CMSC858P: Algorithms for Biosequence Analysis
Lecture 1

Instructor: Mihai Pop

TuTh: 2-3:15pm, CSIC 3118

INTRODUCTIONS

- Instructor: Mihai Pop (mpop at umiacs.umd.edu)
  Office hours: Wednesdays 11-12, AVW 3223
- You
- Class webpage:
  http://www.cbcb.umd.edu/confcour/CMSC858P.shtml

Overview of course

- No knowledge of biology required
- String algorithms – relevant not only to biology
- Exact and inexact matching
- Exact and heuristic approaches
- Multiple sequence alignment
- Phylogenetics
- RNA and protein structure

Policies

- Attendance - follow University policy
  - you must claim excused absences in writing
  - written documentation of illness is required (from Dr. not yourselves)
  - if possible inform me prior to the class you will skip
- Disabilities
  - must inform me during the first 2 weeks of the semester if special accommodations necessary
  - request letter from Office of Disability Support Services
- General – communication is key
  - talk to me about any issues whether covered or not by University policies

Grading & workload

- Homework (10%)
- Goal: 5-10 assignments
  - simple
  - small programming assignments
  - “discovery” exercises (find something in public databases or using public software)
- Programming projects (15% + 15%)
  - Project 1 – assigned by instructor (suffix tree)
  - Project 2 – chosen by student
- In-class midterm (25%) & final (35%)
- Late policy: 1 day late – 10 points off; 2 days late – 20 points off; 3 days late – 0 points

Academic Honesty

http://www.studenthonorcouncil.umd.edu/code.html

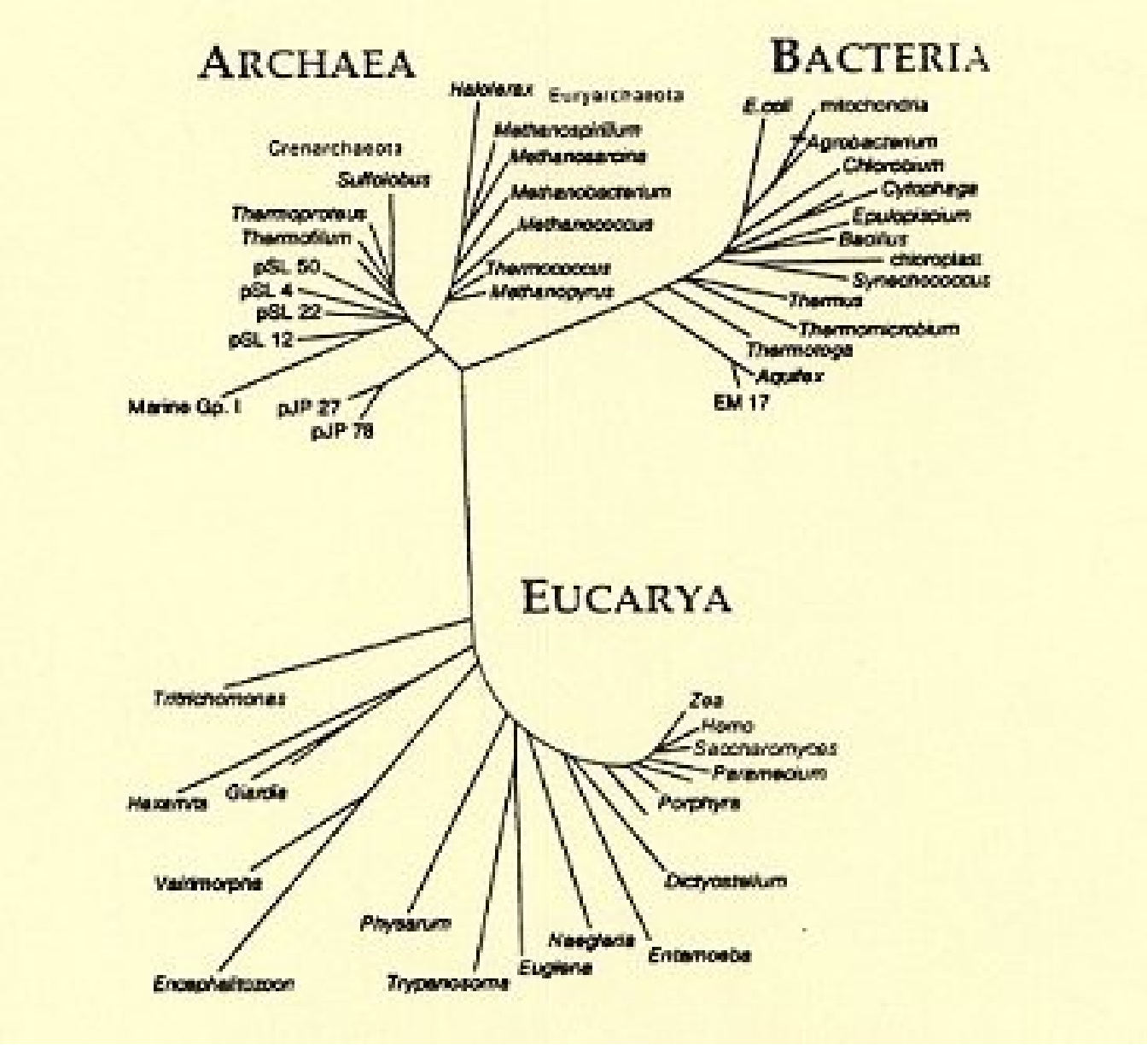
- No cheating on homeworks/projects/exams
- No making up data/results
- No copying of other people's code
- You can work together on homeworks/projects but WRITE THE ANSWER BY YOURSELF

I pledge on my honor that I have not given or received any unauthorized assistance on this examination.

