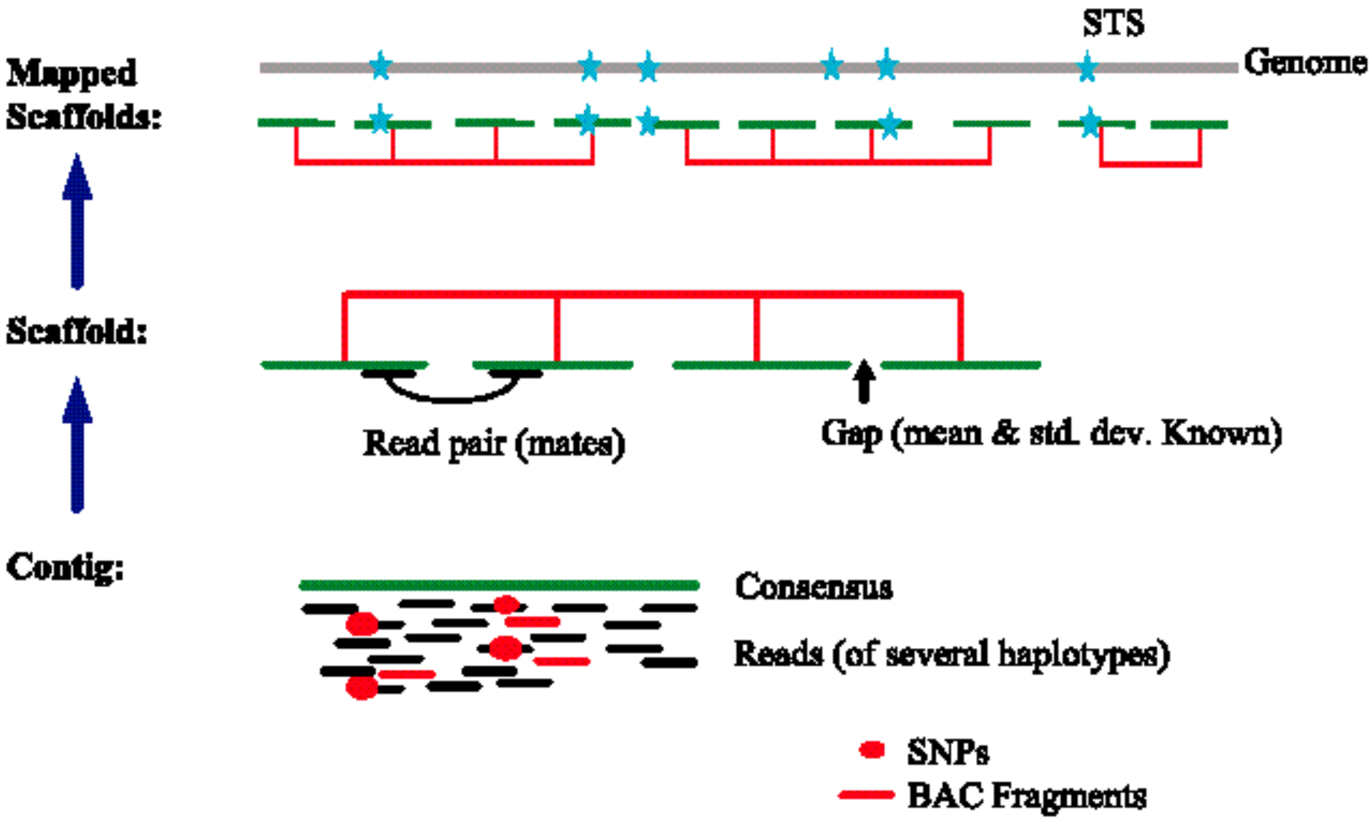


**FIGURE 4.10**

*Target sequence with inverted repeat. The region marked $\overline{X}$ is the reverse complement of the region marked X.*

# Shotgun Sequencing



From http://www.tulane.edu/~biochem/lecture/723/humgen.html

# Human Genome Project

- Many videos available on youtube.com, dnatube.com, and elsewhere.
- Find some and watch them.

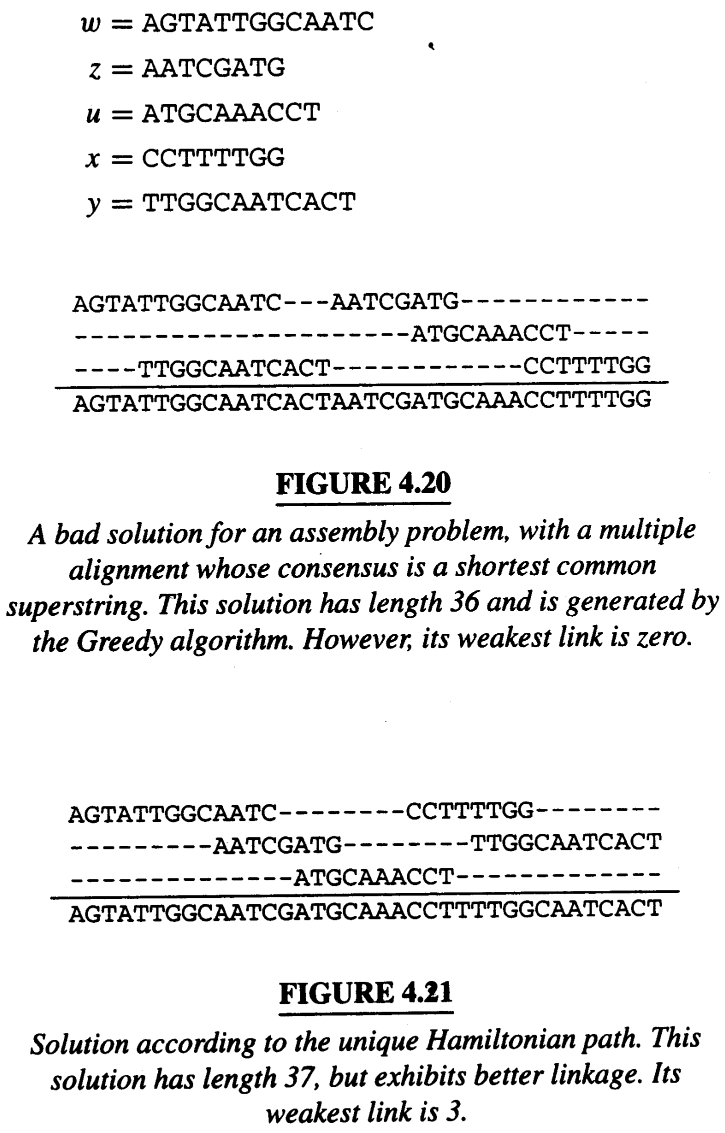
# Assembly: Simple Example

- ACCGT, CGTGC, TTAC, TACCGT
- Total length = ~10
-
  - --ACCGT--
  - ----CGTGC
  - TTAC-----
  - -TACCGT–
  - TTACCGTGC

# 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

# Assembly: Complications



$w =$ AGTATTGGCAATC

$z =$ AATCGATG

$u =$ ATGCAAACCT

$x =$ CCTTTTGG

$y =$ TTGGCAATCACT

```
AGTATTGGCAATC---AATCGATG------------
---------------------ATGCAAACCT-----
----TTGGCAATCACT------------CCTTTTGG
------------------------------------
AGTATTGGCAATCACTAATCGATGCAAACCTTTTGG
```

**FIGURE 4.20**

*A bad solution for an assembly problem, with a multiple alignment whose consensus is a shortest common superstring. This solution has length 36 and is generated by the Greedy algorithm. However, its weakest link is zero.*

```
AGTATTGGCAATC--------CCTTTTGG--------
---------AATCGATG--------TTGGCAATCACT
--------------ATGCAAACCT-------------
-------------------------------------
AGTATTGGCAATCGATGCAAACCTTTTGGCAATCACT
```

**FIGURE 4.21**

*Solution according to the unique Hamiltonian path. This solution has length 37, but exhibits better linkage. Its weakest link is 3.*

# Assembly: Complications

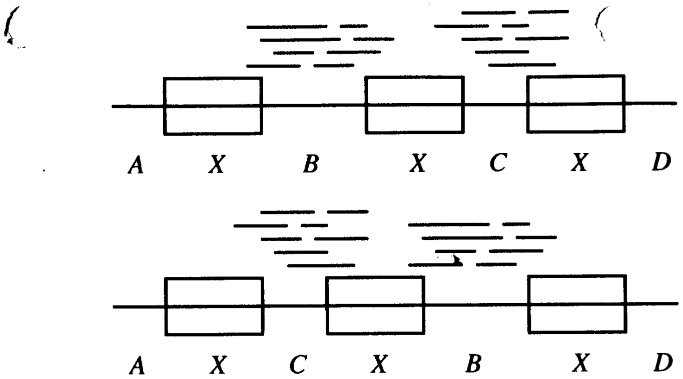


**FIGURE 4.8**

*Target sequence leading to ambiguous assembly because of repeats of the form XXX.*

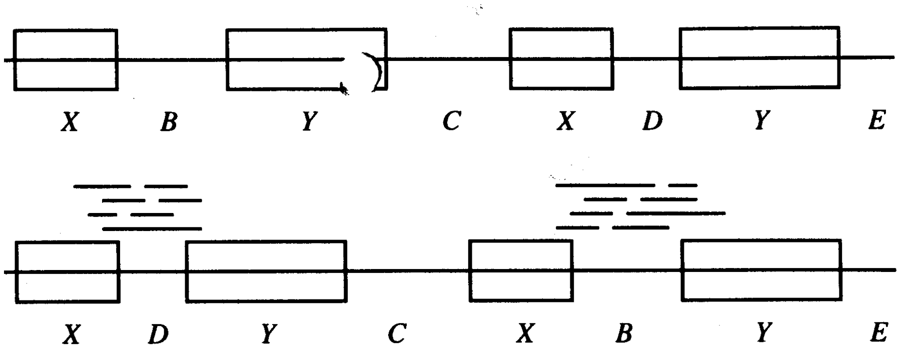


**FIGURE 4.9**

*Target sequence leading to ambiguous assembly because of repeats of the form XYXY.*

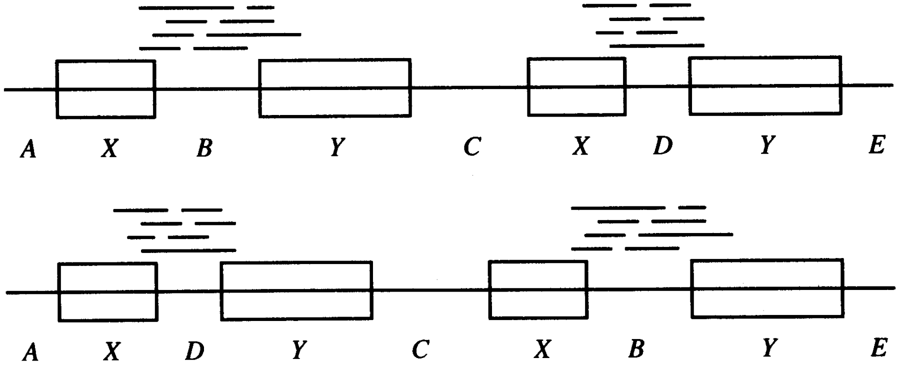


**FIGURE 4.9**

*Target sequence leading to ambiguous assembly because of repeats of the form XYXY.*

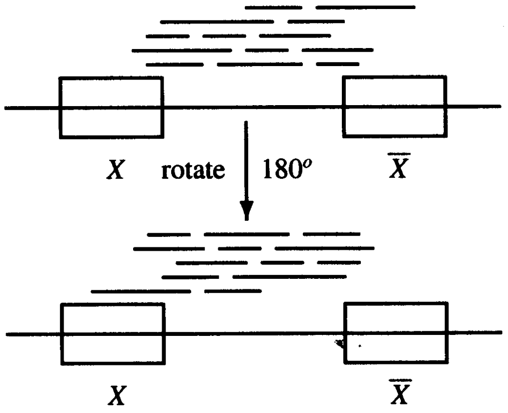


**FIGURE 4.10**

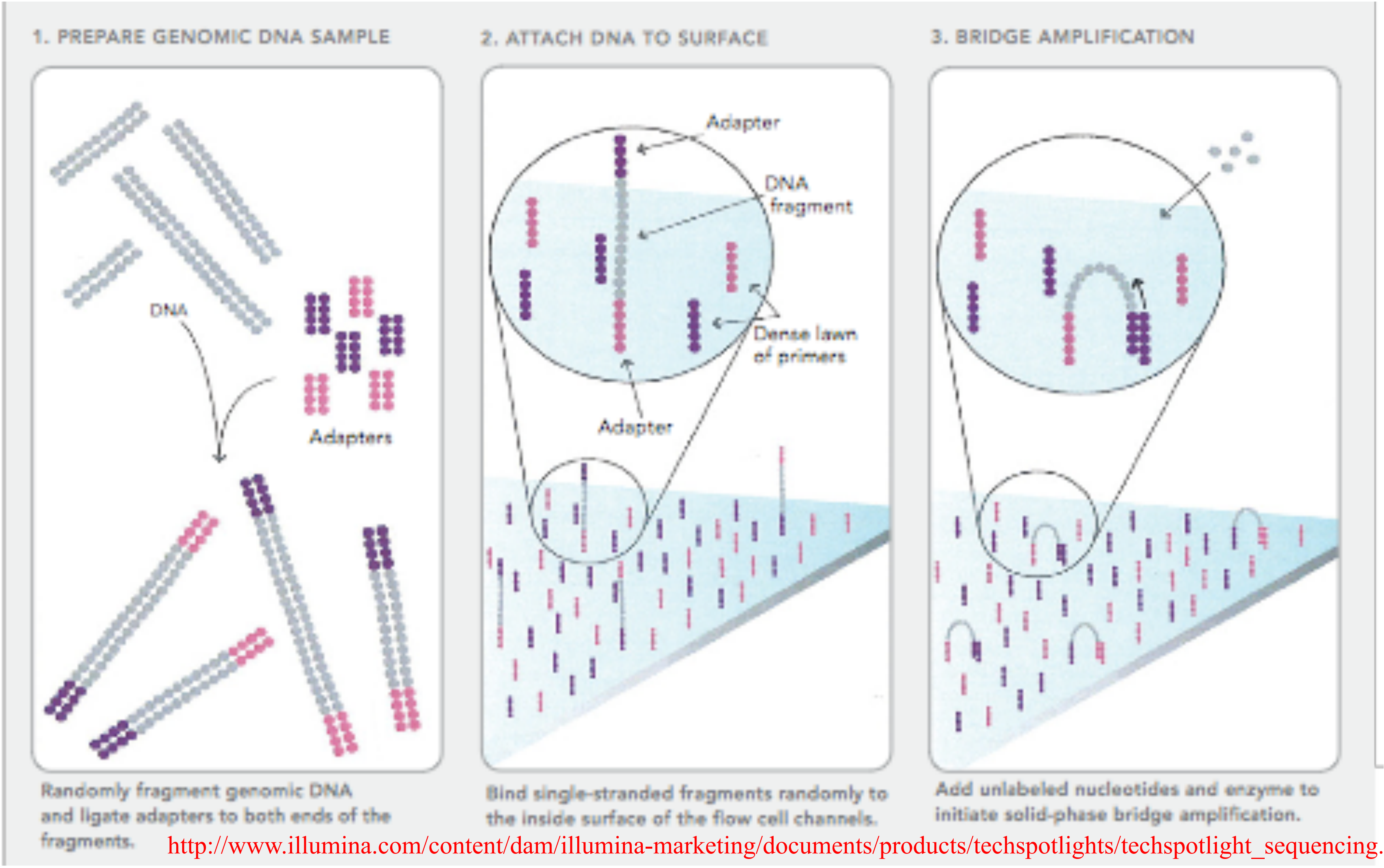
*Target sequence with inverted repeat. The region marked $\overline{X}$ is the reverse complement of the region marked X.*

# Next Generation Sequencing

# History of NGS

- 1977: Sanger Method (70Kbp/run)
- Sequencing by Hybridization (SBH); Dual end sequencing; Chromosome Walking (see page 5-6 of Pevzner's text);
- 1987: Automated Sequencer (AB Prism)
- 1996: Capillary Sequencer (ABI 310)
- 2005: 454 Sequencing (GS 20; 60Mbp/run)
- 2006: Solexa Sequencing (Illumina; 600Mbp/run)
- 2007 : SOLiD (AB)
- 2009 : Helicos single molecule sequencer
- 2011 : Ion Torrent (PGM)
- 2011 : Pacific Biosciences single molecule sequencer
- 2012 : Oxford Nanopore Tech. ultra long single mol. reads

