Molecular Biology Fundamentals

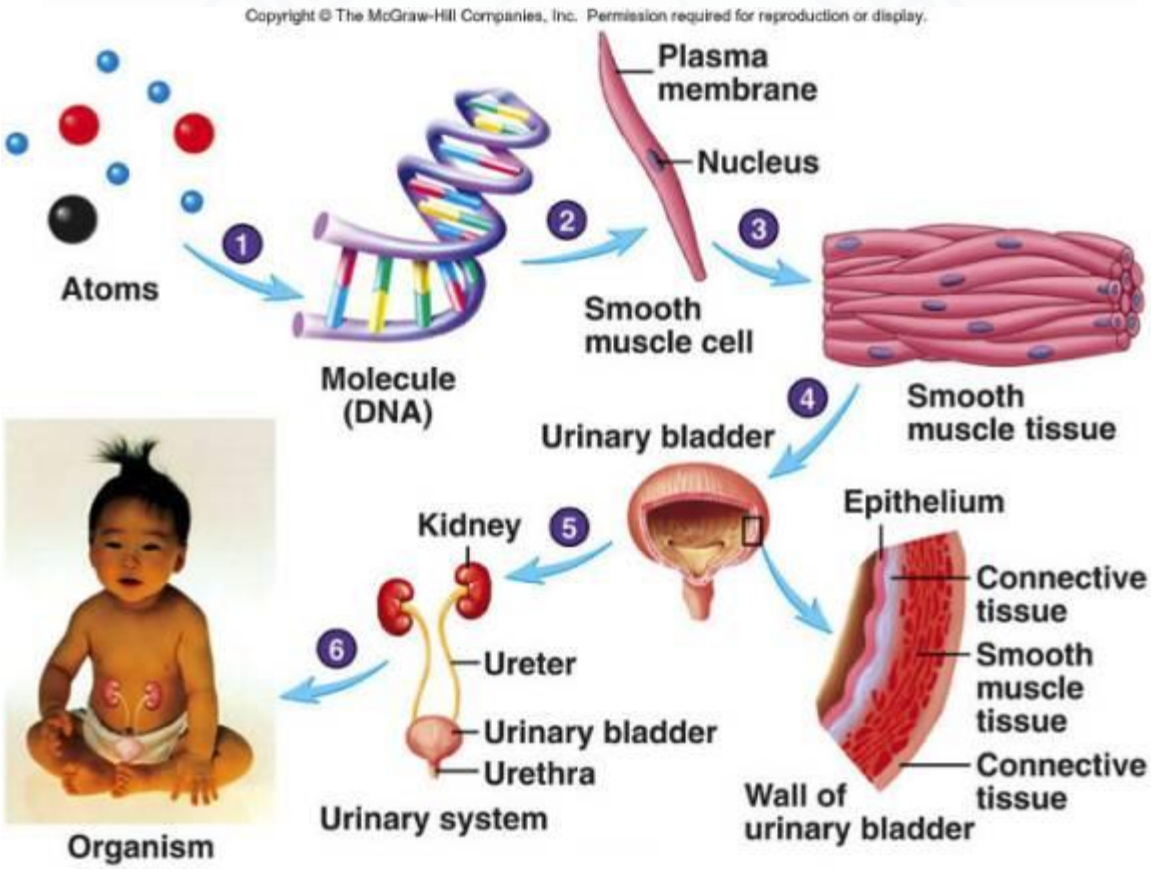
Dr. Fayyaz ul Amir Afsar Minhas
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

Life

- Life is a characteristic distinguishing physical entities having signaling and self-sustaining processes from those that do not
- Organism is the smallest contiguous unit of life
- Biology is the branch of science that studies life

D. E. Koshland, "The Seven Pillars of Life," *Science*, vol. 295, no. 5563, pp. 2215–2216, Mar. 2002.

What are we made of?

