CMSC423: Bioinformatic Algorithms, Databases and Tools

Mihai Pop

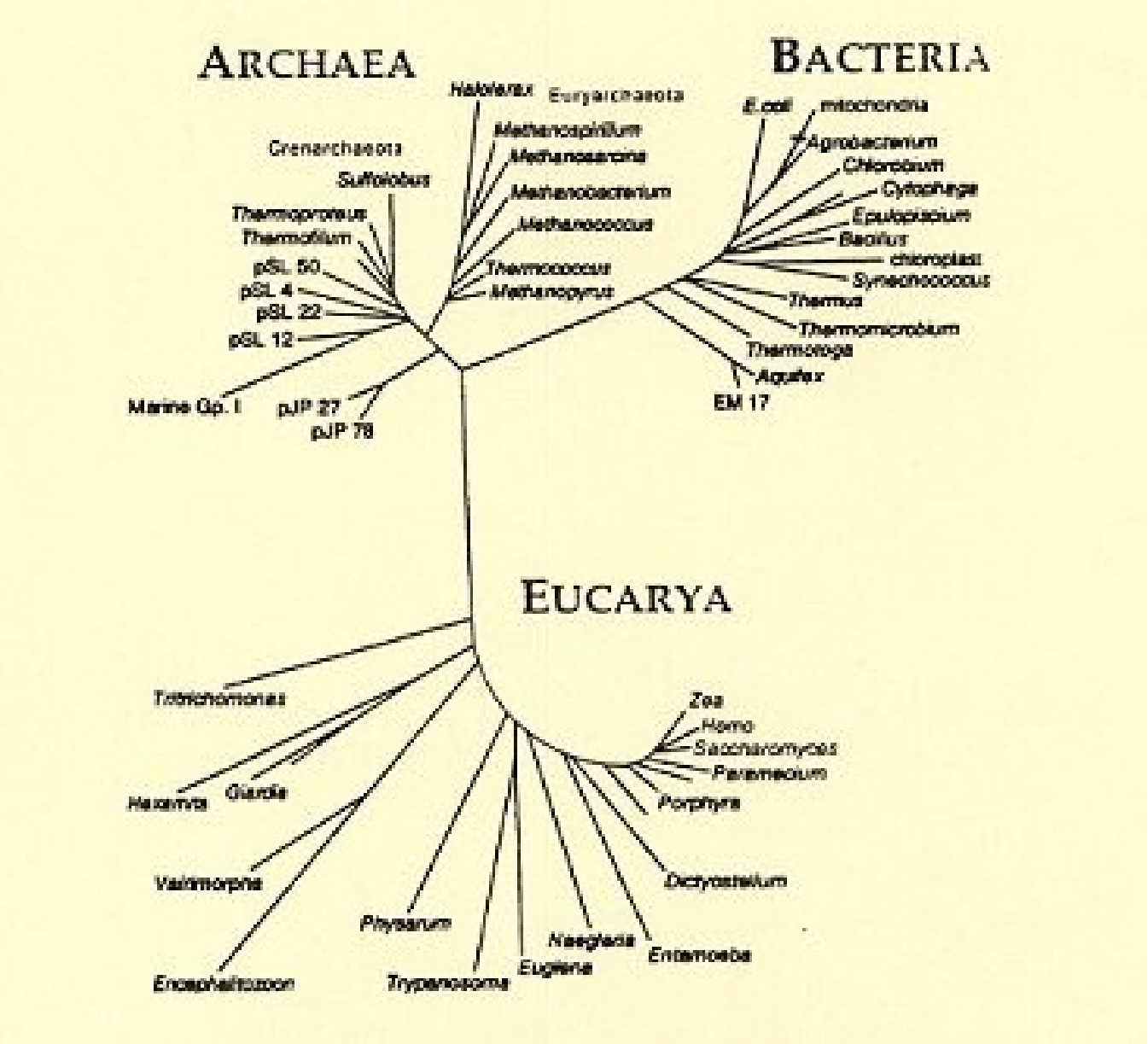
Molecular biology primer

Admin...

- Have you tried your glue accounts?
- Issues/concerns/questions about class and policies?

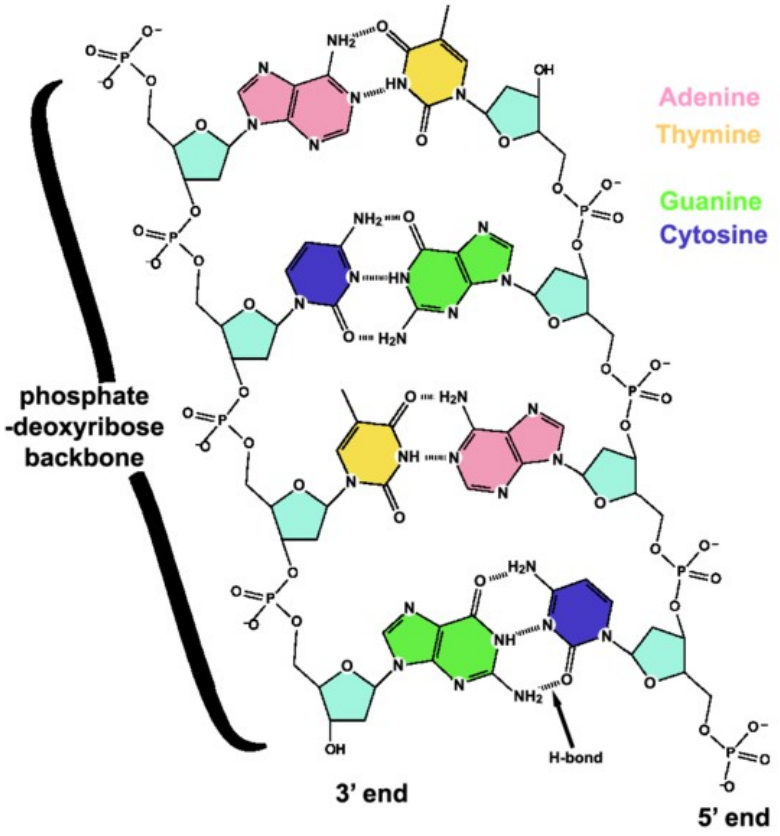
- Reading “assignment” - Chapter 1 in the book.

