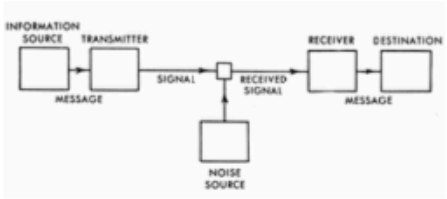
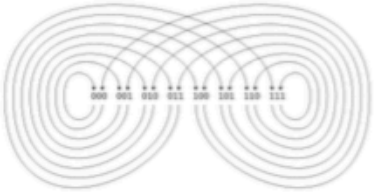
Part 1: Basics of Genomics

Olgica Milenkovic
University of Illinois, Urbana-Champaign

North American School of Information Theory, Texas, 2018

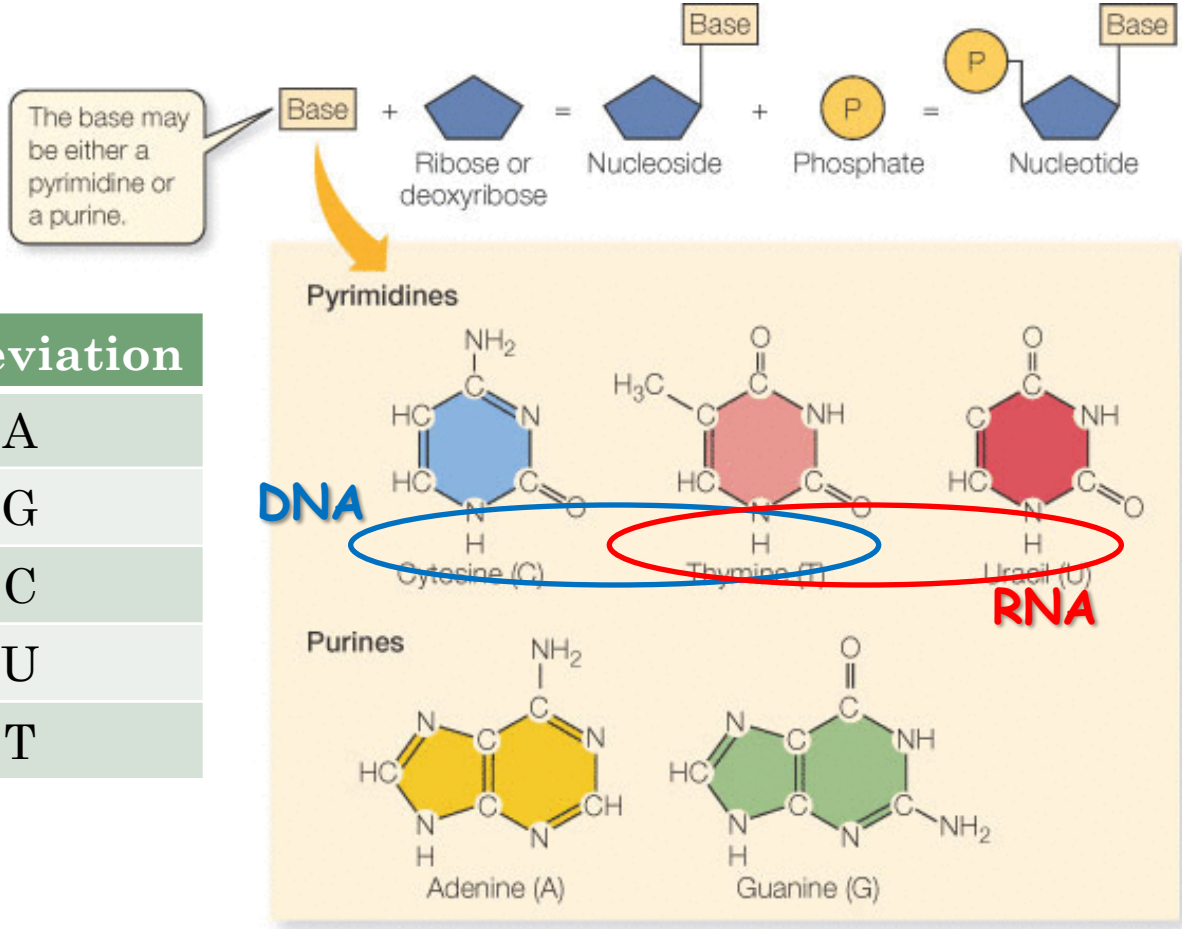
May 2018





Nucleic Acids

- **DNA** (Deoxyribonucleic Acid) and **RNA** (Ribonucleic Acid): information storage molecules made up of "nucleotides".

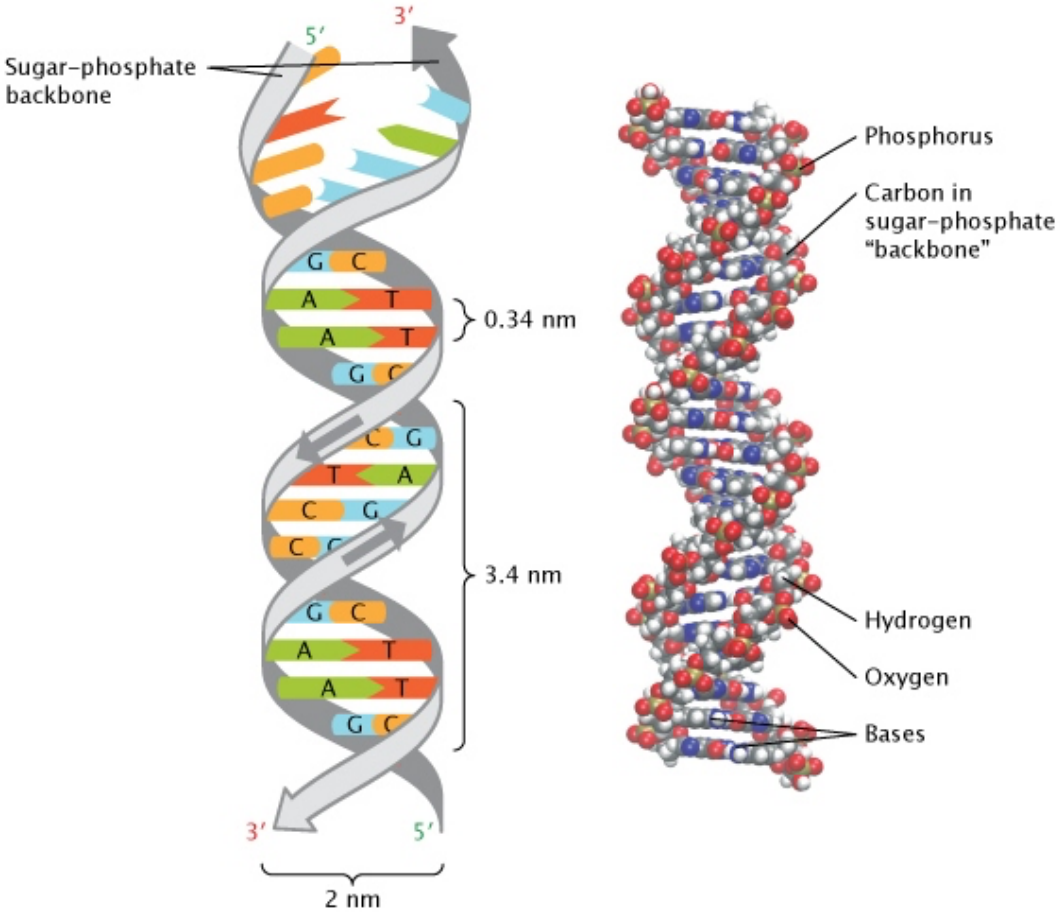


Base	Abbreviation
Adenine	A
Guanine	G
Cytosine	C
Uracil	U
Thymine	T

Source: Nature

Nucleic Acids

- **DNA** (Deoxyribonucleic Acid) and **RNA** (Ribonucleic Acid): information storage molecules made up of "nucleotides".



Source: Nature

Discovering the structure of DNA

- Chargaff's Rules (1949)
  - Amount of each dNTP varies between organisms, but [dA]=[dT] and [dC]=[dG] in ALL organisms

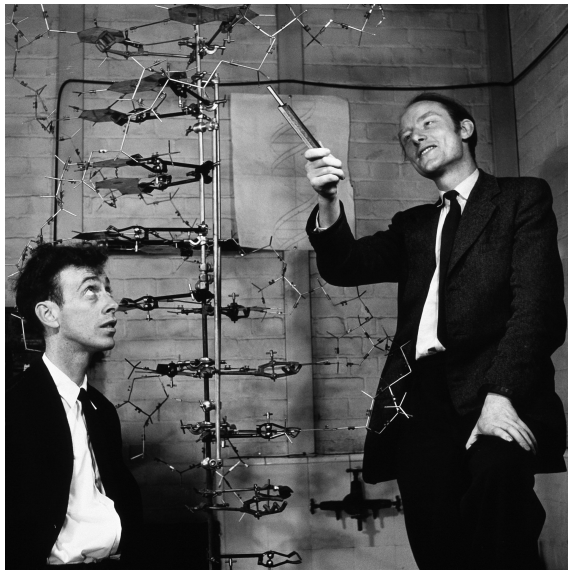
Table 3-2 Data Leading to the Formulation of Chargaff's Rules

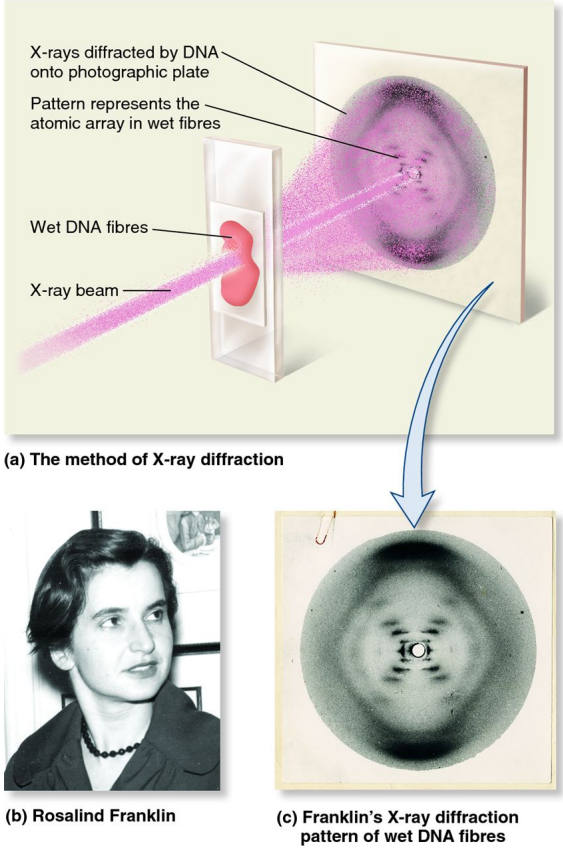
Source	Adenine to Guanine	Thymine to Cytosine	Adenine to Thymine	Guanine to Cytosine	Purines to Pyrimidines
Ox	1.29	1.43	1.04	1.00	1.1
Human	1.56	1.75	1.00	1.00	1.0
Hen	1.45	1.29	1.06	0.91	0.99
Salmon	1.43	1.43	1.02	1.02	1.02
Wheat	1.22	1.18	1.00	0.97	0.99
Yeast	1.67	1.92	1.03	1.20	1.0
Hemophilus influenzae	1.74	1.54	1.07	0.91	1.0
E-coli K2	1.05	0.95	1.09	0.99	1.0
Avian tubercle bacillus	0.4	0.4	1.09	1.08	1.1
Serratia marcescens	0.7	0.7	0.95	0.86	0.9
Bacillus schatz	0.7	0.6	1.12	0.89	1.0

SOURCE: After E. Chargaff et al., *J. Biol. Chem.* 177 (1949).

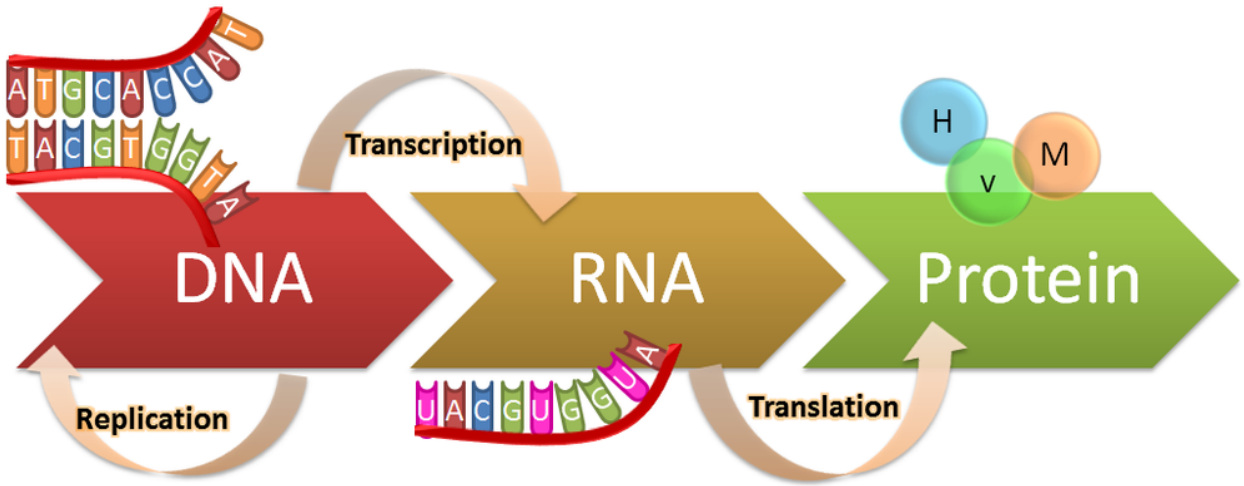
Discovering the structure of DNA

- Rosalind Franklin & Maurice Wilkins
  - X-ray diffraction suggested helix of uniform width with stacked bases, with sugar-phosphate on outside.
- James Watson & Francis Crick
  - Postulated double-helix model.