# Illumina's Sequencing-by-Synthesis



1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

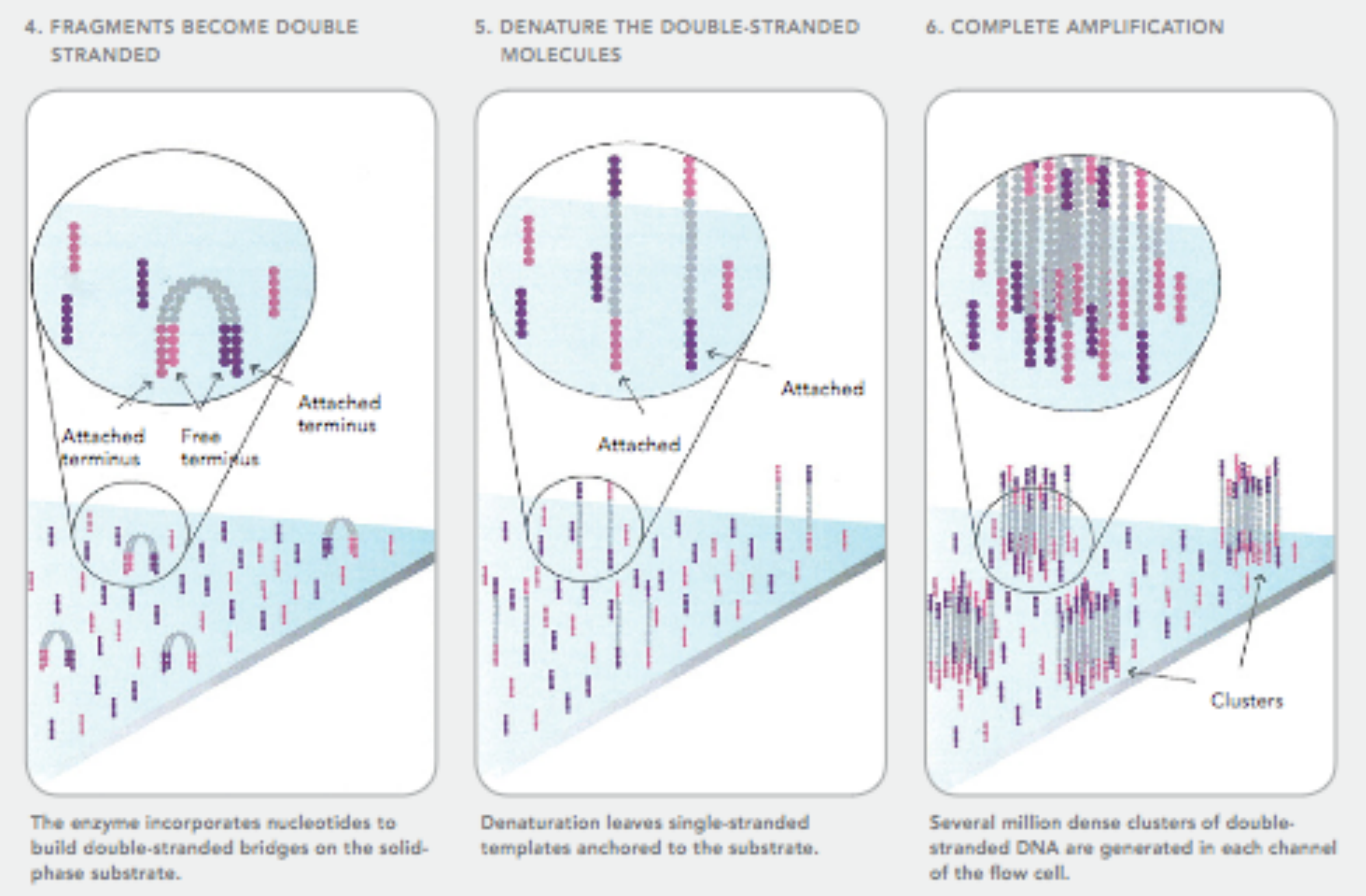
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

http://www.illumina.com/content/dam/illumina-marketing/documents/products/techspotlights/techspotlight_sequencing.pdf

# Solexa Sequencing



4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

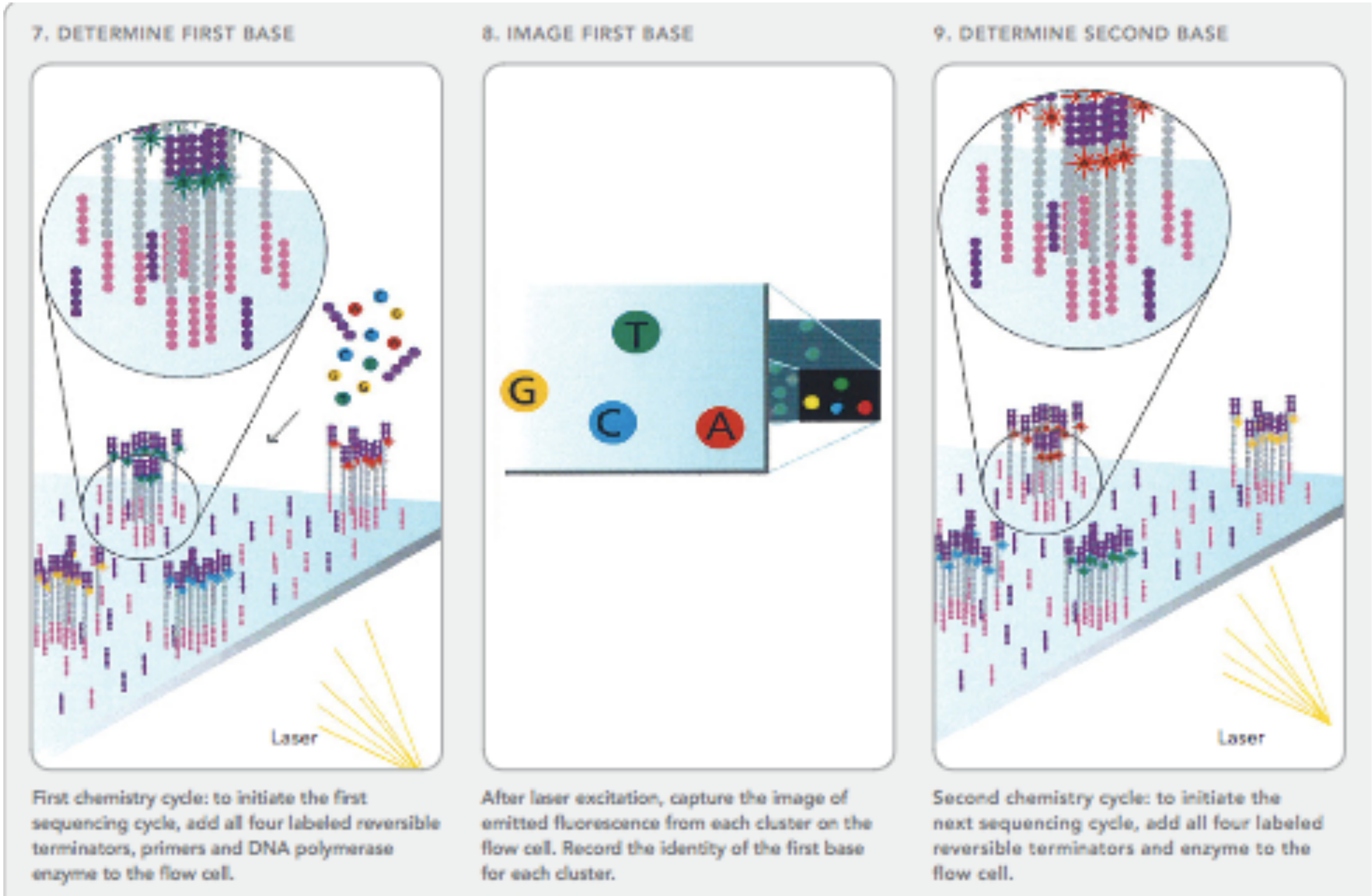
5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

# Solexa Sequencing



7. DETERMINE FIRST BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

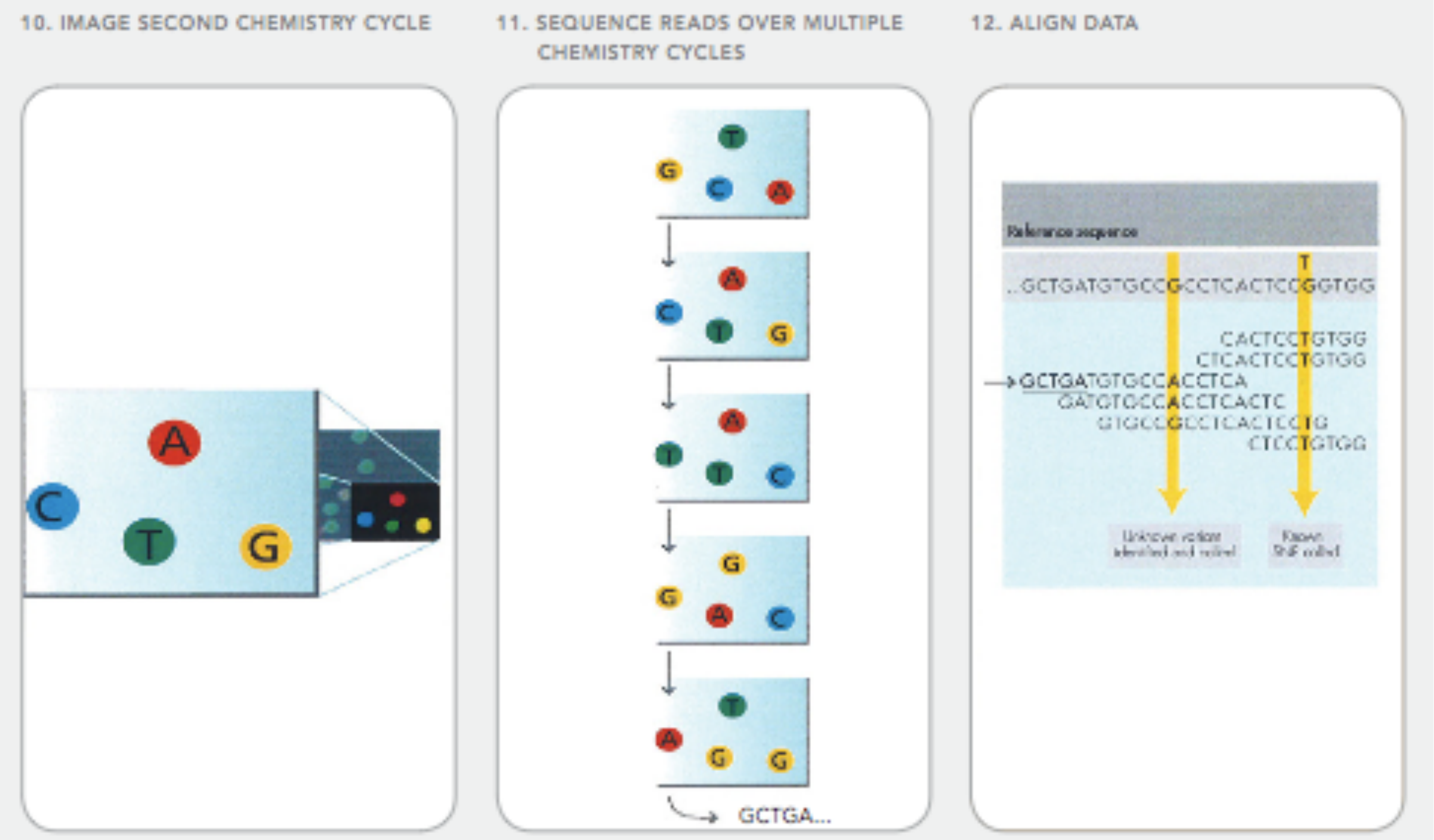
8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

# Solexa Sequencing



10. IMAGE SECOND CHEMISTRY CYCLE

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

12. ALIGN DATA

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

Align data, compare to a reference, and identify sequence differences.

# Ion Torrent Sequencer

- Harness power of semiconductor technology
- During nucleotide synthesis, a proton is released
- This can be detected by measuring pH, not fluorescence
- The dNTPs are flowed over the surface in a predetermined sequence & the ligations are detected

# PacBio Sequencing

- Single molecule technology
- Extraordinarily long reads
- Non-trivial error, but unbiased

# Assemblers

- ❑ TIGR Assembler (TIGR)
- ❑ Phrap (U Washington)
- ❑ Celera Assembler (Celera Genomics)
- ❑ Arachne (Broad Institute of MIT & Harvard)
- ❑ Phusion (Sanger Center)
- ❑ Atlas (Baylor College of Medicine)

# Applications of Sequencing

- ❑ Sequencing
- ❑ Resequencing
- ❑ SNP detection
- ❑ RNA-Seq
- ❑ CHiP-Seq
- ❑ Metagenomics

# Basic Assembler

- **Read**: sequenced fragment; **Contig**: contiguous segment. How to assemble a contig?

```
TCGAGTTAAGCTTTAG
 CGAGTTAAGCTTTAGC
   AGTTAAGCTTTAGCCT
    GTTAAGCTTTAGCCTA
       AGCTTTAGCCTAGGGC
         GCTTTAGCCTAGGCAG
            ...
```

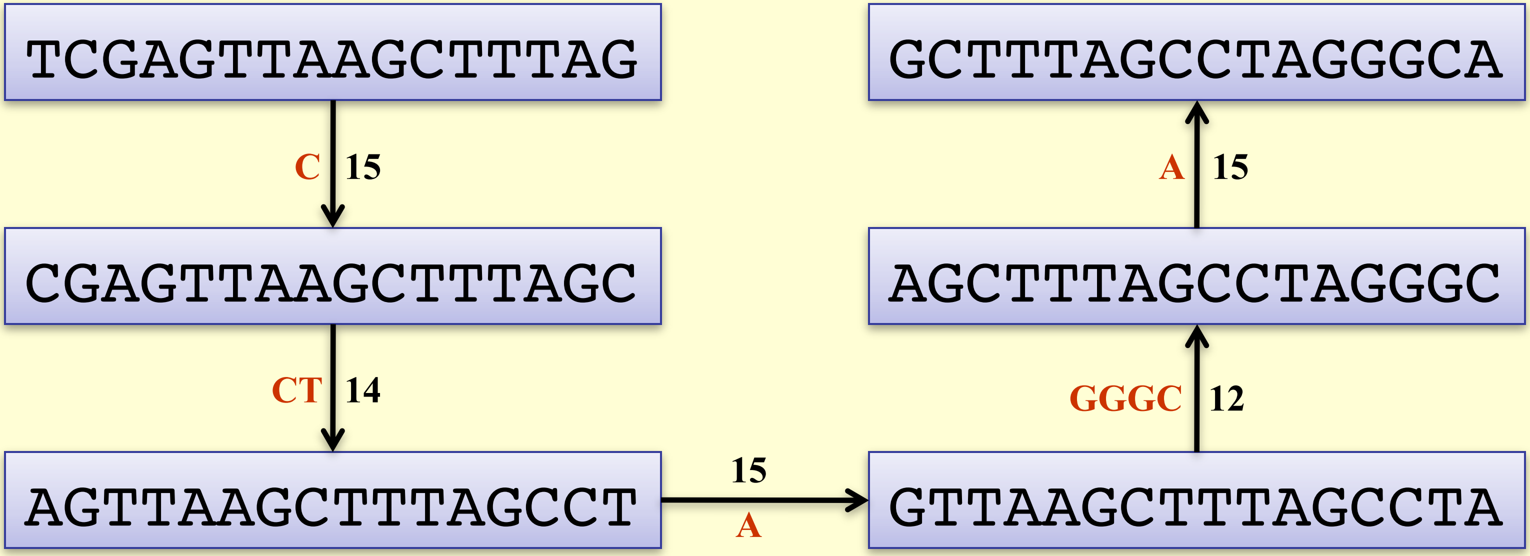
```
AGCTTTAGCCTAGGGC
AGTTAAGCTTTAGCCT
CGAGTTAAGCTTTAGC
GCTTTAGCCTAGGCAG
GTTAAGCTTTAGCCTA
TAAGCTTTAGCCTAGG
TCGAGTTAAGCTTTAG
```

**Problem**: Need to try every pair of reads!

# Reduce to Graph Problem

❑ How to assemble a contig?

- Node ⟷ Read
- Edge between Nodes ⟷ Overlapping Reads
- **Problem**: Find a path through each node in graph.



**Issues**: Problem is NP-Complete
# nodes = # reads
# of edges ≤ k(# nodes)

# String graph

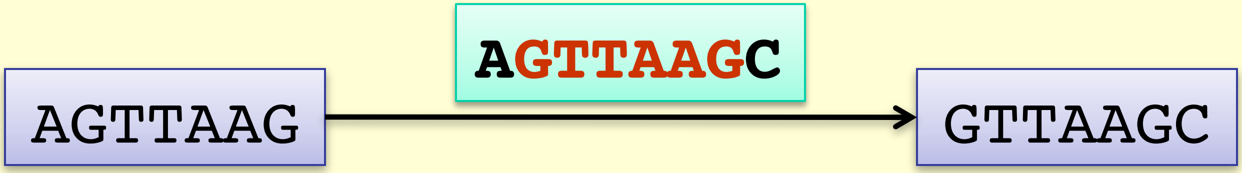
- Combine nodes that form paths into strings

# A better solution

- Take each read and chop it into k-mers.
- Represent k-mers by nodes in a graph and edges between k-mers that overlap in k-1 bases.
- **Consequence**:
  - Number of nodes = $4^k$ ;
  - Number of edges = $k4^k$ ;
- **Issues**:
  - Problem (i.e., find path through all vertices) remains NP-Complete

# A more efficient solution

- Represent every possible (k-1)-mer by a node.
- Edges connect 2 nodes if they share k-2 bases.
- Label each edge by k-mer.



- Problem:
  - Find a path through each edge in the graph
- The Eulerian path problem is **NOT** NP-Complete. It can be solved in linear time!