The tree of life



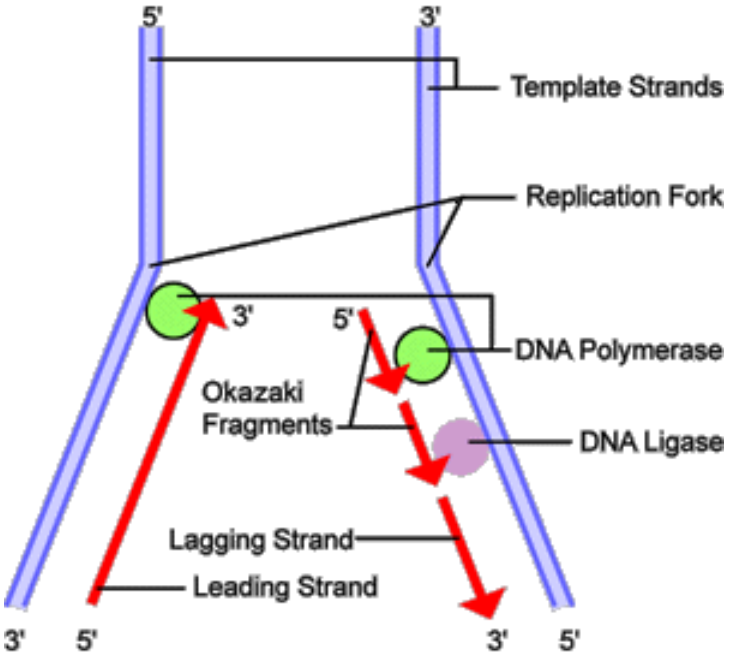
http://www.fossilmuseum.net/Tree_of_Life/Domains_Archaea_Bacteria/

DNA – the code of life



- Purines A, G, caffeine
- Pyrimidines C, T
- Sugar backbone (ticker tape)

- Double-stranded – allows replication



pictures from wikipedia

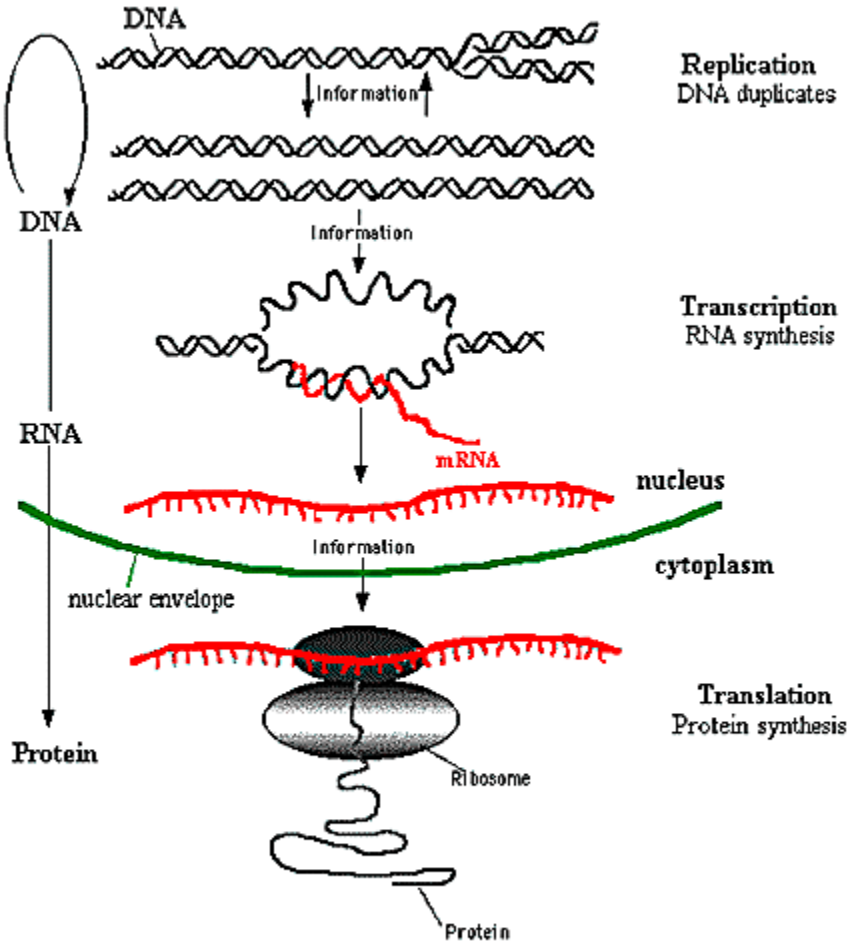
DNA in the computer

- FASTA/multi-FASTA file format

```
>gi|110227054|gb|AE004091.2| Pseudomonas aeruginosa PAO1, complete
   genome
TTTAAAGAGACCGGCGATTCTAGTGAAATCGAACGGGCAGGTCAATTTCCAACCAGCGATGACGTAATAG
ATAGATACAAGGAAGTCATTTTTCTTTTAAAGGATAGAAACGGTTAATGCTCTTGGGACGGCGCTTTTCT
GTGCATAACTCGATGAAGCCCAGCAATTGCGTGTTTCTCCGGCAGGCAAAAGGTTGTCGAGAACCGGTGT
CGAGGCTGTTTCCTTCCTGAGCGAAGCCTGGGGATGAACGAGATGGTTATCCACAGCGGTTTTTTCCACA
CGGCTGTGCGCAGGGATGTACCCCCTTCAAAGCAAGGGTTATCCACAAAGTCCAGGACGACCGTCCGTCG
```