The Central Dogma

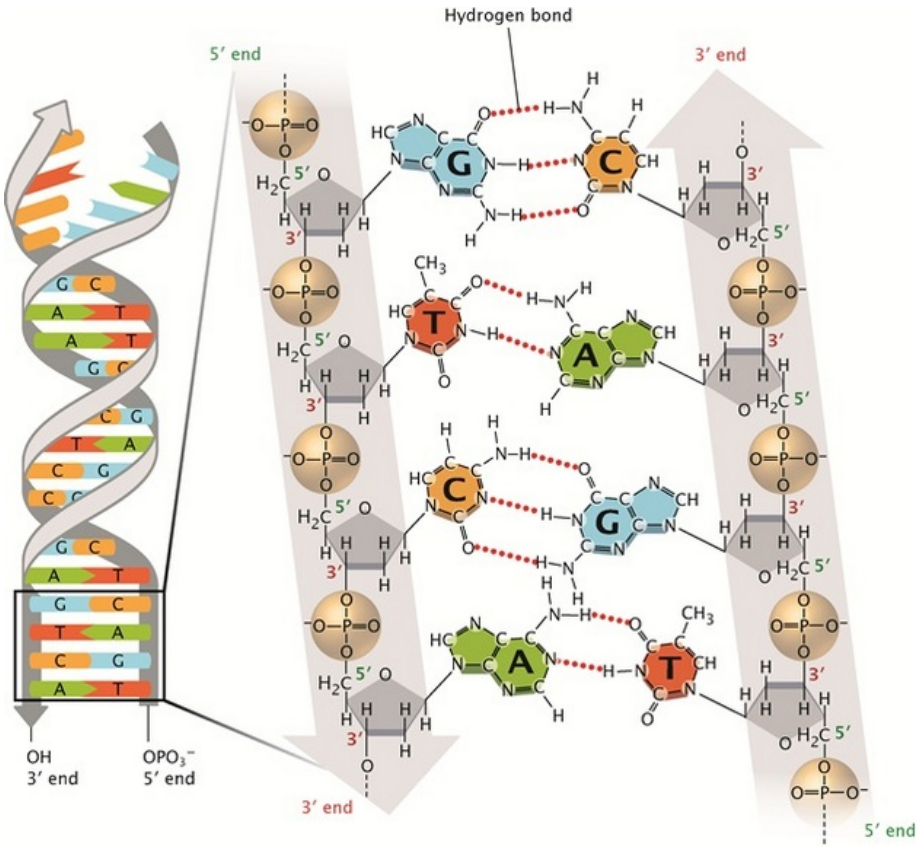


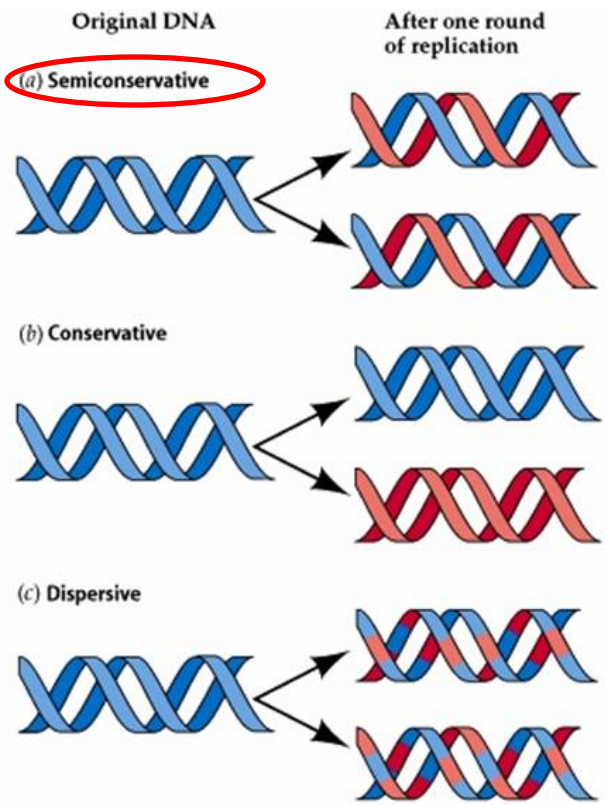
Source: Genius Media Group

- DNA encodes genetic information that "directs" the cell how to make proteins and RNAs.
- Information carried in the nucleotide sequence is copied into an RNA (TRANSCRIPTION).
- Information in RNA is used to build proteins (TRANSLATION).

DNA Replication

- Complementary base pairing (A-T, C-G).
- Semiconservative model of DNA replication.

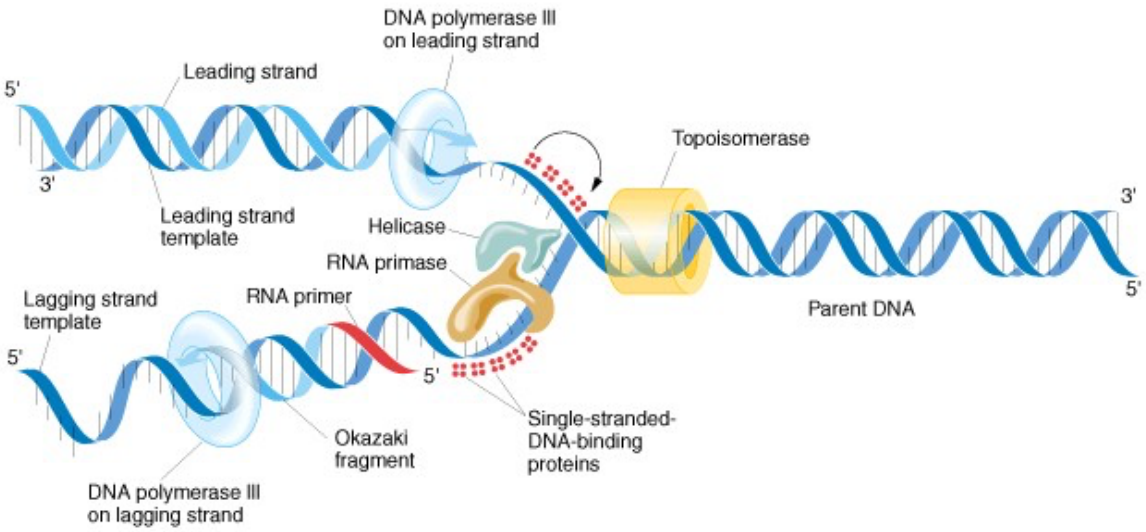




Source: Nature

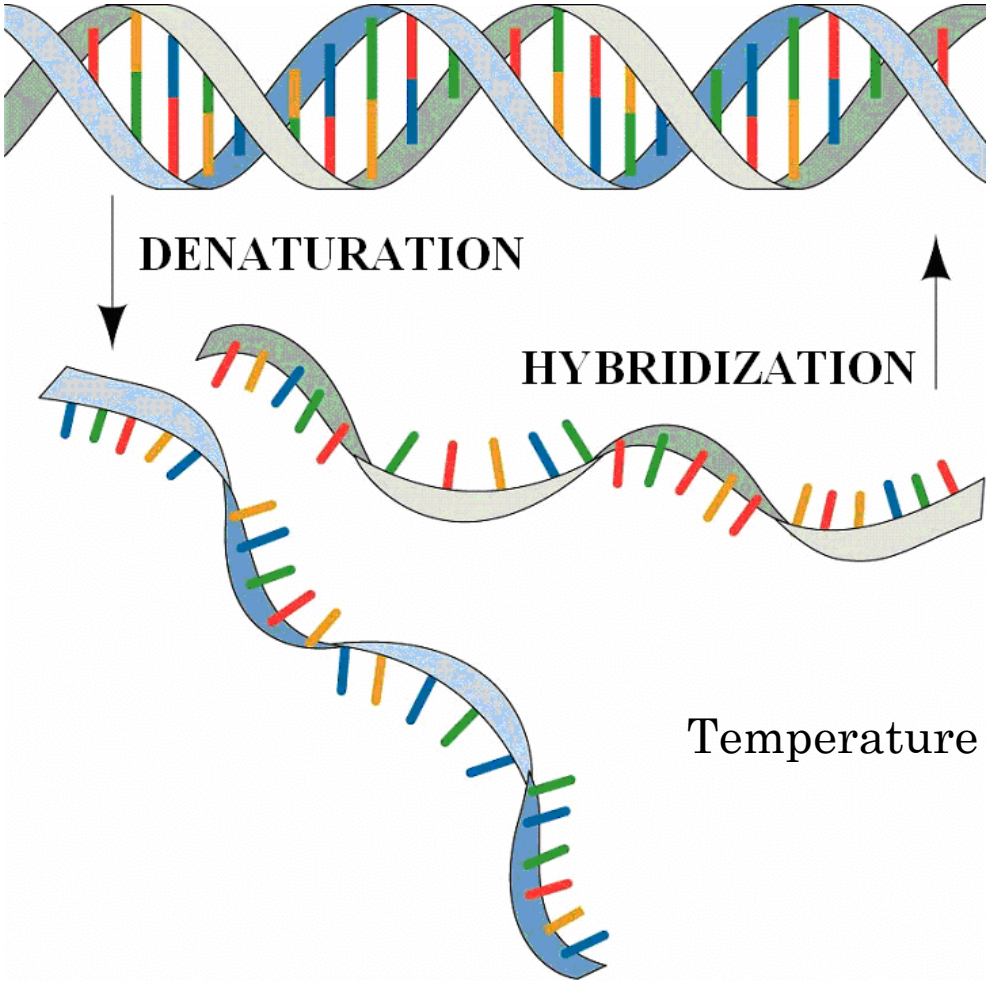
DNA Replication Requires Primers!

- "A **primer** is a short strand of RNA or **DNA** (generally about 18-22 bases) that serves as a starting point for **DNA** synthesis. It is required for **DNA replication** because the enzymes that catalyze this process, **DNA** polymerases, can only add new nucleotides to an existing strand of **DNA**."
- Source: Wikipedia ☺.



Source: Nature

DNA Hybridizes and Denatures

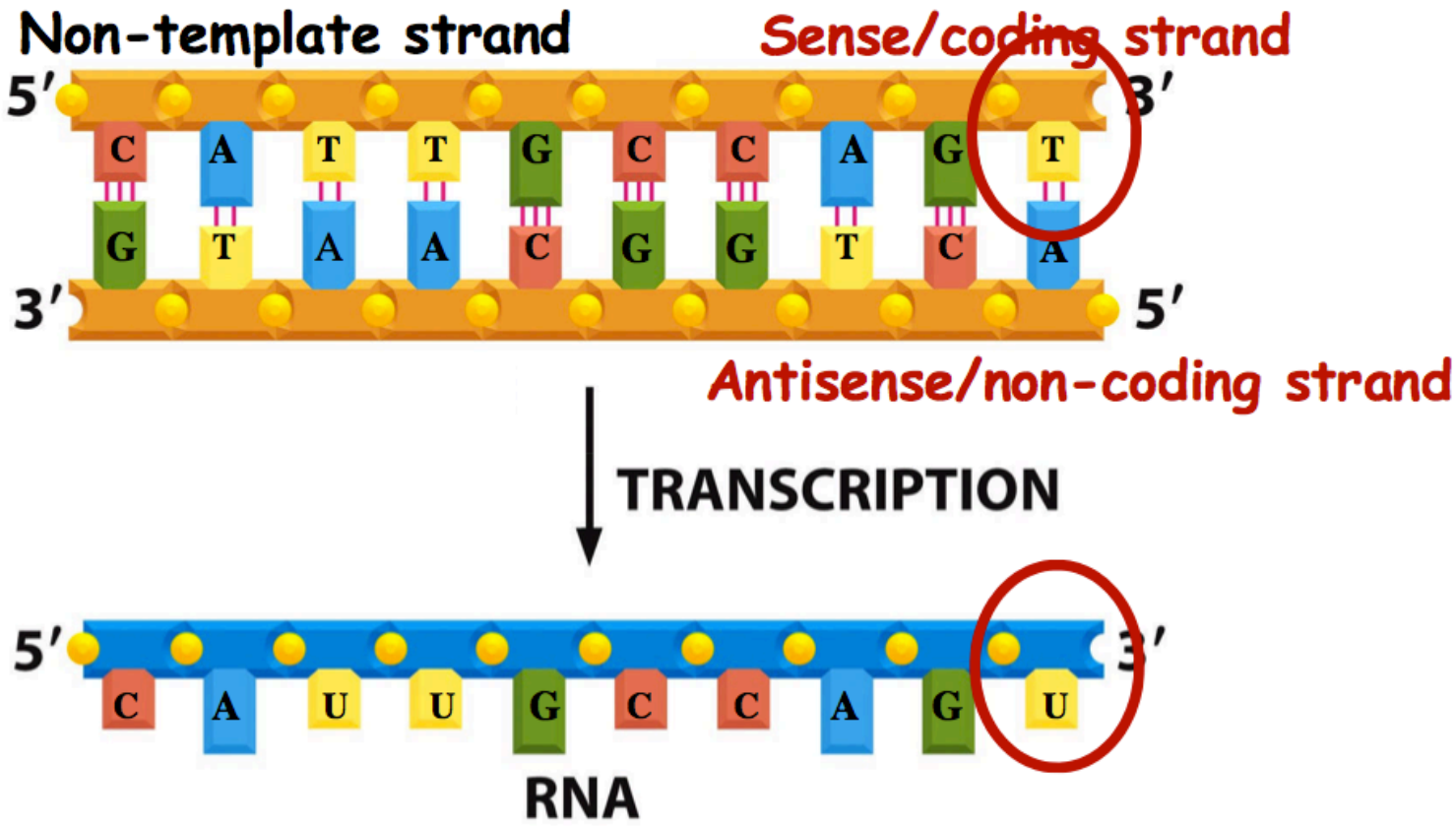


Temperature controlled

Source: Nature

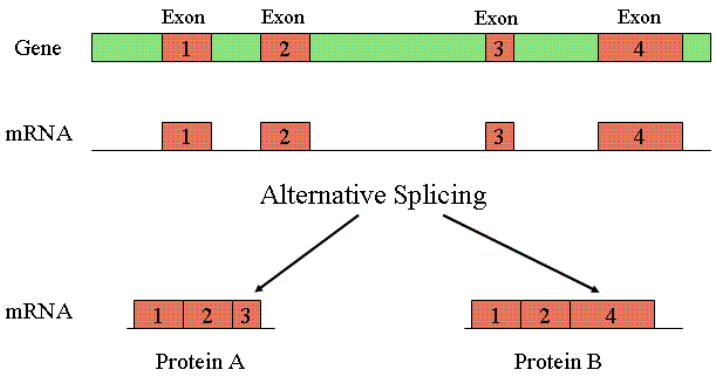
Transcription (DNA → RNA)

- RNA (usually) single-stranded.
- T → U.