# Sources of Assembly Errors

- Errors in reads – caused by technology
  - Error in base calls, color calls (SOLID Technology), or repeated base calls (454 Technology)
- Missing reads – sequencing bias
- Read orientation error
  - One or both orientations may occur
  - Not told which ones are present
- Sequence Variations – mixed sample study
  - SNP, cancer, metagenomics studies
- REPEATS
- Combinations of the above

# How to deal with REPEAT Regions

- If no errors or repeat regions, then the graph has a unique path through all the edges.
- **Problem**: REPEAT regions cause branching in graph. If no errors in reads, then the graph has a unique path through all edges, but with some edges traversed more than once.
- How to identify REPEAT regions:
  - Higher coverage of repeat regions
  - Branching of nodes

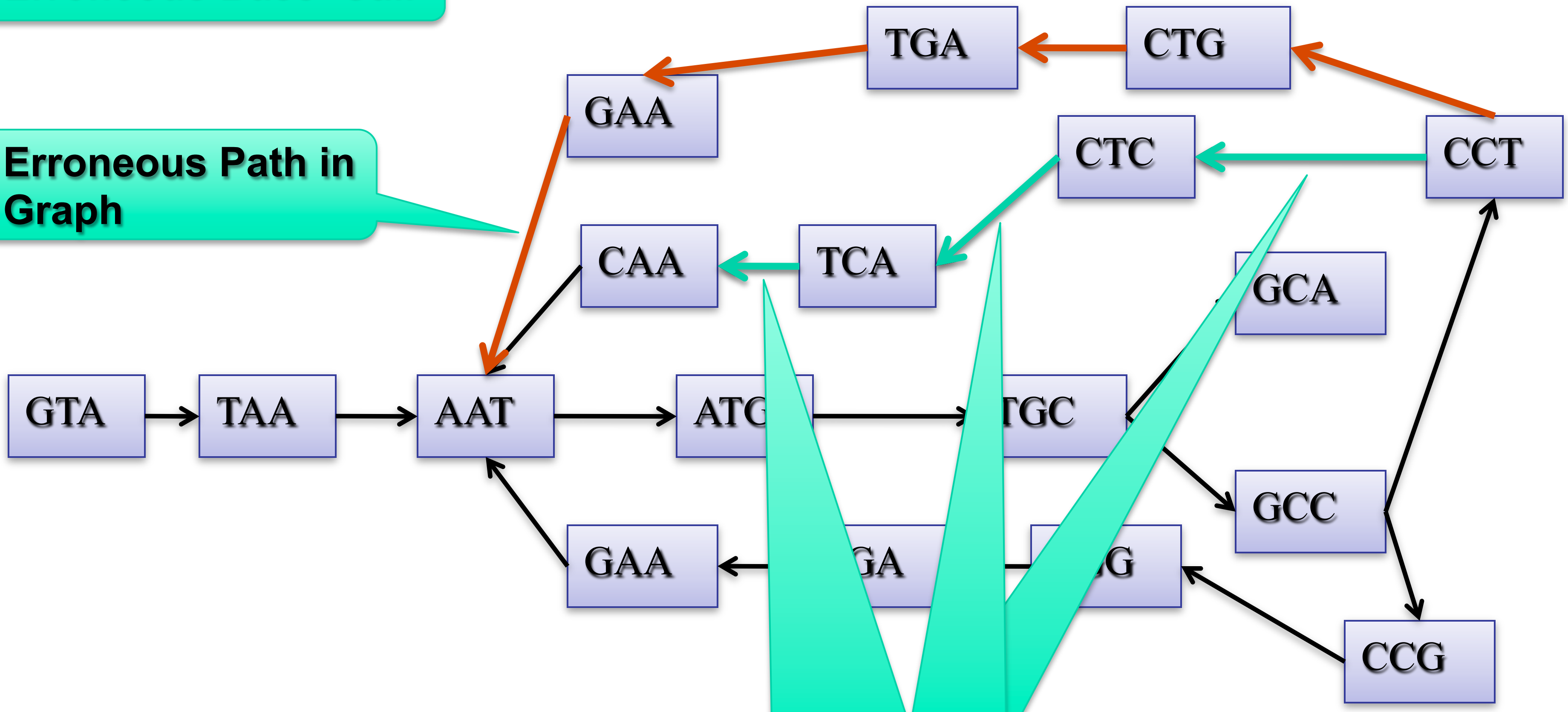
# Sources of Assembly Errors

- Errors in reads – caused by technology
  - Error in base calls, color calls (SOLID Technology), or repeated base calls (454 Technology)
- Missing reads – sequencing bias
- Read orientation error
  - One or both orientations may occur
  - Not told which ones are present
- Sequence Variations – mixed sample study
  - SNP, cancer, metagenomics studies
- Combinations of the above

# *Handling Read Error*

GTAATGCCTCAATGCCGGAATGCA

CTGAA



Erroneous Base Call

Erroneous Path in Graph

Potential Missing Edges in Graph

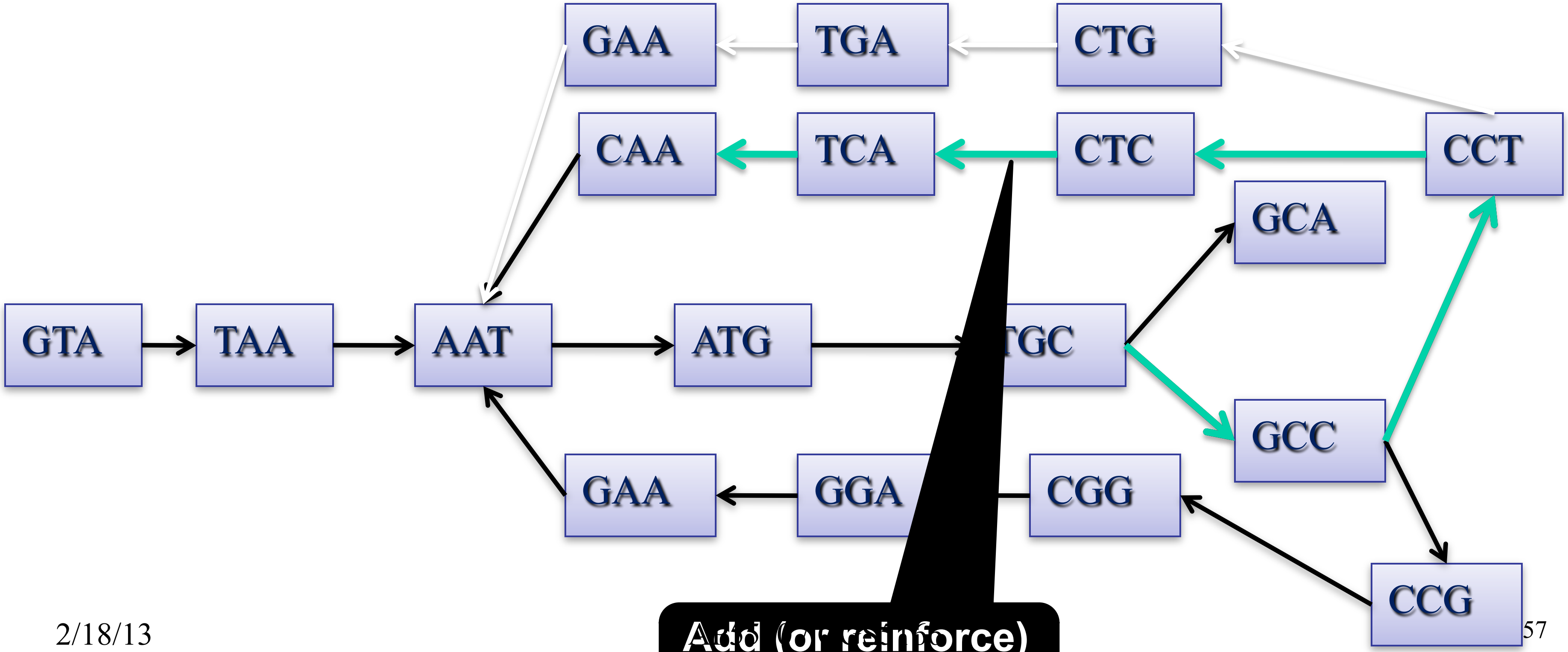
**Well conserved regions in related genomes**

TGCCTCAA

TGCCTCAA

GTAATGCCTCAATGCCGGAATGCA

CTGAA



**Add (or reinforce) path in graph**

# Ideas

- Start with k-mer graph or string graph or overlap graph or contig (Velvet) graph
  - Advantages/disadvantages of each?
- Place highly conserved reads or regions on this graph
- Identify missing nodes/edges/paths

# When is a genome assembly done?

- Almost never perfectly! Great cost in time, effort, and money.
  - Currently 92% of human genome is done to 99.99% accuracy [Schmutz et al., Nature 429, 365-368]
  - More likely to complete with bacterial and viral genomes, but they evolve much faster.
- Hard part with bacterial genomes are genomic rearrangements
- Often enough to get gene content to perform comparative genomics
- Tools to compare gene content
  - CEGMA – Eukaryote
  - CheckM – Bacterial; https://peerj.com/preprints/554.pdf
- Useful papers
  - Salzberg et al., Genome Res, 2012
  - Vezzi et al., PLoS ONE, 2012, DOI: 10.1371/journal.pone.0031002
  - Gurevich et al., Bioinformatics, 29(8): 1072-75, 2013
  - Shengguan et al., PLoS ONE, 2013, DOI: 10.1371/journal.pone.0069890

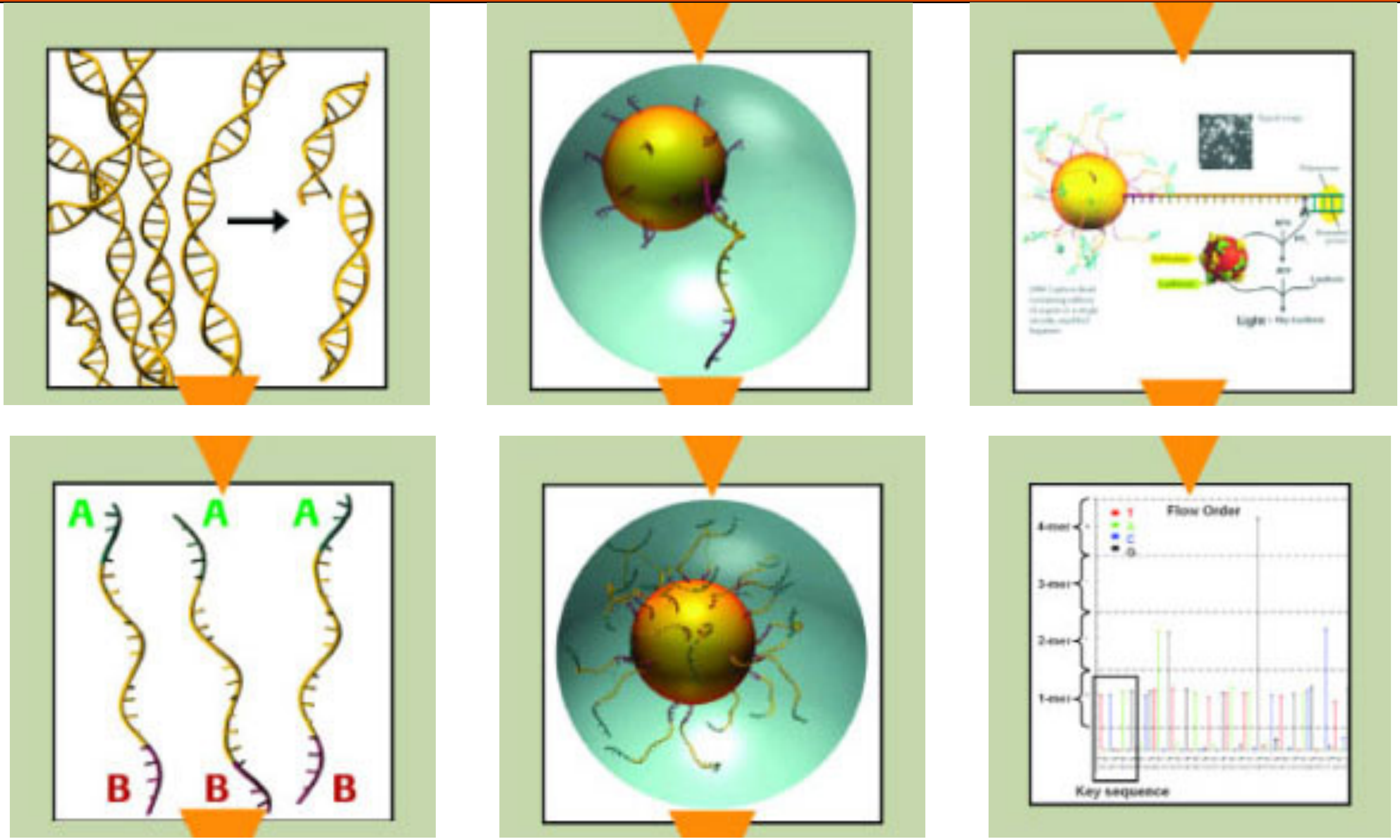
# N50 measure

- https://www.broad.harvard.edu/crd/wiki/index.php/N50
- Statistical measure of "average length" of a set of sequences.
- Used widely in evaluating assemblies.
- N50 length is defined as the length N for which 50% of all bases in the sequences are in a sequence of length L < N.
- N50 is a weighted median statistic such that 50% of entire assembly is contained in contigs or scaffolds equal to or larger than this value
- Given list of lengths L. Create another list L' , which is identical to L, except that every element n in L has been replaced with n copies of itself. Then the median of L' is the N50 of L.
- **Example**:
  - Let L = {2, 2, 2, 3, 3, 4, 8, 8},
  - L' consists of six 2's, six 3's, four 4's, and sixteen 8's; the N50 of L is the median of L' , which is 6.
  - Alternatively, sum = 32, halfSum = 16. You need the two 8's to sum up to 16

# 454 Sequencing: New Sequencing Technology

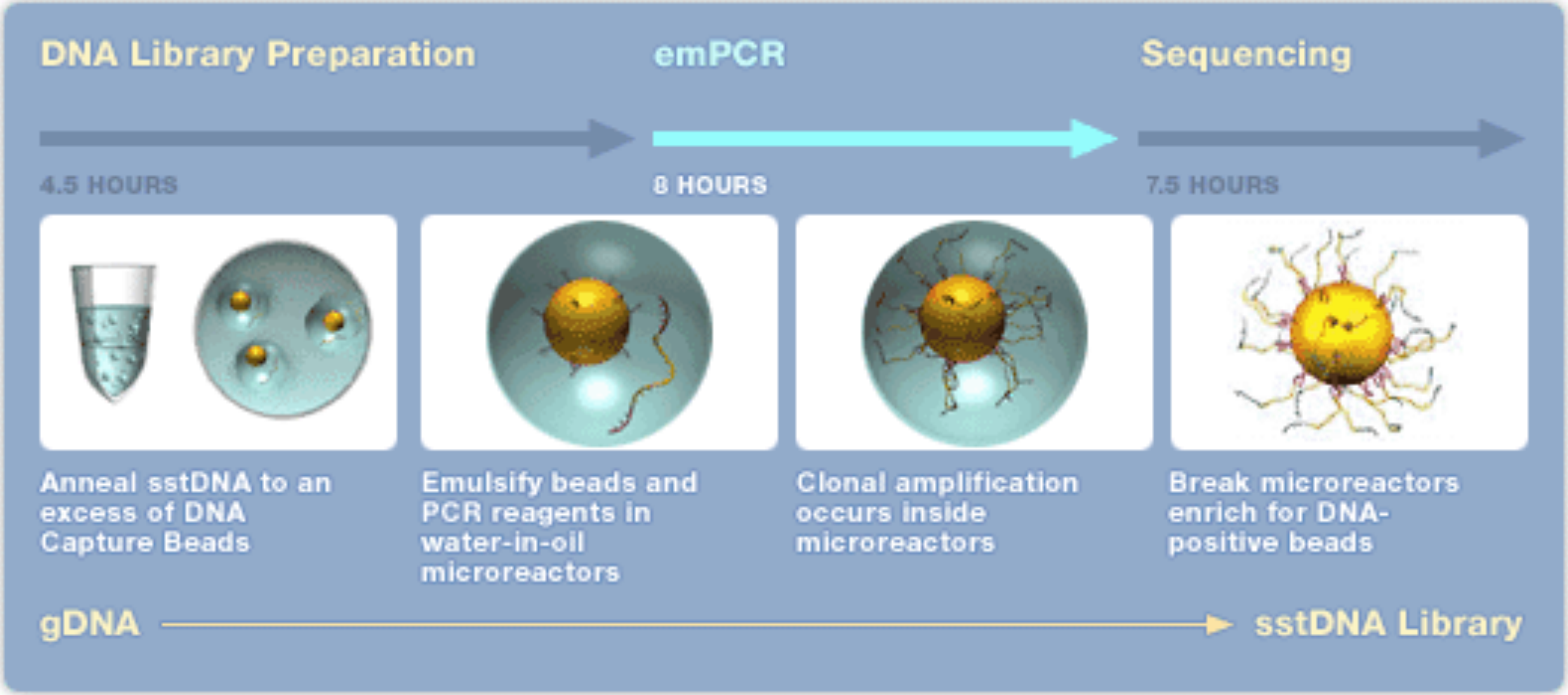
- This technology, started in 2005 and is now being phased out
- 454 Life Sciences, Roche
- Fast (20 million bases per 4.5 hour run)
- Low cost (lower than Sanger sequencing)
- Simple (entire bacterial genome in days with one person -- without cloning and colony picking)
- Convenient (complete solution 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.