Advice: how to do well in the class

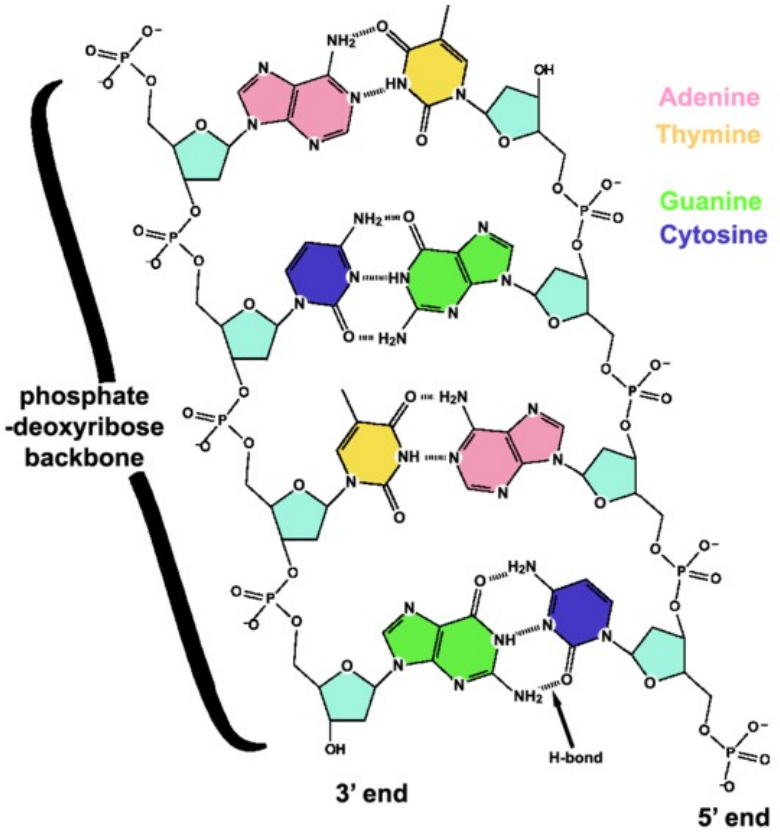
- Start early on assignments – at least read the assignment after class
- Ask questions – during class, exams, office hours, using email (I'm available most time by email)
- Be inquisitive – follow up on topics discussed in class: Google, Wikipedia
- Be social – get to know some biologists – learn what they do, what they are interested in
- Get to know your colleagues

The tree of life



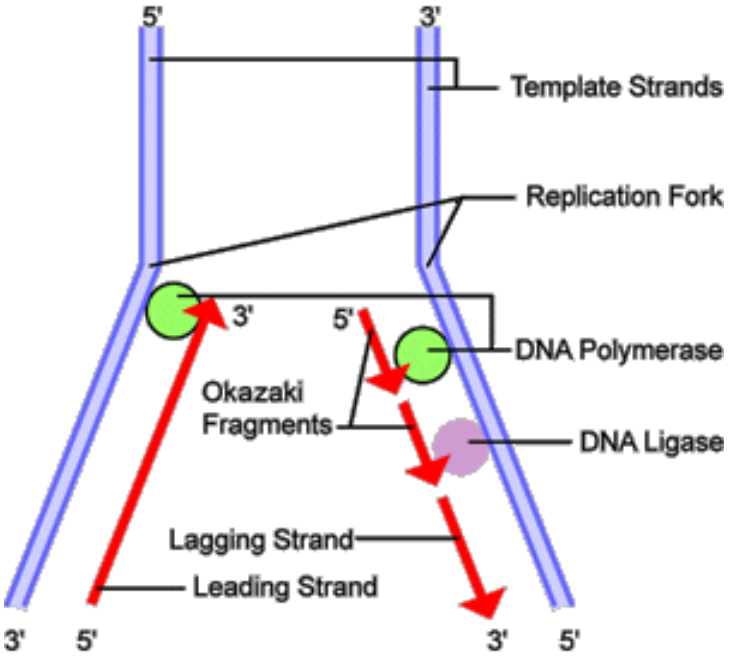
http://www.fossilmuseum.net/Tree_of_Life/Domains_Archaea_Bacteria/

DNA – the code of life



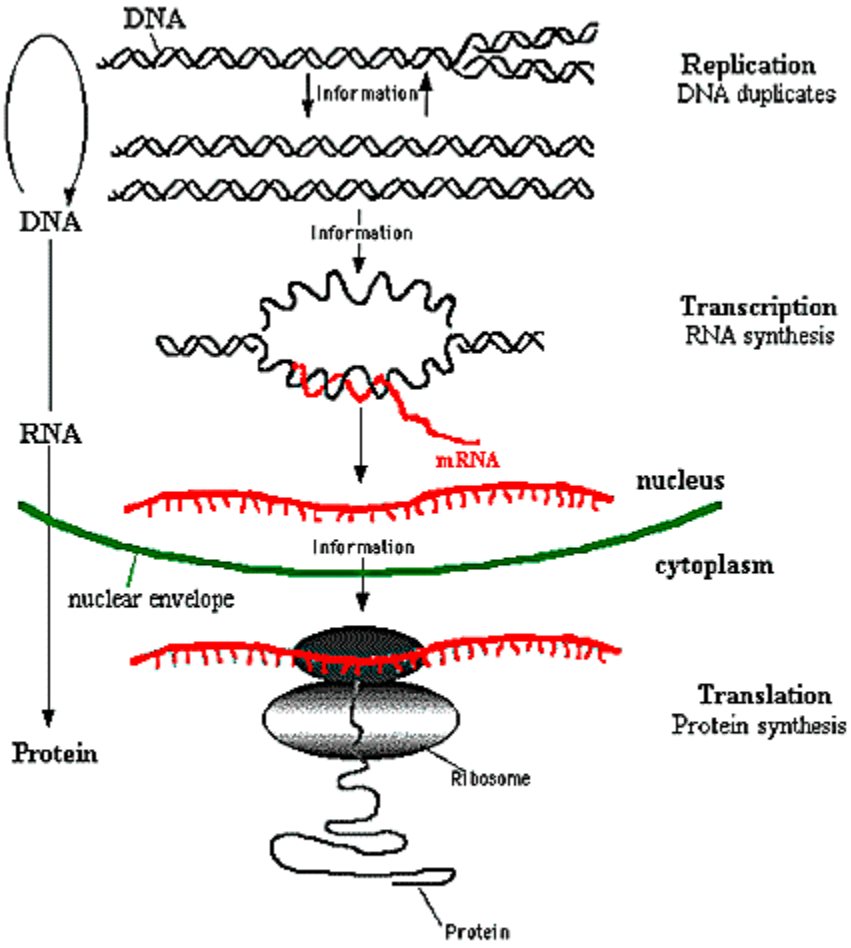
- Purines A, G, caffeine
- Pyrimidines C, T
- Sugar backbone (ticker tape)

