CS681: Advanced Topics in Computational Biology

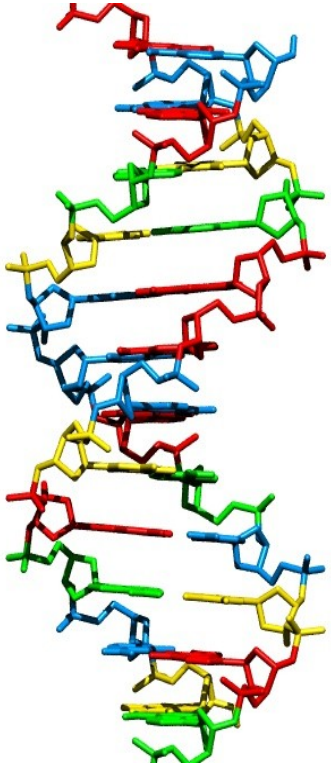
Week 3, Lecture 1

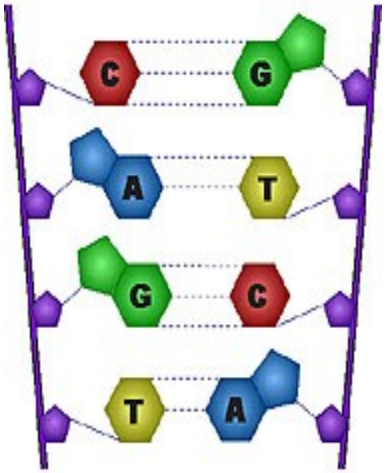
Can Alkan
EA224
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

DNA sequencing

How we obtain the sequence of nucleotides of a species





...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAAATATT...

DNA Sequencing

GENERAL CONCEPTS AND CAPILLARY (SANGER) SEQUENCING

DNA Sequencing

Goal:

Find the complete sequence of A, C, G, T's in DNA

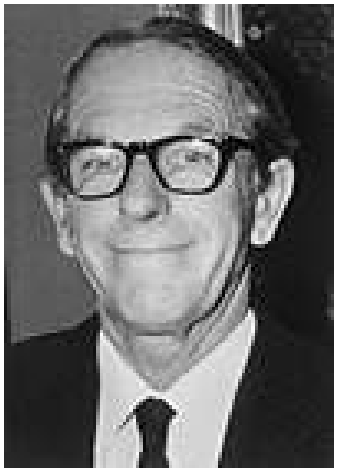
Challenge:

There is no machine that takes long DNA as an input, and gives the complete sequence as output

DNA Sequencing: History

Sanger method (1977): labeled ddNTPs terminate DNA copying at random points.

Gilbert method (1977): chemical method to cleave DNA at specific points (G, G+A, T+C, C).



Both methods generate labeled fragments of varying lengths that are further electrophoresed.

History of DNA Sequencing

Adapted from Eric Green, NIH; Adapted from Messing & Llaca, *PNAS* (1998)

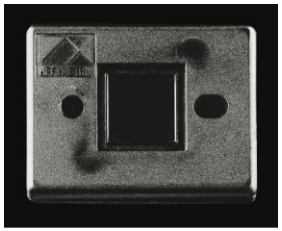
Efficiency (bp/person/year)		
1	1870	Miescher: Discovers DNA
	1940	Avery: Proposes DNA as 'Genetic Material'
		Watson & Crick: Double Helix Structure of DNA
	1953	Holley: Sequences Yeast $tRNA^{Ala}$
15	1965	Wu: Sequences λ Cohesive End DNA
150	1970	
1,500		Sanger: Dideoxy Chain Termination Gilbert: Chemical Degradation
	1977	Messing: M13 Cloning
15,000	1980	
25,000		Hood et al.: Partial Automation
50,000	1986	
200,000	1990	• Cycle Sequencing • Improved Sequencing Enzymes • Improved Fluorescent Detection Schemes
50,000,000	2002	
100,000,000,000	2009	• Next Generation Sequencing •Improved enzymes and chemistry •New image processing

Sequencing by Hybridization (SBH): History

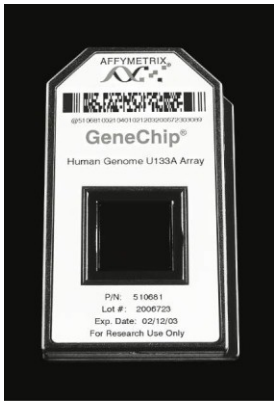
- **1988:** SBH suggested as an an alternative sequencing method.
- **1991:** Light directed polymer synthesis developed by Steve Fodor and colleagues.
- **1994:** Affymetrix develops first 64-kb DNA microarray

First microarray prototype **(1989)**

First commercial DNA microarray prototype w/16,000 features **(1994)**



500,000 features per chip **(2002)**



How SBH Works

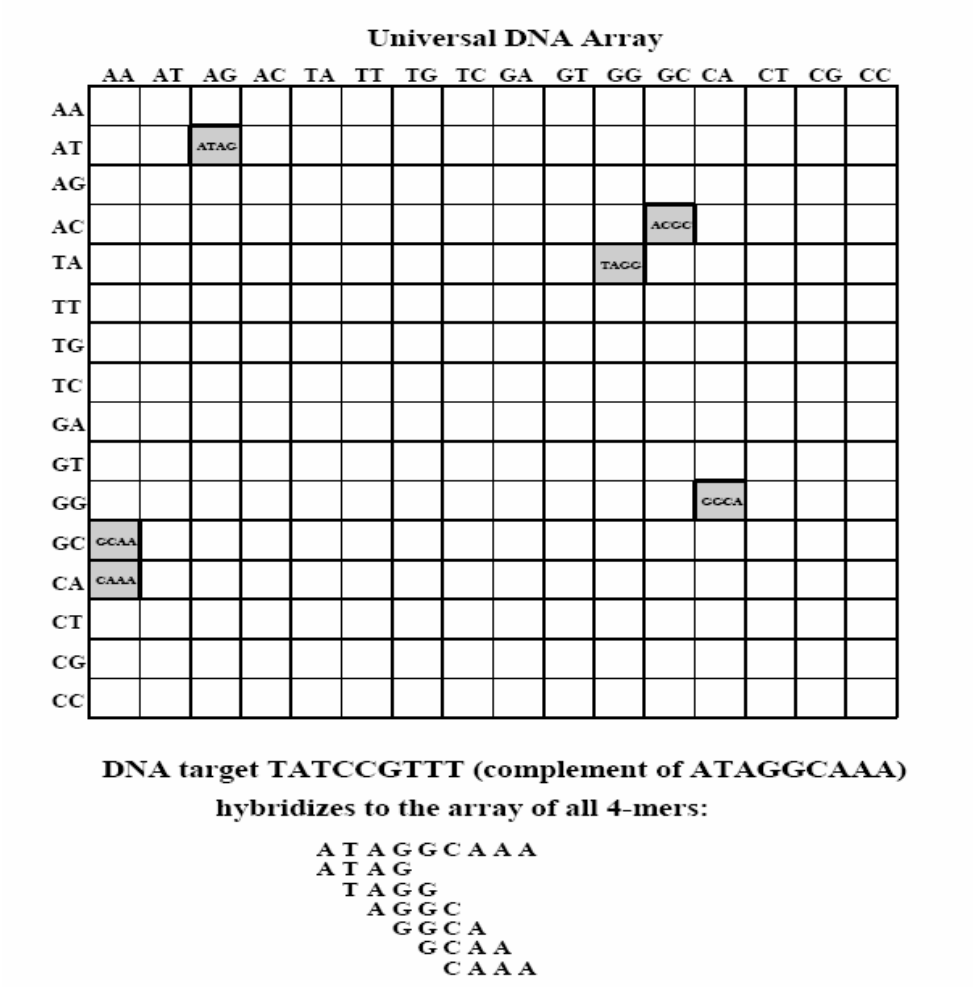
- Attach all possible DNA probes of length *l* to a flat surface, each probe at a distinct and known location. This set of probes is called the DNA array.
- Apply a solution containing fluorescently labeled DNA fragment to the array.
- The DNA fragment hybridizes with those probes that are complementary to substrings of length *l* of the fragment.