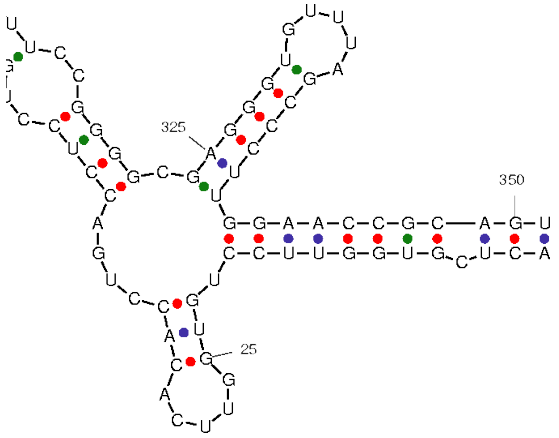
- Parsers easy to write, also available in a variety of software libraries

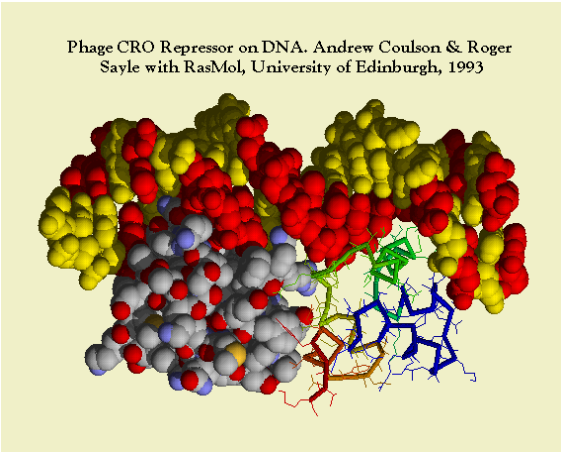
Central dogma



The Central Dogma of Molecular Biology

AGGTACGCGTACCTGACAGG





http://www.accessexcellence.org/RC/VL/GG/central.html

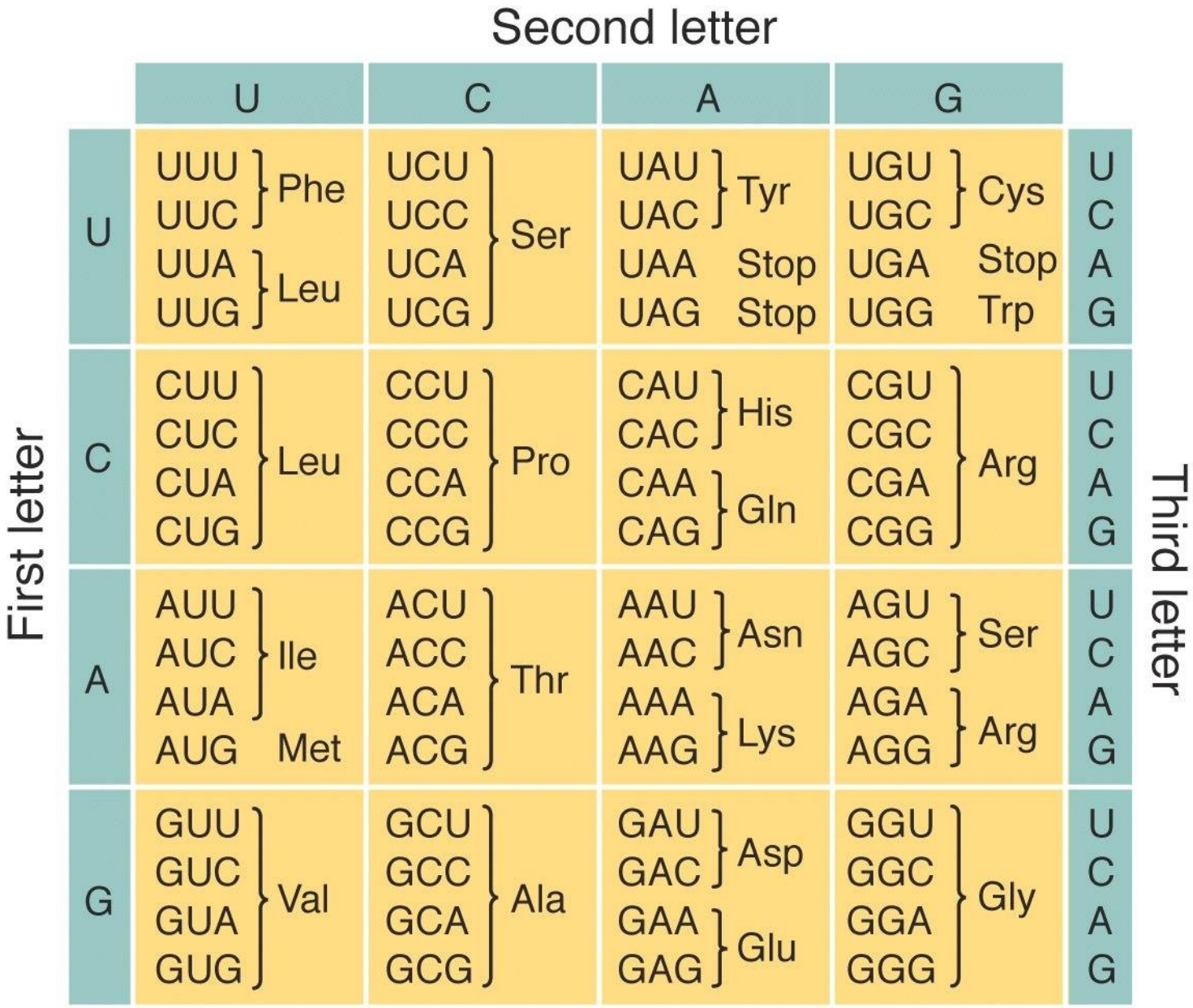
Genes, transcription, translation

- DNA – RNA - Thymine replaced by Uracil (T-U)
- The transcribed segments are called genes

ACCGUACC**AUG**UUA...AUAGGC**UGA**GCA

- AUG – start codon (also amino-acid Methionine)
- UAA, UAG, UGA – stop codons
- Genes are read in sets of 3 nucleotides during translation – $4^3 = 64$ possible combinations
- Each combination codes for one of 20 amino-acids – the building blocks for proteins