Translation (RNA → Protein)

- Alternative Splicing in eukaryotes.



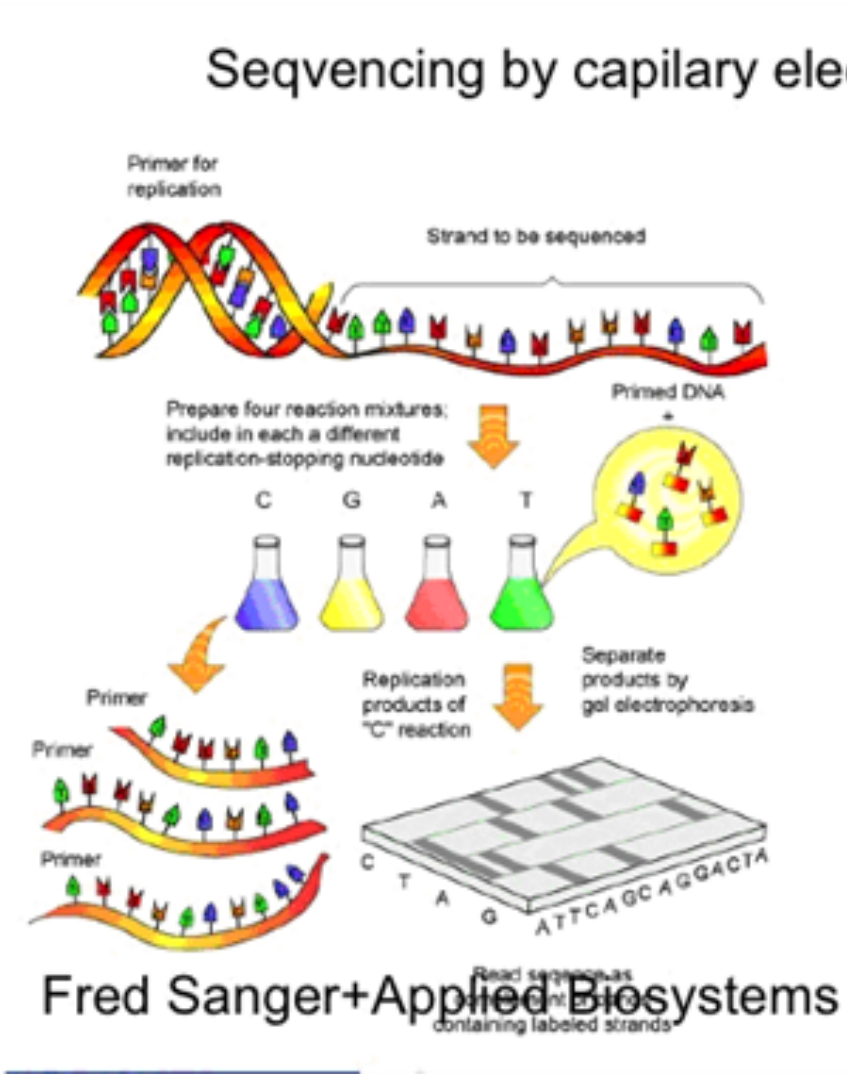
Source: NCBI

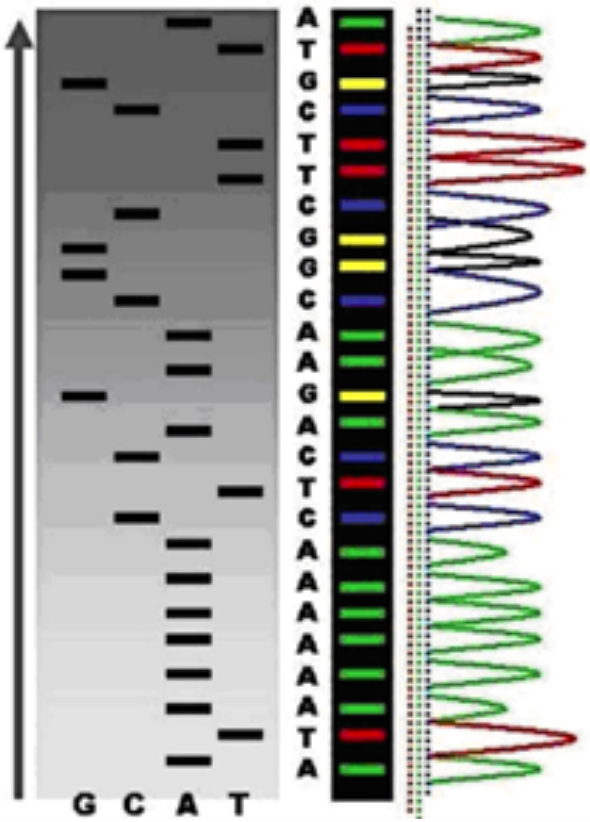
- Codon table of Amino Acids (protein)
  - Degenerate (redundant).

2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

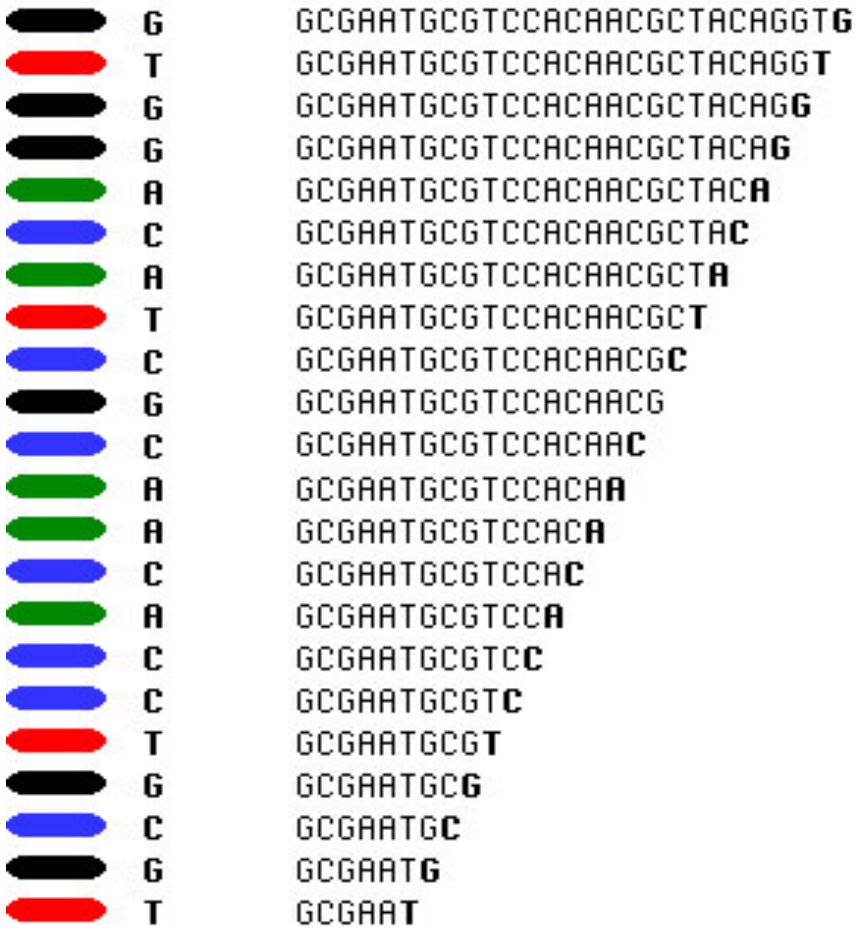
Reading DNA: Sanger Sequencing





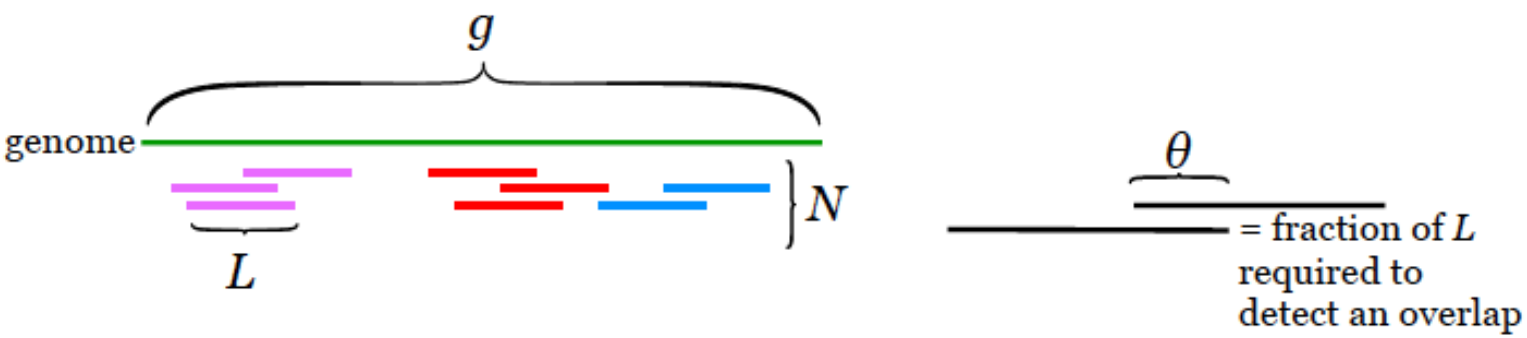
Reading DNA: Sanger Sequencing

Gel:



Reading DNA: Coverage and Read Length

How many reads to we need to be sure we cover the whole genome?



An ***island*** is a contiguous group of reads that are connected by overlaps of length $\geq \theta L$.
(Various colors above)

Want: Expression for expected # of islands given N, g, L, θ.

From C. Kingsford lecture notes

Reading DNA: Coverage and Read Length

$\lambda := N/g$ = probability a read starts at a given position (assuming random sampling)

Pr(k reads start in an interval of length x)

x trials, want k "successes," small probability λ of success

Expected # of successes = λx

Poisson approximation to binomial distribution:

$$\Pr(k \text{ reads in length } x) = e^{-\lambda x}\frac{(\lambda x)^k}{k!}$$

Expected # of islands = $N \times$ Pr(read is at rightmost end of island)

$$= N \times \Pr(0 \text{ reads start in } (1\text{-}\theta)L)$$

$$= Ne^{-\lambda(1-\theta)L}\frac{\lambda^0}{0!} \quad \text{(from above)}$$

$$= Ne^{-\lambda(1-\theta)L}$$

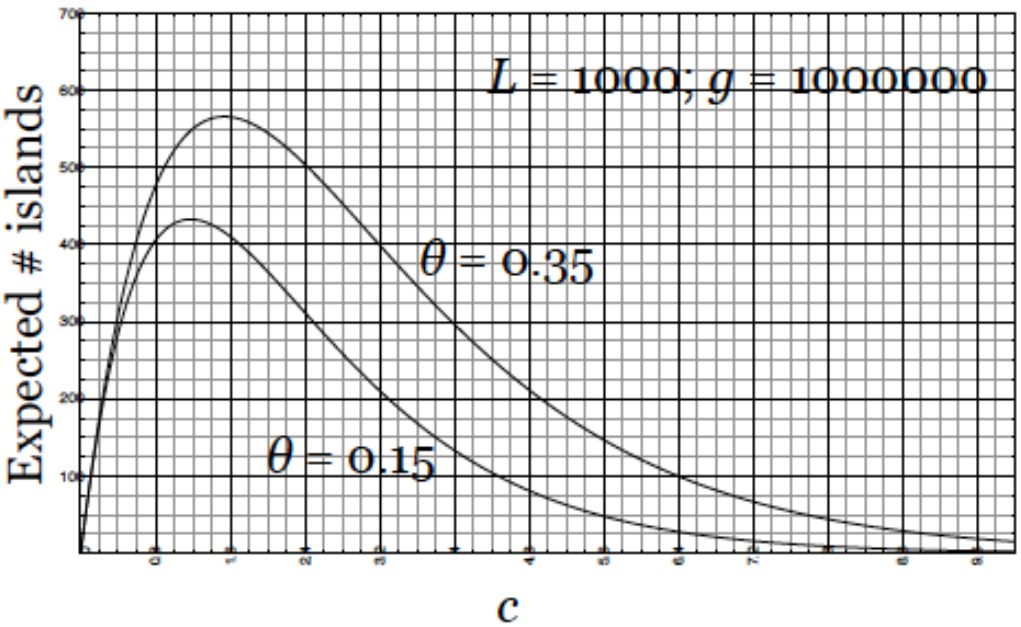
$$= Ne^{-(1-\theta)LN/g} \quad \leftarrow LN/g \text{ is called the } \textbf{coverage } c.$$