- Double-stranded – allows replication



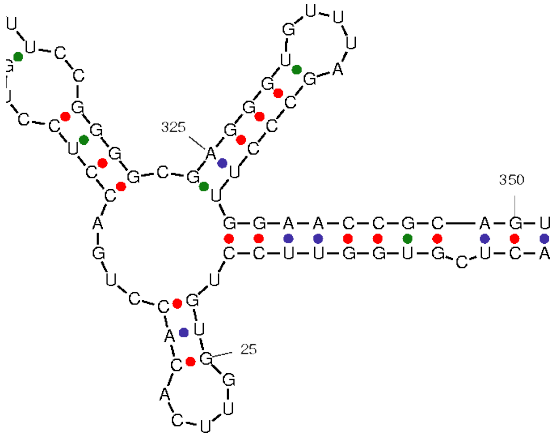
pictures from wikipedia

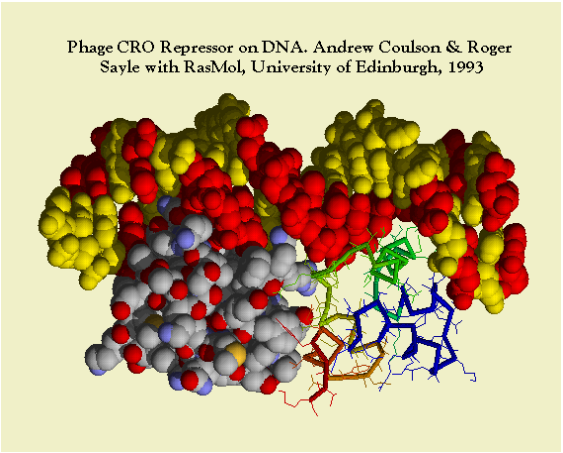
Central dogma



The Central Dogma of Molecular Biology

AGGTACGCGTACCTGACAGG





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

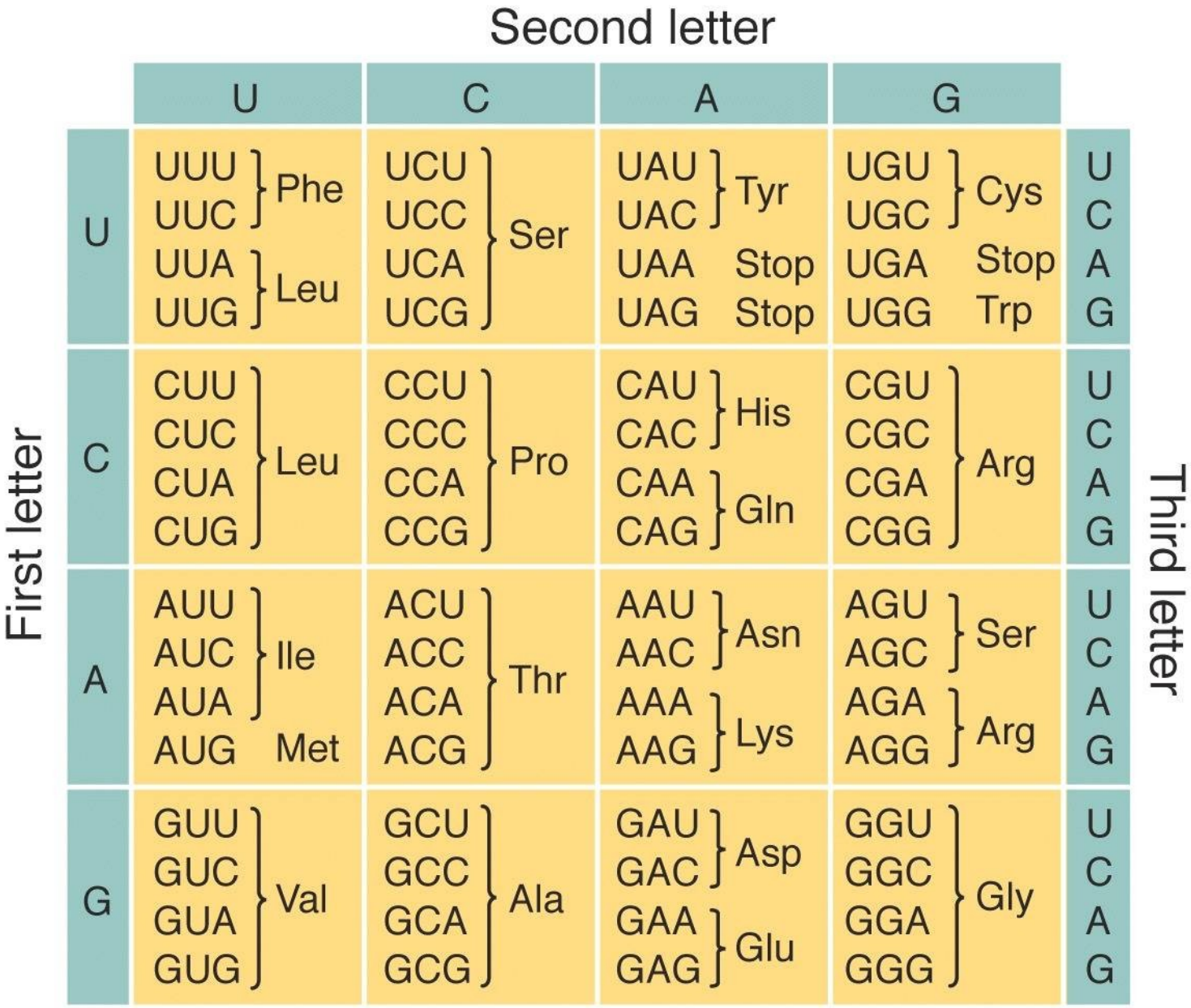
Genes, transcription, translation

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