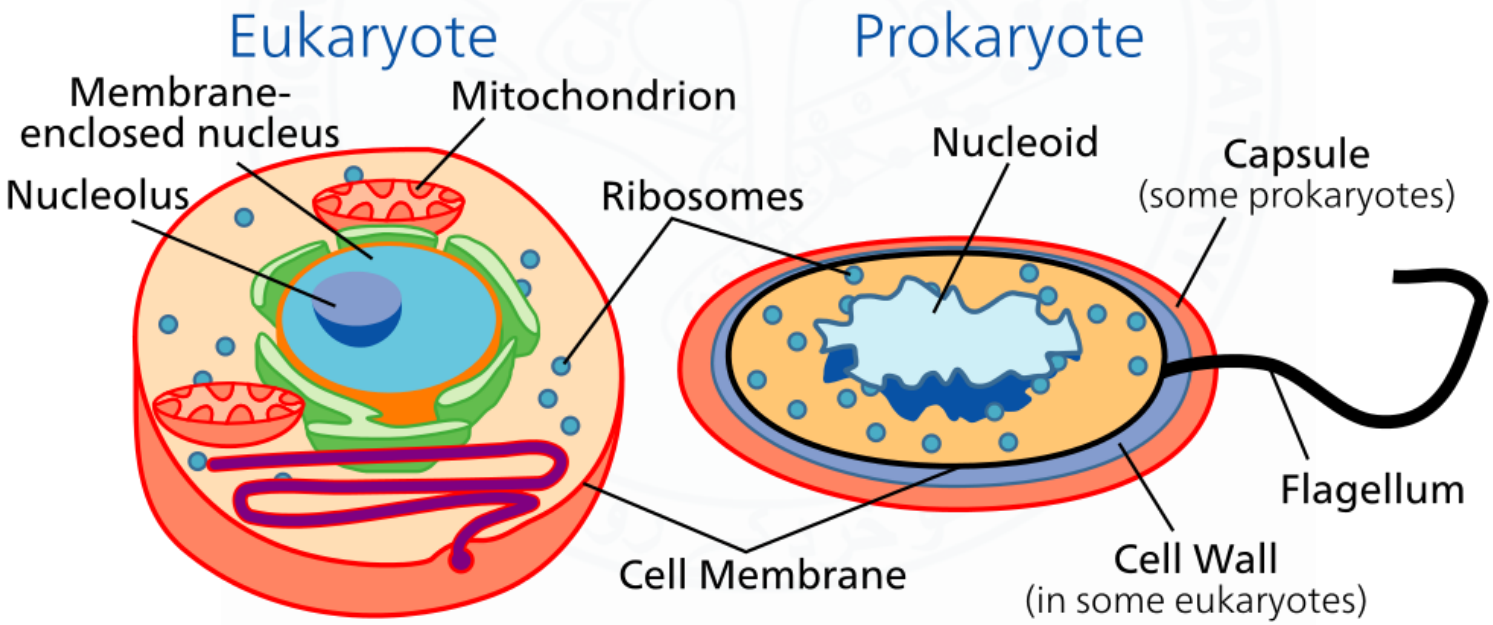


The Cell

- Size: 1-100 microns
- Number: 100 Trillion in human
- Types:

With nucleus: Plants, Animals, Fungi

Without nucleus: Bacteria & Archaea



Read 'Cell' in Wikipedia

Differences

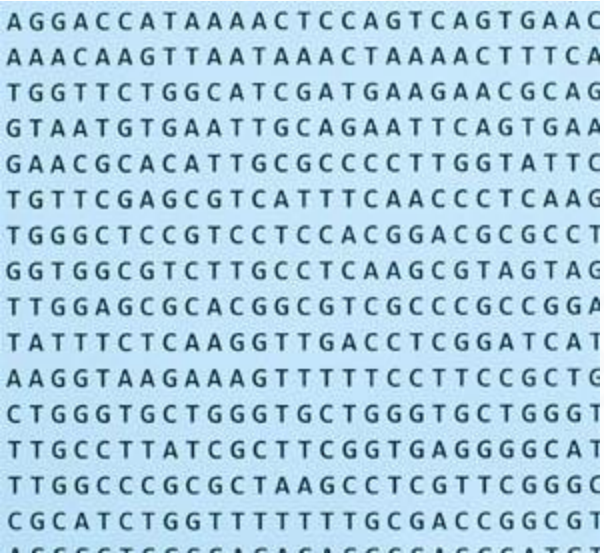
Prokaryotes	Eukaryotes
Single cell	Single or multi cell
No nucleus	Nucleus
No organelles	Organelles
One piece of circular DNA	Chromosomes
No mRNA post transcriptional modification	Exons/Introns splicing

Cell: Bio-molecular Composition

	Prokaryotes	Eukaryotes
Water	70%	70%
DNA	0.25%	1%
RNA	1%	6%
Proteins	18%	15%
Lipids (fat)	5%	2%
Carbohydrates	2%	2%
Metabolites	4%	4%