How SBH Works (cont'd)

- Using a spectroscopic detector, determine which probes hybridize to the DNA fragment to obtain the *l*–mer composition of the target DNA fragment.

- Apply the combinatorial algorithm (below) to reconstruct the sequence of the target DNA fragment from the *l* – mer composition.

Hybridization on DNA Array



l-mer composition

- ***Spectrum (s, l)*** - *unordered* multiset of all possible *(n – l + 1) l*-mers in a string *s* of length *n*
- The order of individual elements in *Spectrum (s, l)* does not matter
- For *s* = TATGGTGC all of the following are equivalent representations of *Spectrum (s, 3):*

  {TAT, ATG, TGG, GGT, GTG, TGC}

  {ATG, GGT, GTG, TAT, TGC, TGG}

  {TGG, TGC, TAT, GTG, GGT, ATG}

Different sequences – the same spectrum

- Different sequences may have the same spectrum:

  Spectrum(GTATCT,2)=

  Spectrum(GTCTAT,2)=

  {AT, CT, GT, TA, TC}

The SBH Problem

- Goal: Reconstruct a string from its *l*-mer composition
- Input: A set *S*, representing all *l*-mers from an (unknown) string *s*
- Output: String *s* such that *Spectrum (s,l)* = *S*

l-mer composition

- ***Spectrum (s, l)*** - *unordered* multiset of all possible *(n – l + 1) l*-mers in a string *s* of length *n*
- The order of individual elements in *Spectrum (s, l)* does not matter
- For *s* = TATGGTGC all of the following are equivalent representations of *Spectrum (s, 3):*

  {TAT, ATG, TGG, GGT, GTG, TGC}

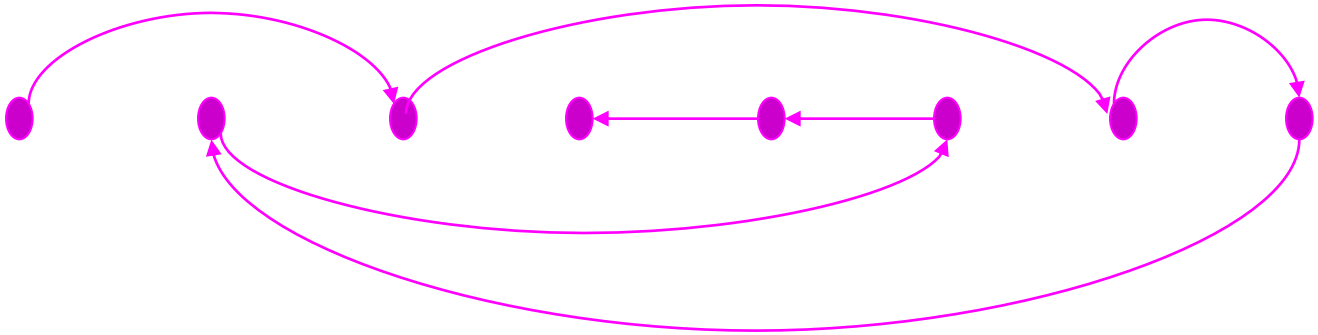
  {ATG, GGT, GTG, TAT, TGC, TGG}

  {TGG, TGC, TAT, GTG, GGT, ATG}

SBH: Hamiltonian Path Approach

S = { ATG AGG TGC TCC GTC GGT GCA CAG }

H ATG AGG TGC TCC GTC GGT GCA CAG



ATG CAG GTC C

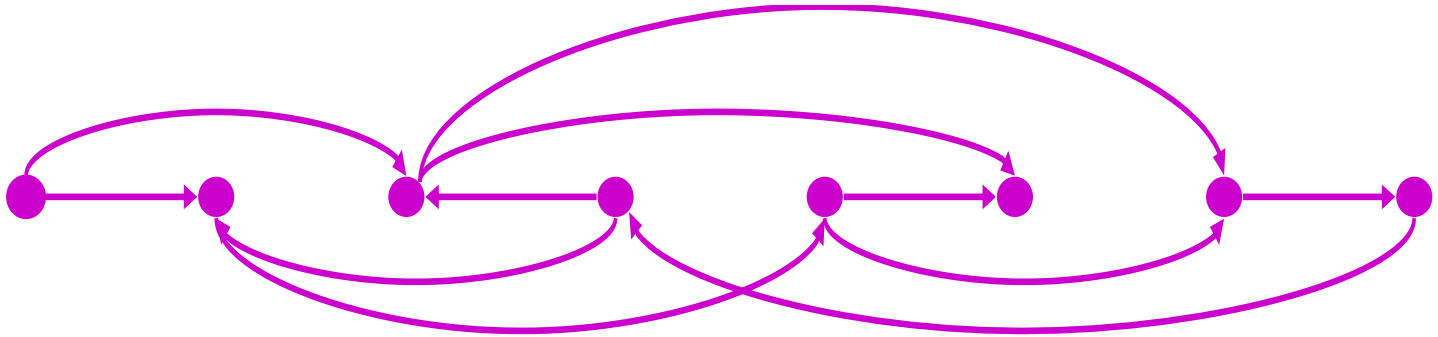
Path visited every VERTEX once

SBH: Hamiltonian Path Approach

A more complicated graph:

S = { ATG TGG TGC GTG GGC GCA GCG CGT }

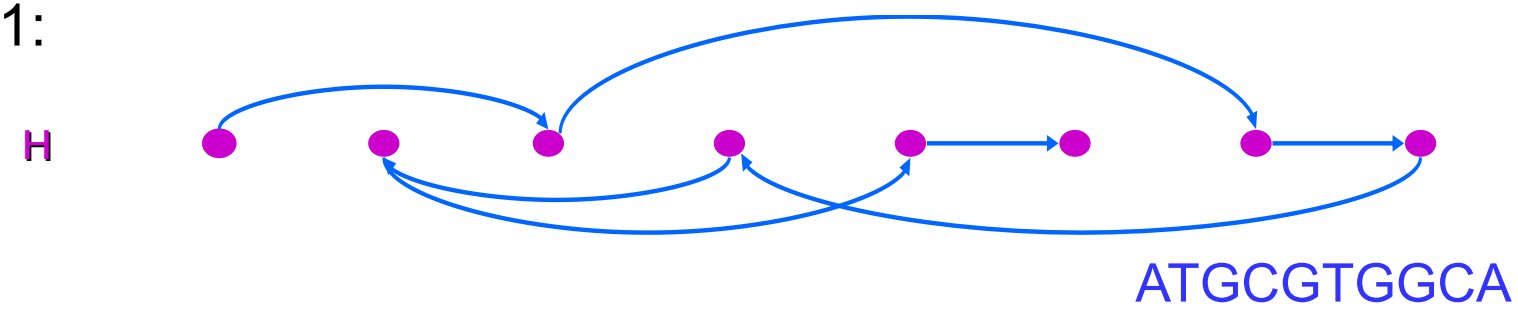
H



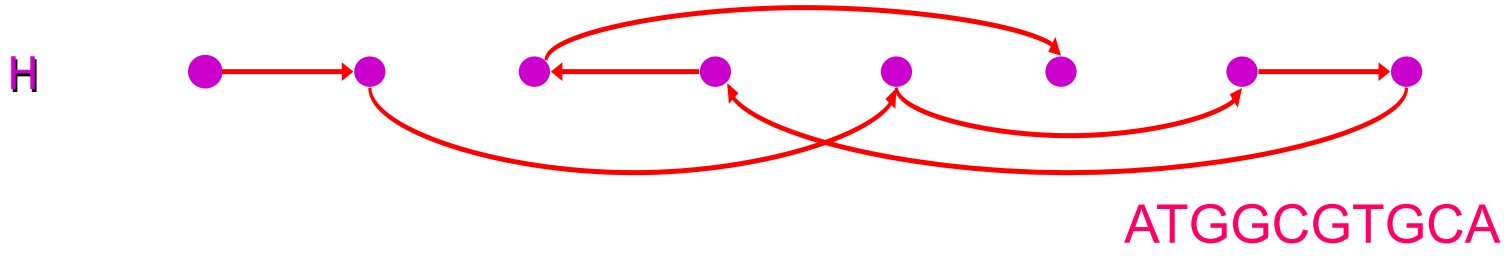
SBH: Hamiltonian Path Approach

S = { ATG TGG TGC GTG GGC GCA GCG CGT }

Path 1:



Path 2:

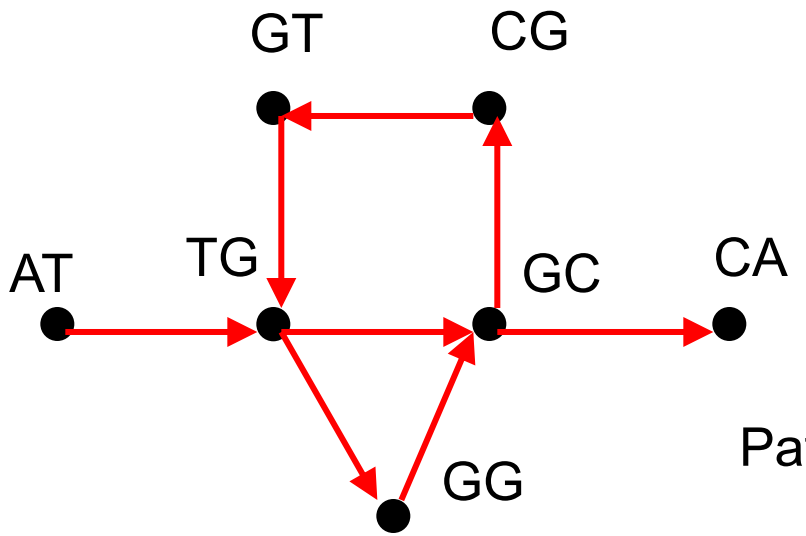


SBH: Eulerian Path Approach

S = { ATG, TGC, GTG, GGC, GCA, GCG, CGT }

Vertices correspond to (*l* – 1) – mers : { AT, TG, GC, GG, GT, CA, CG }

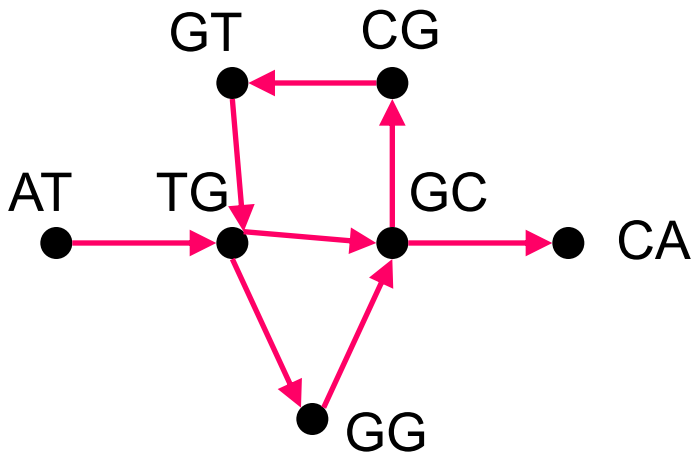
Edges correspond to *l* – mers from *S*



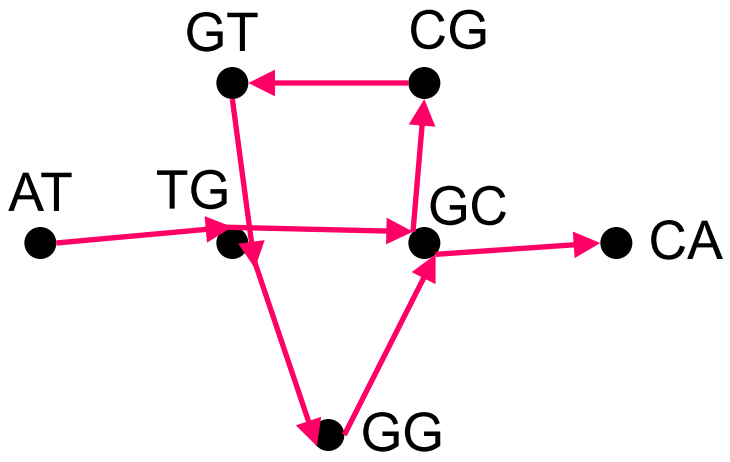
Path visited every EDGE once

SBH: Eulerian Path Approach

S = { AT, TG, GC, GG, GT, CA, CG } corresponds to two different paths:



ATGGCGTGCA

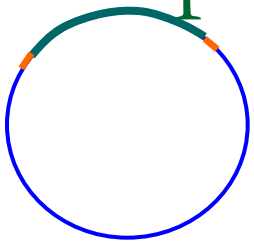


ATGCGTGGCA

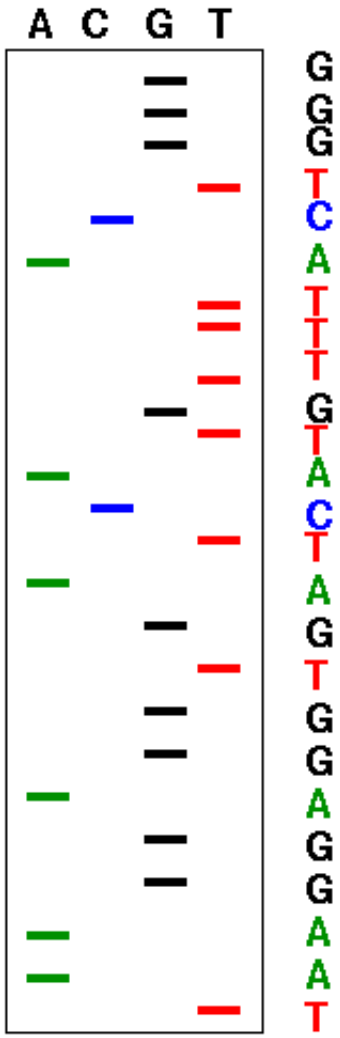
Some Difficulties with SBH

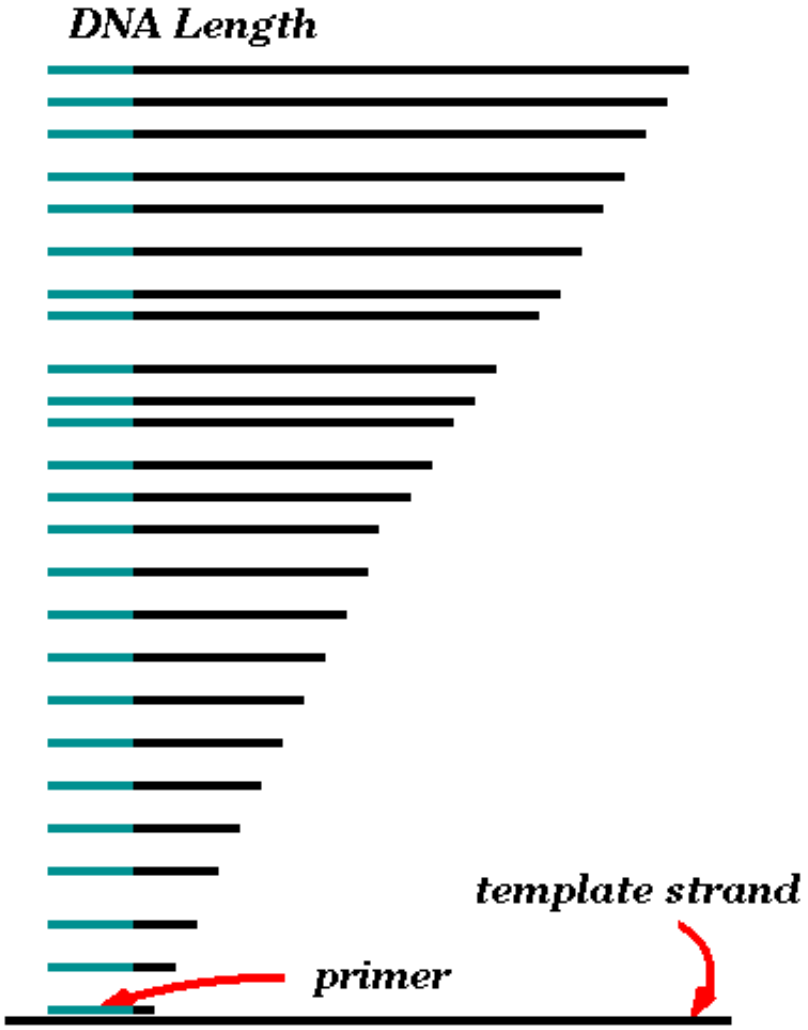
- **Fidelity of Hybridization:** difficult to detect differences between probes hybridized with perfect matches and 1 or 2 mismatches
- **Array Size:** Effect of low fidelity can be decreased with longer *l*-mers, but array size increases exponentially in *l.* Array size is limited with current technology.
- **Practicality:** SBH is still impractical.
- **Practicality again**: Although SBH is still impractical, it spearheaded expression analysis and SNP analysis techniques