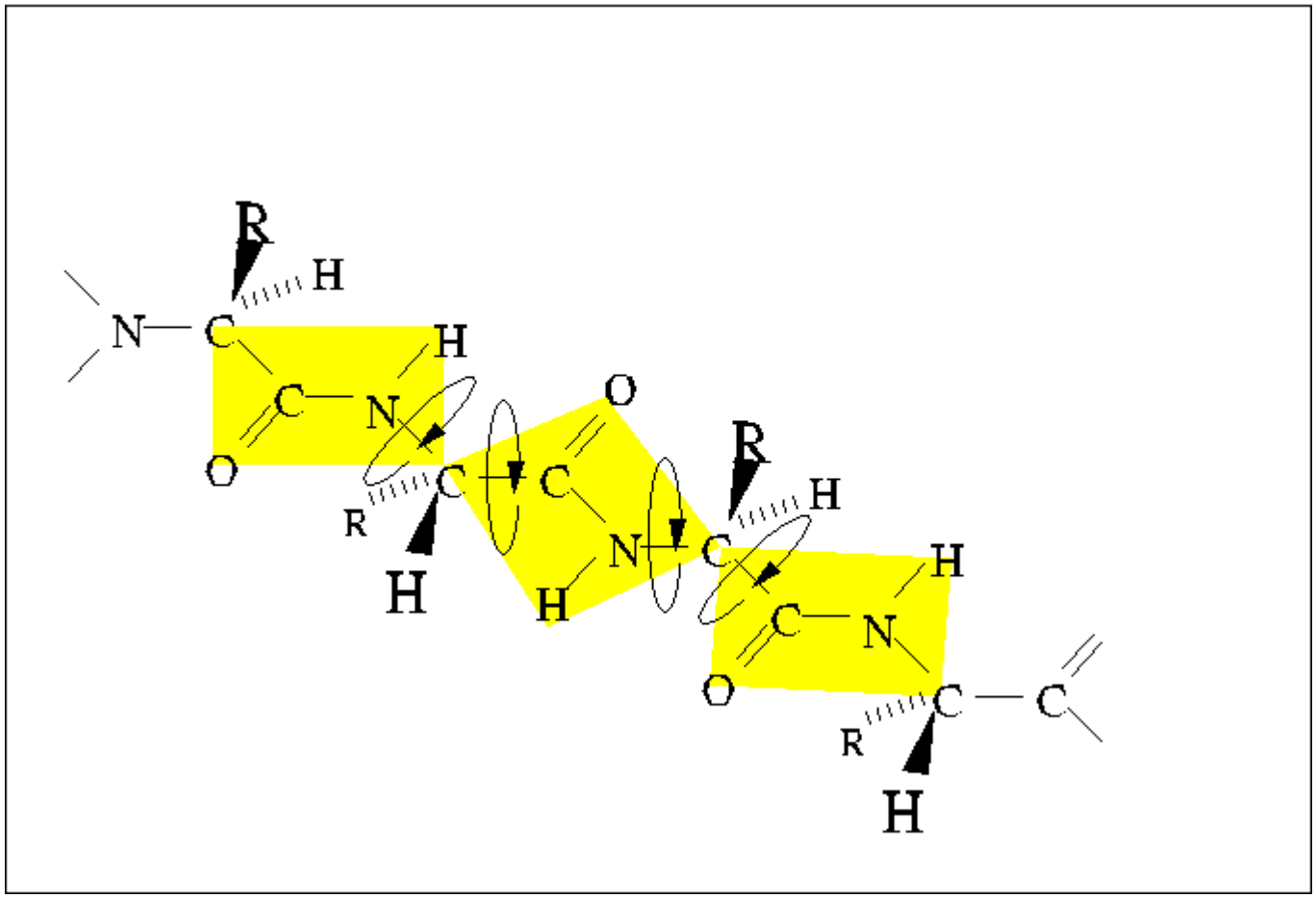
ACCGUACC**AUG**UUA...AUAGGC**UGA**GCA

- 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

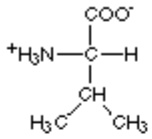


Protein structure

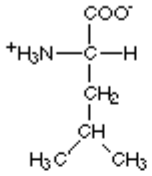


http://www.tulane.edu/~biochem/med/second.htm

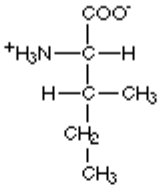
Amino acids with hydrophobic side groups