From C. Kingsford lecture notes

Reading DNA: Coverage and Read Length

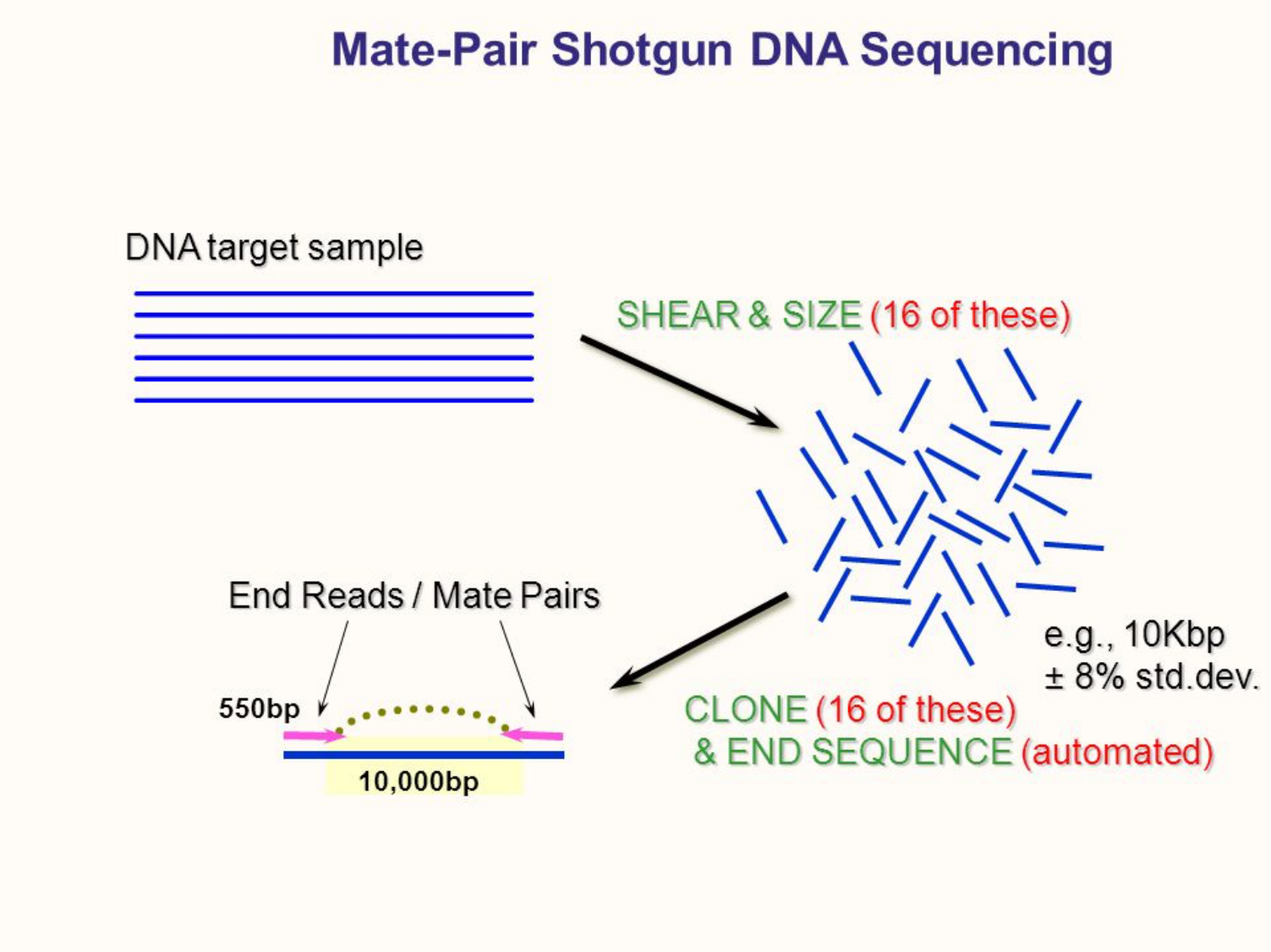
Rewrite to depend more directly on the things we can control: c and θ

$$\text{Expected \# of islands} = Ne^{-(1-\theta)LN/g}$$
$$= Ne^{-(1-\theta)c}$$
$$= \frac{L/g}{L/g} Ne^{-(1-\theta)c}$$
$$= \frac{g}{L} ce^{-(1-\theta)c}$$



From C. Kingsford lecture notes

Reading DNA: Lessons



DNA Sequencing: Illumina Platforms

50 – 600Gb
2 – 11 days
2 x 100bp max

10 – 180Gb
7 – 40 hours
2 x 150bp max

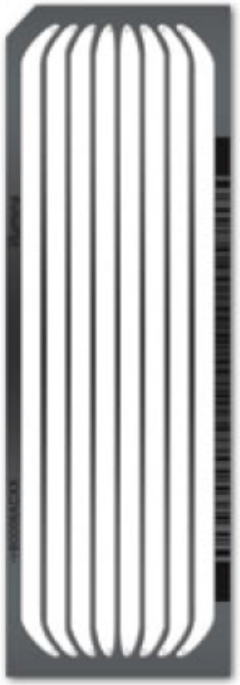
Larger projects, fewer runs

Smaller projects, quick results

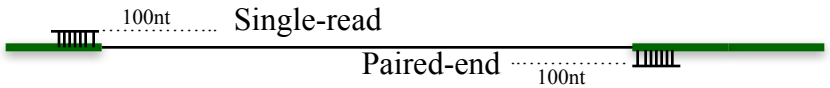
Courtesy of Alvaro Hernandez, UIUC

ILLUMINA HISEQ 2000

Flowcell
(8 lanes)



ILLUMINA LIBRARY FRAGMENT:



* Reads (sequences) = 100nt in length

***Number of reads per lane**:

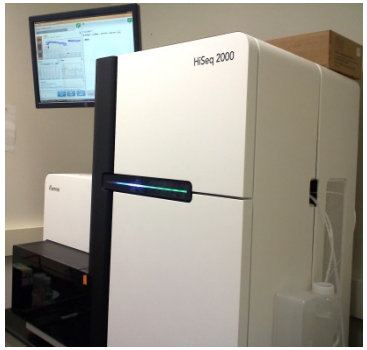
* 150 to 200 million single-reads (15 to 20 GB)

* 300 to 400 million paired-reads (30 to 40 GB)

- Run length: 7 days for single-reads,
  13 days for paired-reads

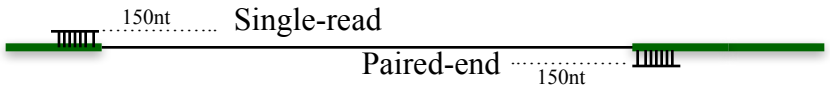
Courtesy of Alvaro Hernandez, UIUC

ILLUMINA HISEQ 2500



Flowcell
(2 lanes)

ILLUMINA LIBRARY FRAGMENT:



*Reads (sequences) = 150-300nt in length

Number of reads per lane:

* 100 to 150 million single-reads (15 to 22 GB)

* 200 to 300 million paired-reads (30 to 45 GB)

- Run length: <1 day for single-reads,
  2 days for paired-reads

Courtesy of Alvaro Hernandez, UIUC

DNA Sequencing: Illumina NovaSeq

NovaSeq 5000 and 6000
cost **$850,000** and **$985,000 (2017)**

In "Rapid Run Mode," the **Illumina HiSeq** 2500 instrument is capable of generating approximately 150 millions **reads** passing filter **per** lane, or up to 300 million **reads** passing filter **per** lane for paired-end sequencing.

NovaSeq: Output size up to 6Tb, up to 6 billion reads per run, length of reads 2x150

Courtesy of Alvaro Hernandez, UIUC

File Format: FASTA

- Text-based format for nucleotide or peptide sequences
- Line 1: description
  - “>” symbol
  - Sequence identifier

Database	Format
GenBank	gb | *accession* | *locus*
EMBL Data Library	emb | *accession* | *locus*
DDBJ, DNA Database of Japan	dbj | *accession* | *locus*
NBRF PIR	pir | | *entry*
Protein Research Foundation	prf | | *name*
SWISS-PROT	sp | *accession* | *entry name*
Brookhaven Protein Data Bank	pdb | *entry* | *chain*
Patents	pat | *country* | *number*
GenInfo Backbone Id	bbs | *number*
General database identifier	gnl | *database* | *identifier*
NCBI Reference Sequence	ref | *accession* | *locus*
Local Sequence Identifier	lcl | *identifier*

  - Description [optional]

File Format: FASTA

- Line 2: sequence data (1 or more lines)
  - Protein or nucleic acids sequences
    - Amino acids: A-Z, * (translation stop), - (gap)
    - Nucleic acids: A, C, G, T, U, R, Y, K, M, S, W, B, D, H, V, N (any A C G T U), X (masked), - (gap)
  - Each line less than 80 characters
- File extensions

Extension	Meaning	Notes
fasta (.fas)	Generic fasta	Can be .fa, .seq, .fsa
fna	Fasta nucleic acid	Generic nucleic acids
ffn	FASTA nucleotide coding regions	Coding regions for a genome
faa	Fasta amino acid	.mpfa: multiple protein fasta
frn	FASTA non-coding RNA	Non-coding RNA regions

File Format: FASTA

- Example: “random.fna”

```
>SEQUENCE_1
TGGCAATCTTGCTTCTGTTTACGGCTGGCATAGTTACGACA
GGTCTTTTTCT
>SEQUENCE_2
CCGGTTTCTTCAACCTTAGTTCTGGTAGCAGAATCAAGATA
CATGTTTTCGT
>SEQUENCE_3
GACGGCGTCAGCTGCAACAACTGTGCGCGCCATTGCCCTG
CCGGGGCGATC
```

- Example: “NP_852610.1”

```
>gi|31563518|ref|NP_852610.1| microtubule-associated
proteins 1A/1B light chain 3A isoform b [Homo sapiens]
MKMRFFSSPCGKAAVDPADRCKEVQQIRDQHPSKIPVIIERY
KGEKQLPVLDKTKFLVPDHVNMSELVKI
```

File Format: FASTQ

- Text-based format for sequences and quality scores
- Line 1
  - "@" symbol
  - Sequence identifier
  - Description [optional]
- Line 2
  - Raw sequence letters
- Line 3
  - "+" symbol
  - Sequence identifier
  - Description [optional]
- Line 4
  - Quality values for sequences in Line 2

File Format: FASTQ

- Quality value Q is an integer-valued function of p, the probability that the corresponding base call is incorrect
- Phred quality score:
  - $Q_{sanger} = -10\, log_{10} p$
- Quality values in increasing order of quality (ASCII):

```
!"#$%&'()*+,-
./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]
^_`abcdefghijklmnopqrstuvwxyz{|}~
```

File Format: FASTQ

- Illumina sequence identifiers
  - Example: @HWUSI-EAS100R:6:73:941:1973#0/1

HWUSI-EAS100R	Unique instrument name
6	Flowcell lane
73	Tile number within flowcell lane
941	'x'-coordinate of the cluster within tile
1973	'y'-coordinate of the cluster within tile
#0	Index number for multiplexed sample
/1	Member of a pair

- File extensions
  - .fq, .fastq

File Format: FASTQ

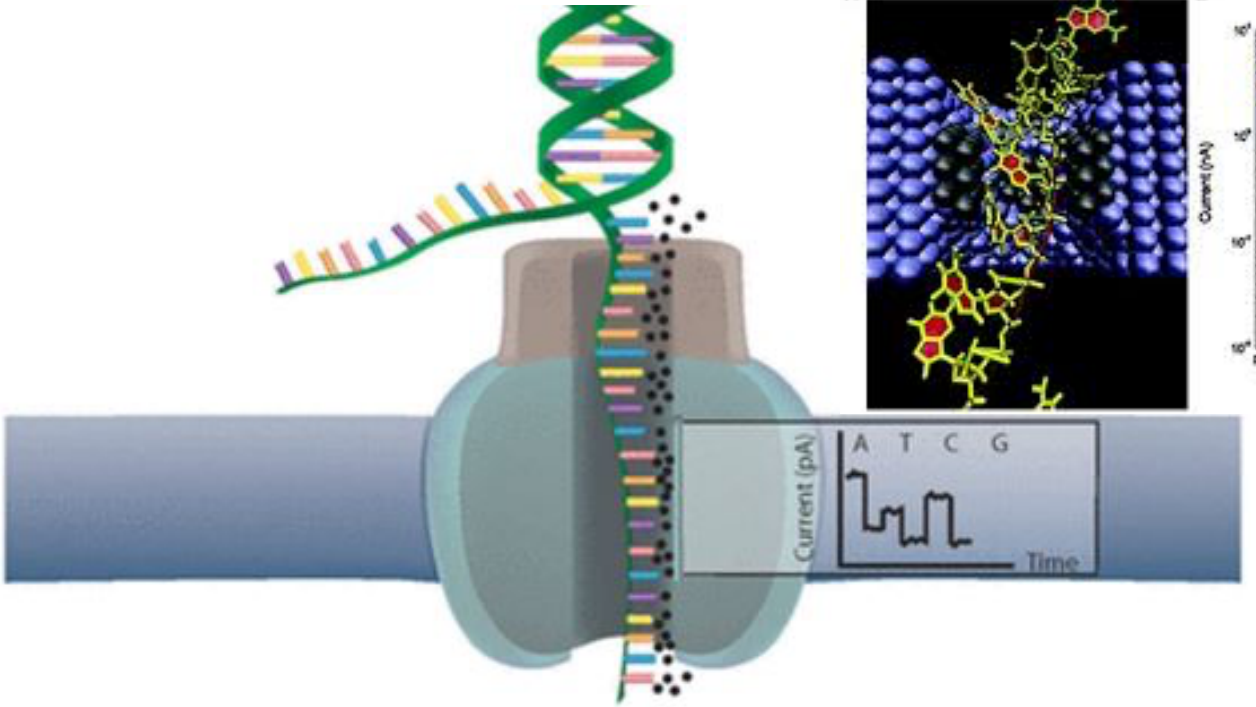
- Example (Illumina): “sample1.fq”