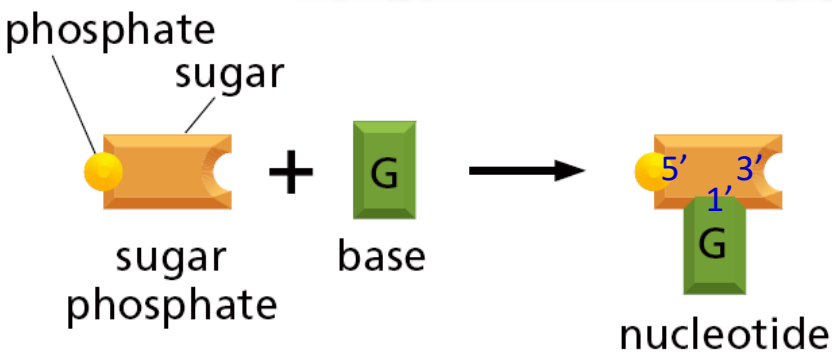
Genome: The 'Program'

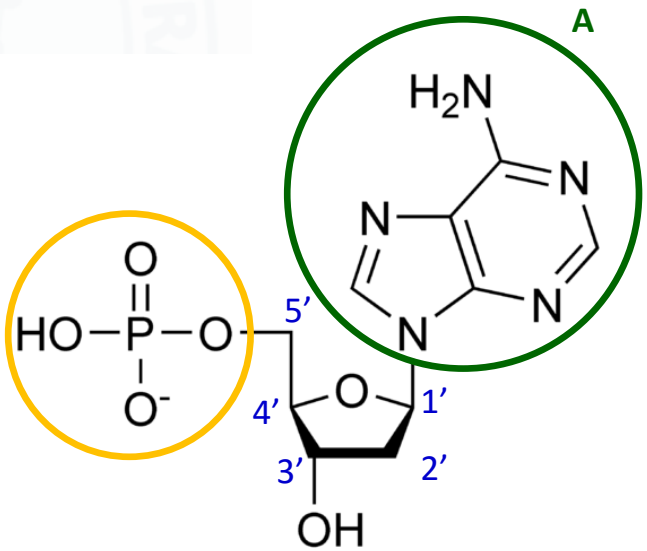
- Genome is the genetic material of an organism
- Deoxyribonucleic acid (DNA)
  - Encodes these genetic instructions



Building blocks of DNA

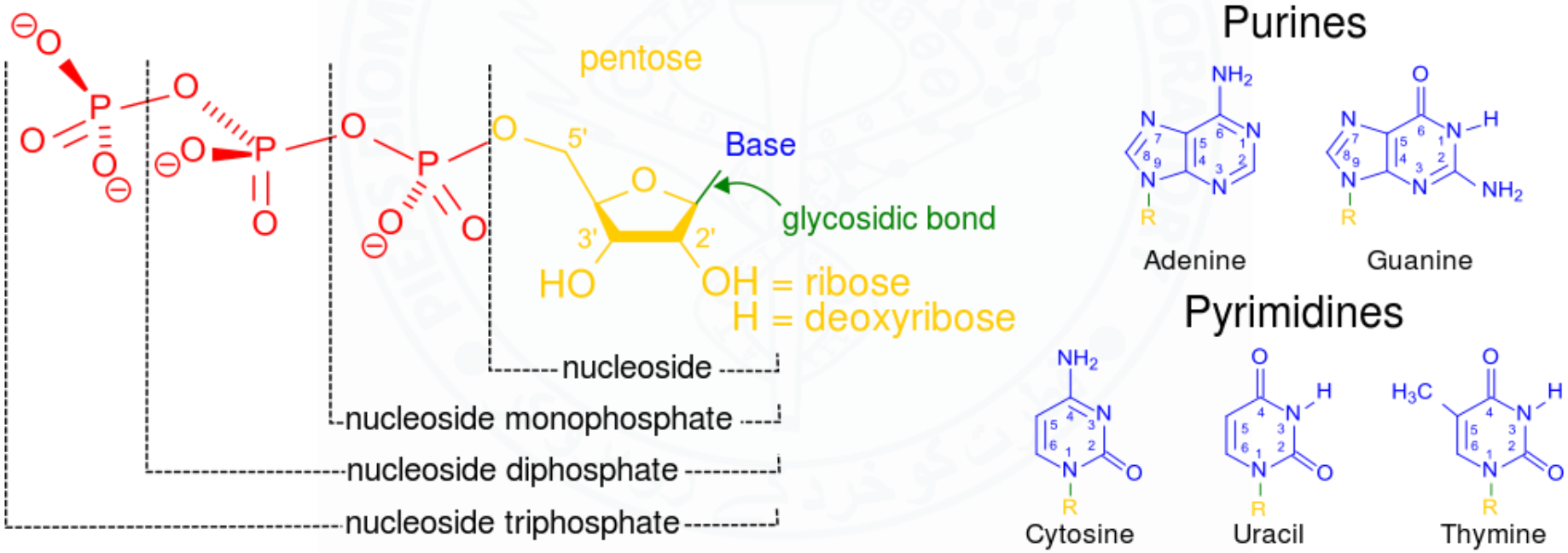
- Bases also known as nucleobases form the basic building block of DNA
  - 4 types: A, T, G, C
- These bases form <u>nucleotides</u>