Amino-acid translation table



Genes/proteins in the computer

```
>gi|15596155|ref|NP_249649.1| basic amino acid,
MKVMKWSAIALAVSAGSTQFAVADAFVSDQAEAKGFIEDSSLDLLLRNYYFNRDGKSGSGDRVDWTQGFL
TTYESGFTQGTVGFGVDAFGYLGLKLDGTSDKTGTGNLPVMNDGKPRDDYSRAGGAVKVRISKTMLKWGE
MQPTAPVFAAGGSRLFPQTATGFQLQSSEFEGLDLEAGHFTEGKEPTTVKSRGELYATYAGETAKSADFI
GGRYAITDNLSASLYGAELEDIYRQYYLNSNYTIPLASDQSLGFDFNIYRTNDEGKAKAGDISNTTWSLA
AAYTLDAHTFTLAYQKVHGDQPFDYIGFGRNGSGAGGDSIFLANSVQYSDFNGPGEKSWQARYDLNLASY
GVPGLTFMVRYINGKDIDGTKMSDNNVGYKNYGYGEDGKHHETNLEAKYVVQSGPAKDLSFRIRQAWHRA
NADQGEGDQNEFRLIVDYPLSIL
```

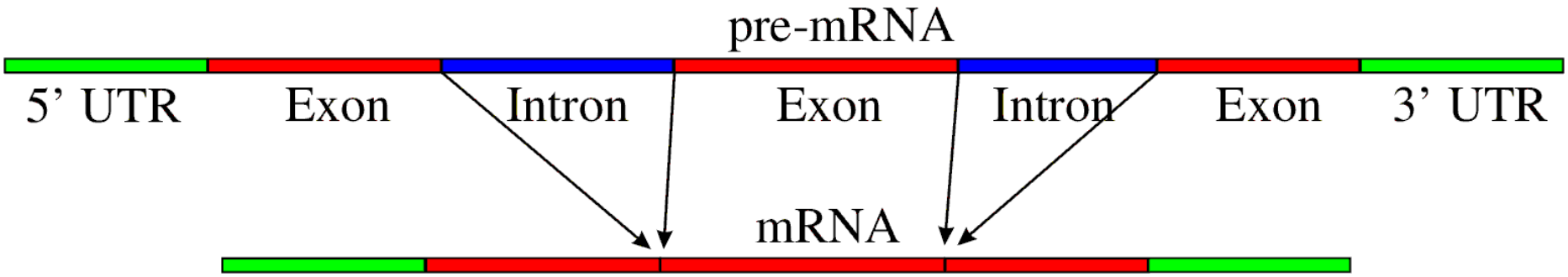
- Same FASTA/multi-FASTA but with bigger alphabet

Genes/proteins in the computer

```
gene            complement(1043983..1045314)
                /gene="oprD"
                /locus_tag="PA0958"
CDS             complement(1043983..1045314)
                /gene="oprD"
                /locus_tag="PA0958"
                /note="Product name confidence: Class 1 (Function
                experimentally demonstrated in P. aeruginosa)"
                /codon_start=1
                /transl_table=11
                /product="Basic amino acid, basic peptide and
                imipenem outer membrane porin OprD precursor"
                /protein_id="AAG04347.1"
                /db_xref="GI:9946864"
```

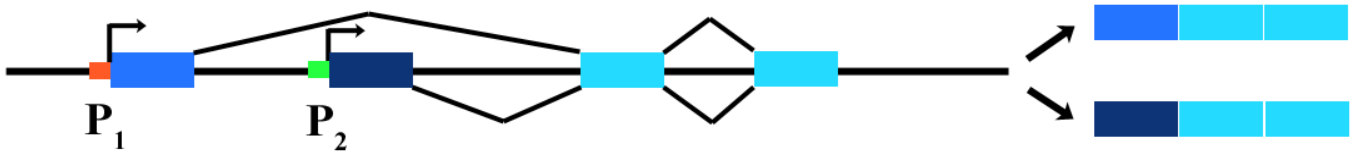
- GenBank file format

Translation – complications

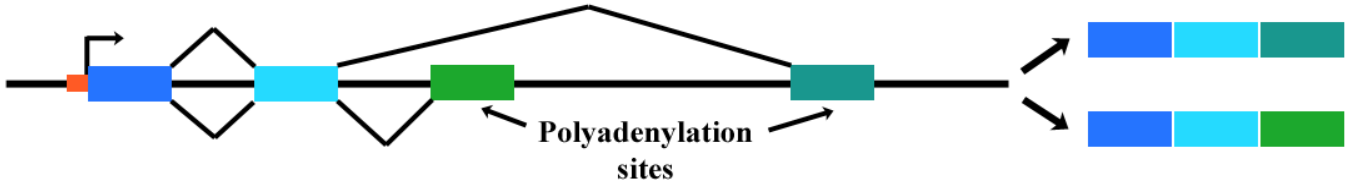


Alternative splicing examples

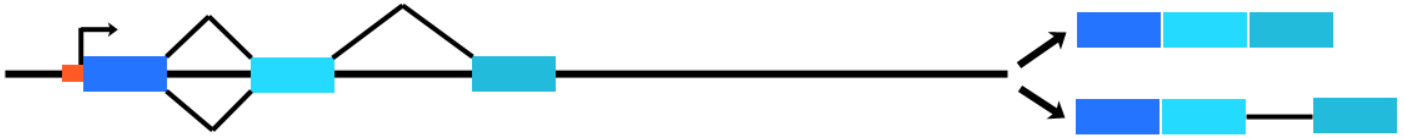
(a) Alternative selection of promoters (e.g., *myosin* primary transcript)



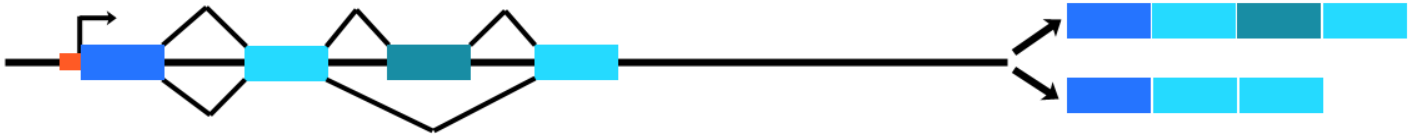
(b) Alternative selection of cleavage/polyadenylation sites (e.g., tropo*myosin* transcript)