# (a) Fragment, (b) add adaptors, (c) “1 fragment, 1 bead”, (d) emPCR on bead, (e) put beads in PicoTiterPlate and start sequencing: “1 bead, 1 read”, and (f) analyze



# emPCR

FIGURE 8



DNA Library Preparation

4.5 HOURS

emPCR

8 HOURS

Sequencing

7.5 HOURS

Anneal sstDNA to an excess of DNA Capture Beads

Emulsify beads and PCR reagents in water-in-oil microreactors

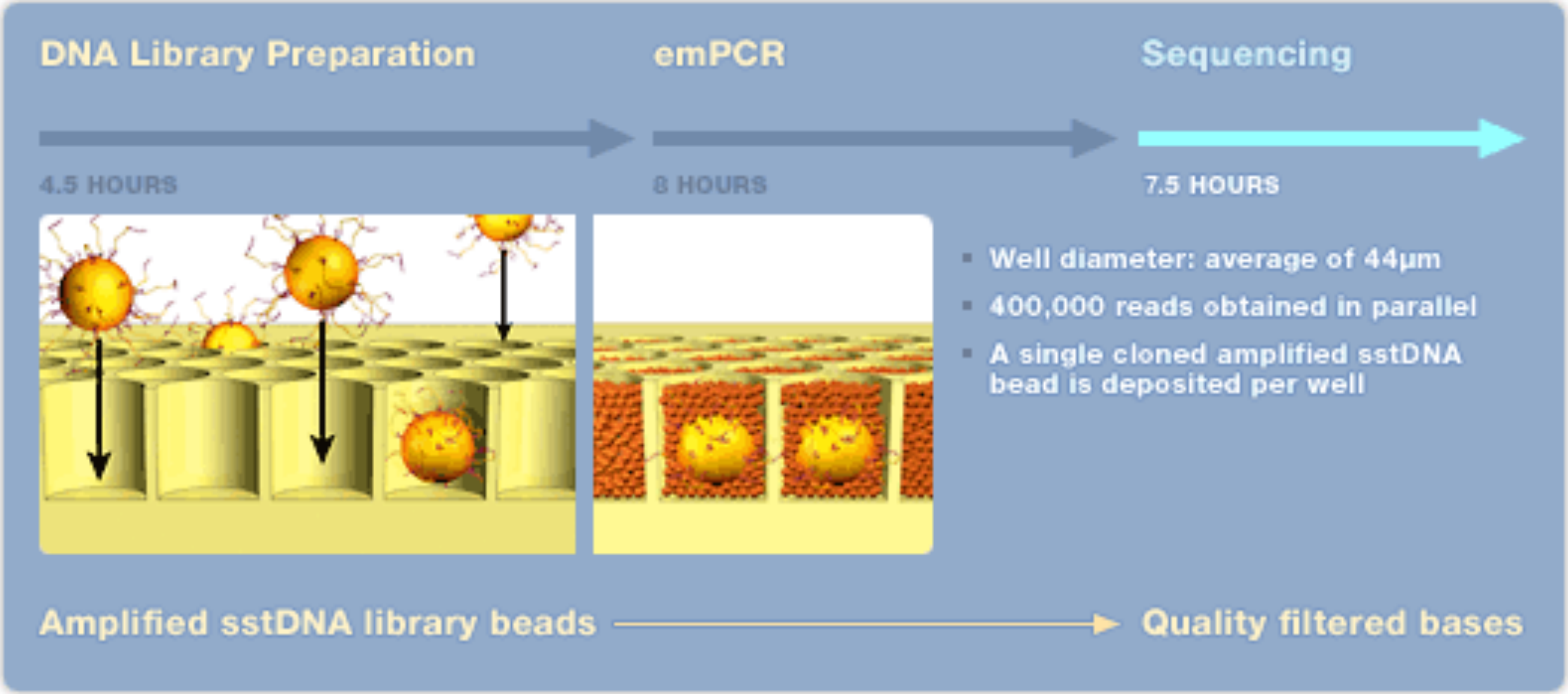
Clonal amplification occurs inside microreactors

Break microreactors enrich for DNA-positive beads

gDNA → sstDNA Library

# Sequencing

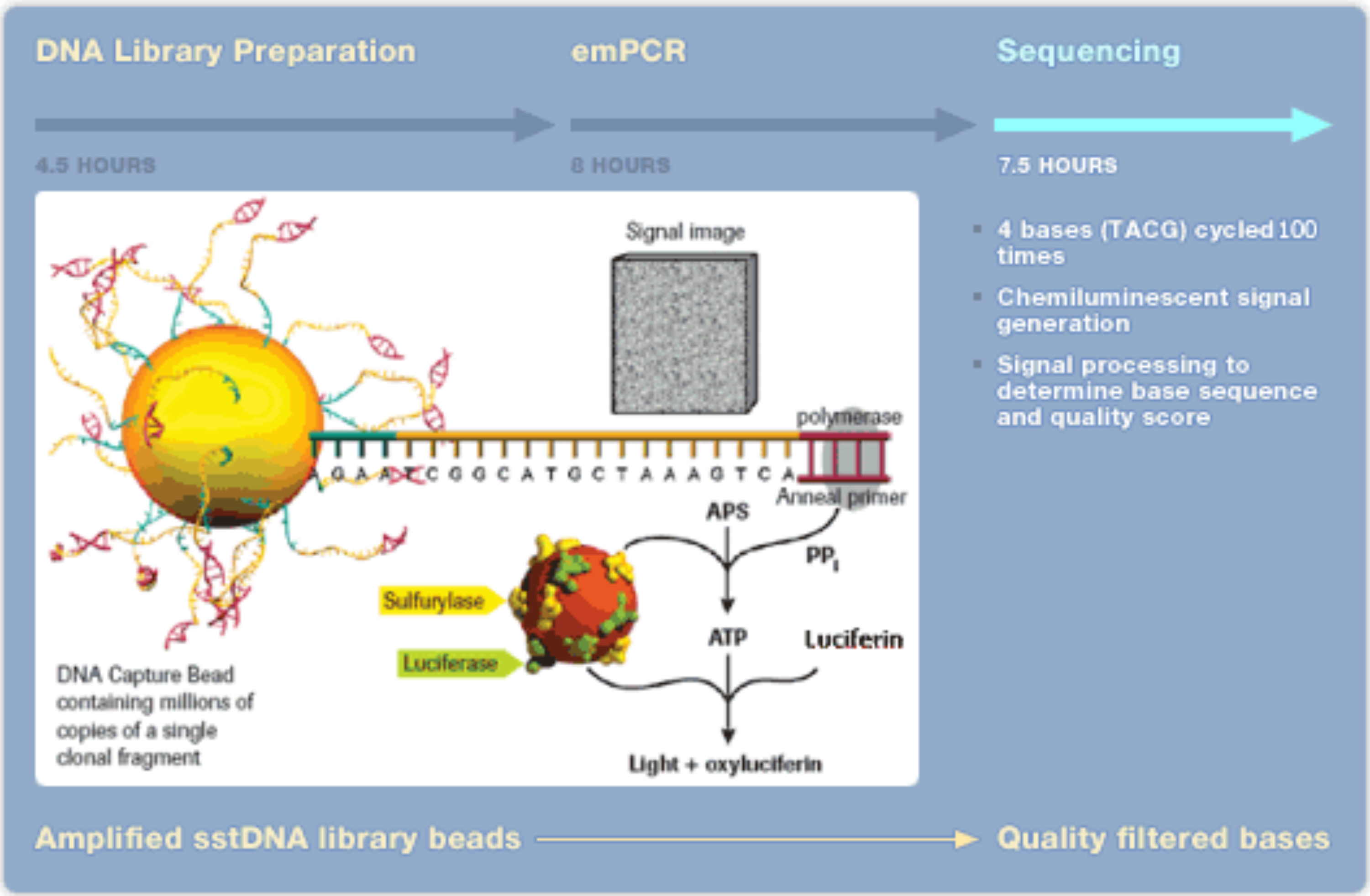
FIGURE 9



- Well diameter: average of 44µm
- 400,000 reads obtained in parallel
- A single cloned amplified sstDNA bead is deposited per well

# Sequencing

FIGURE 10



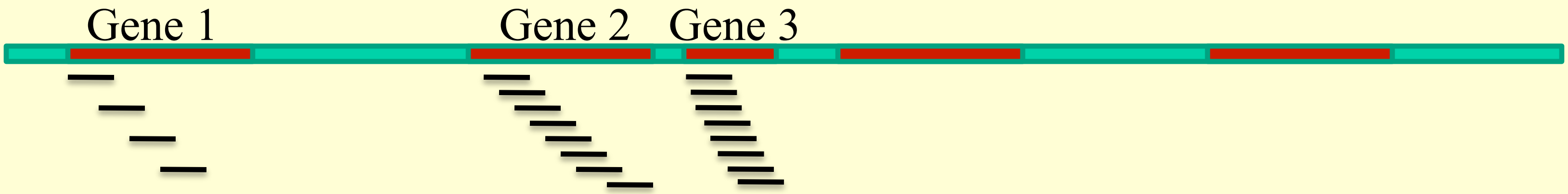
- 4 bases (TACG) cycled 100 times
- Chemiluminescent signal generation
- Signal processing to determine base sequence and quality score

# NGS Applications

# Applications of NGS

- ❑ RNA-Seq
- ❑ ChIP-Seq
- ❑ SNP-Seq
- ❑ Metagenomics
- ❑ Alternative Splicing
- ❑ Copy Number Variations (CNV)
- ❑ …

# RNA-Seq



- Align reads to genes and count
- Assume uniform sampling
  - Count of number of reads mapped per gene is a measure of its expression level
  - Expression of Gene 2 is twice that of Gene 1
  - Expression of Gene 3 is twice that of Gene 2

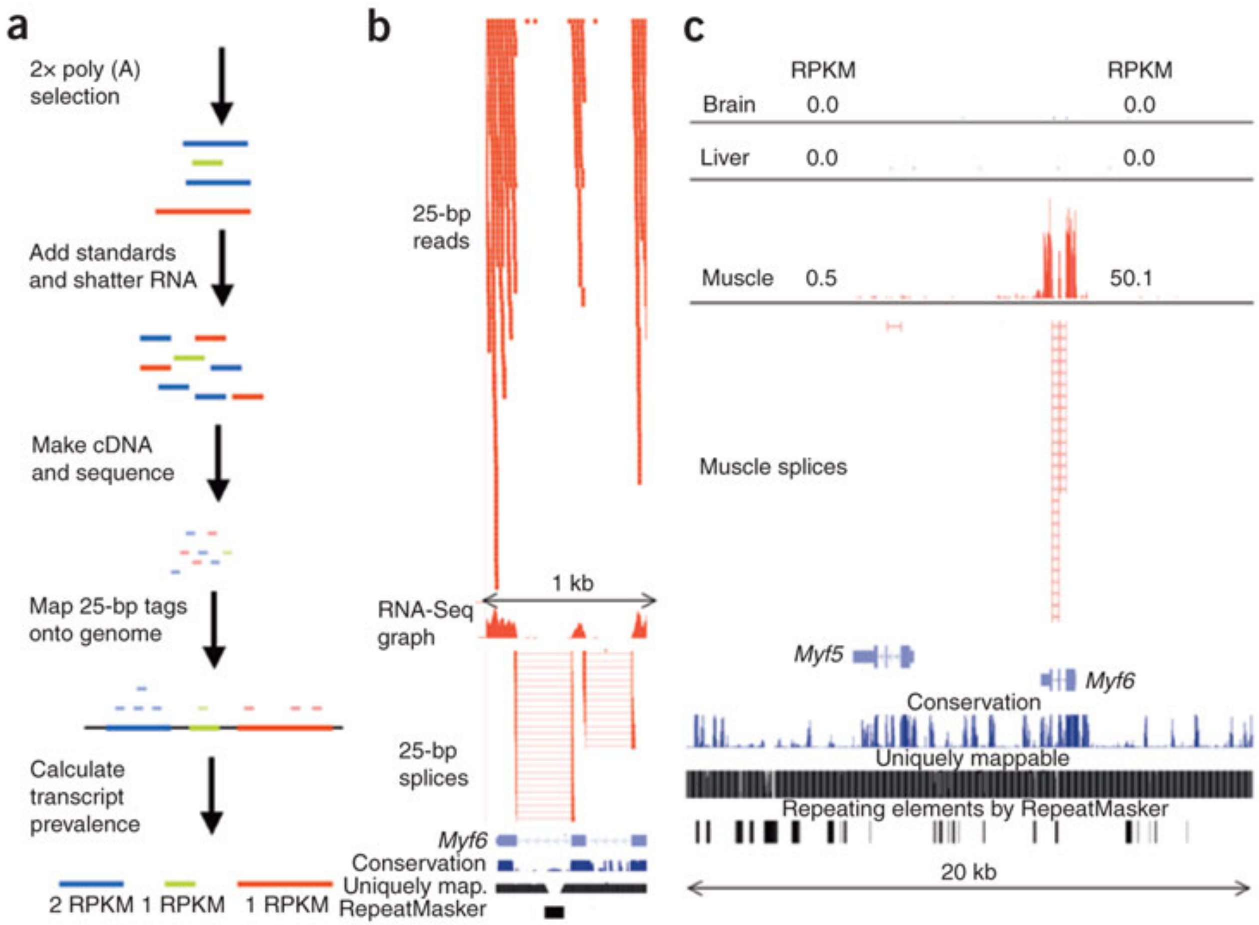
# Expression Level of Gene

- RPKM = Ng / (N X L)
  - Ng = Number of reads mapped to gene
  - N = Total number of mapped reads (in millions)
  - L = Length of gene in KB
  - [Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B., Nat Methods. 2008 Jul;5(7):621-8. **Mapping and quantifying mammalian transcriptomes by RNA-Seq.**]

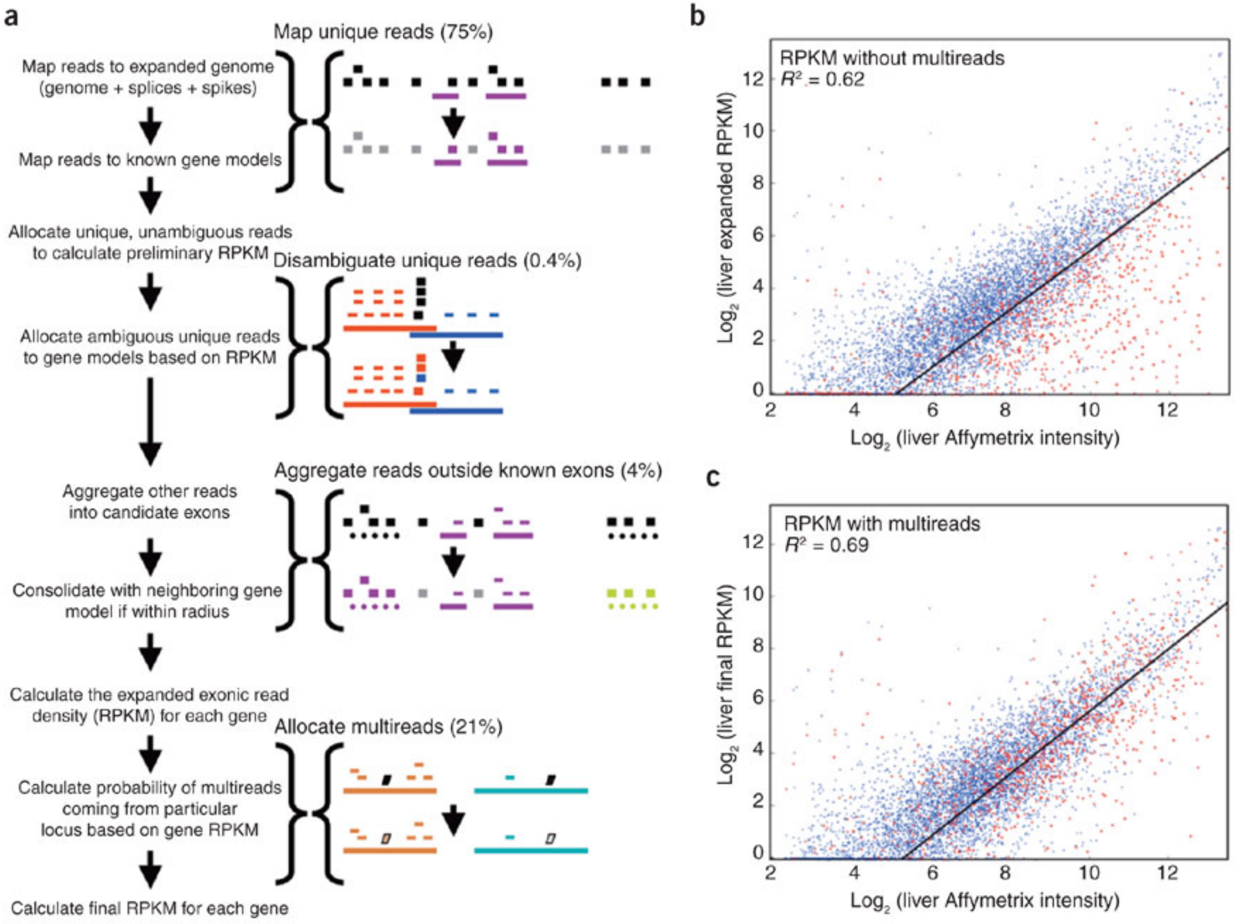
# Complications

- Repeat regions
  - Paralogs and other homologous regions in genes
  - Ambiguities in maps
- Introns and Exons
  - Aligning reads to genome is more complex
- Alternative Splicing
- Transcription start site is upstream of ORFs
- Unknown ORFs and Small RNAs
- Other transcripts