The instruction set: Bases

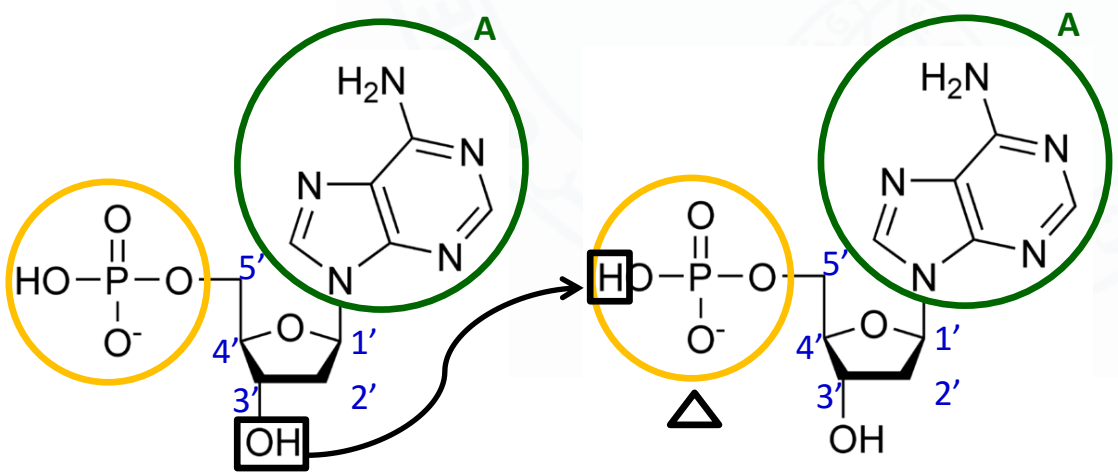
- Nucleobase, Nucleoside & Nucleotide

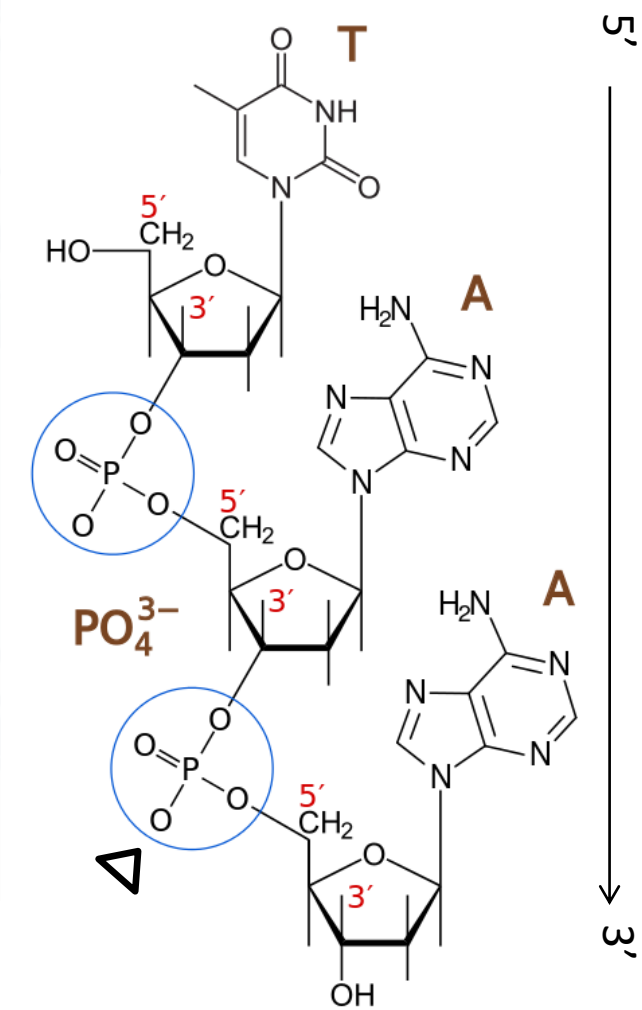


RNA

Building blocks of DNA: Direction

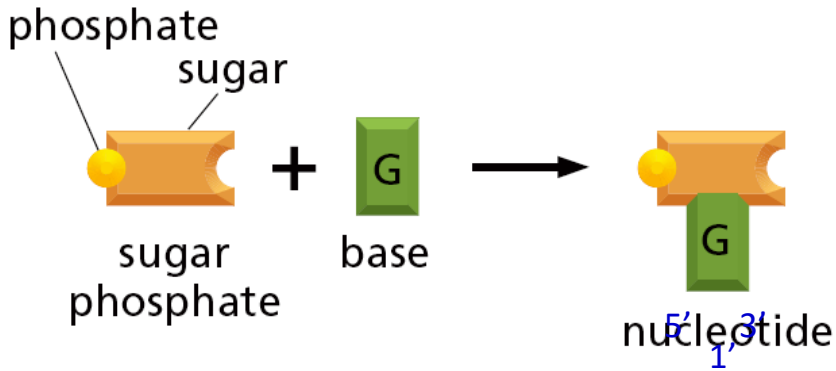
- DNA is a polynucleotide
  - A nucleotide is ligated (joined) to another by covalent bonds as shown below