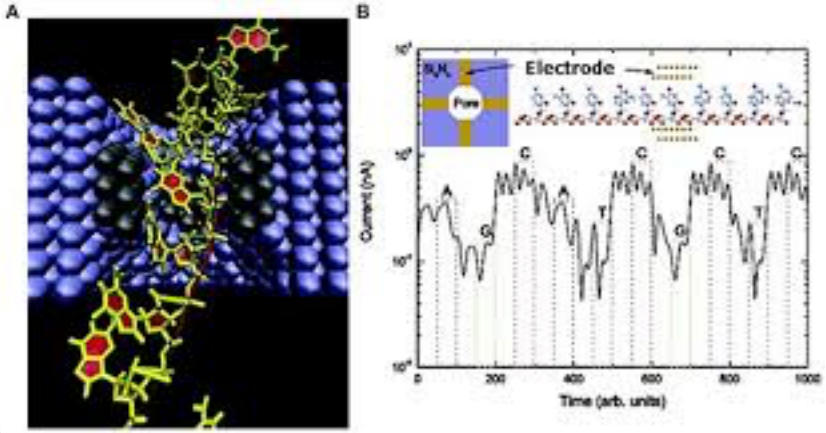
```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNN
TAGTTTCTTGAGATTTGTTGGGGGAGACATTTTTGTGATTGCCTTGAT
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`ddd^dddadd
^BBBBBBBBBBBBBBBBBBBBBBBBB
```

- Example (NCBI read archive)

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Third Generation Sequencers

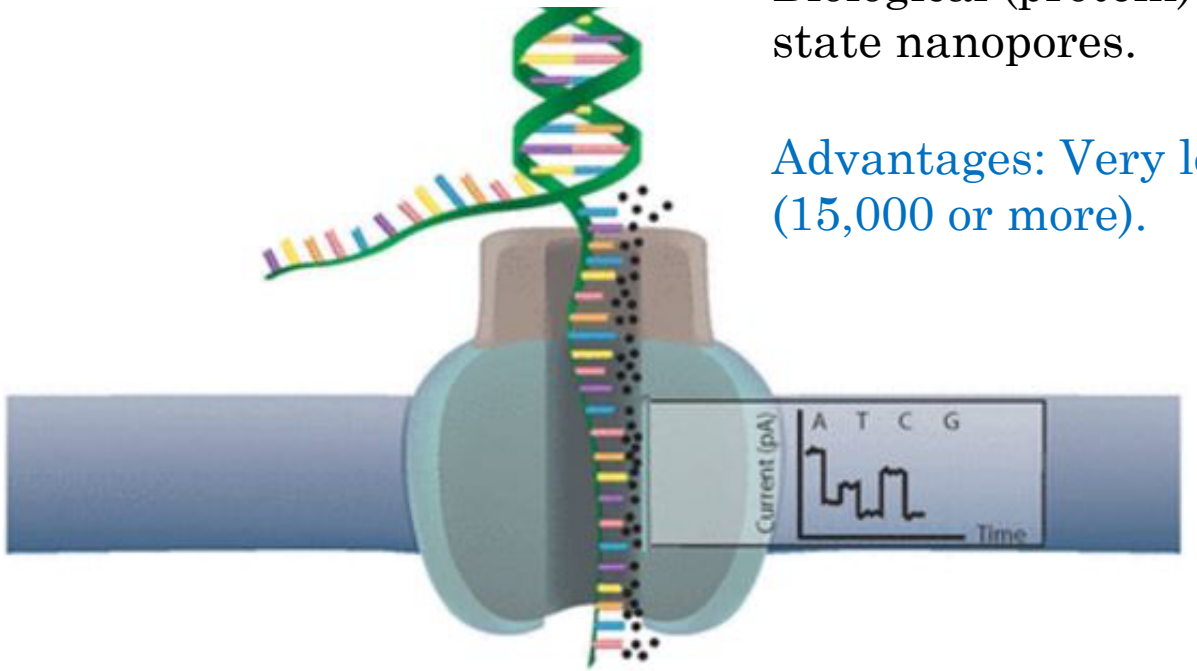




Credit: LabBiotech

Biological (protein) versus solid state nanopores.

Third Generation Sequencers



Credit: LabBiotech

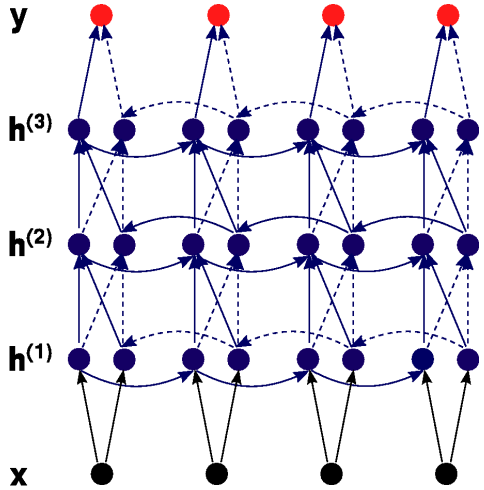
Biological (protein) versus solid state nanopores.

Advantages: Very long reads (15,000 or more).

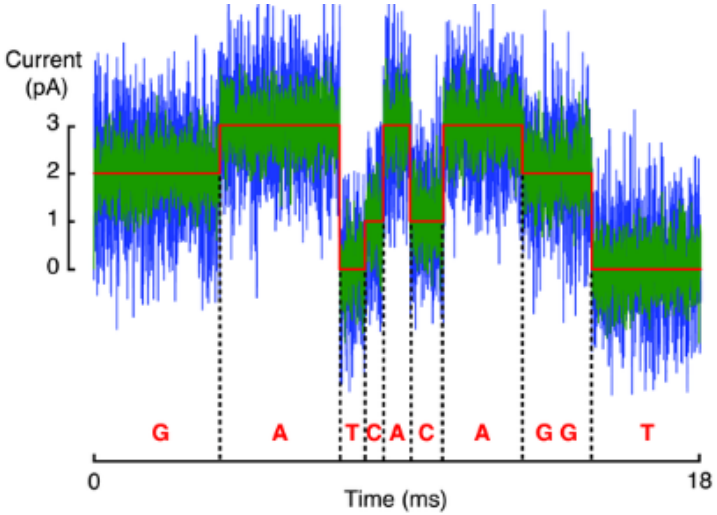
Advantages: Very long reads (15,000 or more).

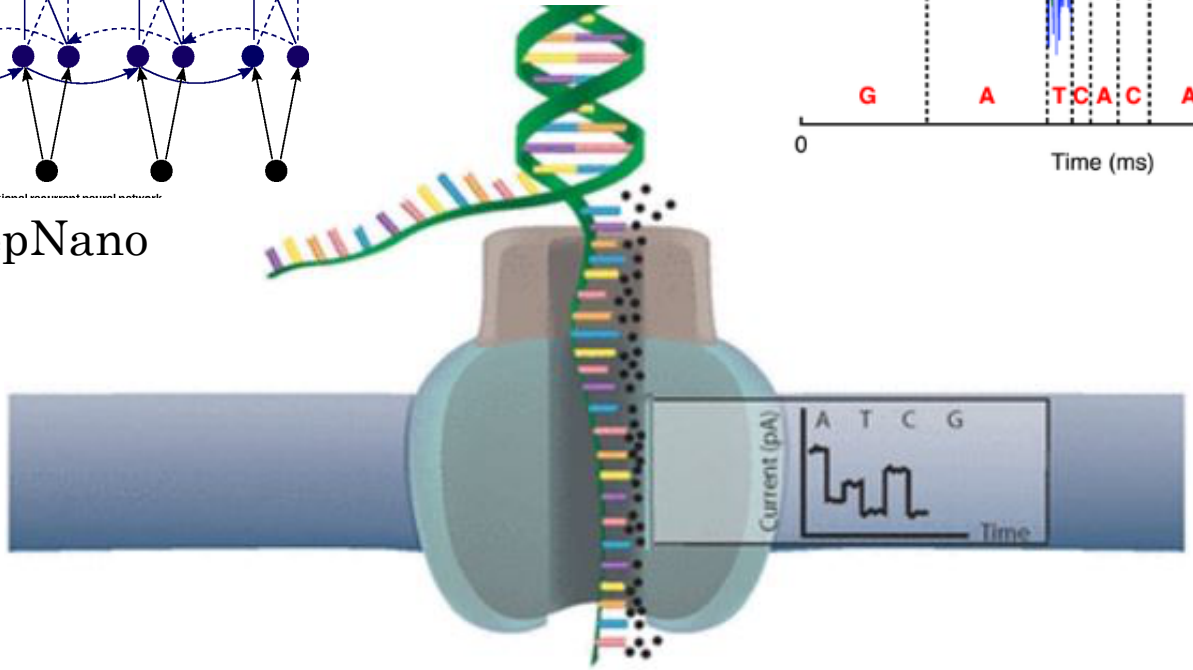
Disadvantages: Many substitution, insertion and deletion errors.