DNA sequencing – gel electrophoresis



1. Start at primer (restriction site)
2. Grow DNA chain
3. Include dideoxynucleotide (modified a, c, g, t)
4. Stops reaction at all possible points
5. Separate products with length, using gel electrophoresis

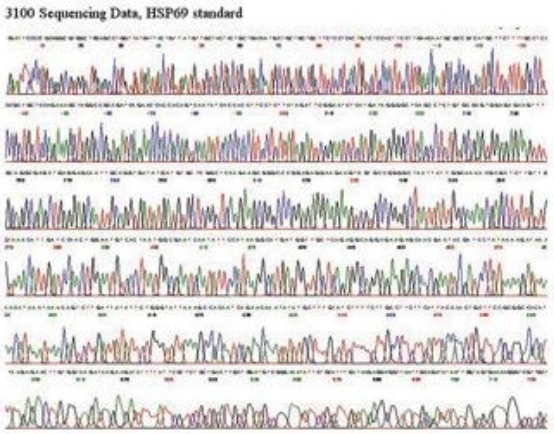




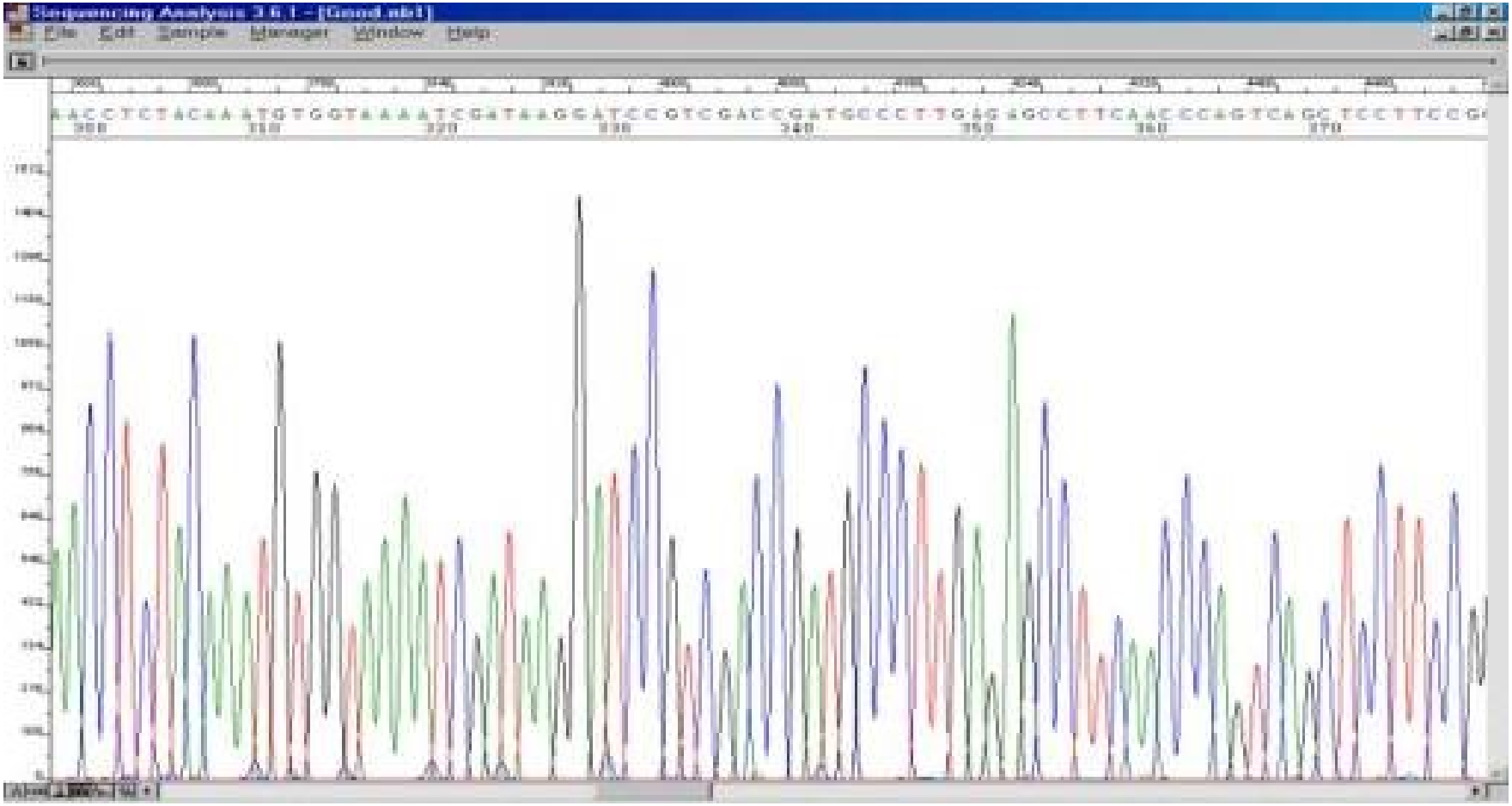
Capillary (Sanger) sequencing

Capillary sequencing (Sanger):