    - 3’ OH connected to phosphate
    - Phosphate is connected at 5’
    - Thus, 3` connects to 5`
  - This gives the DNA direction: 5’ to 3’





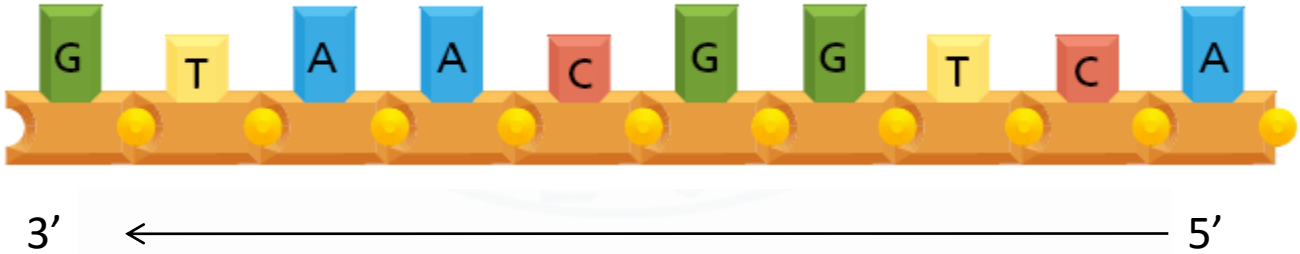
Building blocks of DNA: DNA Strand

- DNA is a polynucleotide
- 3` of one nucleotide is connected to 5` (phosphate) of next