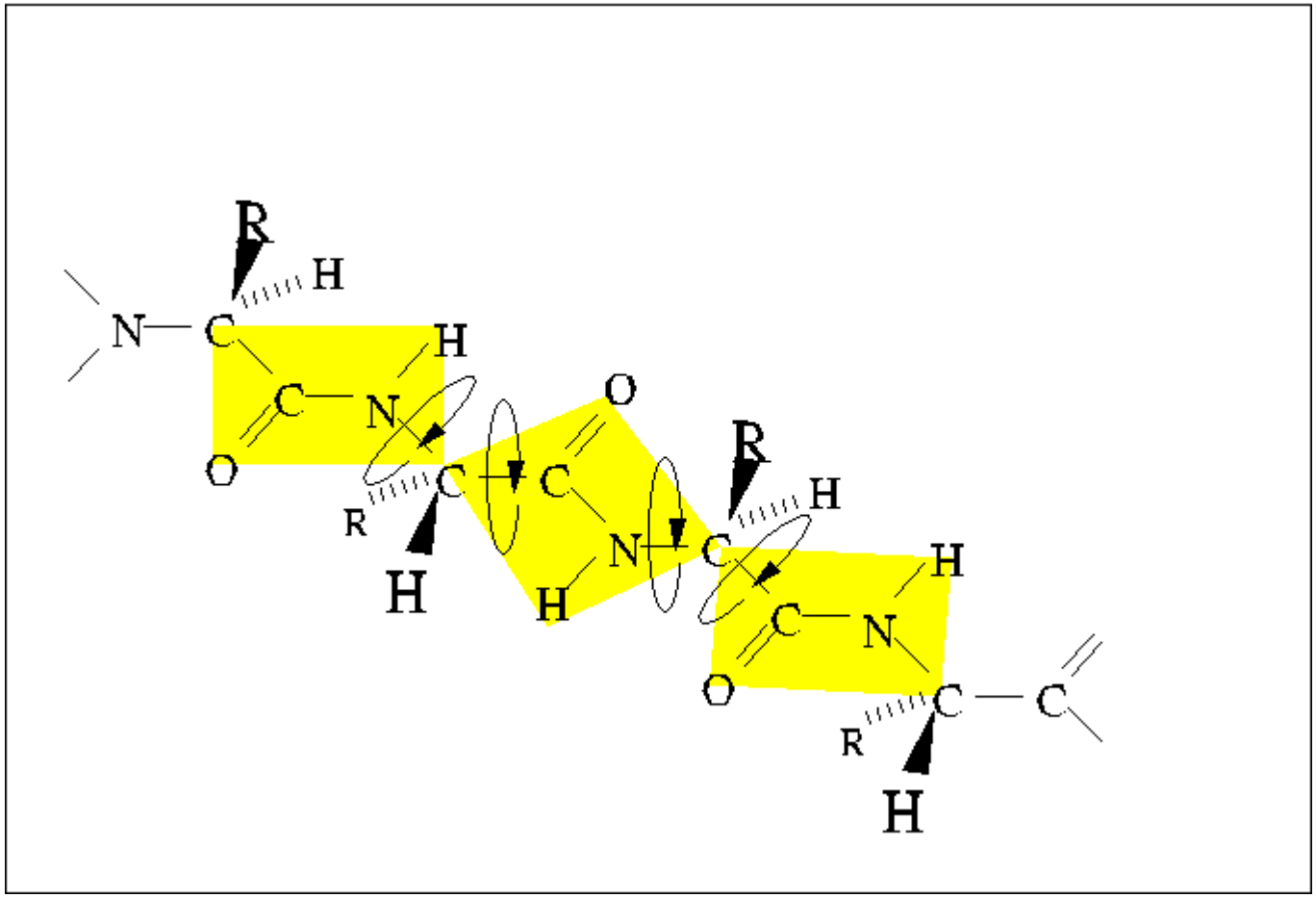
(c) Intron retaining mode (e.g., *transposase* primary transcript)



(d) Exon cassette mode (e.g., *troponin* primary transcript)