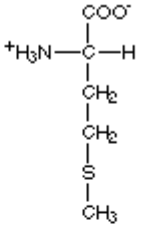
Valine
(val)



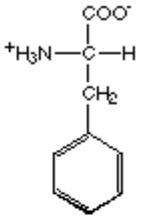
Leucine
(leu)



Isoleucine
(ile)



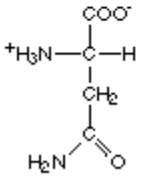
Methionine
(met)



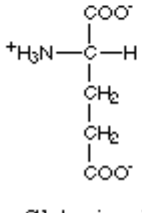
Phenylalanine
(phe)

hate water

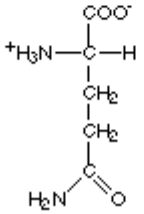
Amino acids with hydrophilic side groups



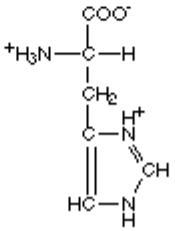
Asparagine
(asn)



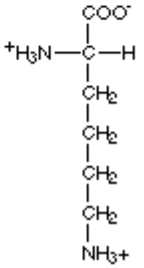
Glutamic acid
(glu)



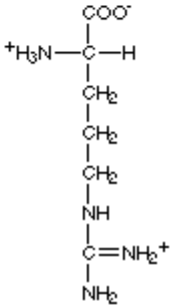
Glutamine
(gln)



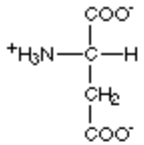
Histidine
(his)



Lysine
(lys)



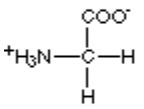
Arginine
(arg)



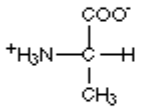
Aspartic acid
(asp)

like water

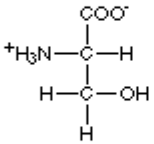
Amino acids that are in between



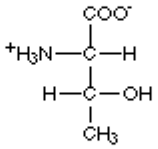
Glycine
(gly)



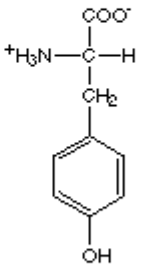
Alanine
(ala)



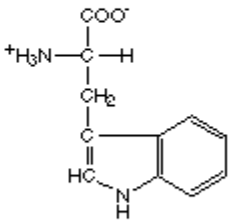
Serine
(ser)



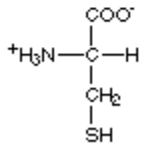
Threonine
(thr)



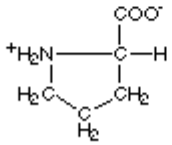
Tyrosine
(tyr)



Tryptophan
(trp)



Cysteine
(cys)

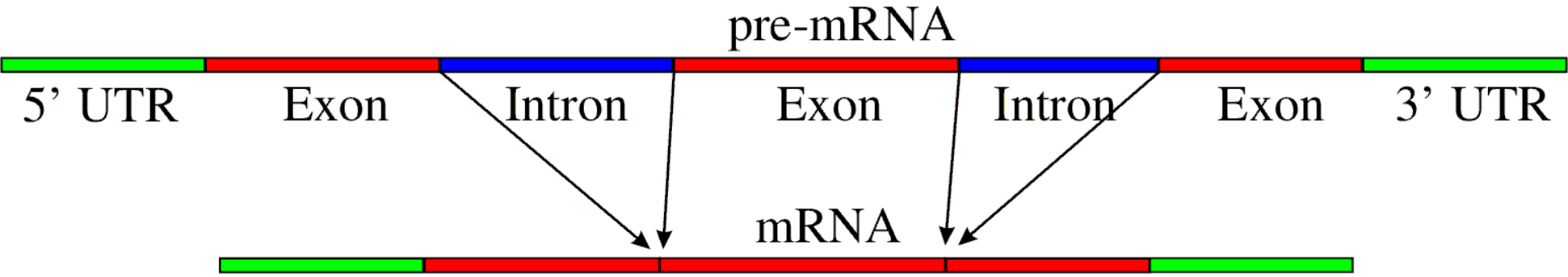


Proline
(pro)

can't decide

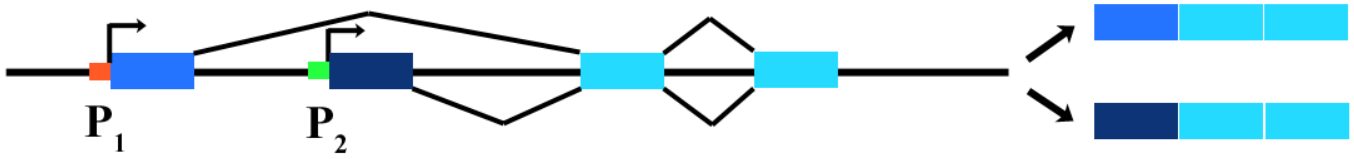
http://web.mit.edu/esgbio/www/lm/proteins/aa/aminoacids.html