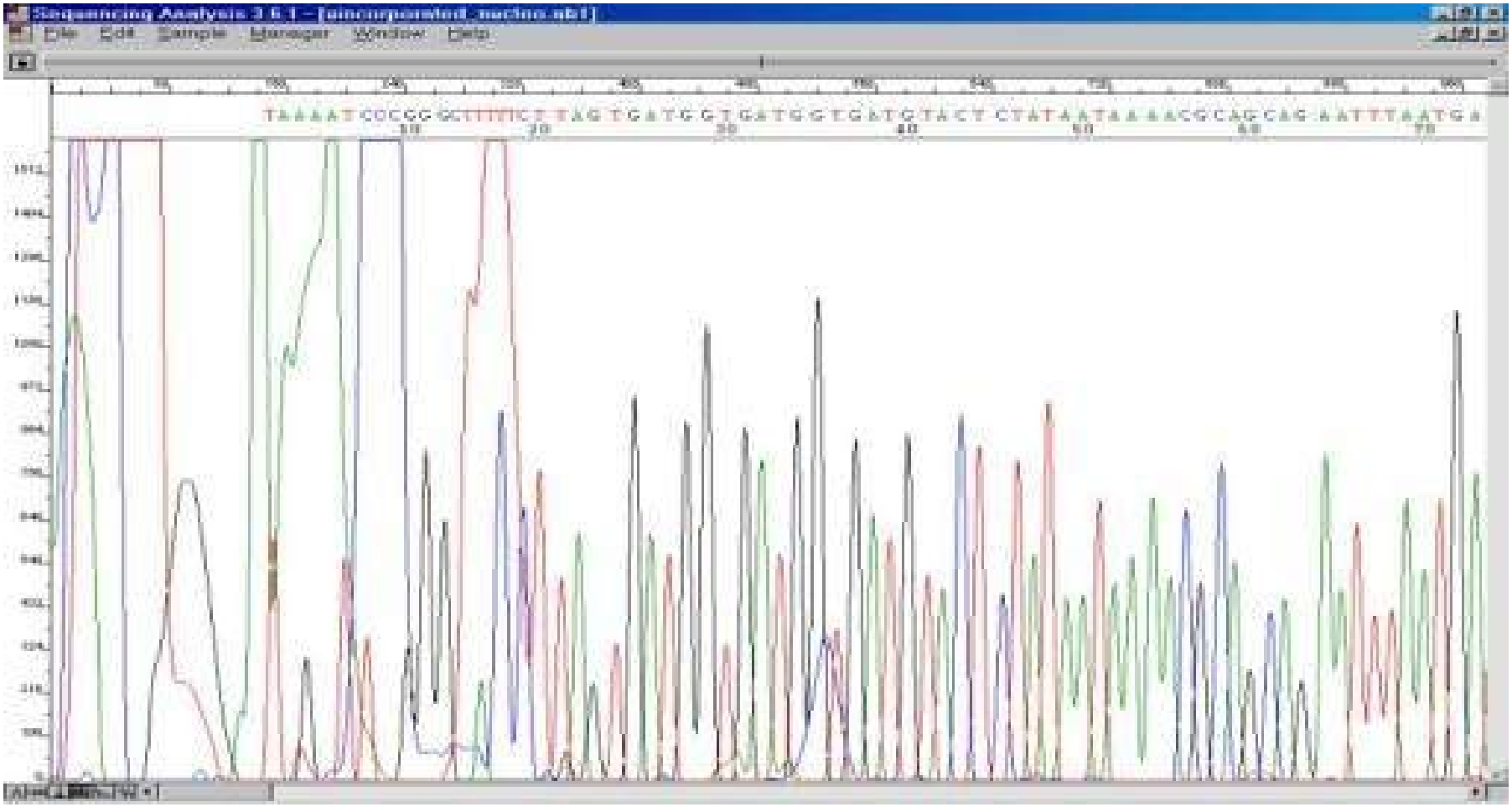
Can only sequence ~1000 letters at a time



Electrophoresis diagrams



Challenging to Read Answer



Reading an electropherogram

1. Filtering
2. Smoothening
3. Correction for length compressions
4. A method for calling the letters – **PHRED**

PHRED – **PH**il's **R**evised **ED**itor (by Phil Green)

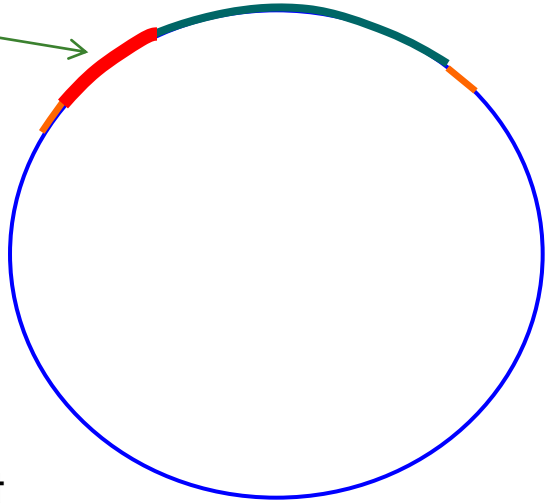
Based on dynamic programming

Several better methods exist, but labs are reluctant to change

Output of PHRED: a read

A **read**: ~1000 nucleotides

```
A  C  G  A  A  T  C  A  G  …A
16 18 21 23 25 15 28 30 32 …21
```



Quality scores: -10*$\log_{10}$Prob(Error)

"FASTQ format": ASCII character that corresponds to *q+33* (or 64)

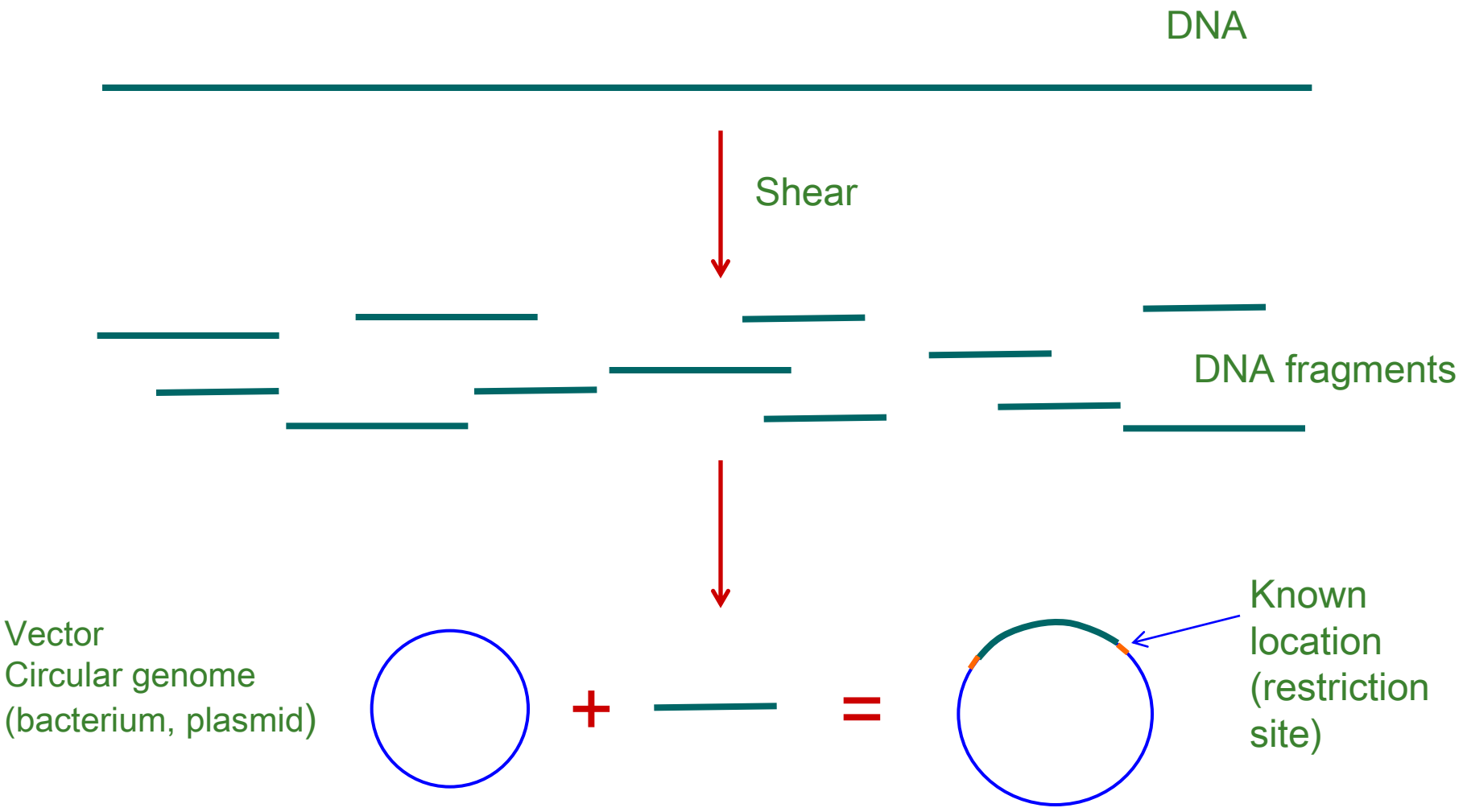
(I = 73; 73-33 = 40 = *q;* q40-> 0.01% error)