Base Calling



DeepNano

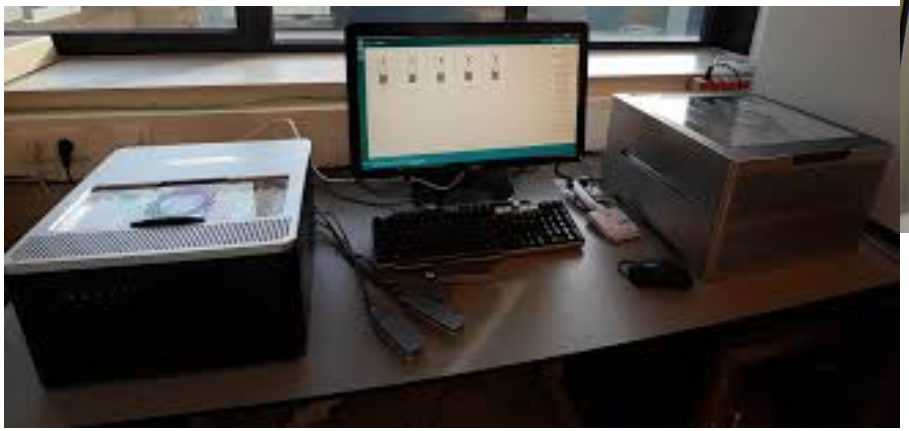




Credit: LabBiotech

MINION AND GRIDION

Cheap(er) and Portable

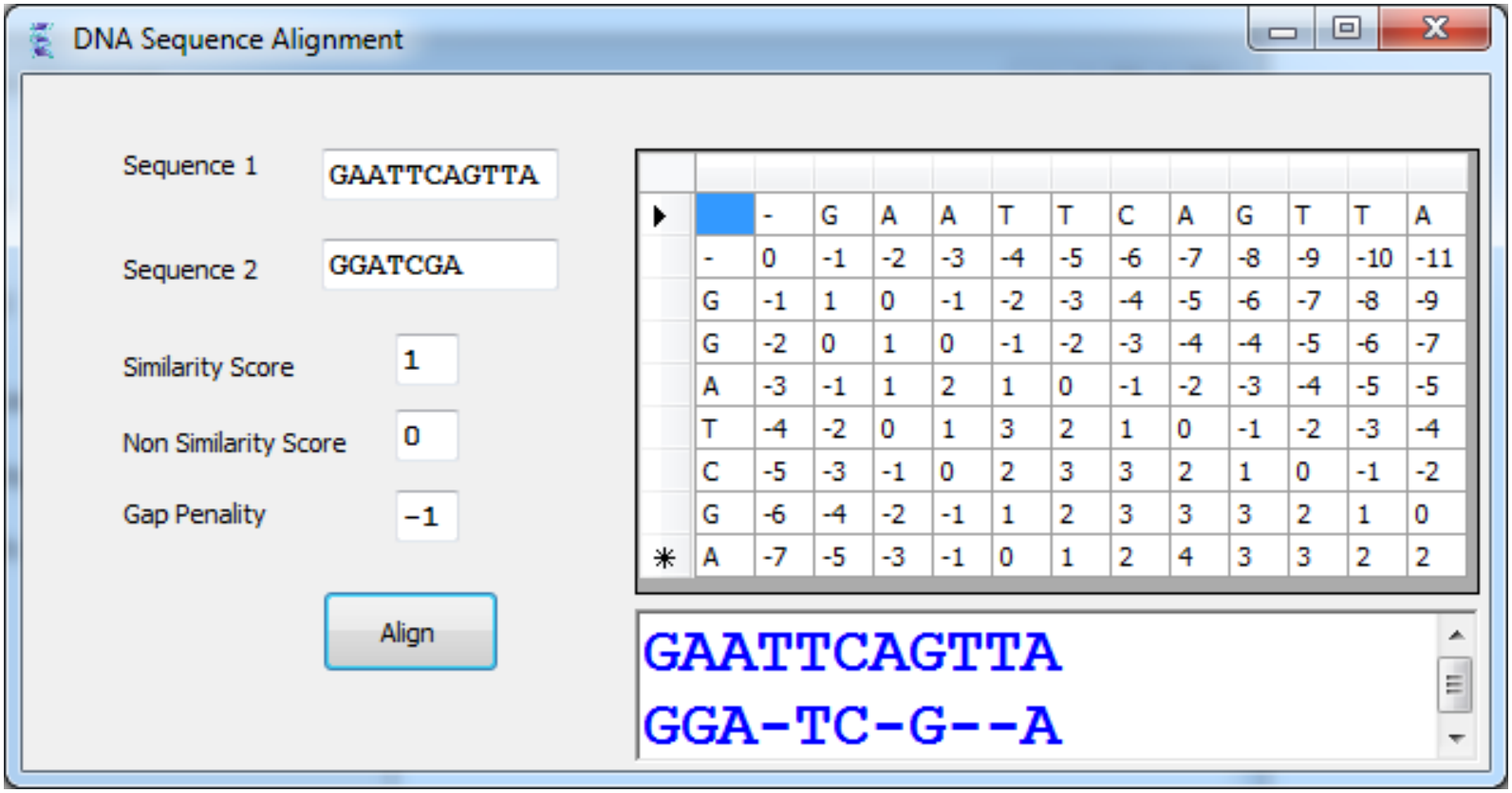


Reconstructing Sequences from Traces Alignment

Position	20	22	23	24	25	26	32	42	45	50	53	56	69	81	93	103	121	122	123	124	147	148	149	153	155	156	170	183	Total
Consensus	A	A	A	A	G	C	A	C	T	A	C	C	C	C	A	A	A	T	T	A	G	T	T	G	C	G	C	G	
AXE	A	-	-	-	-	-	A	C	C	A	C	C	T	C	A	A	-	-	-	-	G	T	T	T	C	G	A	G	13
DW	A	-	-	-	-	-	A	C	C	A	C	C	T	C	A	A	-	-	-	-	G	T	T	G	C	G	A	G	12
FNZ	A	-	-	-	-	-	A	C	C	A	C	C	T	C	A	A	-	-	-	-	G	T	T	G	C	G	A	G	12
NQ	A	-	-	-	-	-	A	C	C	G	C	C	T	C	A	A	-	-	-	-	G	T	T	G	C	G	C	G	12
ZT	A	-	-	-	-	-	A	C	C	G	C	C	T	C	A	A	-	-	-	-	G	T	T	G	C	G	C	G	12
MEM	A	-	-	-	-	-	A	C	C	G	T	C	T	C	A	A	-	-	-	-	G	T	T	G	C	G	C	G	13
GD	A	A	A	A	G	C	A	C	T	A	C	C	C	G	A	A	-	-	-	-	G	C	T	G	T	G	C	G	7
IM	A	A	A	A	G	C	A	C	T	A	C	C	C	G	A	A	-	-	-	-	G	C	T	G	T	G	C	G	7
NRS	A	A	A	A	G	C	A	C	T	A	C	C	C	G	A	A	-	-	-	-	G	C	T	G	T	G	C	G	7
ZPJ	A	A	A	A	G	C	A	C	T	A	C	C	C	G	A	A	-	-	-	-	G	C	T	G	T	G	C	G	7
AB	A	A	A	A	G	C	A	C	T	A	C	C	C	C	A	A	A	T	T	A	G	T	T	G	C	G	C	G	0
XC	A	A	A	A	G	C	A	C	T	A	C	C	C	C	A	A	A	T	T	A	G	T	T	G	C	G	C	G	0
MR	T	A	A	A	G	C	C	C	T	A	C	C	C	C	A	A	A	T	T	A	G	T	T	G	C	G	C	G	2
UBG	T	A	A	A	G	C	C	C	T	A	C	C	C	C	A	A	A	T	T	A	G	T	T	G	C	G	C	G	2
FYC	A	A	A	A	G	C	A	A	T	A	C	T	C	C	T	G	A	T	T	A	G	T	T	G	C	T	C	-	6
JRR	A	A	A	A	G	C	A	A	T	A	C	T	C	C	T	G	A	T	T	A	G	T	T	G	C	T	C	-	6
UM	A	A	A	A	G	C	A	A	T	A	C	T	C	C	T	G	A	T	T	A	G	T	T	G	C	T	C	-	6
WZ	A	A	A	A	G	C	A	A	T	A	C	T	C	C	A	G	A	T	T	A	G	T	T	G	C	T	C	-	5
DGO	A	A	A	A	G	C	A	A	T	A	C	C	C	C	A	A	A	T	T	A	-	-	-	G	C	T	C	-	6
Summary:	17A 2T	13A 6-	13A 6-	13A 6-	13G 6-	13C 6-	17A 2C	14C 5A	13T 6C	16A 3G	18C 1T	15C 4T	13C 6T	15C 4G	16A 3T	15A 4G	9A 10-	9T 10-	9T 10-	9A 10-	18G 1-	14T 4C	18T 1-	18G 1T	15C 4T	14G 5T	16C 3A	14G 5-	135

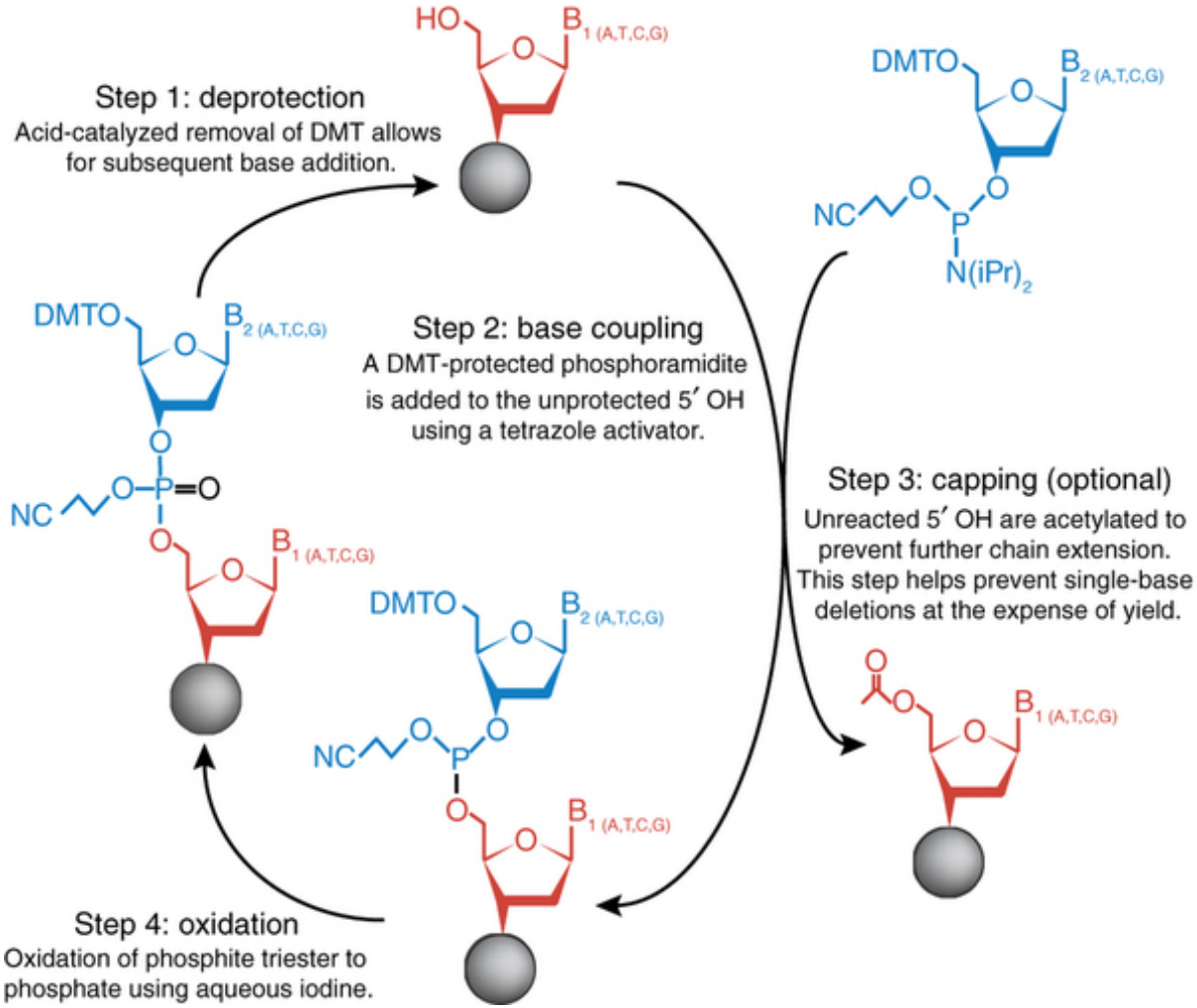
Dynamic programming, greedy algorithms etc
Basic computation of weighted Levenshtein distance
CLUSTAL OMEGA, MUSCLE, etc

Reconstructing Sequences from Traces Alignment



	-	G	A	A	T	T	C	A	G	T	T	A
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
G	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
G	-2	0	1	0	-1	-2	-3	-4	-4	-5	-6	-7
A	-3	-1	1	2	1	0	-1	-2	-3	-4	-5	-5
T	-4	-2	0	1	3	2	1	0	-1	-2	-3	-4
C	-5	-3	-1	0	2	3	3	2	1	0	-1	-2
G	-6	-4	-2	-1	1	2	3	3	3	2	1	0
A	-7	-5	-3	-1	0	1	2	4	3	3	2	2