Translation – complications

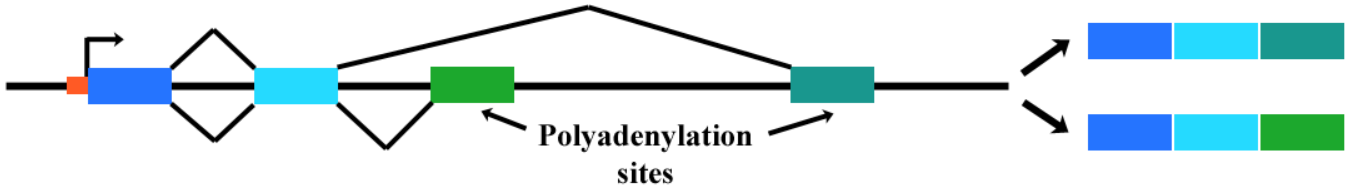


Alternative splicing examples

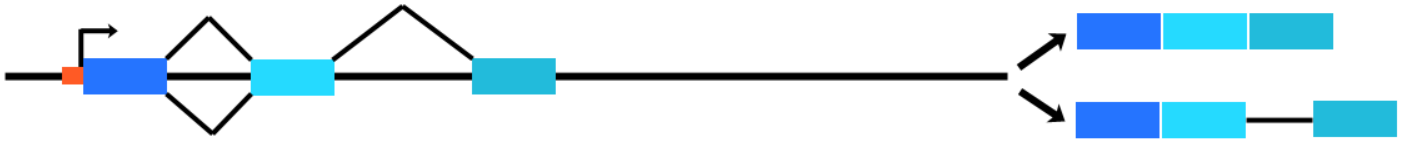
(a) Alternative selection of promoters (e.g., *myosin* primary transcript)



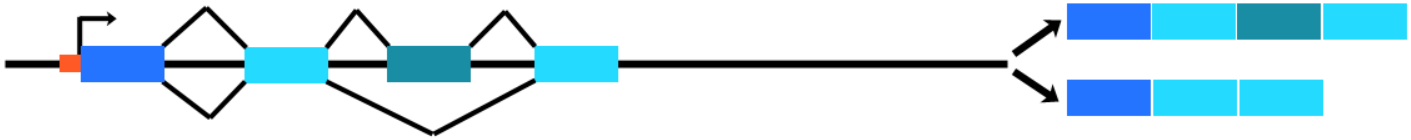
(b) Alternative selection of cleavage/polyadenylation sites (e.g., tropo*myosin* transcript)



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



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



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 many eukaryotes the genes are split into multiple exons, separated by introns: DNA segments that will not get translated
- The protein corresponding to a gene is translated from an RNA representing the concatenation of the exons of the gene

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'

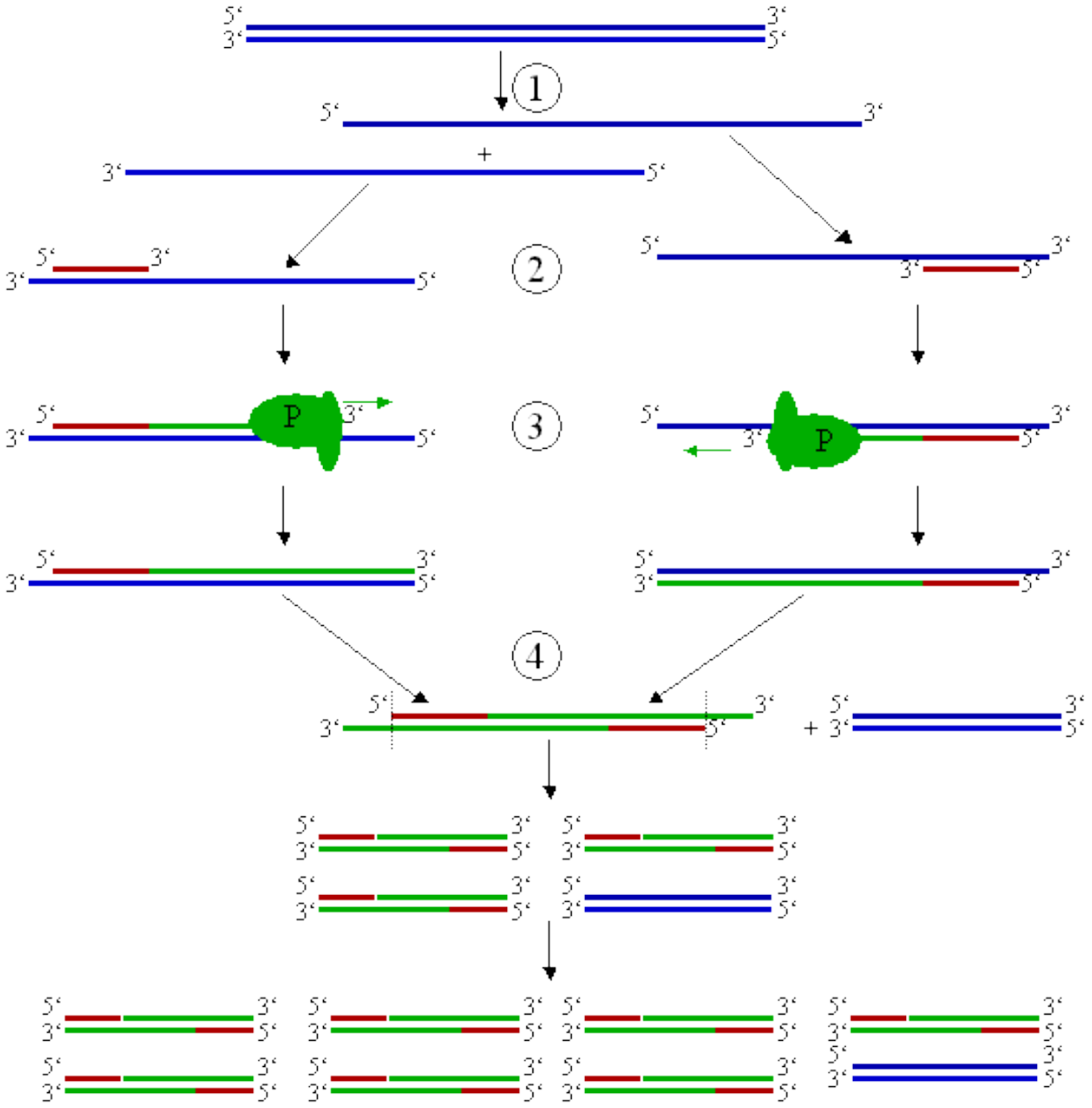
- Attach "things" to DNA (either single or double-strand)