Reads can be obtained from leftmost, rightmost ends of the insert

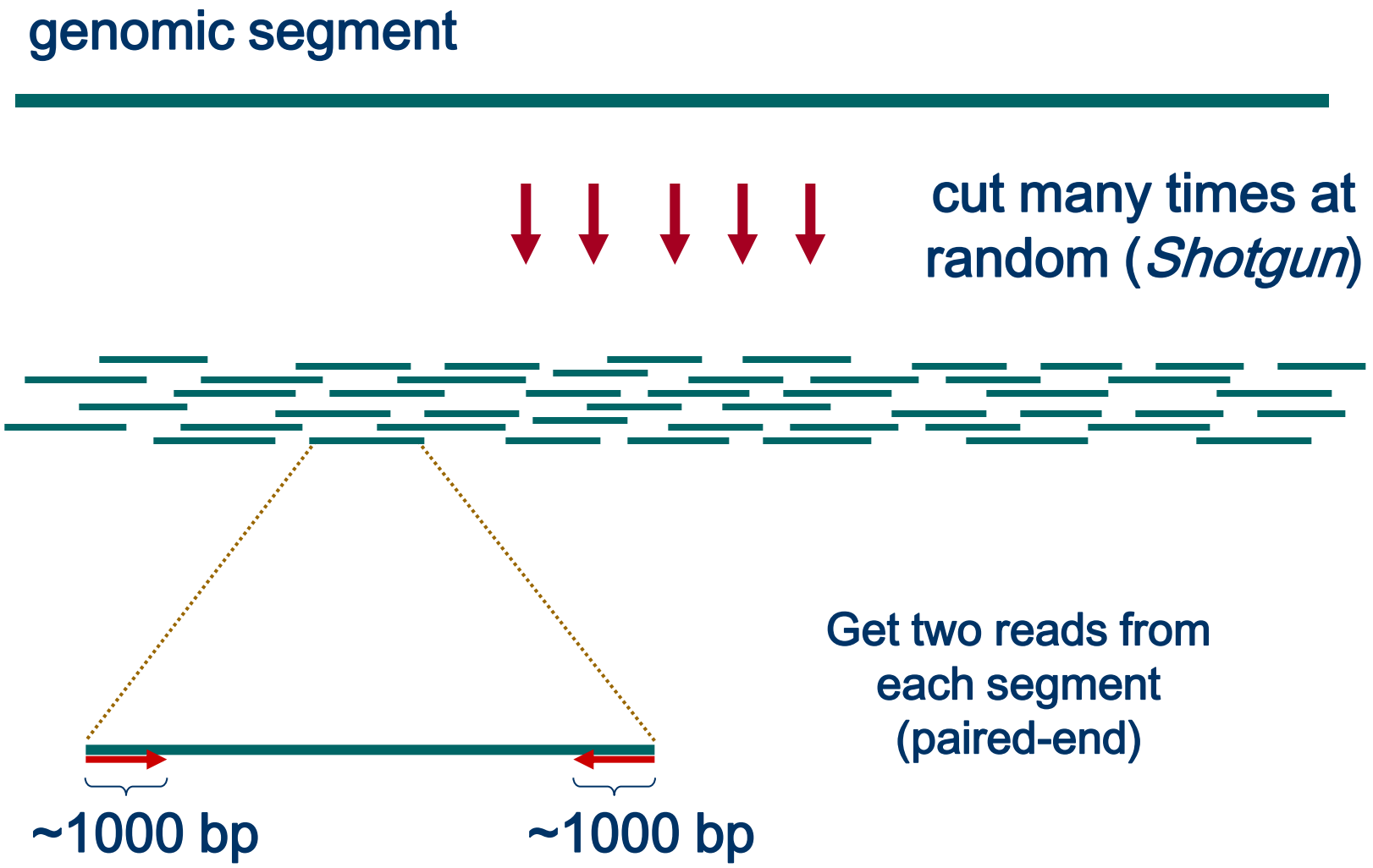
Double-barreled (paired-end, matepair) sequencing:

Both leftmost & rightmost ends are sequenced

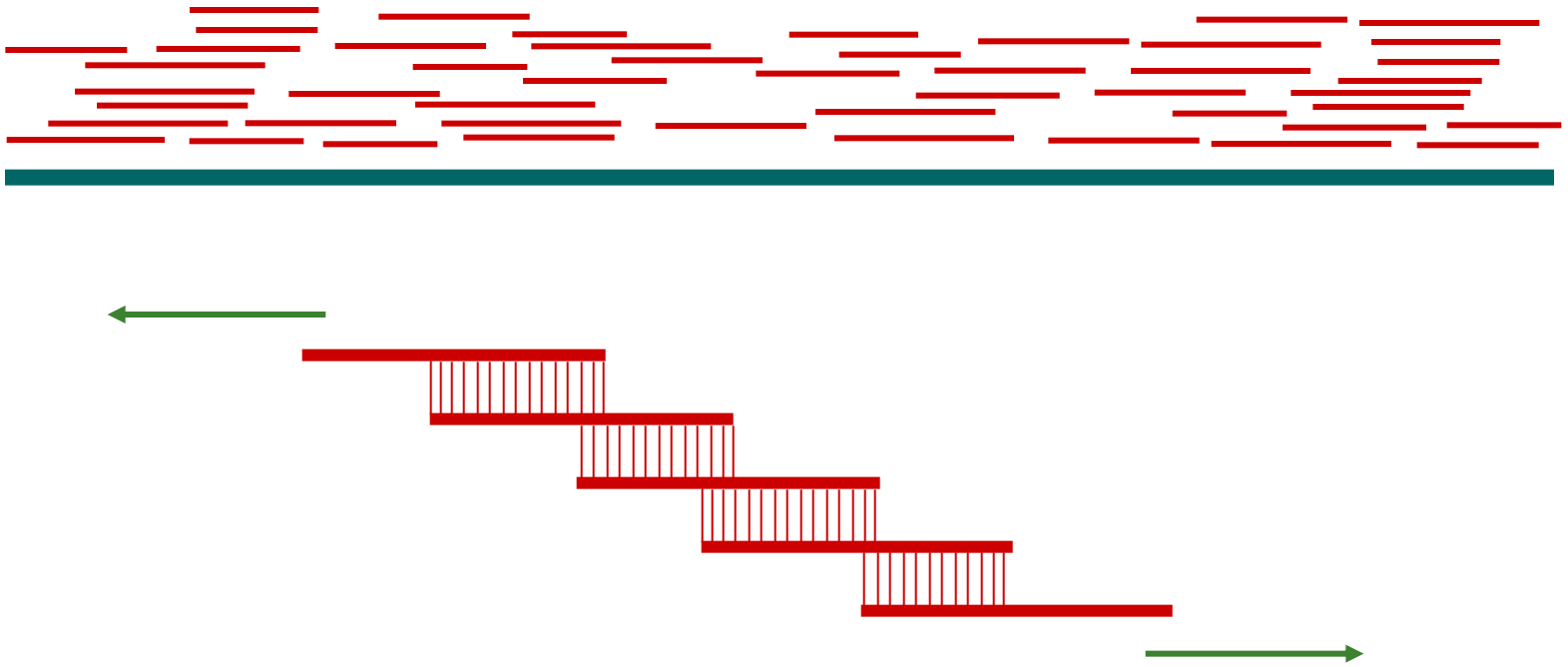
Traditional DNA Sequencing



Double-barreled sequencing



Reconstructing The Sequence



Need to cover region with >7-fold redundancy (7X) if you use Sanger technology

Overlap reads and extend to reconstruct the original genomic region

Definition of Coverage

Length of genomic segment: **L**
Number of reads: **n**
Length of each read: **l**

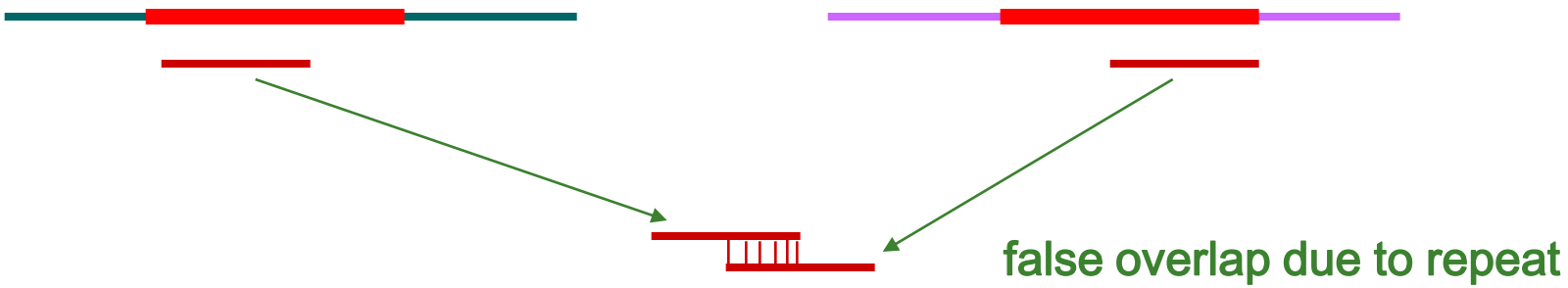
Definition: Coverage **C = n l / L**

How much coverage is enough?