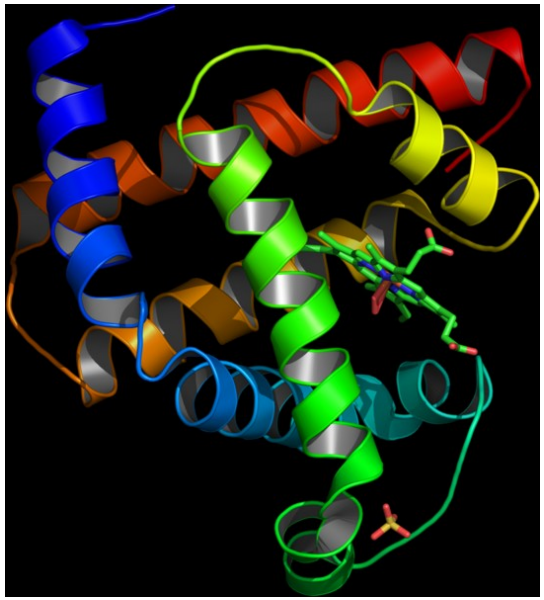
Protein structure

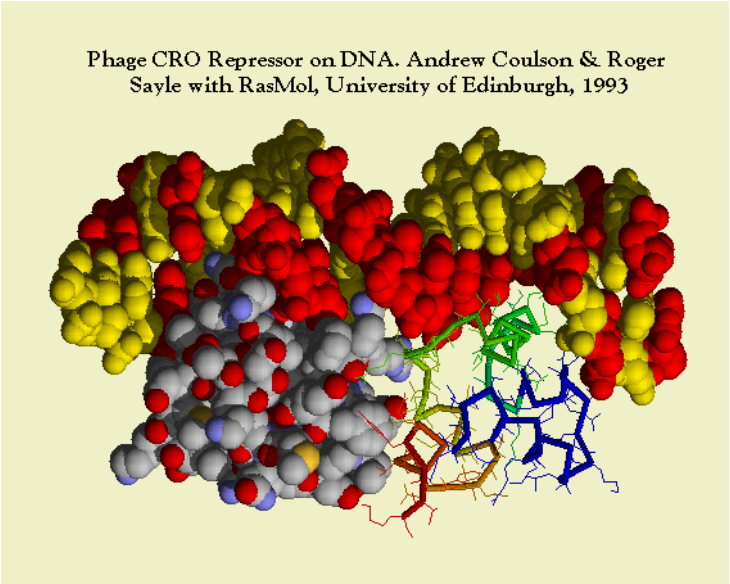


http://www.tulane.edu/~biochem/med/second.htm

Protein structure

- Primary structure – sequence
- Secondary structure – structure motifs
- Tertiary structure – 3D position of atoms
- Quaternary structure – docking of proteins





Protein structure data (PDB format)

```
ATOM      1  N   MET A   1      20.020  28.662  42.801  1.00 51.80  N
ATOM      2  CA  MET A   1      20.598  29.950  42.438  1.00 52.13  C
ATOM      3  C   MET A   1      22.118  29.937  42.576  1.00 47.63  C
ATOM      4  O   MET A   1      22.660  29.623  43.636  1.00 49.97  O
ATOM      5  CB  MET A   1      20.009  31.073  43.293  1.00 51.36  C
ATOM      6  CG  MET A   1      20.331  32.468  42.765  1.00 51.13  C
ATOM      7  SD  MET A   1      21.406  33.373  43.921  1.00103.49  S
ATOM      8  CE  MET A   1      21.129  32.396  45.410  1.00 55.43  C
ATOM      9  N   LEU A   2      22.799  30.285  41.490  1.00 41.99  N
ATOM     10  CA  LEU A   2      24.249  30.178  41.424  1.00 37.25  C
```

RECAP

- DNA is a string formed with letters A, C, T, G (called nucleotides or bases)
- DNA is double-stranded – allows replication: transfer of genetic “code” from parents to offspring
- DNA is naturally oriented from 5’ to 3’ and the two strands are anti-parallel
- If you know the sequence of one strand, you can obtain the sequence of the other by reverse-complementation

```
5′ AGACCTAGTGCACGGCTACTACC 3′

5′ CCATCATCGGCACGTGATCCAGA 3′   Reverse

5′ GGTAGTAGCCGTGCACTAGGTCT 3′   Complement
```

RECAP

- Central Dogma of molecular biology:
  - DNA – RNA (transcription)
  - RNA – Protein (translation)
- The transcribed segments of DNA are called “genes”
- Translation occurs in sets of 3 nucleotides – codons
- Each codon encodes one of 20 amino-acids and 3 stop-codons
- In eukaryotes the genes may be split into multiple exons, separated by introns: DNA segments that will not get translated
- The protein is translated from an RNA representing the concatenation of the exons of the gene

The “new” biology

- DNA is not the only heritable information
  - Epigenetic information: RNA molecules, DNA methylation patterns (affects coiling on DNA on histones)
- Complex regulation patterns
  - Genes turn on other genes
  - Genes inhibit other genes
  - RNA interference – small RNA molecules can destroy specific transcripts (down-regulate production)

Playing with DNA

Biologists can:

- Cut the DNA – restriction enzymes (often palindromes) (Nobel prize – Arber, Nathans, Smith)

5'GAATTC
3'CTTAAG

5'---G
3'---CTTAA

AATTC---3'
G---5'

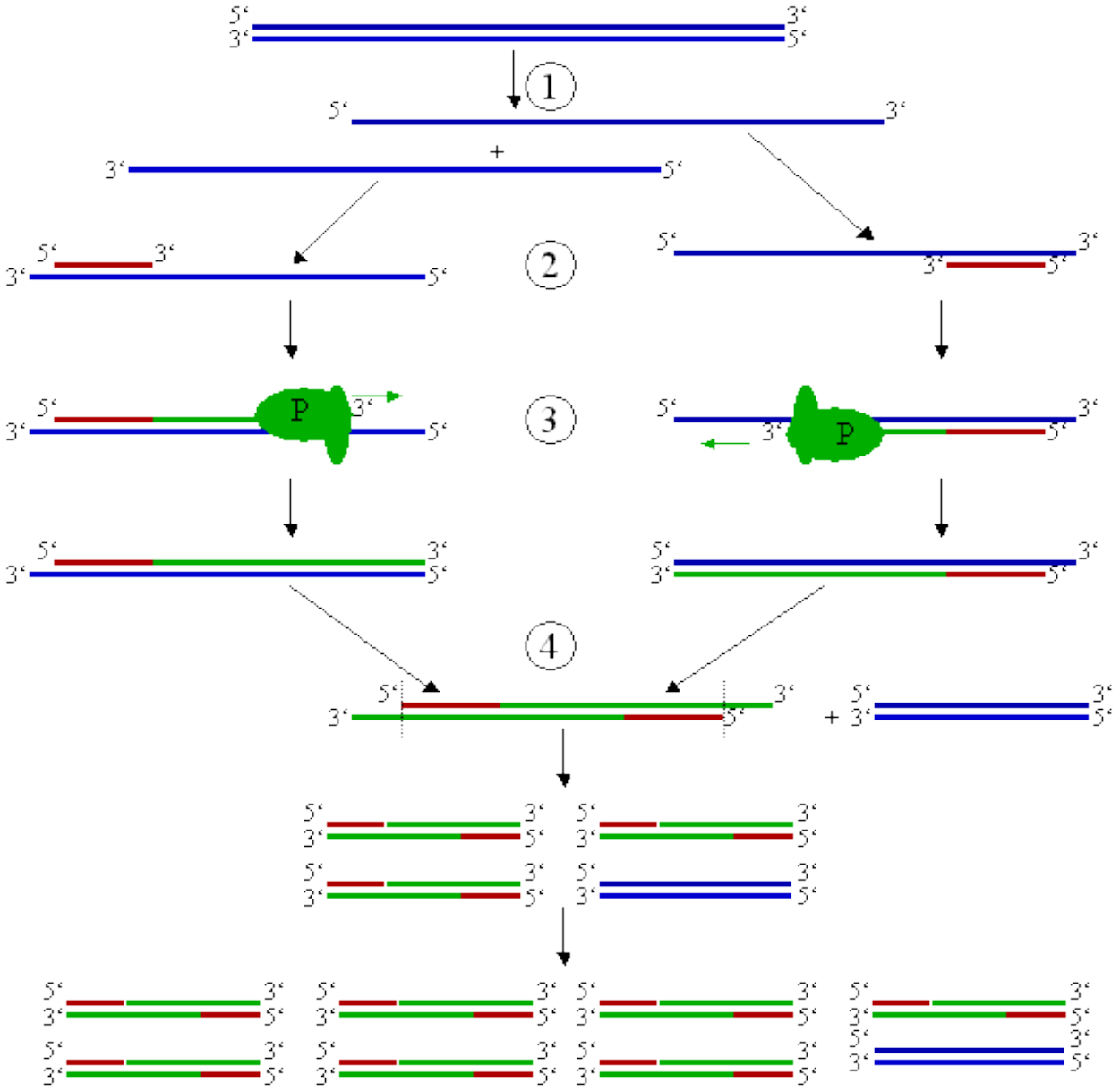
- Attach "things" to DNA (either single or double-strand)