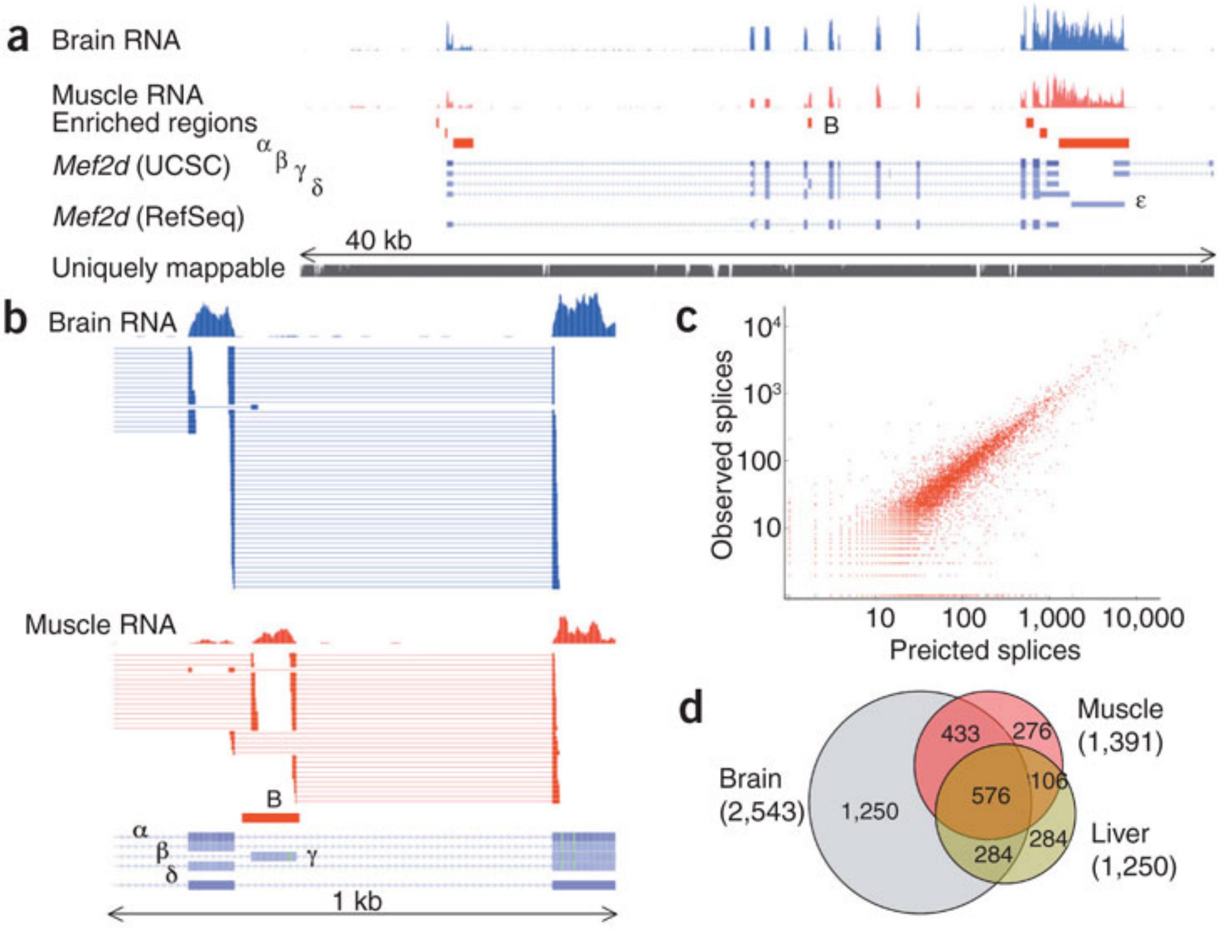
# RNA-Seq Procedure



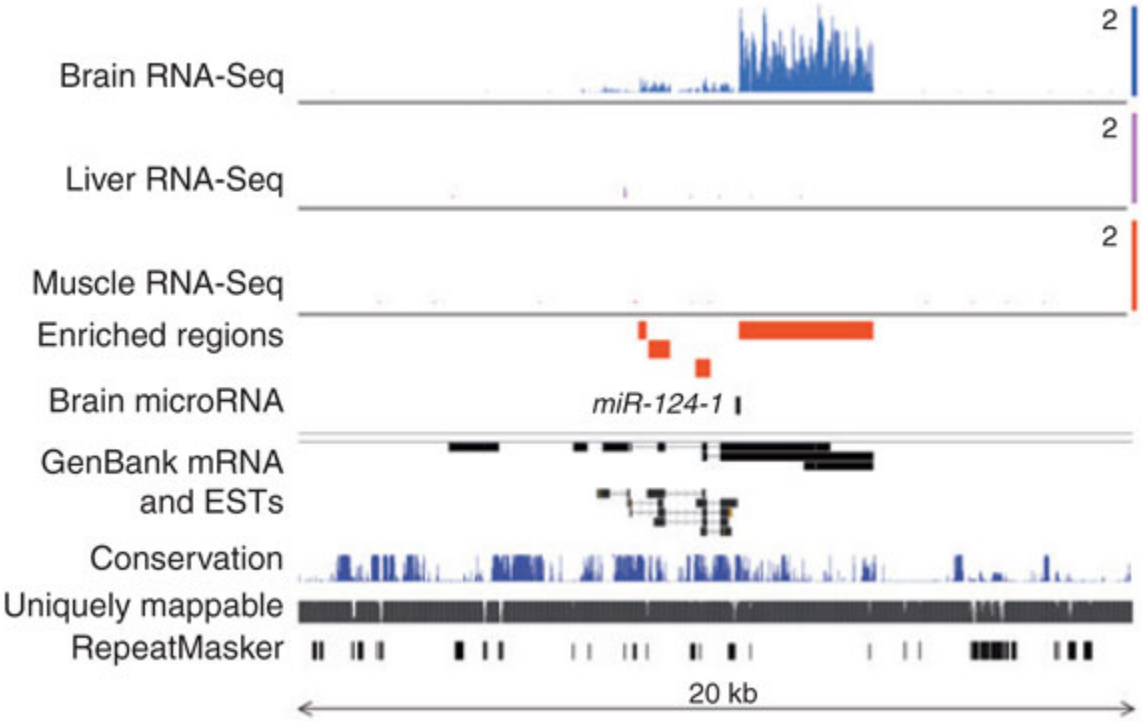
# Mapping Reads to Reference



# Alternative Splicing



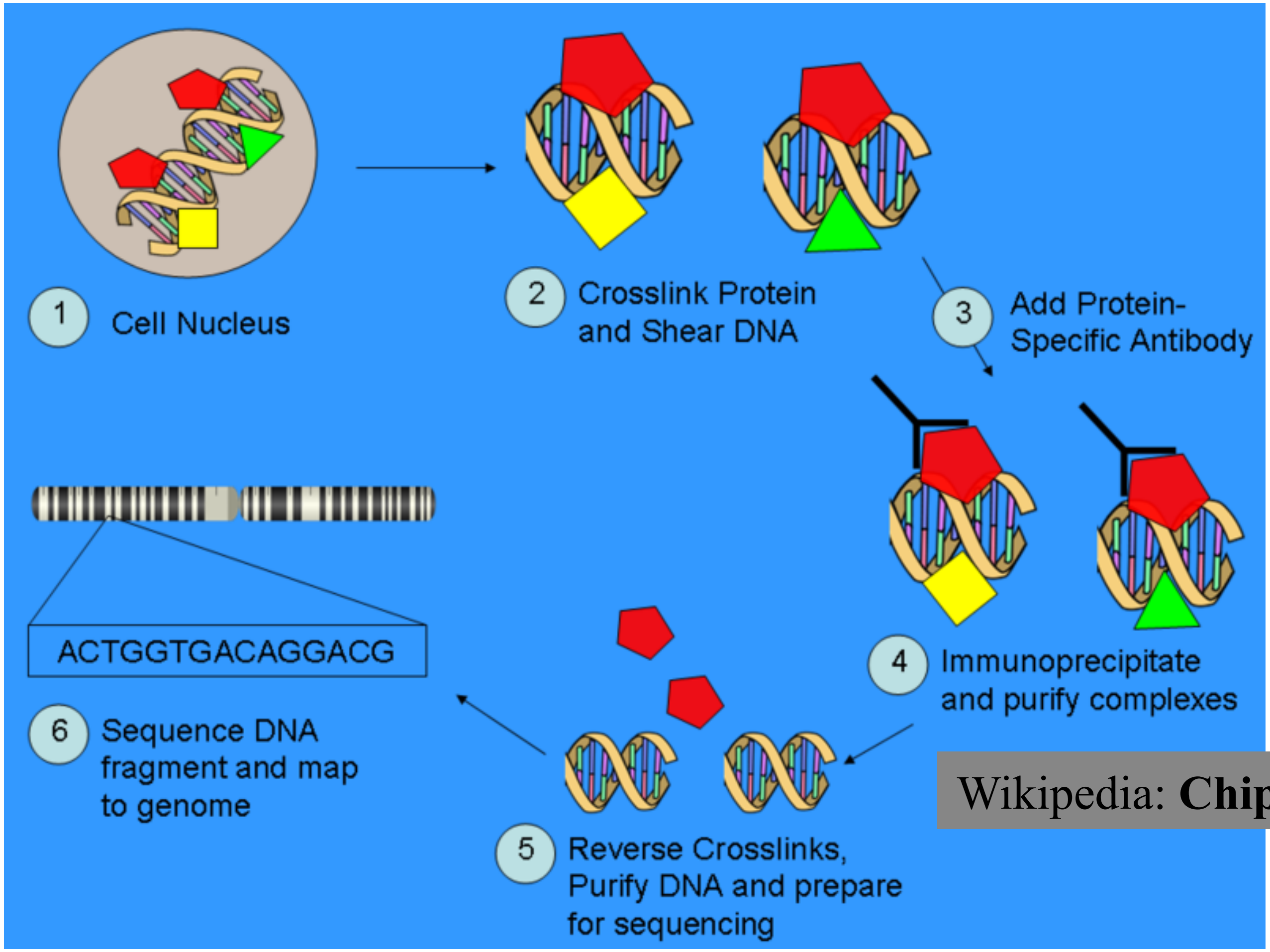
# microRNA


Brain RNA-Seq
2
Liver RNA-Seq
2
Muscle RNA-Seq
2
Enriched regions
Brain microRNA
miR-124-1
GenBank mRNA and ESTs
Conservation
Uniquely mappable
RepeatMasker
20 kb

# Chromatin Immunoprecipitation

- Useful for pinpointing location of TFBS for TF
- High-throughput method to find all binding sites for a specific TF under specific conditions
- Identify sites using
  - ChIP-on-chip (Microarray technique)
  - ChIP-Seq (Sequencing technique)
- Problems: TFs bind to specific TFBS only under specific conditions – hard to predict

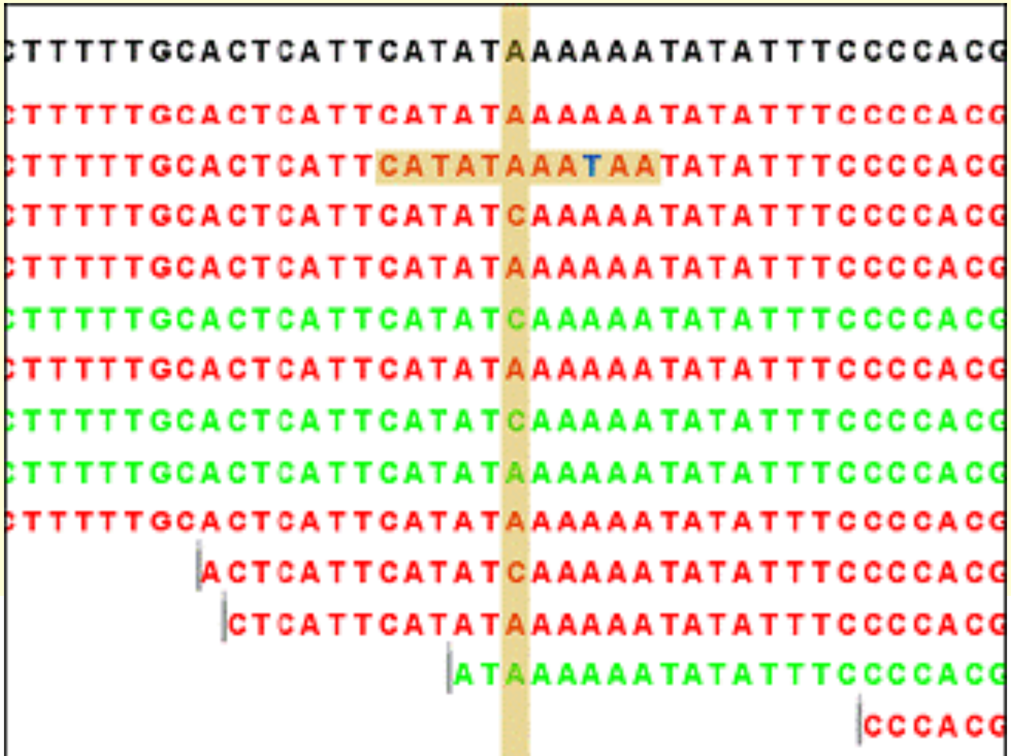
# ChIP-Seq


1 Cell Nucleus
2 Crosslink Protein and Shear DNA
3 Add Protein-Specific Antibody
4 Immunoprecipitate and purify complexes
5 Reverse Crosslinks, Purify DNA and prepare for sequencing
6 Sequence DNA fragment and map to genome
ACTGGTGACAGGACG

Wikipedia: **Chip-Sequencing**

# SNP-Seq

- ❑ Align reads and look for differences
  - Differences to reference
    - Align reads to reference sequence first
  - Differences within reads
  - Differences between samples or sets of reads



```
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAATAATATATTTCCCCACG
CTTTTTGCACTCATTCATATCAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATCAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATCAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
CTTTTTGCACTCATTCATATAAAAAATATATTTCCCCACG
        ACTCATTCATATCAAAAATATATTTCCCCACG
         CTCATTCATATAAAAAATATATTTCCCCACG
                  ATAAAAAATATATTTCCCCACG
                                  CCCACG
```

# Environmental Microbiology

- Conventional methods
  - Culture, then identify
    - Slow, expensive, labor intensive, unculturable microbes
  - PCR-based length heterogeneity studies
- Microarray-based methods
  - Unique probes for organisms (e.g., Virochip)
    - Only works for sequenced regions of known organisms
- NGS-based methods

# Metagenomics

- Detect known pathogens
- Diversity
  - Identity of individual species not needed
- Functional profile of community

# NGS-based method

- ❑ Map reads against appropriate database
- ❑ Identify closest hits for each read
- ❑ Generate contigs
- ❑ Generate abundance information
- ❑ Clustering of reads can be beneficial to estimate abundance

# Profiles and HMMs

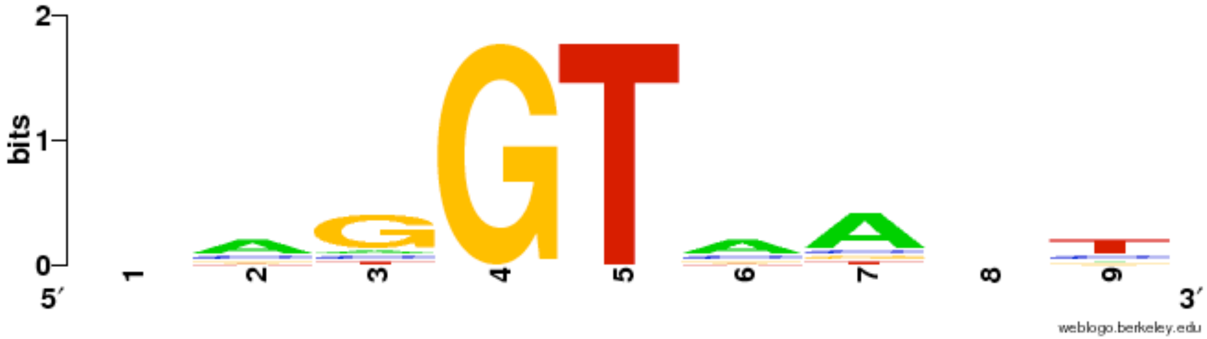
# Pattern: Representations

**GAGGTAAAC**
**TCCGTAAGT**
**CAGGTTGGA**
**ACAGTCAGT**
**TAGGTCATT**
**TAGGTACTG**
**ATGGTAACT**
**CAGGTATAC**
**TGTGTGAGT**
**AAGGTAAGT**

**TAGGTAAGT**

- Alignments
- Consensus Sequences
- Logo Formats
- …

TAGGTAAGT



# Profiles

GAGGTAAAC

TCCGTAAGT

CAGGTTGGA

ACAGTCAGT

TAGGTCATT

TAGGTACTG

ATGGTAACT

CAGGTATAC

TGTGTGAGT

AAGGTAAGT

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | 3 | 6 | 1 | 0 | 0 | 6 | 7 | 2 | 1 |
| C | 2 | 2 | 1 | 0 | 0 | 2 | 1 | 1 | 2 |
| G | 1 | 1 | 7 | 10 | 0 | 1 | 1 | 5 | 1 |
| T | 4 | 1 | 1 | 0 | 10 | 1 | 1 | 2 | 6 |

Frequency Matrix

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | .3 | .6 | .1 | 0 | 0 | .6 | .7 | .2 | .1 |
| C | .2 | .2 | .1 | 0 | 0 | .2 | .1 | .1 | .2 |
| G | .1 | .1 | .7 | 1 | 0 | .1 | .1 | .5 | .1 |
| T | .4 | .1 | .1 | 0 | 1 | .1 | .1 | .2 | .6 |

Relative Frequencies

# Profiles

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | .3 | .6 | .1 | 0 | 0 | .6 | .7 | .2 | .1 |
| C | .2 | .2 | .1 | 0 | 0 | .2 | .1 | .1 | .2 |
| G | .1 | .1 | .7 | 1 | 0 | .1 | .1 | .5 | .1 |
| T | .4 | .1 | .1 | 0 | 1 | .1 | .1 | .2 | .6 |

Relative Frequencies

|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | 0.14 | 0.72 | -0.61 | -1.43 | -1.43 | 0.72 | 0.86 | -0.16 | -0.61 |
| C | -0.16 | -0.16 | -0.61 | -1.43 | -1.43 | -0.16 | -0.61 | -0.61 | -0.16 |
| G | -0.61 | -0.61 | 0.86 | -0.61 | -1.43 | -0.61 | -0.61 | 0.57 | -0.61 |
| T | 0.38 | -0.61 | -0.61 | -1.43 | 1.19 | -0.61 | -0.61 | -0.16 | 0.72 |

# Profiles

Profile entries:

$$P_{ij} = \ln (f_{ij}/b_i)$$

Zero counts:

$$f_{ij} = (c_{ij}+\alpha b_i)/ (n+\alpha)$$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | .3 | .6 | .1 | 0 | 0 | .6 | .7 | .2 | .1 |
| C | .2 | .2 | .1 | 0 | 0 | .2 | .1 | .1 | .2 |
| G | .1 | .1 | .7 | 1 | 0 | .1 | .1 | .5 | .1 |
| T | .4 | .1 | .1 | 0 | 1 | .1 | .1 | .2 | .6 |