Lander-Waterman model:
Assuming uniform distribution of reads, C=10 results in 1 gapped region /1,000,000 nucleotides

Challenges with Fragment Assembly

- **Sequencing errors**
  ~0.1% of bases are wrong
- **Repeats**



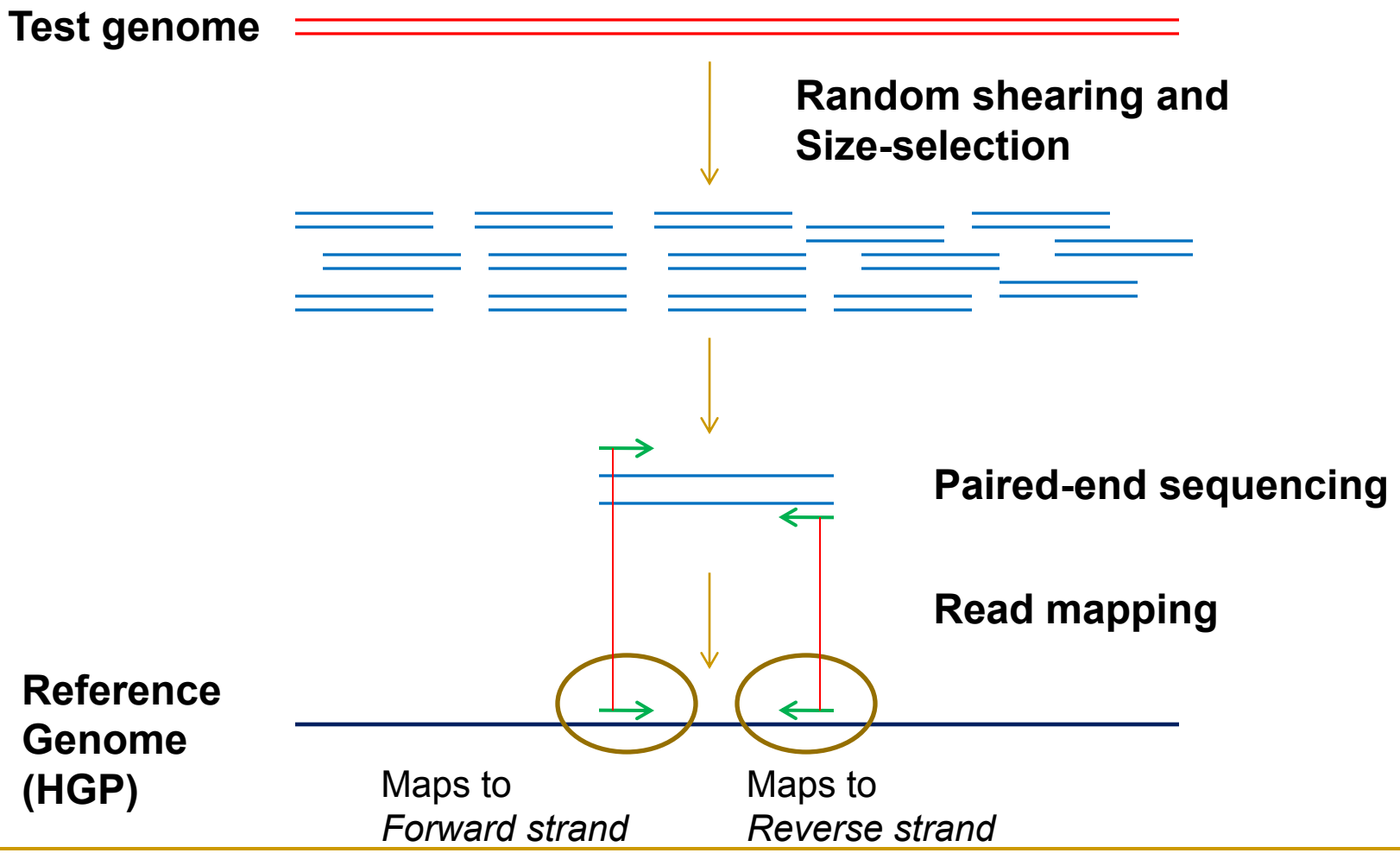
- **Computation: ~ O(N^2) where N = # reads**

Sanger sequencing

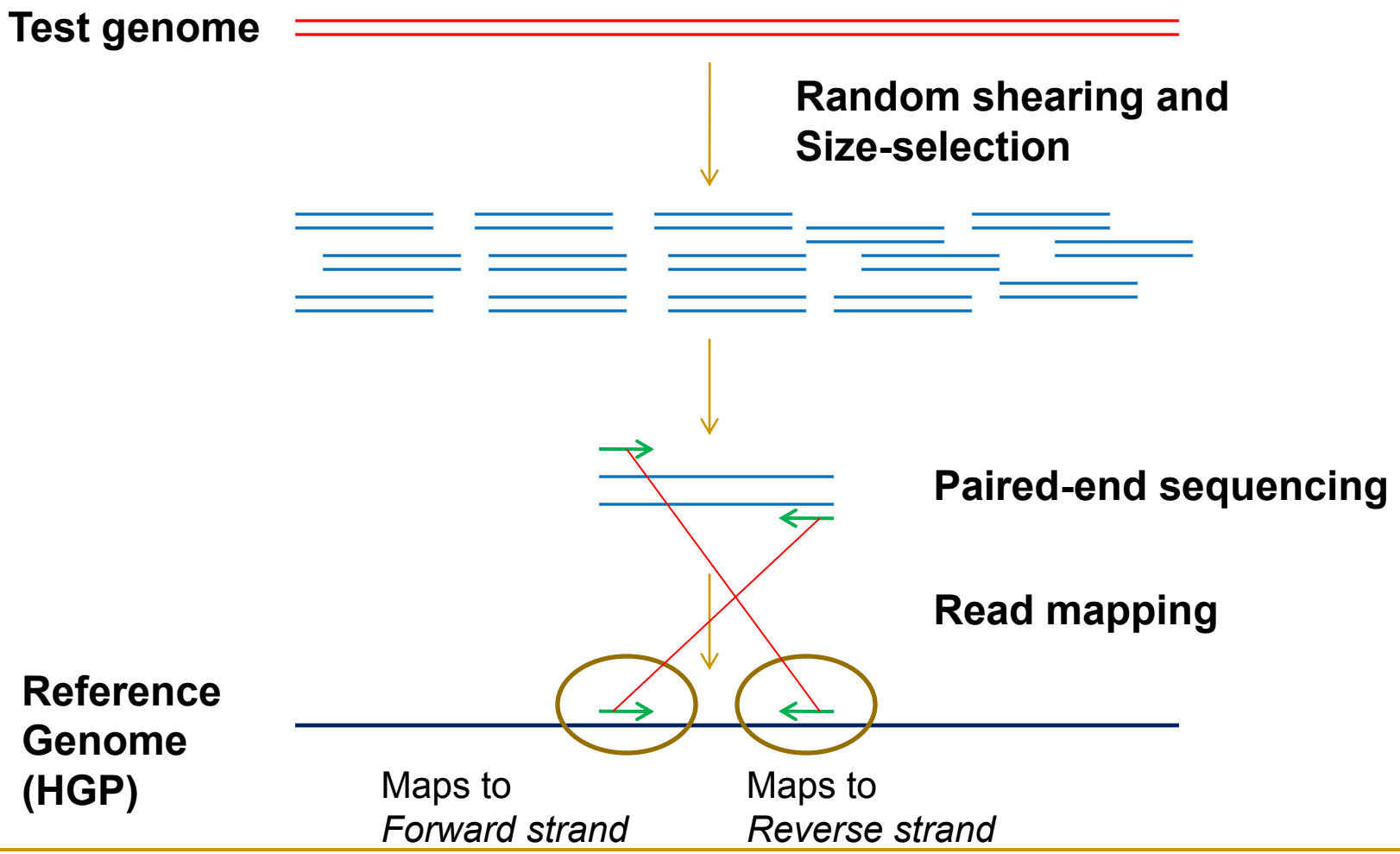
- Advantages
  - Longest read lengths possible today (>1000 bp)
  - Highest sequence accuracy (error < 0.1%)
  - Clone libraries can be used in further processing
- Disadvantages
  - The most expensive technology
    - $1500 per Mb
  - Building and storing clone libraries is hard & time consuming