TAGGCACGTTGCAACTACGGC

TGCAACGT

- "Amplify" DNA – Polymerase Chain Reaction (Nobel prize – Mullis)

Polymerase chain reaction (PCR)



1. Denature

2. Anneal (attach primer)

3. Extend

4. Repeat

How does PCR work?

- 1. Start: 1 double-stranded molecule
- 1. Denature: 2 single-stranded molecules
- 1. Anneal: 2 single-stranded molecules with primers attached
- 1. Extend: 2 double-stranded molecules – one “long” (L) strand and one “short” (S) (terminated at a primer)

- 2. Start: 2 double-stranded molecules: L+S, L+S
- 2. Denature: 2 x L strands, 2 x S strands
- 2. Anneal: all strands with primers attached
- 2. Extend: 2 double-stranded molecules: L+S, L+S, 2 double-stranded molecules: S+SS, S+SS
  SS – strand terminated at both ends with a primer

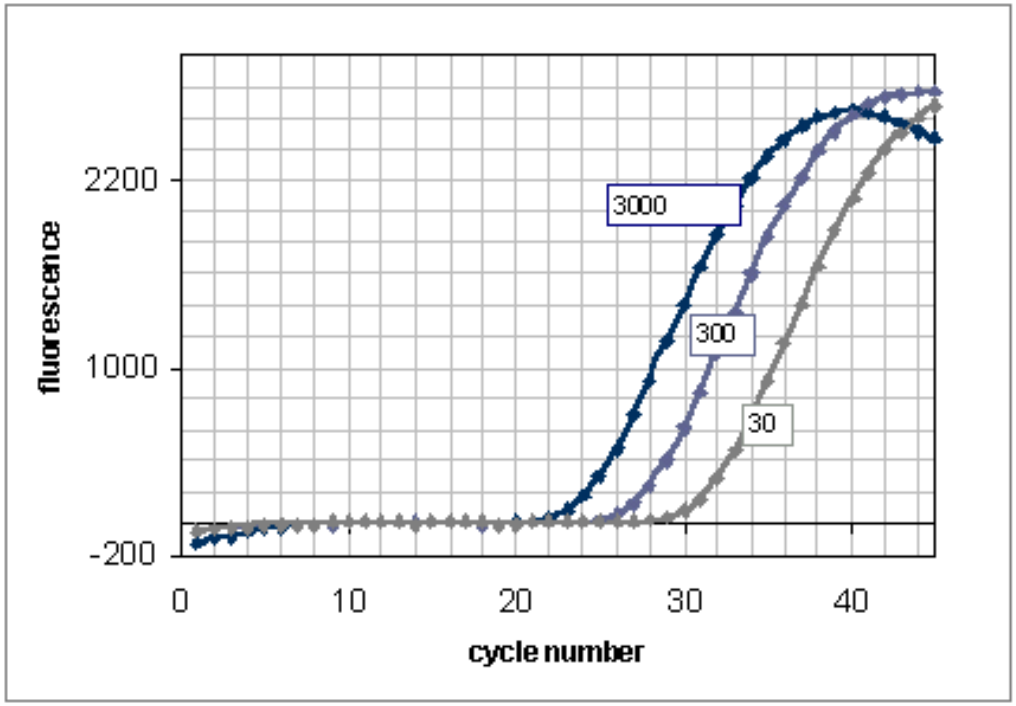
PCR Recurrences

- L_n, S_n, SS_n - # of strands of each type at cycle n
- $L_n = L_{n-1} = 2$
- $S_n = S_{n-1} + L_{n-1} = S_{n-1} + 2 = 2 * (n - 1) = O(n)$
- $SS_n = S_{n-1} + 2 * SS_{n-1} = O(2^n)$

- The sequence between the primers (SS) is amplified exponentially – will quickly overtake the solution

Quantitative PCR

- Measure # of PCR cycles needed to reach a certain concentration of DNA – depends on initial # of molecules
- Used in diagnostics: e.g. is this a random Anthrax spore from the environment or lots of spores from an attack



http://www.dxsgenotyping.com/technology_main.htm

DNA sequencing

- Most techniques “trick” the polymerase into revealing the sequence
- The traditional method – Sanger sequencing – based on “terminator” bases – prevent the polymerase from extending the DNA
- Sanger sequencing is essentially PCR + terminator bases
- Other methods “spy” on the polymerase as it incorporates nucleotides