Relative Frequencies

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | 0.14 | 0.72 | -0.61 | -1.43 | -1.43 | 0.72 | 0.86 | -0.16 | -0.61 |
| C | -0.16 | -0.16 | -0.61 | -1.43 | -1.43 | -0.16 | -0.61 | -0.61 | -0.16 |
| G | -0.61 | -0.61 | 0.86 | 1.19 | -1.43 | -0.61 | -0.61 | 0.57 | -0.61 |
| T | 0.38 | -0.61 | -0.61 | -1.43 | 1.19 | -0.61 | -0.61 | -0.16 | 0.72 |

http://coding.plantpath.ksu.edu/profile/

# CpG Islands

- ❑Regions in DNA sequences with increased occurrences of substring "CG"
- ❑Rare: typically C gets methylated and then mutated into a T.
- ❑Often around promoter or "start" regions of genes
- ❑Few hundred to a few thousand bases long

# Problem 1:

- Input: Small sequence S
- Output: Is S from a CpG island?
  - Build Markov models: M+ and M —
  - Then compare

# Markov Models

| + | A | C | G | T |
|---|---|---|---|---|
| A | 0.180 | 0.274 | 0.426 | 0.120 |
| C | 0.171 | 0.368 | 0.274 | 0.188 |
| G | 0.161 | 0.339 | 0.375 | 0.125 |
| T | 0.079 | 0.355 | 0.384 | 0.182 |

| — | A | C | G | T |
|---|---|---|---|---|
| A | 0.300 | 0.205 | 0.285 | 0.210 |
| C | 0.322 | 0.298 | 0.078 | 0.302 |
| G | 0.248 | 0.246 | 0.298 | 0.208 |
| T | 0.177 | 0.239 | 0.292 | 0.292 |

# How to distinguish?

- Compute

$$S(x) = \log\left(\frac{P(x \mid M+)}{P(x \mid M-)}\right) = \sum_{i=1}^{L} \log\left(\frac{p_{x(i-1)xi}}{m_{x(i-1)xi}}\right) = \sum_{i=1}^{L} r_{x(i-1)xi}$$

| r=p/m | A | C | G | T |
|---|---|---|---|---|
| A | -0.740 | 0.419 | 0.580 | -0.803 |
| C | -0.913 | 0.302 | 1.812 | -0.685 |
| G | -0.624 | 0.461 | 0.331 | -0.730 |
| T | -1.169 | 0.573 | 0.393 | -0.679 |

**Score(GCAC)**
**= .461-.913+.419**
**< 0.**
**GCAC not from CpG island.**

**Score(GCTC)**
**= .461-.685+.573**
**> 0.**
**GCTC from CpG island.**

Problem 1:

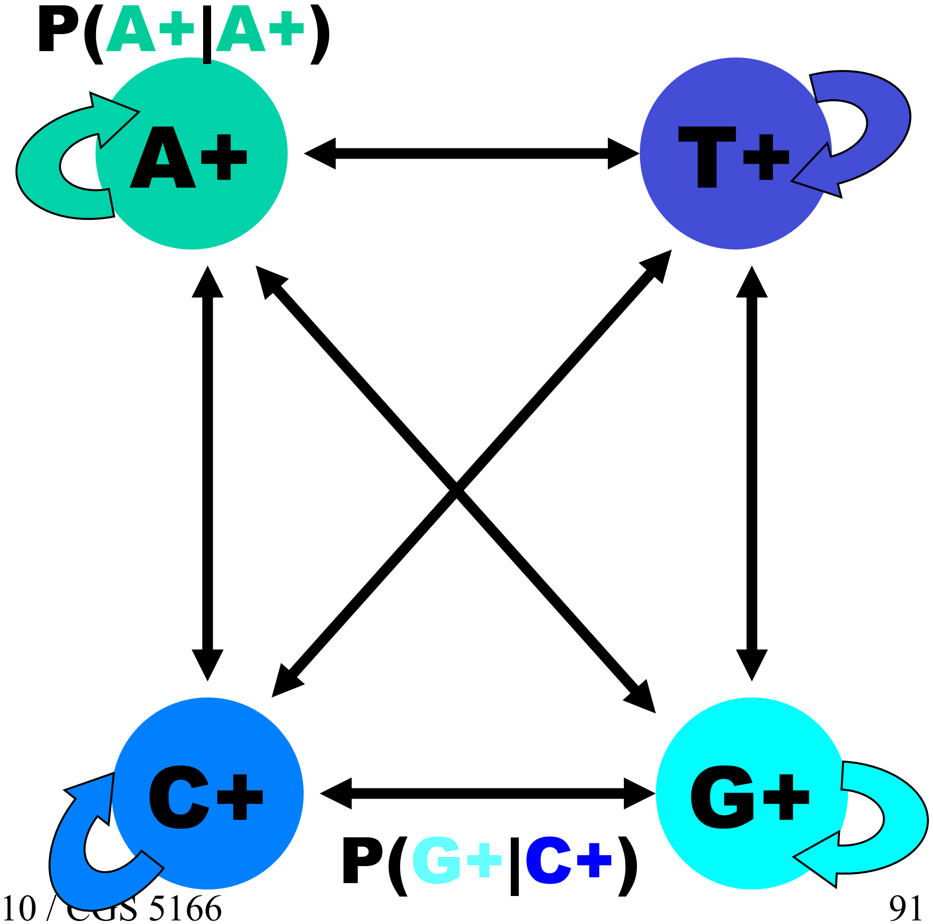
- Input: Small sequence S
- Output: Is S from a CpG island?
  - Build Markov Models: M+ & M-
  - Then compare

Problem 2:

- Input: Long sequence S
- Output: Identify the CpG islands in S.
  - Markov models are inadequate.
  - Need Hidden Markov Models.

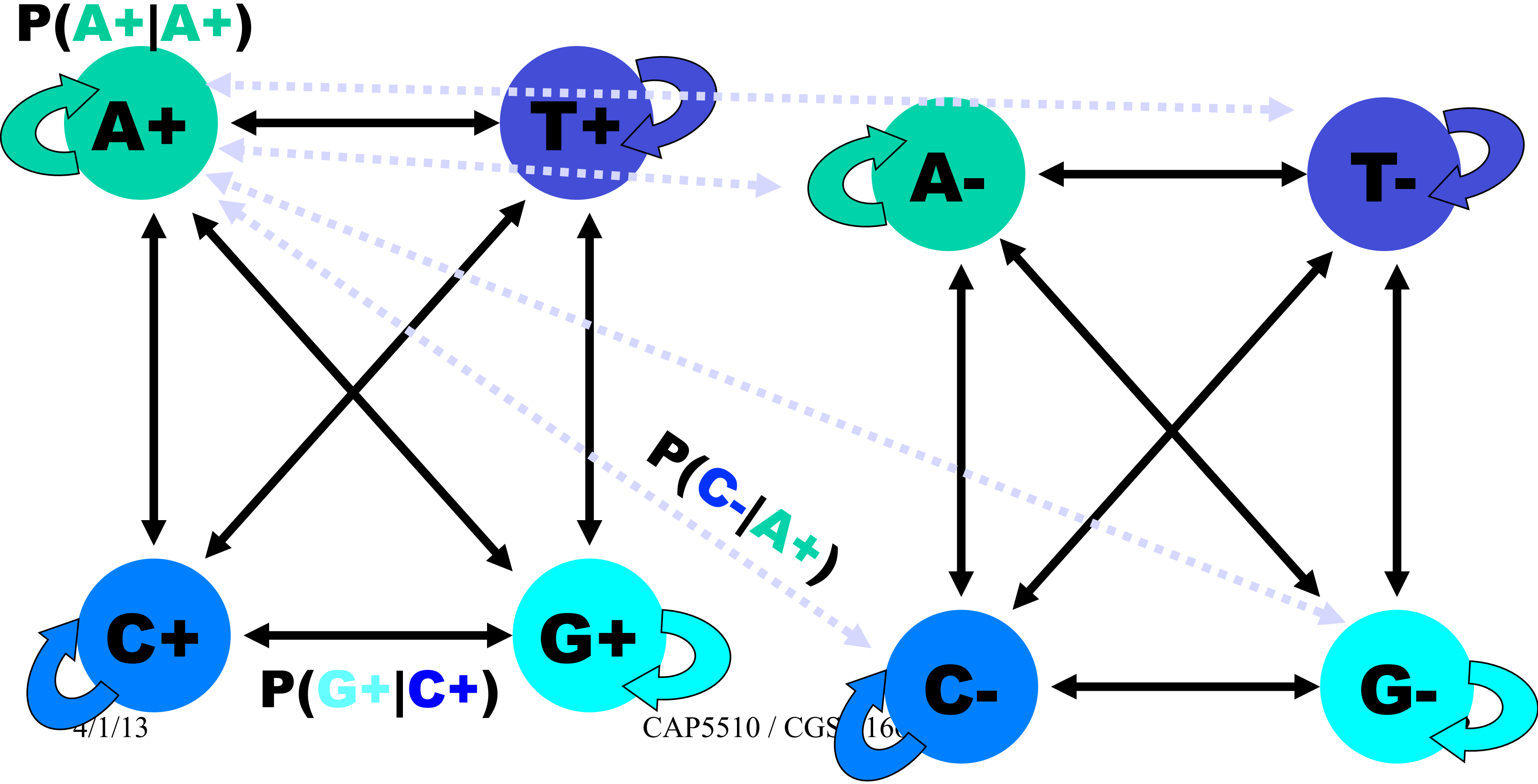
# Markov Models

| + | A | C | G | T |
|---|---|---|---|---|
| A | 0.180 | 0.274 | 0.426 | 0.120 |
| C | 0.171 | 0.368 | 0.274 | 0.188 |
| G | 0.161 | 0.339 | 0.375 | 0.125 |
| T | 0.079 | 0.355 | 0.384 | 0.182 |



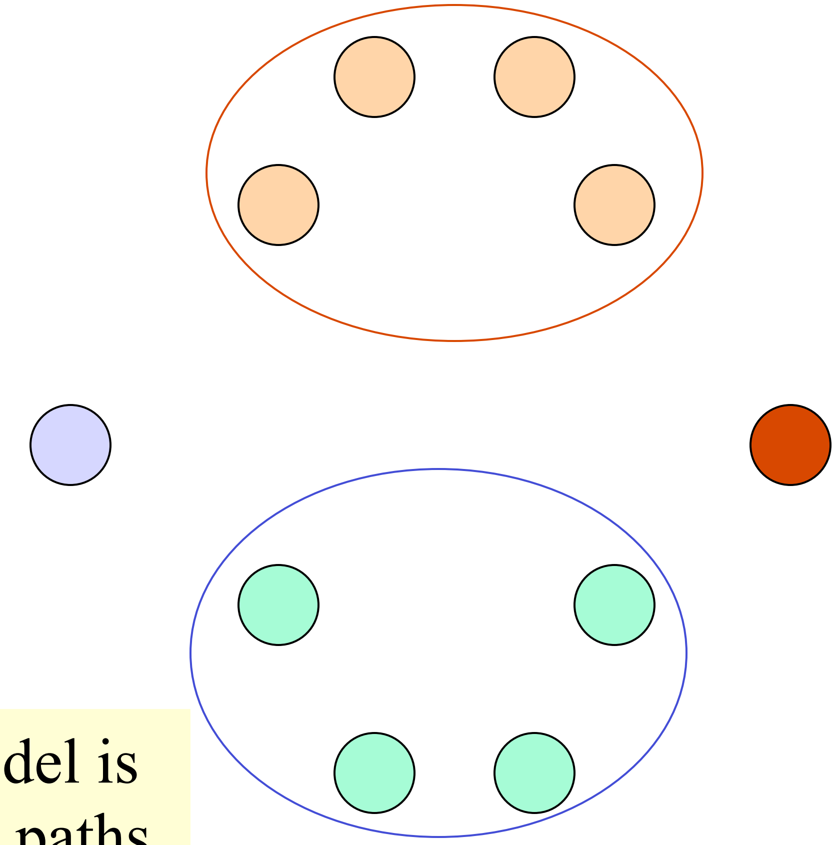
# CpG Island + in an ocean of –
# First order Hidden Markov Model

MM=16, HMM= 64 transition probabilities (adjacent bp)



# Hidden Markov Model (HMM)

- States
- Transitions
- Transition Probabilities
- Emissions
- Emission Probabilities



- What is hidden about HMMs?

Answer: The path through the model is hidden since there are many valid paths.

# How to Solve Problem 2?

❑ Solve the following problem:

Input: Hidden Markov Model M,

parameters Θ, emitted sequence S

Output: Most Probable Path Π

How: Viterbi's Algorithm (Dynamic Programming)

**Define Π[i,j] = MPP for first j characters of S ending in state i**

**Define P[i,j] = Probability of Π[i,j]**

- **Compute state i with largest P[i,j].**

# Profiles

Profile entries:

$$P_{ij} = \ln (f_{ij}/b_i)$$

Zero counts:

$$f_{ij} = (c_{ij}+\alpha b_i)/(n+\alpha)$$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | .3 | .6 | .1 | 0 | 0 | .6 | .7 | .2 | .1 |
| C | .2 | .2 | .1 | 0 | 0 | .2 | .1 | .1 | .2 |
| G | .1 | .1 | .7 | 1 | 0 | .1 | .1 | .5 | .1 |
| T | .4 | .1 | .1 | 0 | 1 | .1 | .1 | .2 | .6 |

Relative Frequencies

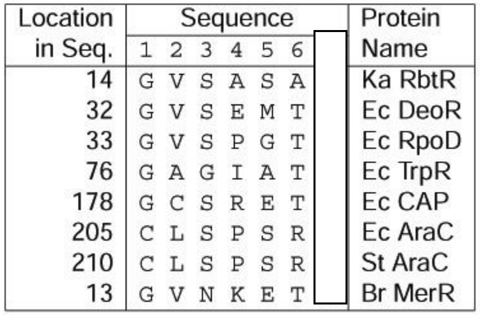
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | 0.14 | 0.72 | -0.61 | -1.43 | -1.43 | 0.72 | 0.86 | -0.16 | -0.61 |
| C | -0.16 | -0.16 | -0.61 | -1.43 | -1.43 | -0.16 | -0.61 | -0.61 | -0.16 |
| G | -0.61 | -0.61 | 0.86 | 1.19 | -1.43 | -0.61 | -0.61 | 0.57 | -0.61 |
| T | 0.38 | -0.61 | -0.61 | -1.43 | 1.19 | -0.61 | -0.61 | -0.16 | 0.72 |

Profiles; Position Weight Matrix (PWM); Position-Specific Scoring Matrix (PSSM)

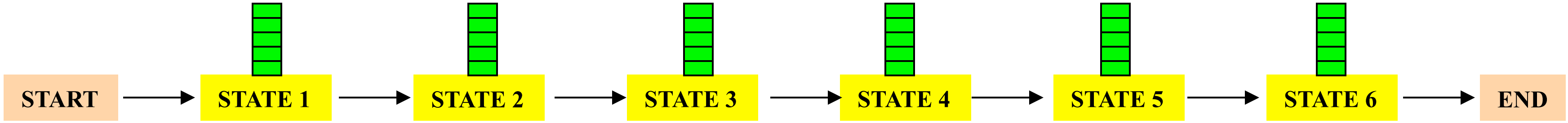
http://coding.plantpath.ksu.edu/profile/

# Profile HMMs

PROFILE METHOD, [M. Gribskov et al., '90]

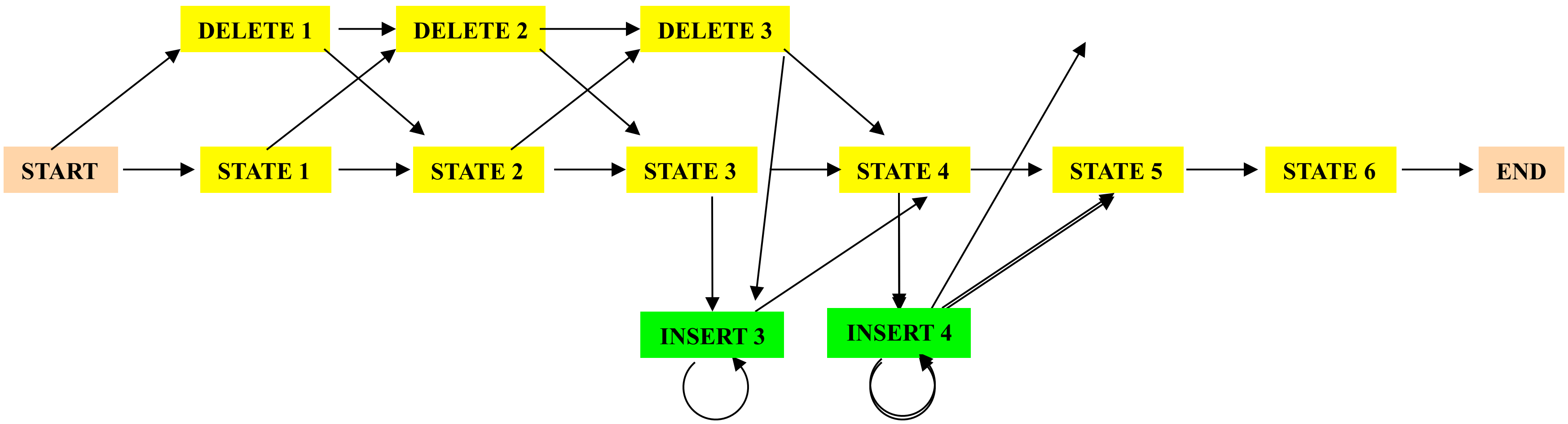


| Location | Sequence | | | | | | | Protein |
|---|---|---|---|---|---|---|---|---|
| in Seq. | 1 | 2 | 3 | 4 | 5 | 6 | | Name |
| 14 | G | V | S | A | S | A | | Ka RbtR |
| 32 | G | V | S | E | M | T | | Ec DeoR |
| 33 | G | V | S | P | G | T | | Ec RpoD |
| 76 | G | A | G | I | A | T | | Ec TrpR |
| 178 | G | C | S | R | E | T | | Ec CAP |
| 205 | C | L | S | P | S | R | | Ec AraC |
| 210 | C | L | S | P | S | R | | St AraC |
| 13 | G | V | N | K | E | T | | Br MerR |