NEXT GENERATION SEQUENCING

WGS revisited



WGS revisited



NGS Technologies

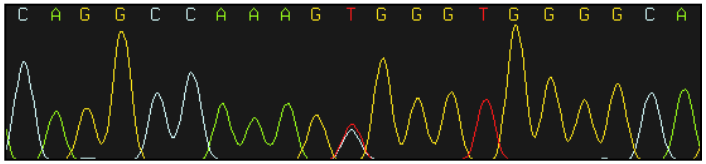
- 454 Life Sciences: the first, acquired by Roche
  - *Pyrosequencing*
- Illumina (Solexa): current market leader
  - *GAIIx, HiSeq2000, MiSeq, HiSeq2500*
  - *Sequencing by synthesis*
- Applied Biosystems:
  - *SOLiD: "color-space reads"*

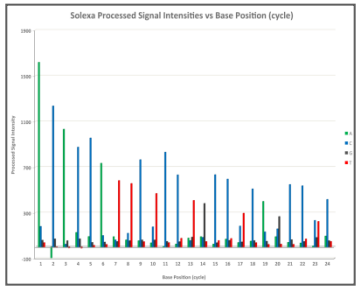
Features of NGS data

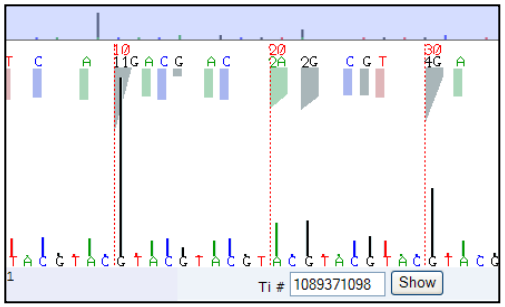
- Short sequence reads
  - –~500 bp: 454 (Roche)
  - – 35 – 150 bp Solexa(Illumina), SOLiD(AB)
- Huge amount of sequence per run
  - –Gigabases per run (600 Gbp for Illumina/HiSeq2000)
- Huge number of reads per run
  - •Up to billions
- Bias against high and low GC content (most platforms)
  - •GC% = (G + C) / (G + C + A + T)
- Higher error (compared with Sanger)
  - –Different error profiles

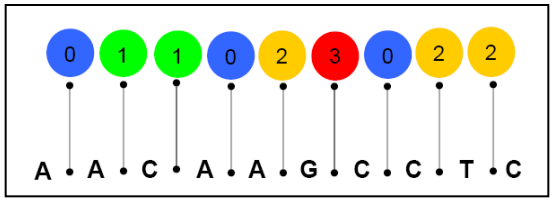
Next Gen: Raw Data

- **Machine Readouts are different**









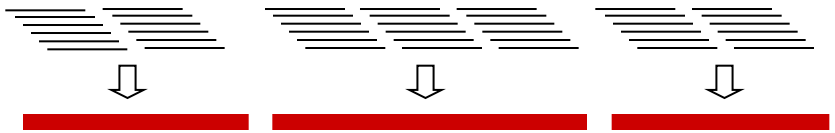
- **Read length, accuracy, and error profiles are variable.**

- **All parameters change rapidly as machine hardware, chemistry, optics, and noise filtering improves**

Current and future application areas

Genome re-sequencing: somatic mutation detection, organismal SNP discovery, mutational profiling, structural variation discovery

De novo genome sequencing

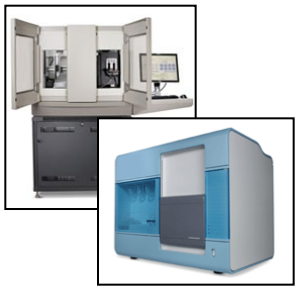


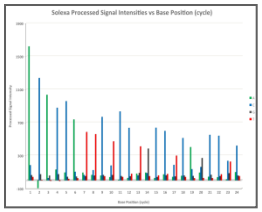
Sequencing is becoming an alternative to microarrays for:

- **DNA-protein interaction analysis (CHiP-Seq)**
- **novel transcript discovery**
- **quantification of gene expression**
- **epigenetic analysis (methylation profiling)**

Fundamental informatics challenges

1. Interpreting machine readouts – base calling, base error estimation

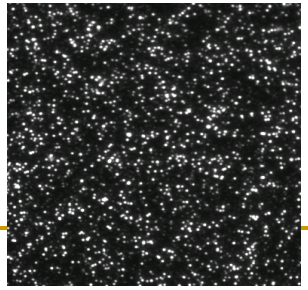




```
>B_TITR_1_1_668_35 TIME: Tue Feb 20 02:26:06 2007
ATATCGGATGACACAATATGGGAGGTTGAC
>B_TITR_1_2_843_403 TIME: Tue Feb 20 02:26:06 2007
TGTAGCTTTTCATGACAATTTTATAGGTGT
>B_TITR_1_1_668_35
27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 27 6
>B_TITR_1_2_843_403
26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 26 24
>B_TITR_1_3_618_922
```

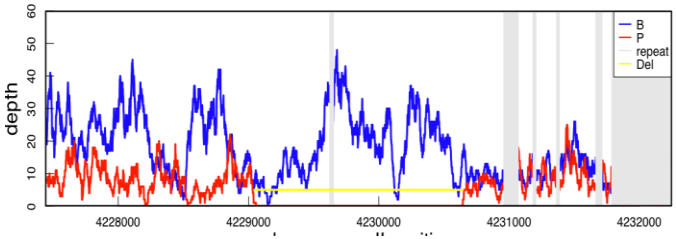
2. Data visualization

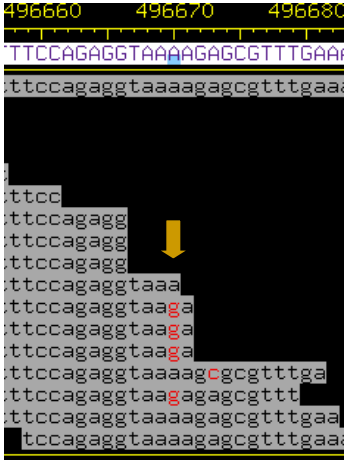
3. Data storage & management
Gzip compressed raw data for one human genome > 100 GB



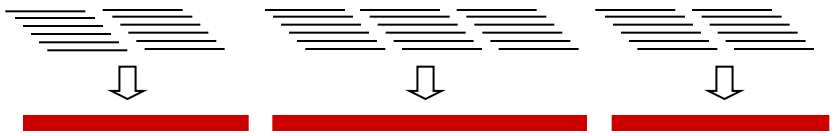
Informatics challenges (cont'd)

4. SNP, indel, and structural variation discovery





5. *De novo* Assembly



What can we use them for?

	SANGER	454	Solexa	AB SOLiD
De novo assembly	Fragmented	Fragmented	Heavily Fragmented	Heavily Fragmented
SNP Discovery	Yes	Yes	>95% of human	>95% of human
Larger events	Yes	Yes	Yes	Yes
Transcript profiling (rare)	No	Maybe	Yes	Yes

Week 3, Lectures 2-3

CURRENT PLATFORMS & DATA COMPRESSION