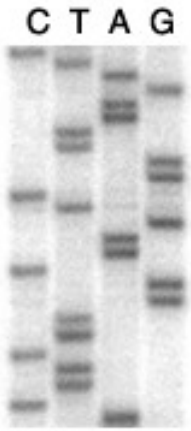
Sanger sequencing

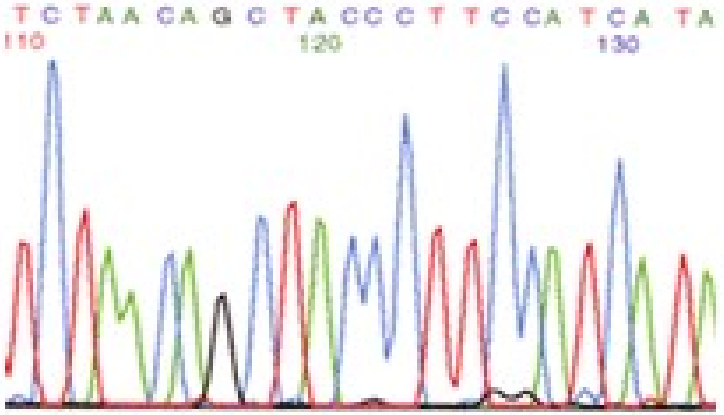
Sanger, F, Coulson AR. *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* J.Mol.Biol. 94 (1975)

```
      A   A
  C   G     G
G   T
      G
```

```
TCTAATAGA
AGATTATCTAACAGCTACCCTTCCATCA
```

```
TCTAATT
TCTAATTA
TCTAATTAG
TCTAATTAGA
TCTAATTAGAT
```



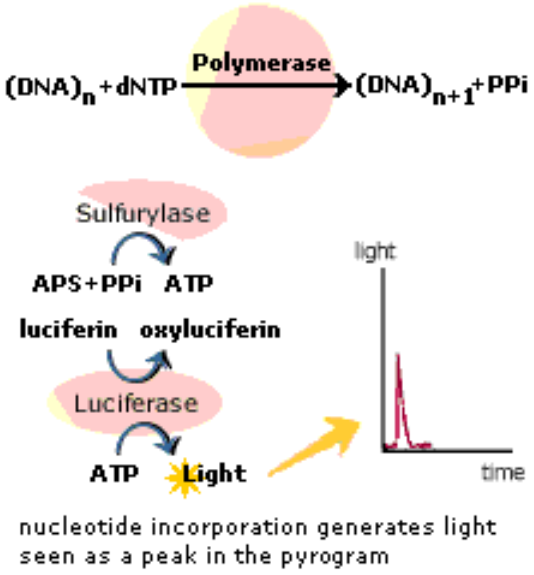


pictures from http://www.uvm.edu/~cgep/Education/Sequence.html

The future of sequencing

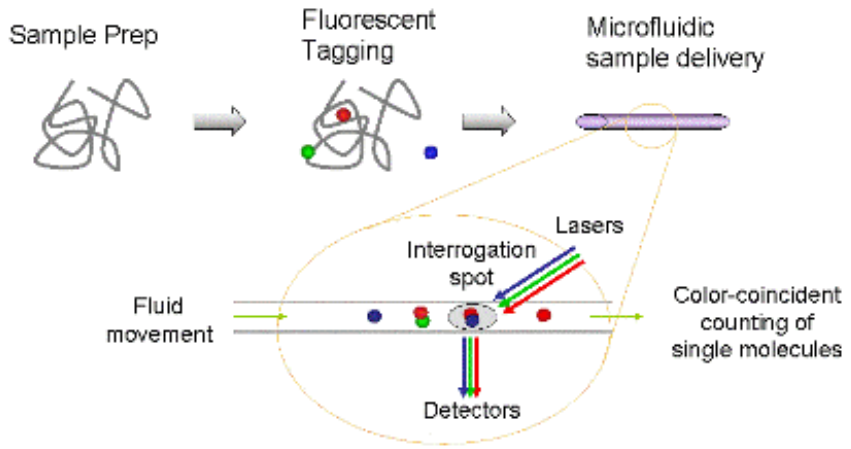
- Single molecule sequencing - current technology requires many copies of DNA being sequenced - requires DNA amplification
- Massively-parallel sequencing - 100k sequencing reactions occuring at the same time

Sequencing by synthesis



TCTAATAGA
AGATTATCTAACAGCTACCCTTCCATCA

Micro-fluidics

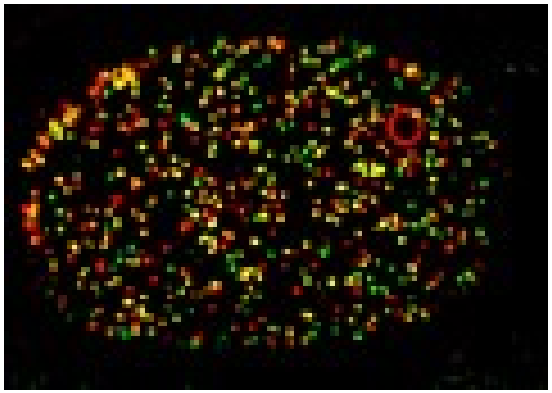


http://www.genetics.ucla.edu/sequencing/pyro.php

http://www.usgenomics.com

The future of sequencing

Massively parallel sequencing



http://arep.med.harvard.edu/

- each spot is a molecule or amplified from one molecule
- image processing used to track molecules during sequencing by synthesis
- often micro-fluidics/lab-on-a-chip used

The evolution of DNA sequencing

Since	Technology	Read length	Throughput/run	Throughput/ hour	cost/run
1977-	Sanger sequencing	> 1000bp	4hr 400-500 kbp	100 kbp	$200
2005-	454 pyrosequencing	250-400bp	4hr 100-500 Mbp	25-100 Mbp	$13,000
2006-	Illumina/Solexa	50-100bp	3 days 2-3 Gbp	25-40 Mbp	$3,000
2007-	ABI SOLiD	35-50bp	3 days 6-20 Gbp	75-250 Mbp	est. $3-5,000
2008-	Helicos **single molecule**	25-50 bp	8 days 10 Gbp	~50 Mbp est. 1Gbp/hour	~$18,000
TBA (2010)	Pacific Biosciences **single molecule**	100-200 kbp	?	?	?