TAGGCACGTTGCAACTACGGC

TGCAACGT

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

Polymerase chain reaction (PCR)



1. Denature

2. Anneal (attach primer)

3. Extend

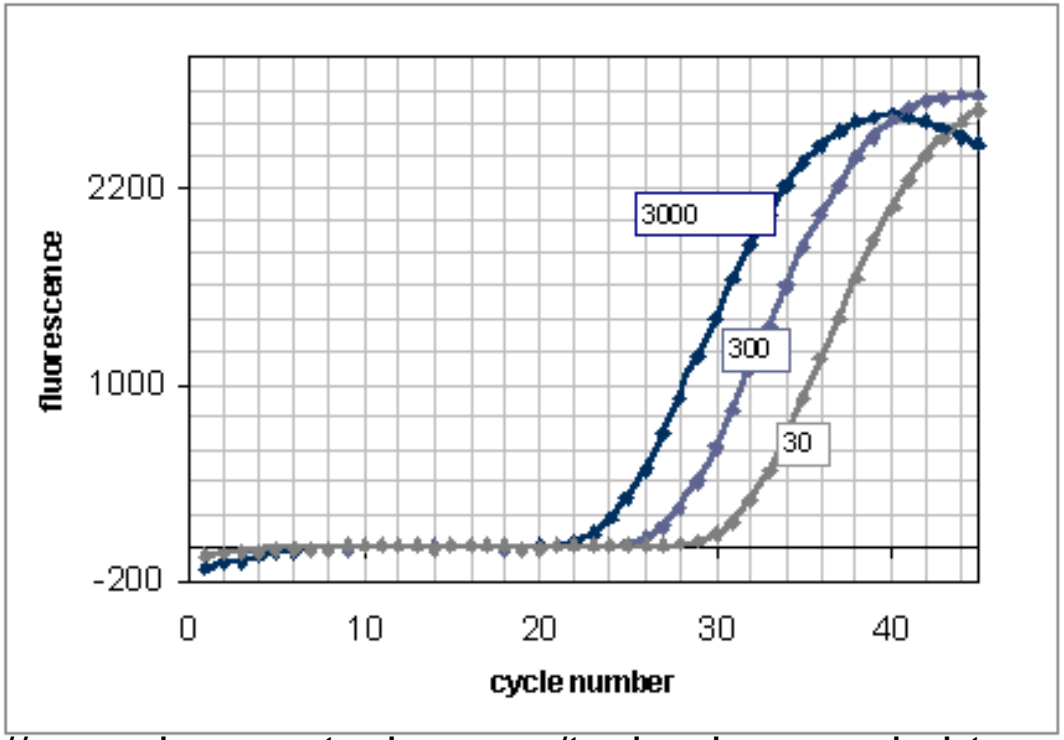
4. Repeat

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

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

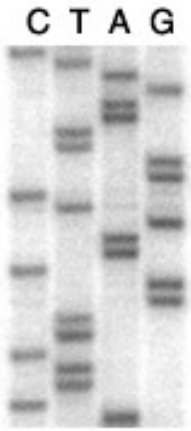
Sanger sequencing

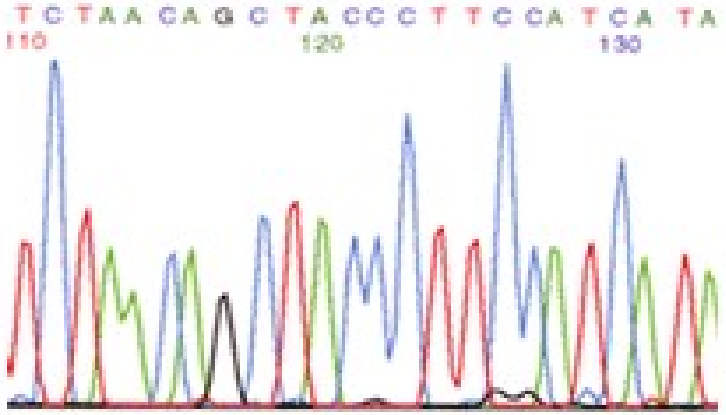
Sanger, F, Coulson AR. *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* J.Mol.Biol. 94 (1975)

```
      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

- Roche/454 Life Sci. – approx. 60-100 Mbp, 250 bp reads / 4 hr
- Illumina/Solexa – approx. 1-2 Gbp, 30-40 bp reads / 3 day run
- Applied Biosystems/SOLiD – approx 1 Gbp, 25-35 bp reads
- Helicos – single molecule sequencing ~ 1Gbp/hour, 30-40 bp