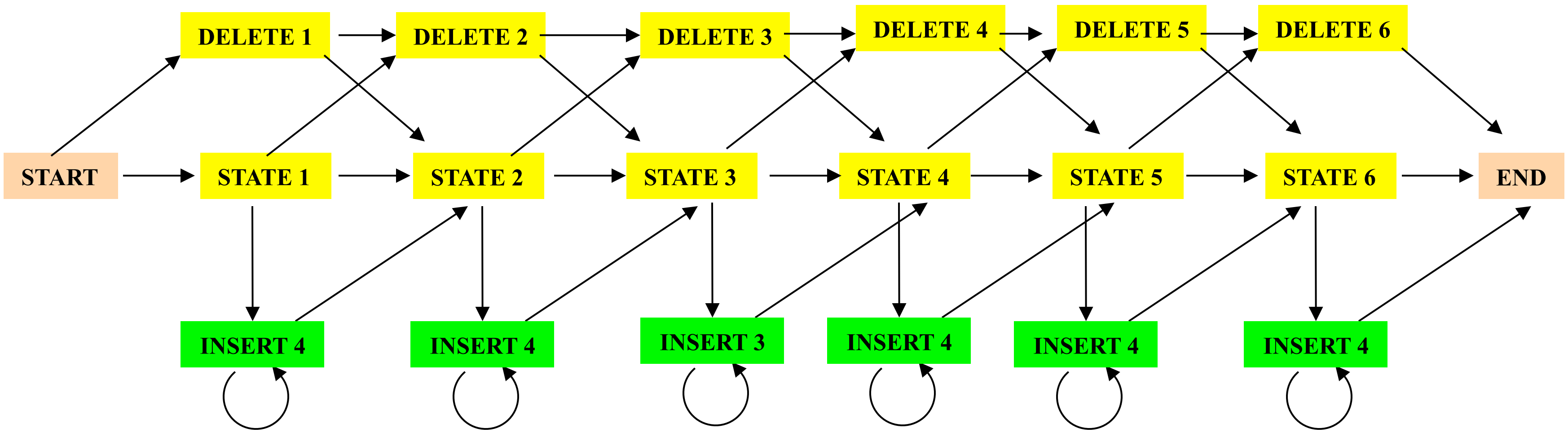


# Profile HMMs with InDels

- Insertions
- Deletions
- Insertions & Deletions

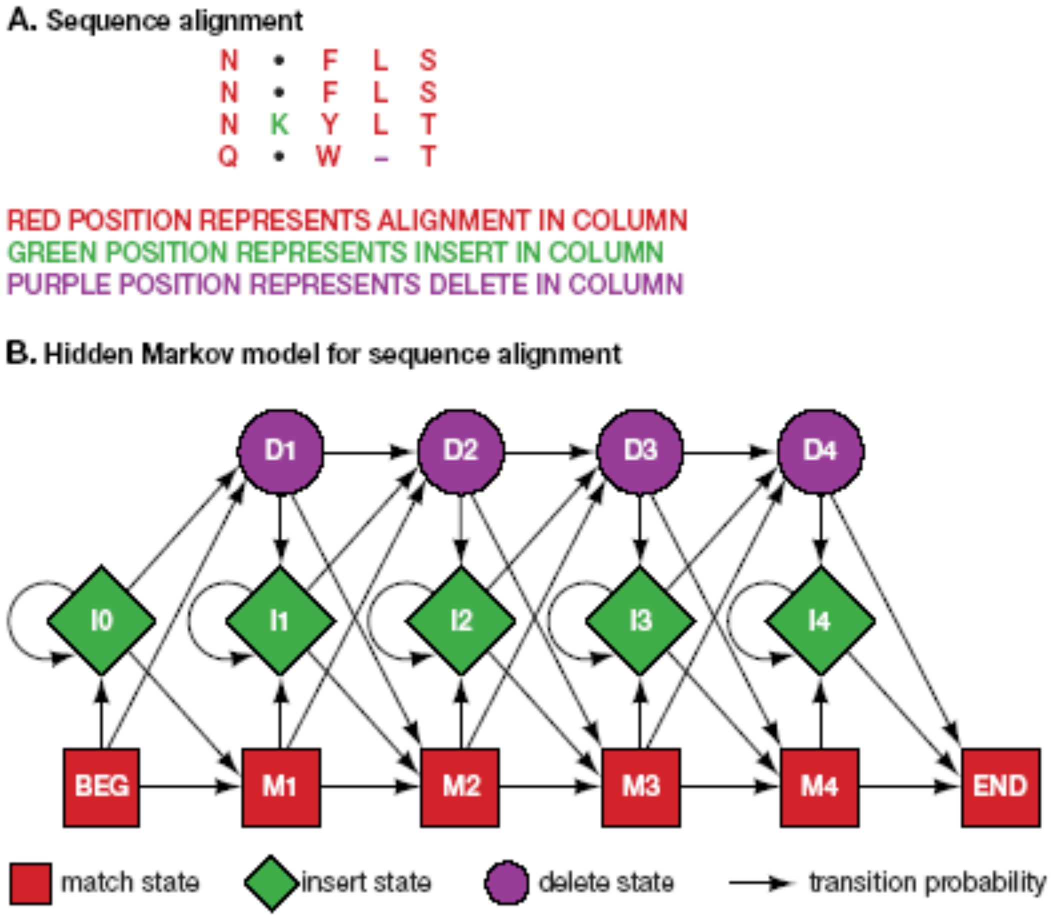


# Profile HMMs with InDels



Missing transitions from DELETE j to INSERT j and from INSERT j to DELETE j+1.

# HMM for Sequence Alignment



**A.** Sequence alignment

N • F L S
N • F L S
N K Y L T
Q • W – T

RED POSITION REPRESENTS ALIGNMENT IN COLUMN
GREEN POSITION REPRESENTS INSERT IN COLUMN
PURPLE POSITION REPRESENTS DELETE IN COLUMN

**B.** Hidden Markov model for sequence alignment

FIGURE 5.16. Relationship between the sequence alignment and the hidden Markov model of the alignment (Krogh et al. 1994). This particular form for the HMM was chosen to represent the sequence, structural, and functional variation expected in proteins. The model accommodates the identities, mismatches, insertions, and deletions expected in a group of related proteins. (*A*) A section of an msa. The illustration shows the columns generated in an msa. Each column may include matches and mismatches (*red* positions), insertions (*green* positions), and deletions (*purple* positions). (*B*) The HMM. Each column in the model represents the possibility of a match, insert, or delete in each column of the alignment in *A*. The HMM is a probabilistic representation of a section of the msa. Sequences can be generated from the HMM by starting at the beginning state labeled BEG and then by following

# Profile HMM Software

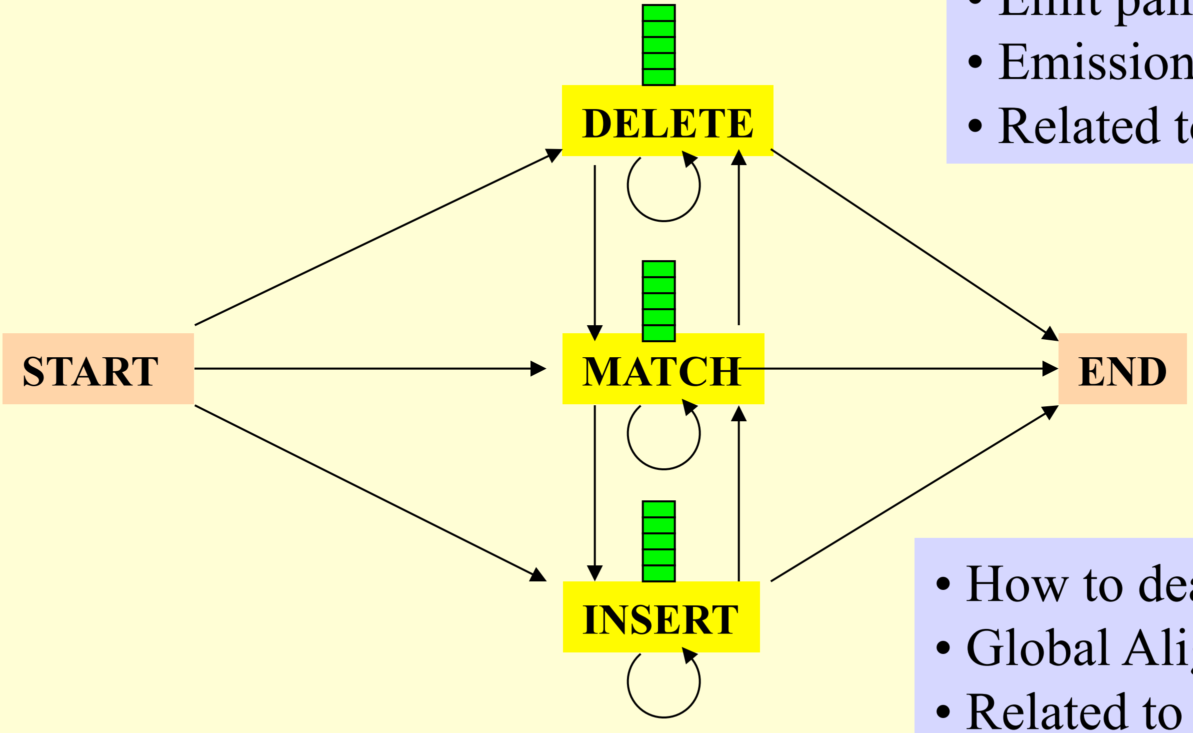
| | |
|---|---|
| HMMER | http://hmmer.wustl.edu/ |
| SAM | http://www.cse.ucsc.edu/research/compbio/sam.html |
| PFTOOLS | http://www.isrec.isb-sib.ch/ftp-server/pftools/ |
| HMMpro | http://www.netid.com/html/hmmpro.html |
| GENEWISE | http://www.ebi.ac.uk/Wise2/ |
| PROBE | ftp://ftp.ncbi.nih.gov/pub/neuwald/probe1.0/ |
| META-MEME | http://metameme.sdsc.edu/ |
| BLOCKS | http://www.blocks.fhcrc.org/ |
| PSI-BLAST | http://www.ncbi.nlm.nih.gov/BLAST/newblast.html |

- Read more about Profile HMMs at
  - http://www.csb.yale.edu/userguides/seq/hmmer/docs/node9.html

# How to model Pairwise Sequence Alignment

**LEAPVE**

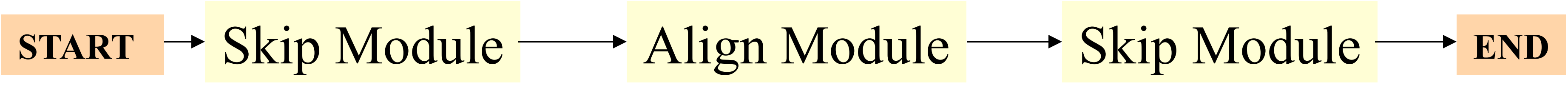
**LAPVIE**



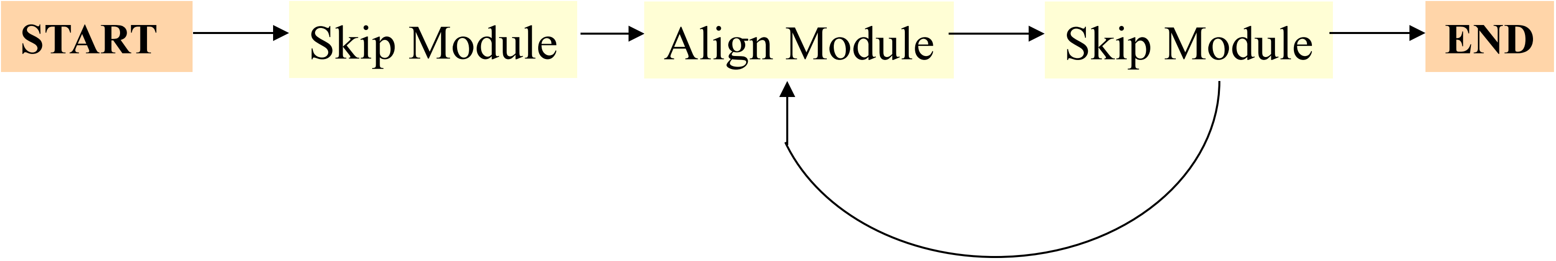
Pair HMMs
- Emit pairs of synbols
- Emission probs?
- Related to Sub. Matrices

- How to deal with InDels?
- Global Alignment? Local?
- Related to Sub. Matrices

# How to model Pairwise Local Alignments?

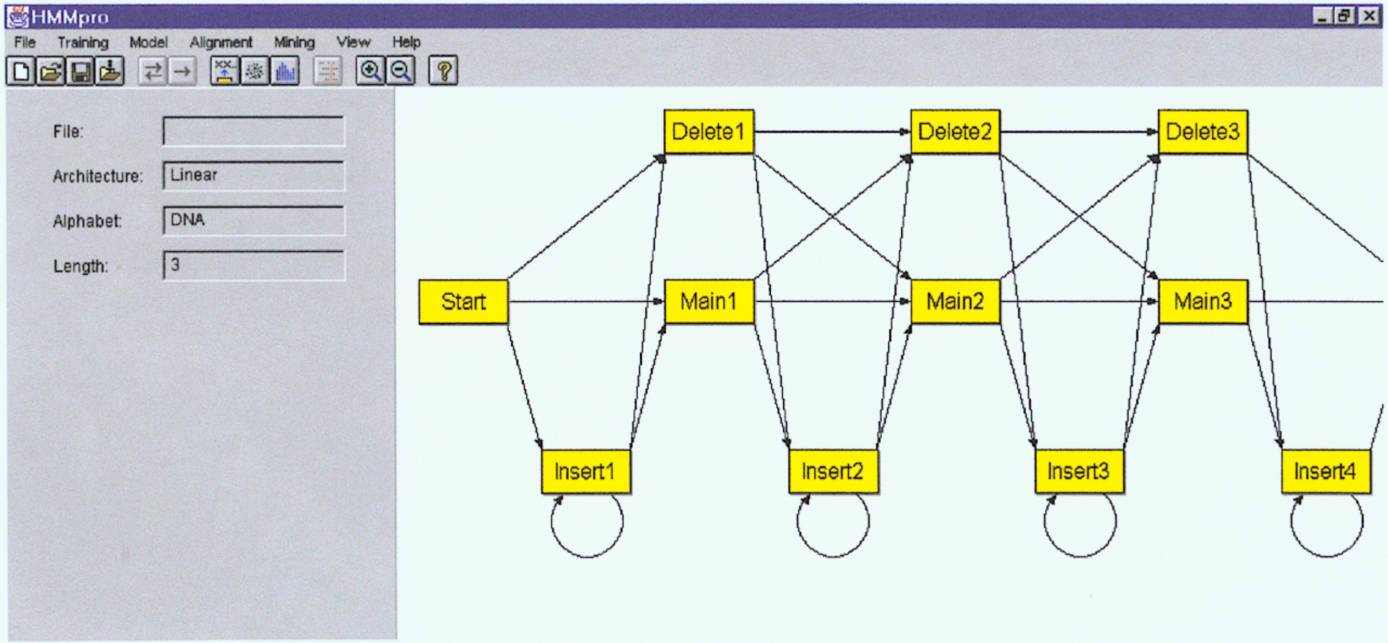


## How to model Pairwise Local Alignments with gaps?



# Standard HMM architectures

## Linear Architecture



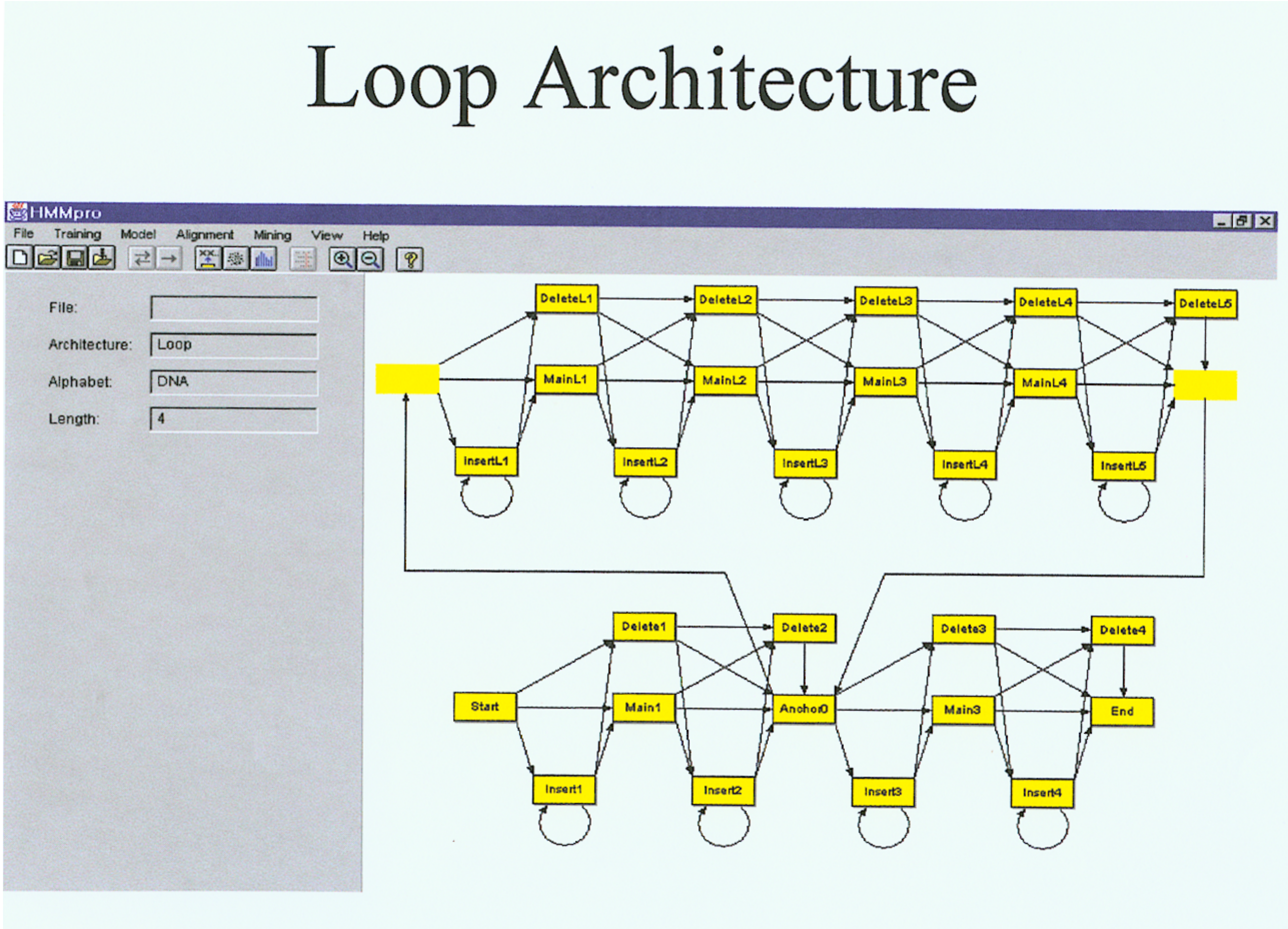
File:

Architecture: Linear

Alphabet: DNA

Length: 3

# Standard HMM architectures



## Loop Architecture

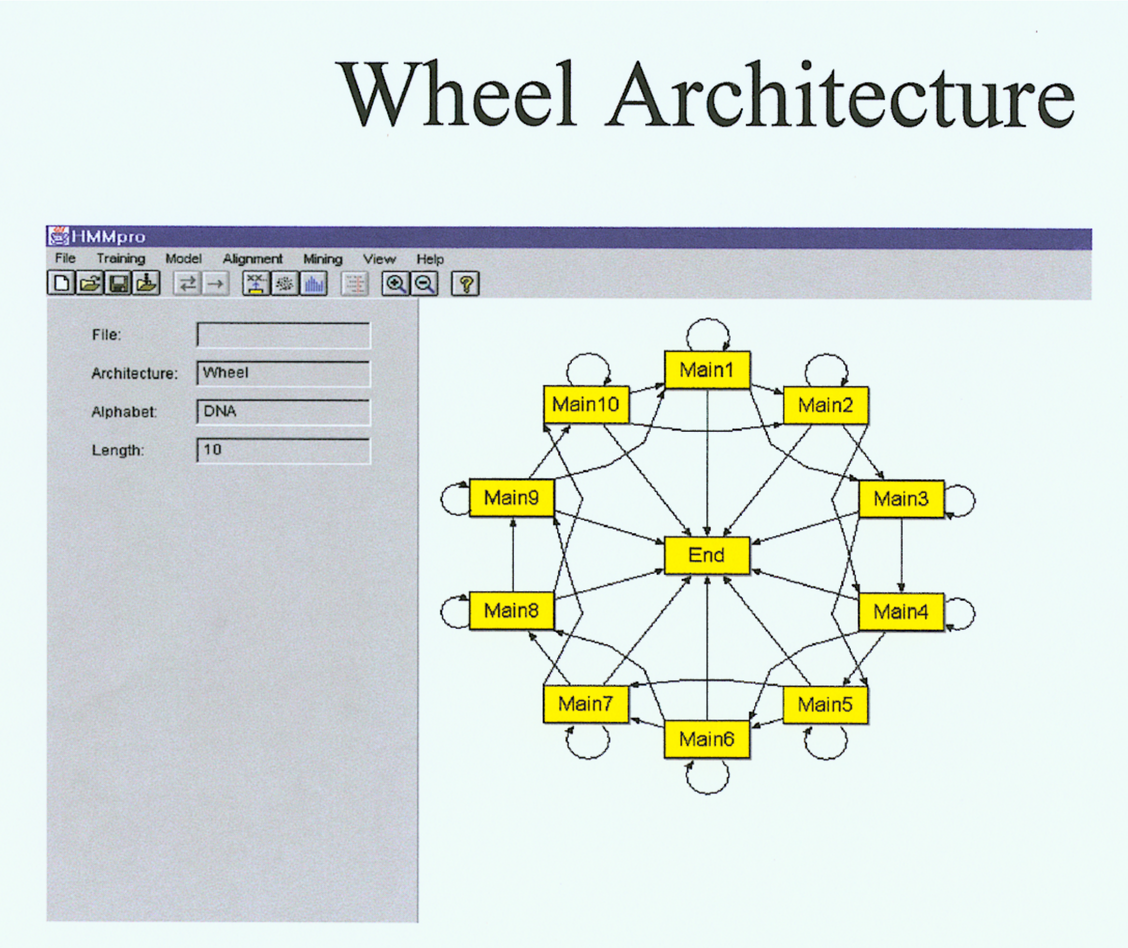
File:

Architecture: Loop

Alphabet: DNA

Length: 4

# Standard HMM architectures



Wheel Architecture

File:

Architecture: Wheel

Alphabet: DNA

Length: 10

## Problem 3: LIKELIHOOD QUESTION

- Input: Sequence S, model M, state i
- Output: Compute the probability of reaching state i with sequence S using model M
  - Backward Algorithm (DP)

## Problem 4: LIKELIHOOD QUESTION

- Input: Sequence S, model M
- Output: Compute the probability that S was emitted by model M
  - Forward Algorithm (DP)

## Problem 5: LEARNING QUESTION

- Input: model structure M, Training Sequence *S*
- Output: Compute the parameters Θ
- Criteria: ML criterion
  - maximize P(*S* | M, Θ)    HOW???

## Problem 6: DESIGN QUESTION

- Input: Training Sequence *S*
- Output: Choose model structure M, and compute the parameters Θ
  - No reasonable solution
  - Standard models to pick from

# Iterative Solution to the LEARNING QUESTION (Problem 5)

- ❑ Pick initial values for parameters $\Theta_0$
- ❑ Repeat
  - Run training set $S$ on model $M$
  - Count # of times transition $i \Rightarrow j$ is made
  - Count # of times letter x is emitted from state $i$
  - Update parameters $\Theta$
- ❑ Until (some stopping condition)

# Entropy

- Entropy measures the variability observed in given data.

$$E = -\sum_{c} p_c \log p_c$$

- Entropy is useful in multiple alignments & profiles.
- Entropy is max when uncertainty is max.

# G-Protein Couple Receptors

- Transmembrane proteins with 7 α-helices and 6 loops; many subfamilies
- Highly variable: 200-1200 aa in length, some have only 20% identity.
- [Baldi & Chauvin, '94] HMM for GPCRs
- HMM constructed with 430 match states (avg length of sequences) ; Training: with 142 sequences, 12 iterations

# GPCR - Analysis

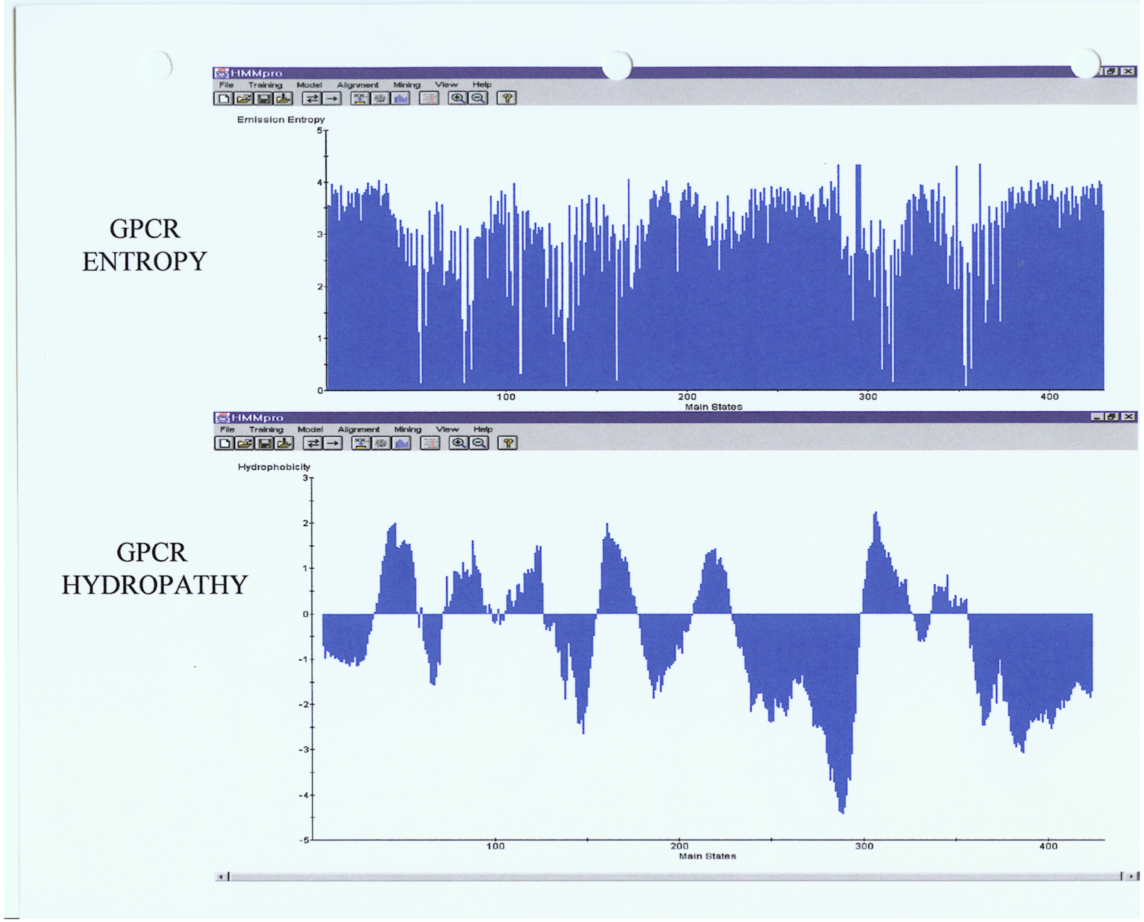
- Compute main state entropy values

$$H_i = -\sum_a e_{ia} \log e_{ia}$$

- For every sequence from test set (142) & random set (1600) & all SWISS-PROT proteins
  - Compute the negative log of probability of the most probable path π

$$Score(S) = -\log\left(P(\pi \mid S, M)\right)$$

# GPCR Analysis



# Entropy

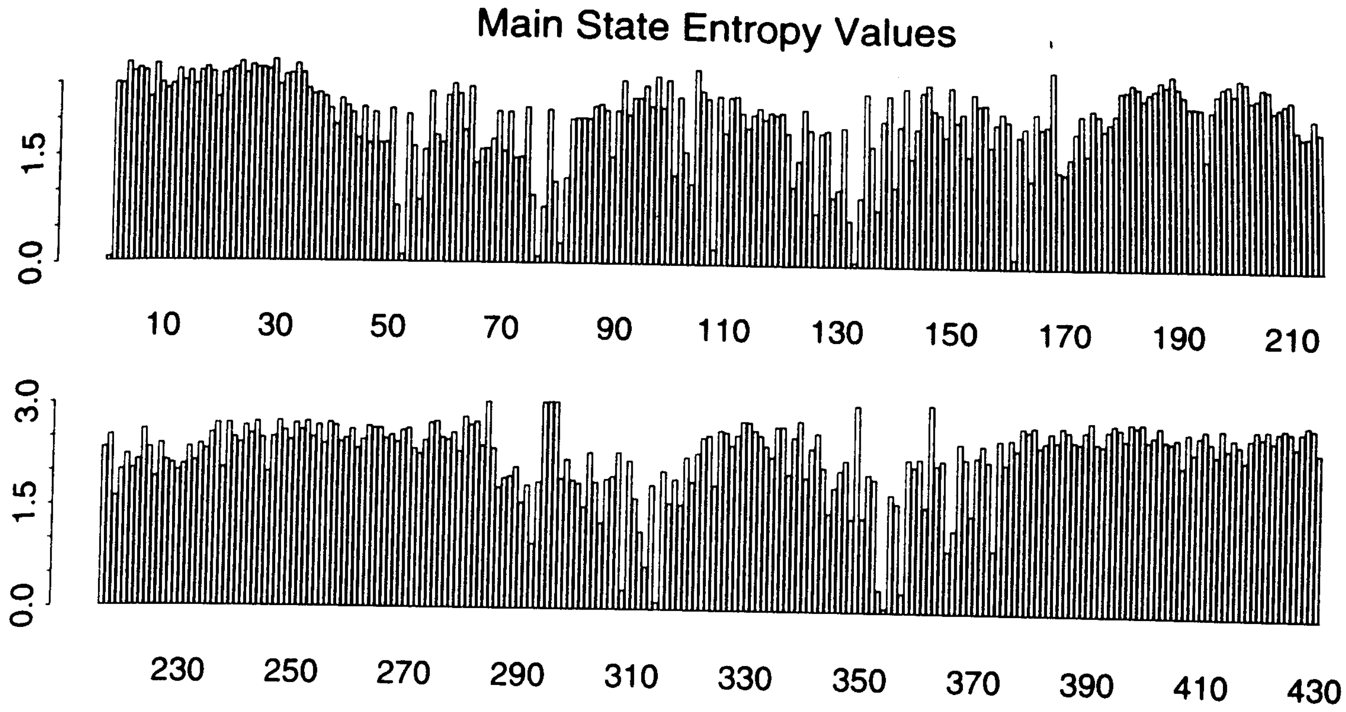


Figure 8.1: Entropy Profile of the Emission Probability Distributions Associated with the Main States of the HMM After 12 Cycles of Training.

# GPCR Analysis (Cont'd)

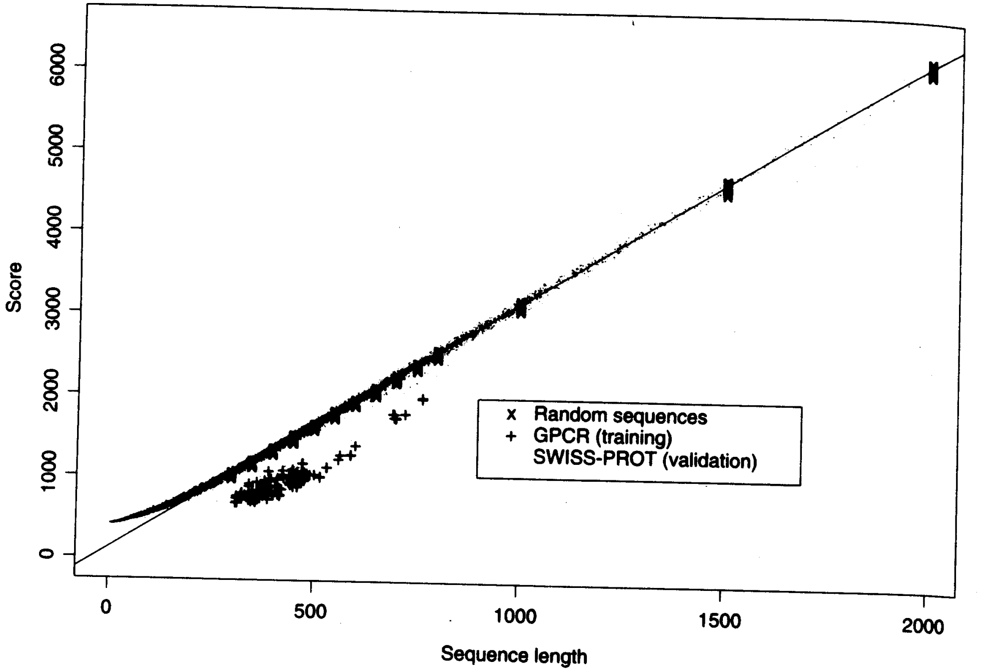


Figure 8.2: Scores (Negative Log-likelihoods of Optimal Viterbi Paths). Represented sequences consist of 142 GPCR training sequences, all sequences from the SWISS-PROT database of length less than or equal to 2000, and 220 randomly generated sequences with same average composition as the GPCRs of length 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800 (20 at each length). The regression line was obtained from the 220 random sequences. The horizontal distances in the histogram correspond to( -malized scores (6).

# Applications of HMM for GPCR

- Bacteriorhodopsin
  - Transmembrane protein with 7 domains
  - But it is not a GPCR
  - Compute score and discover that it is close to the regression line. Hence not a GPCR.
- Thyrotropin receptor precursors
  - All have long initial loop on INSERT STATE 20.
  - Also clustering possible based on distance to regression line.

# HMMs – Advantages

- Sound statistical foundations
- Efficient learning algorithms
- Consistent treatment for insert/delete penalties for alignments in the form of locally learnable probabilities
- Capable of handling inputs of variable length
- Can be built in a modular & hierarchical fashion; can be combined into libraries.
- Wide variety of applications: Multiple Alignment, Data mining & classification, Structural Analysis, Pattern discovery, Gene prediction.

# HMMs – Disadvantages

- ❑Large # of parameters.
- ❑Cannot express dependencies & correlations between hidden states.

# References

- Krogh, Brown, Mian, Sjolander, Haussler, J. Mol. Biol. 235:1501-1531, 1994
- Gribskov, Luthy, Eisenberg, Meth. Enzymol. 183:146-159, 1995
- Gribskov, McLachlan, Eisenberg, Proc Natl. Acad. Sci. 84:4355-4358, 1996.