From Reading to Writing

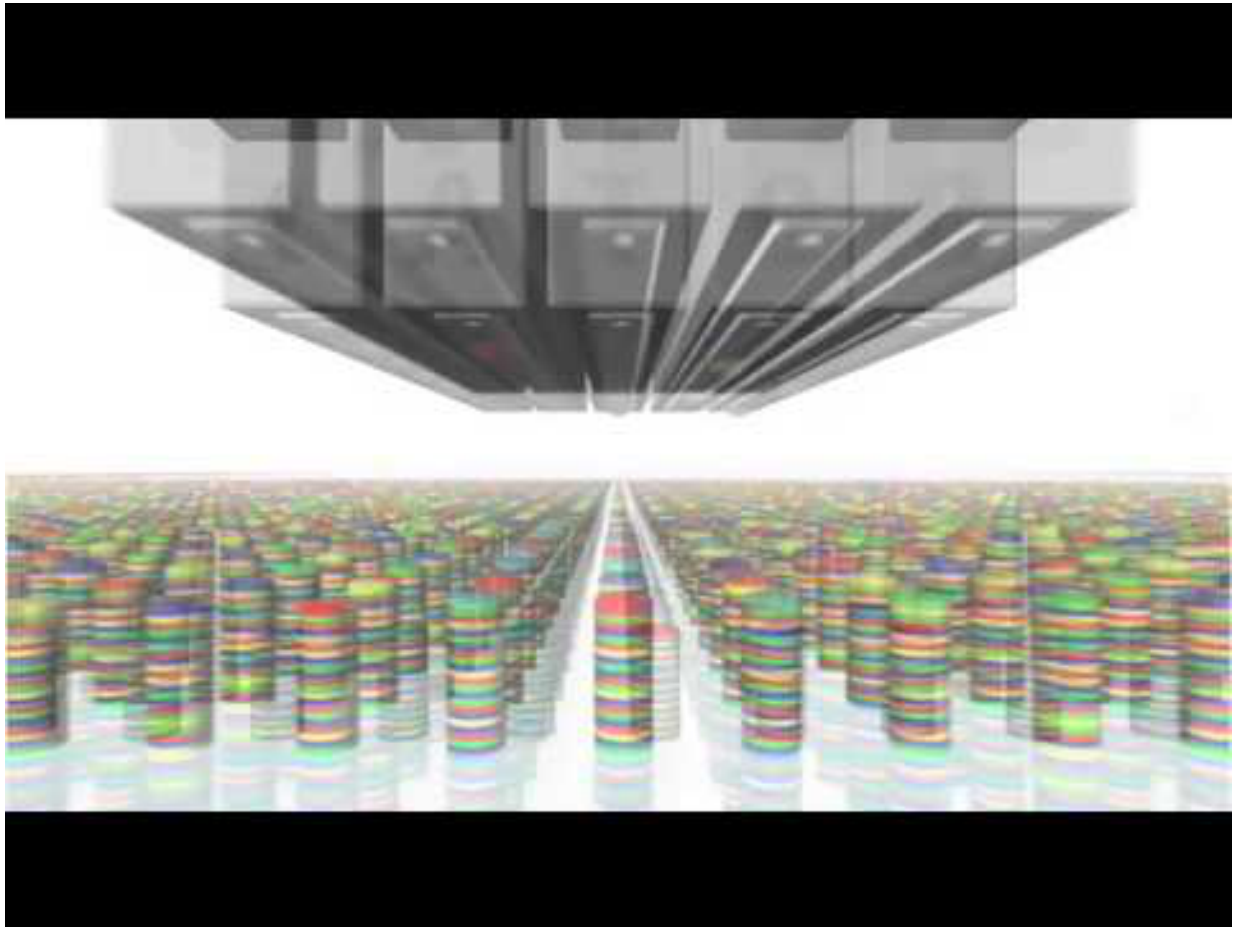


Expensive

Relatively slow but parallelizable

Commercially available from
Agilent
Twist
IDT

From Reading to Writing



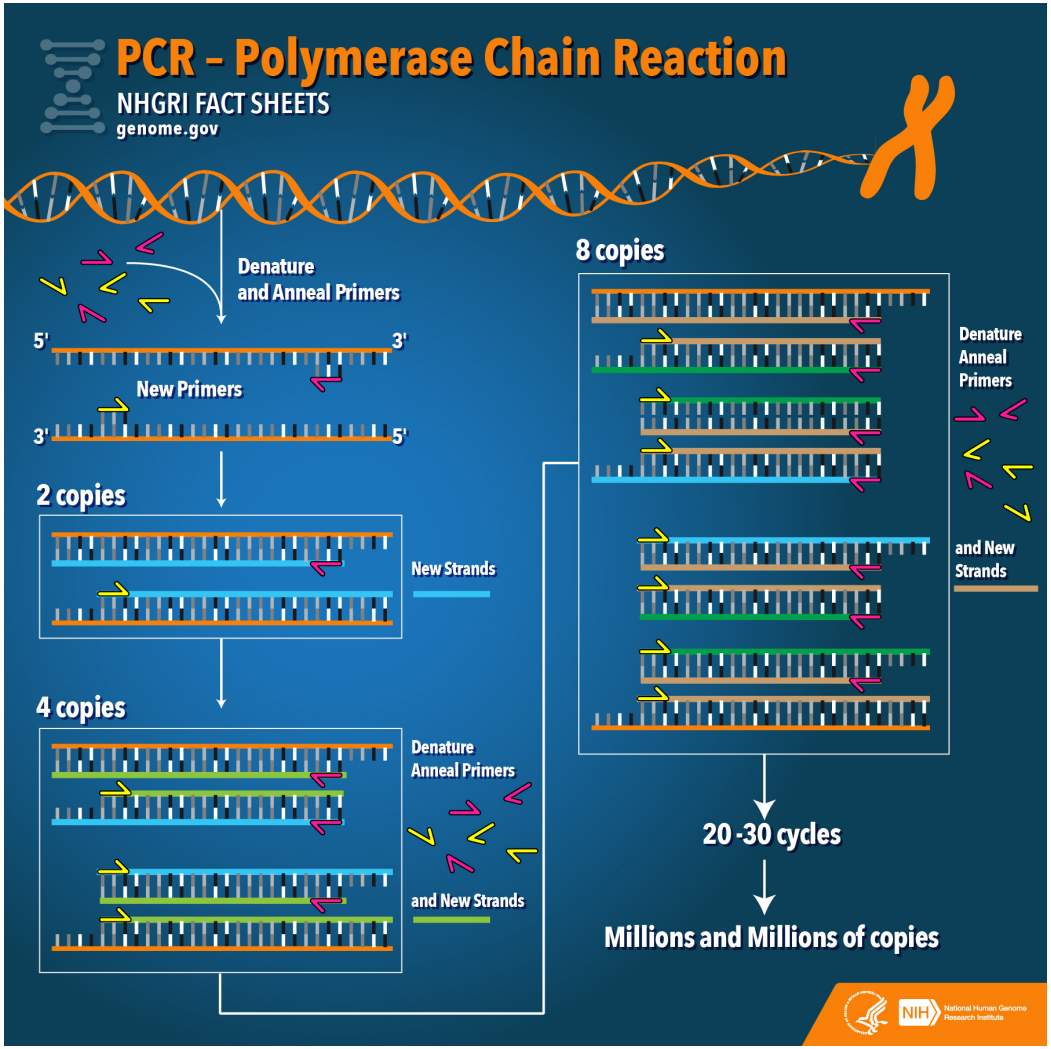
Expensive

Relatively slow but parallelizable

Commercially available from
Agilent
Twist
IDT

Credit: Agilent (INKJet Technology)

From Writing to Copying



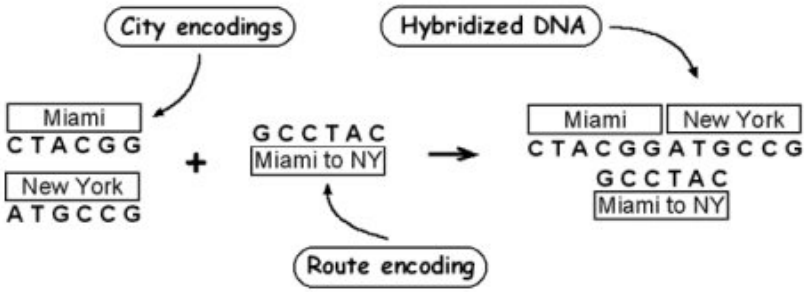
Credit: NIH

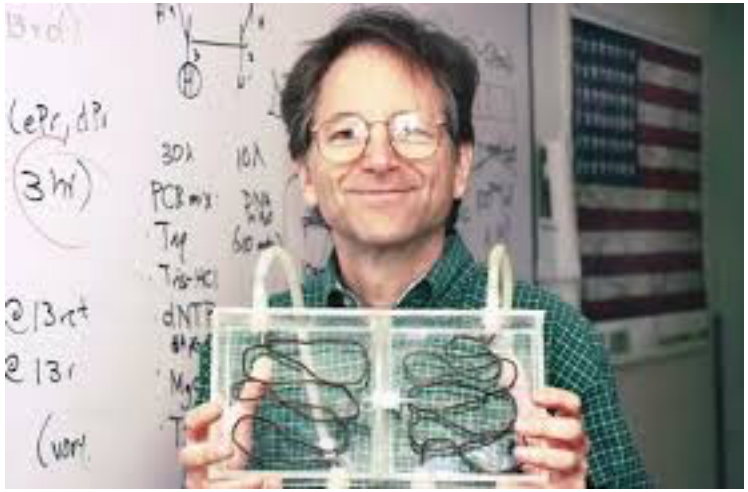
PCR machine

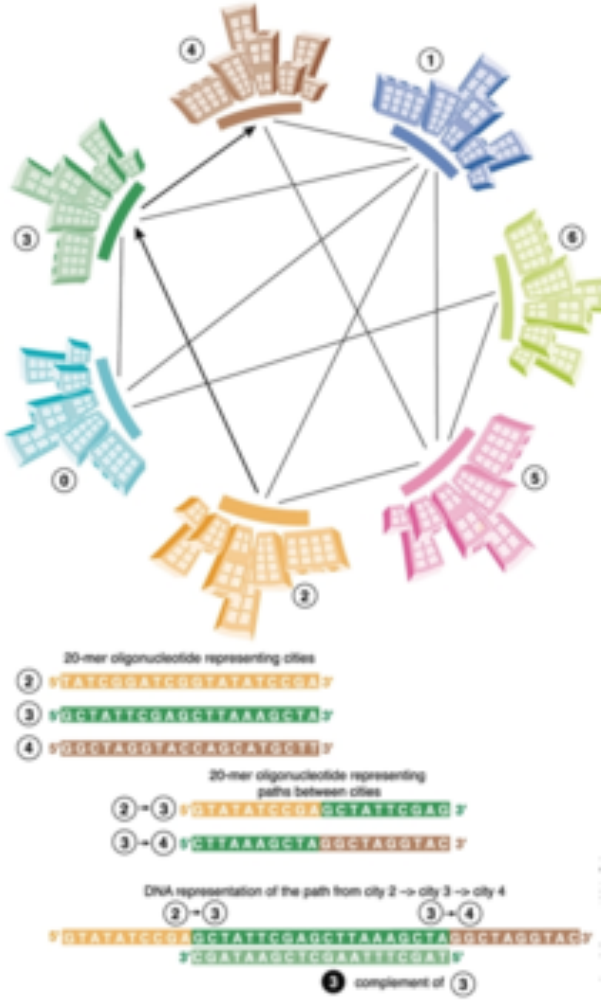
Thermofisher

($200)