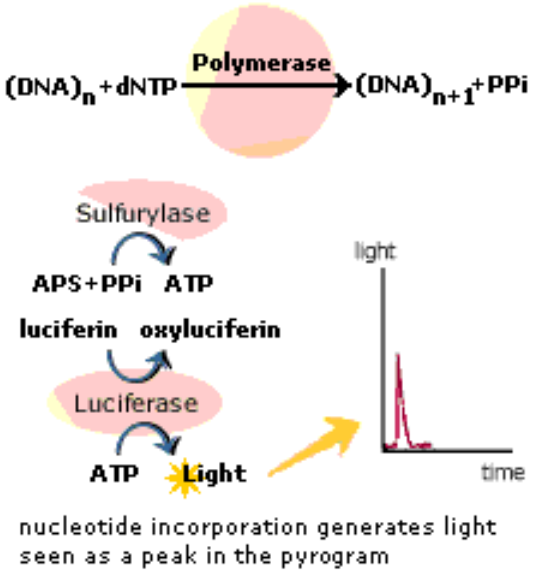
Not yet available:

- nanopore sequencing

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

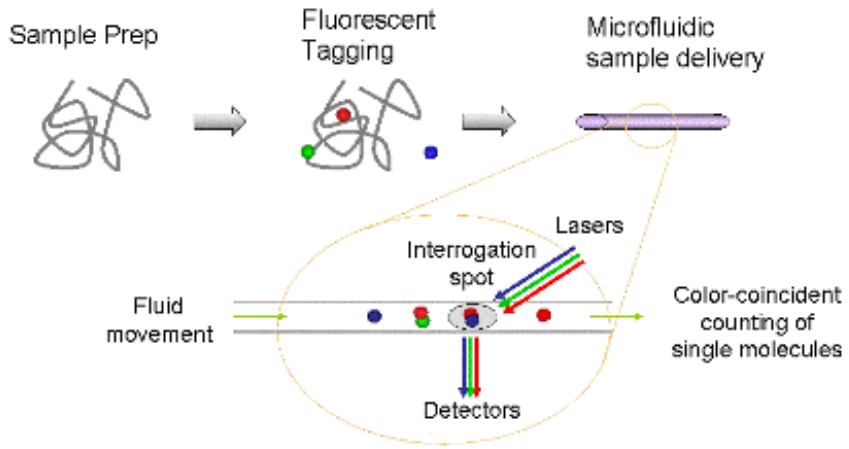
Sequencing by synthesis



TCTAATAGA
AGATTATCTAACAGCTACCCTTCCATCA

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

Micro-fluidics

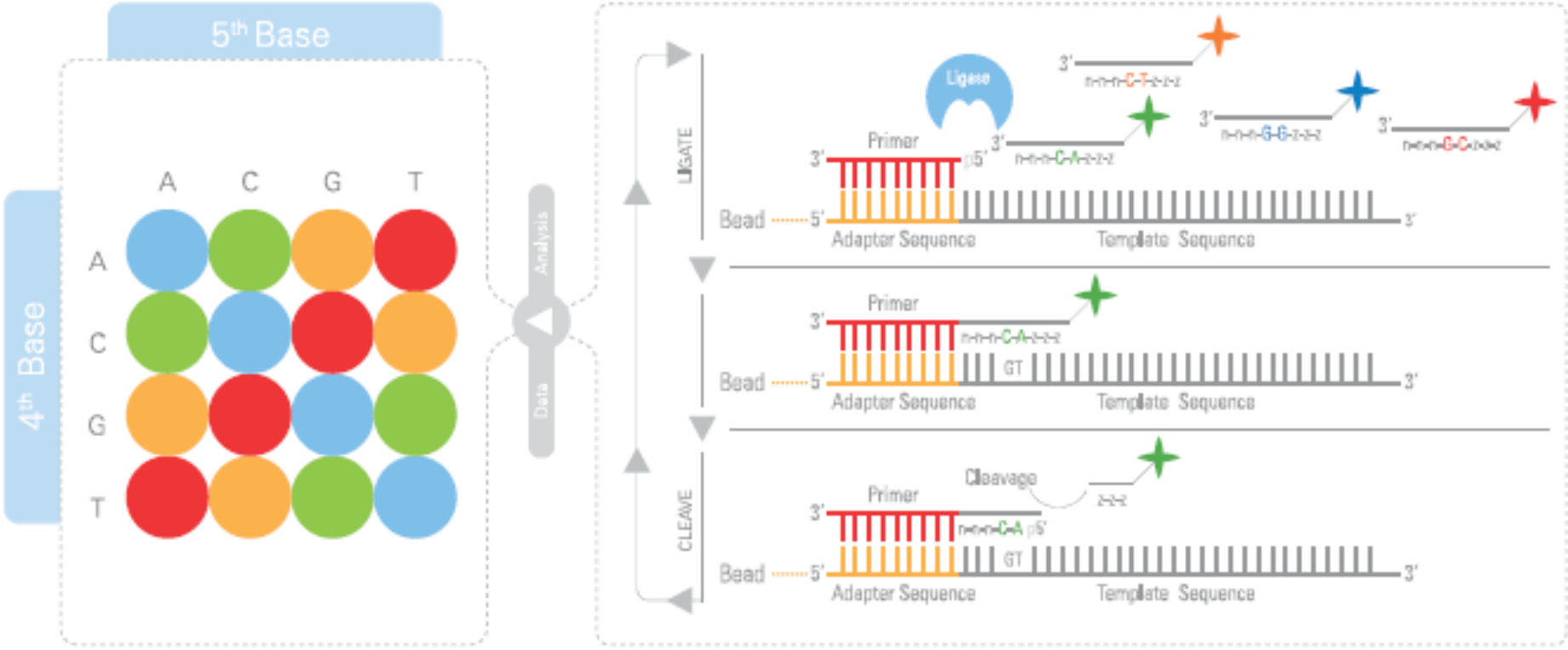


http://www.usgenomics.com

How they work

- Amplify DNA
  - Roche/454 – emulsion PCR on beads (water droplets in oil)
  - Illumina/Solexa – PCR on surface
  - ABI SOLiD – emulsion PCR
- Sequence
  - Roche/454 – pyrosequencing
  - Illumina/Solexa – reversible terminators
  - ABI SOLiD – sequencing by ligation two-color encoding

ABI SOLiD



http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/PDF/SOLiD_Brochure_092707.pdf