And Computing: Adleman's Experiment



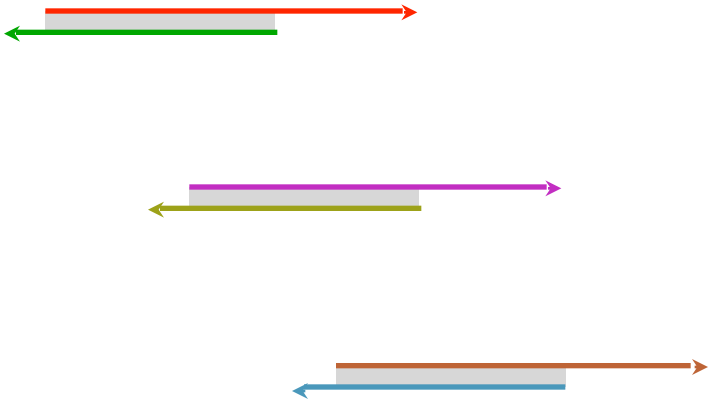




Credit: Scientific American and Len Adleman

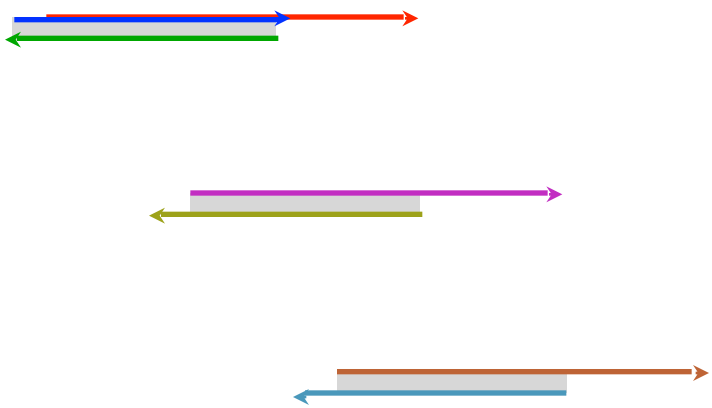
And Computing: Strand Displacement

Input



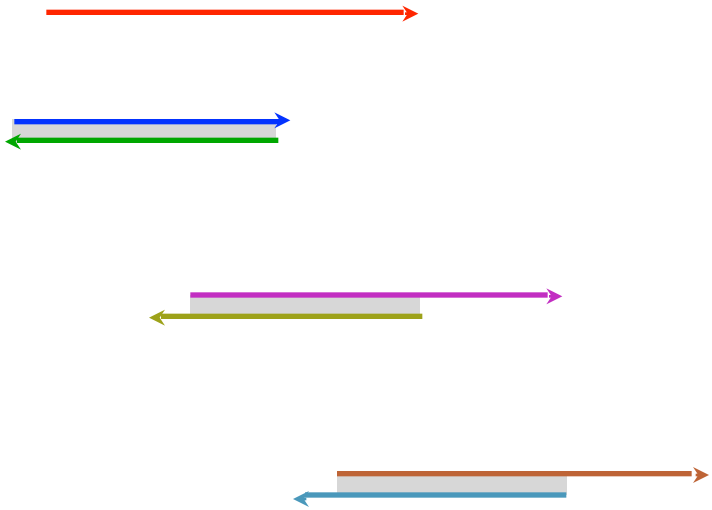
Curtesy of D. Soloveichik

And Computing: Strand Displacement



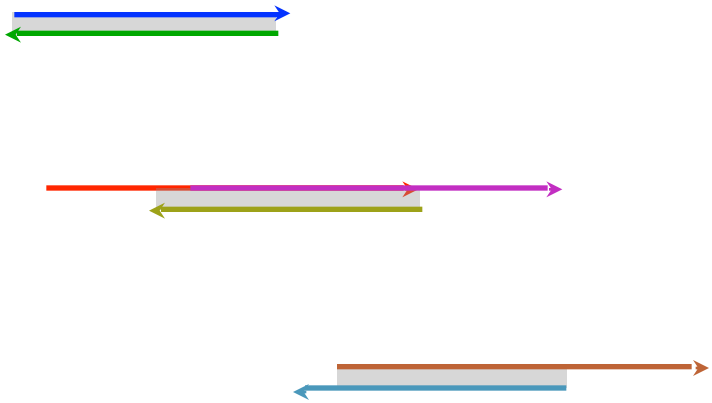
Curtesy of D. Soloveichik

And Computing: Strand Displacement



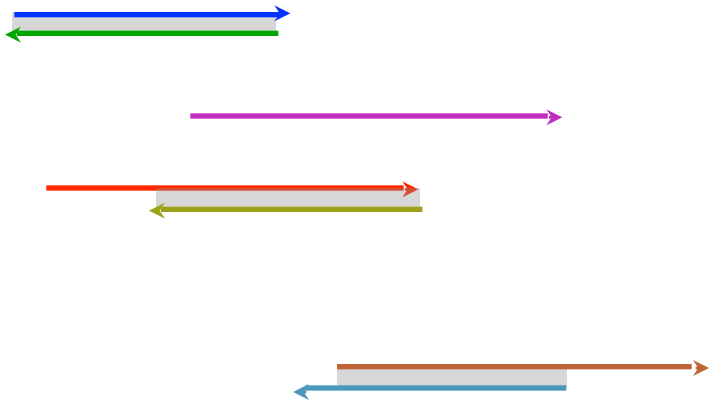
Curtesy of D. Soloveichik

And Computing: Strand Displacement



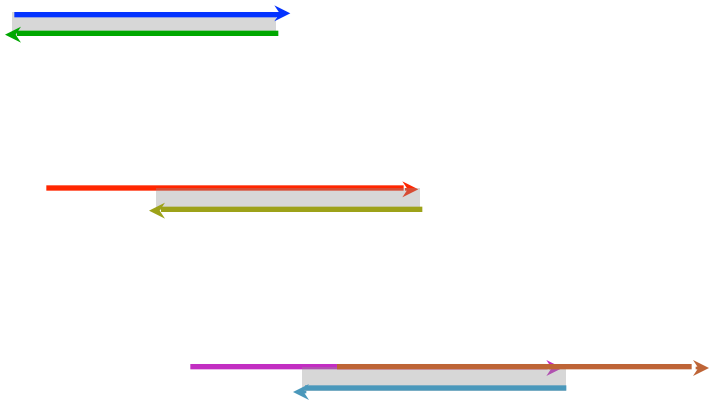
Curtesy of D. Soloveichik

And Computing: Strand Displacement



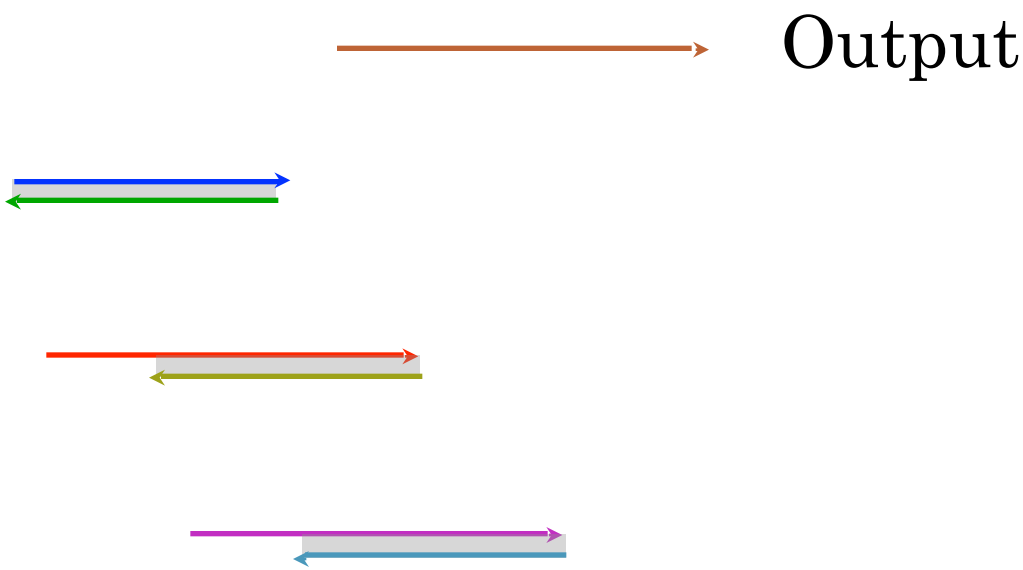
Curtesy of D. Soloveichik

And Computing: Strand Displacement



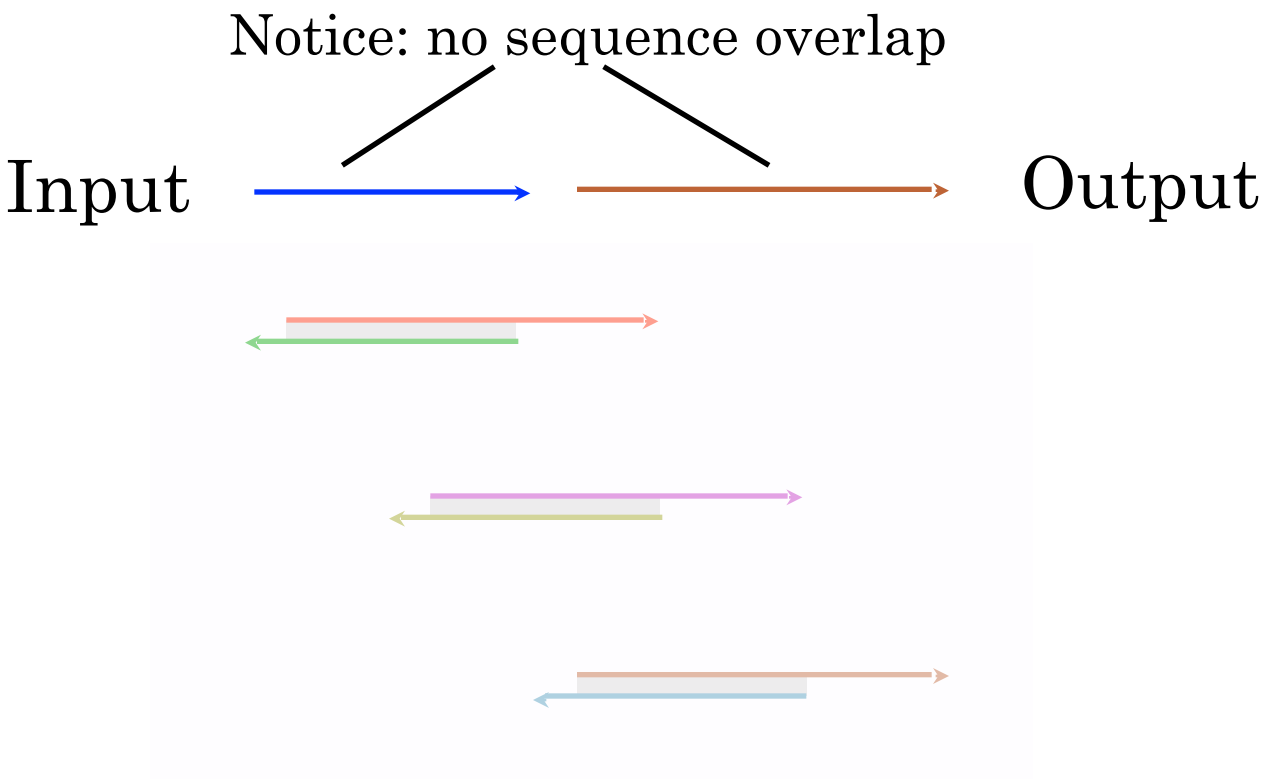
Curtesy of D. Soloveichik

And Computing: Strand Displacement



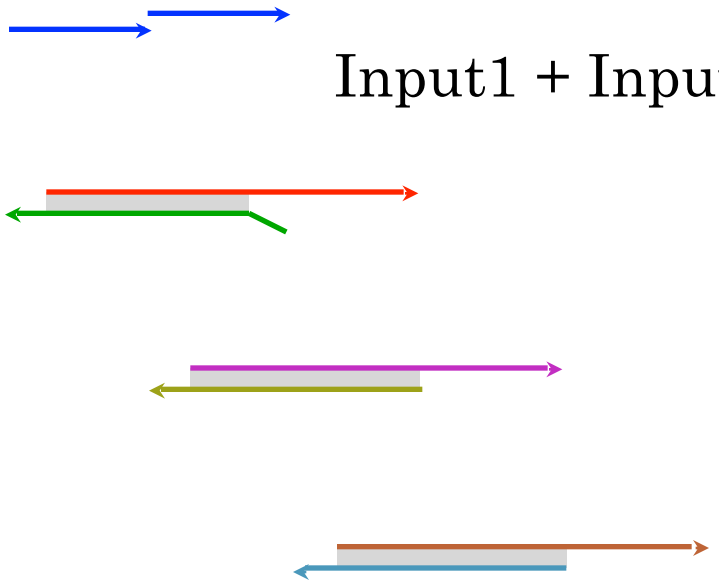
Curtesy of D. Soloveichik

And Computing: Strand Displacement



Curtesy of D. Soloveichik

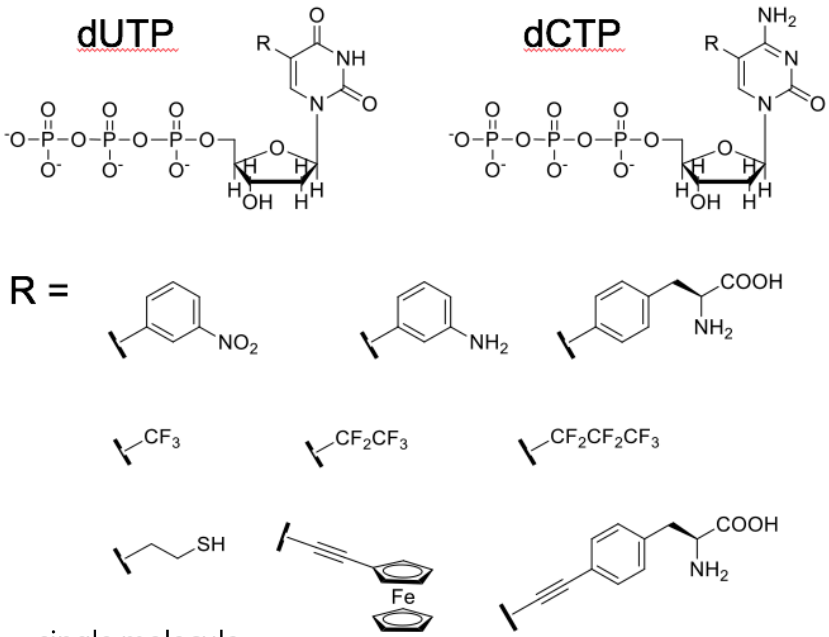
And Computing: Strand Displacement

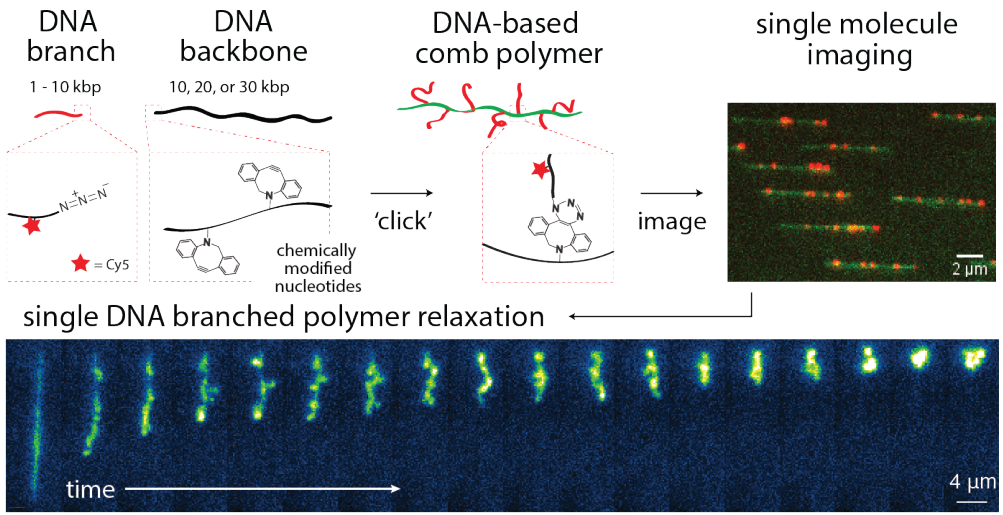


Input1 + Input2 → Output

Curtesy of D. Soloveichik

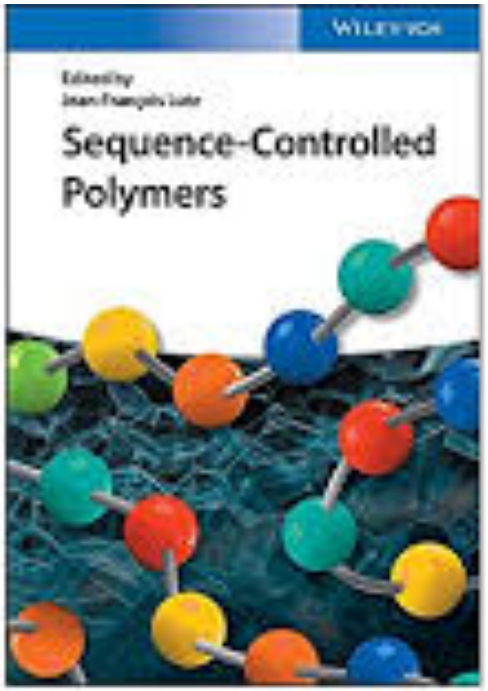
Beyond DNA: Chemically Modified DNA

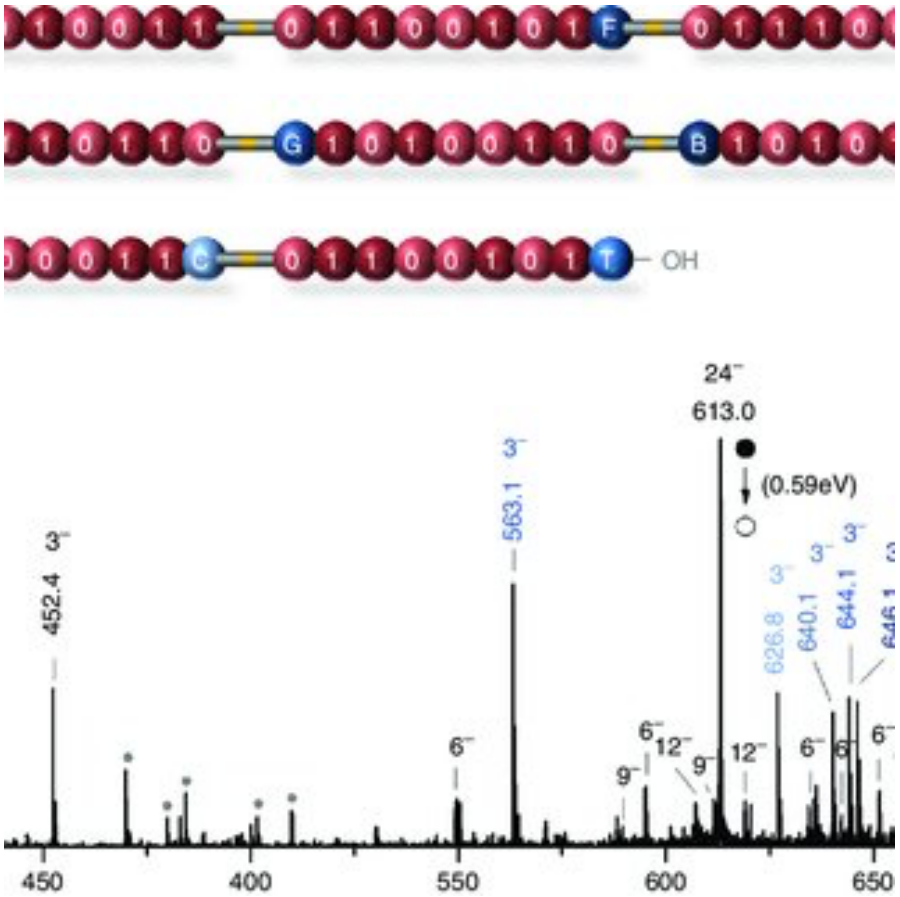




Credit: C. Schroeder

Beyond DNA: Synthetic polymers





Credit: J. F. Lutz