- A strand of DNA

ACTGGCAATG

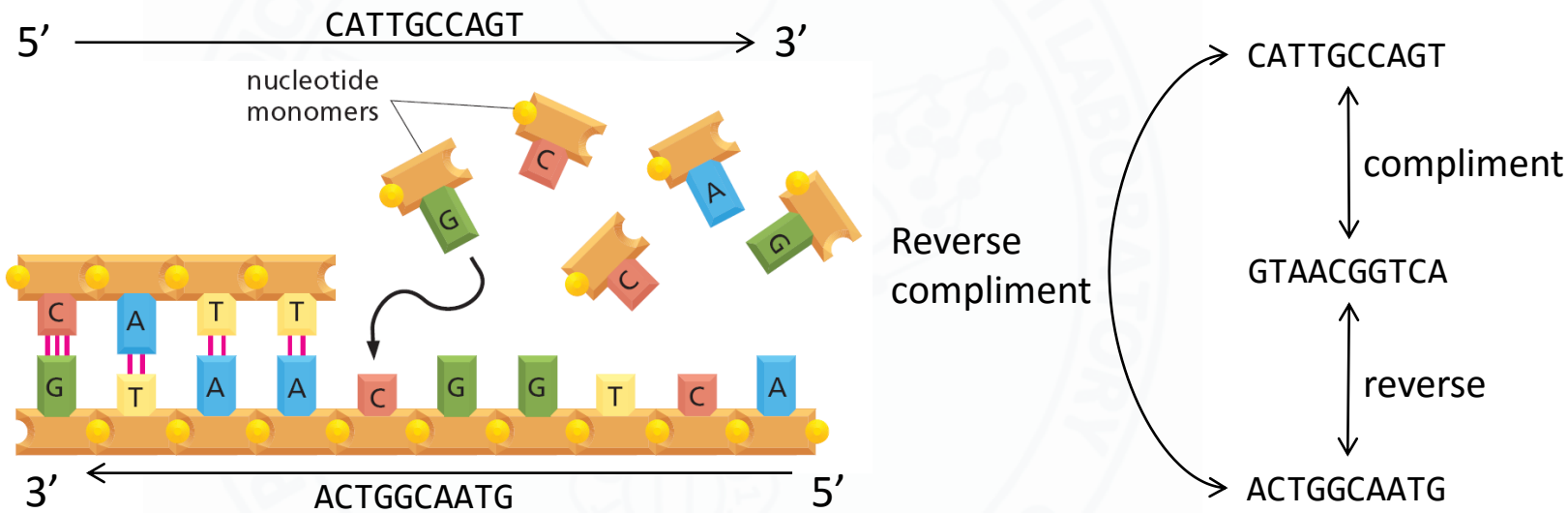


DNA: Why direction?

- The naming convention (5’ to 3’) is important because:
  - nucleic acids can only be synthesized in vivo in the 5'-to-3' direction
  - The relative positions of structures along a strand of nucleic acid, including genes and various protein binding sites, are usually noted as being either upstream (towards the 5'-end) or downstream (towards the 3'-end)
- By convention, single strands of DNA and RNA sequences are written in 5'-to-3' direction

Building blocks of DNA: DNA Strands

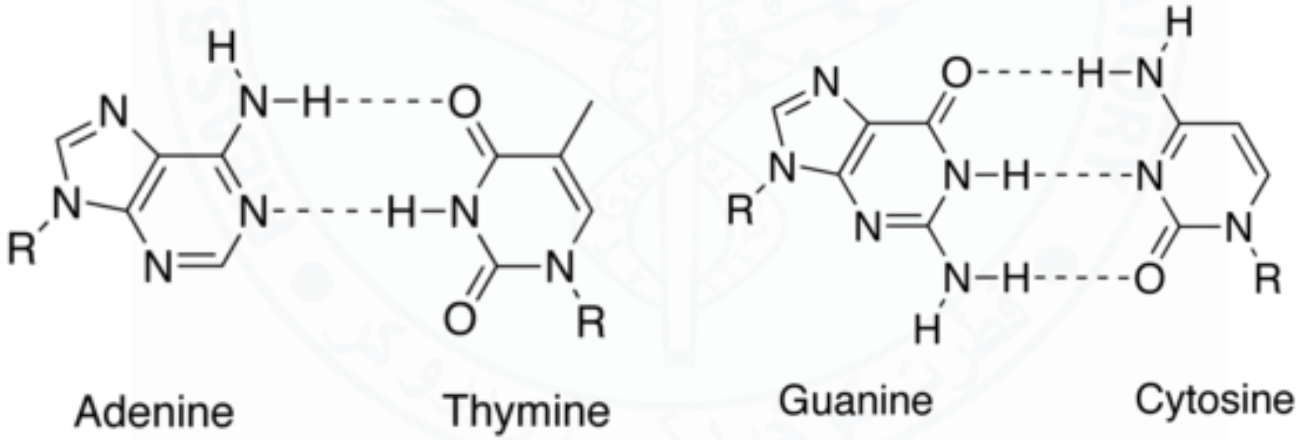
- The sequence of nucleotides in an existing DNA strand controls the sequence in which nucleotides are joined together in a new DNA strand
  - A→T, G→C



- The strand corresponding to an existing DNA strand is called its complimentary strand
  - The two DNA strands are antiparallel to two eachother

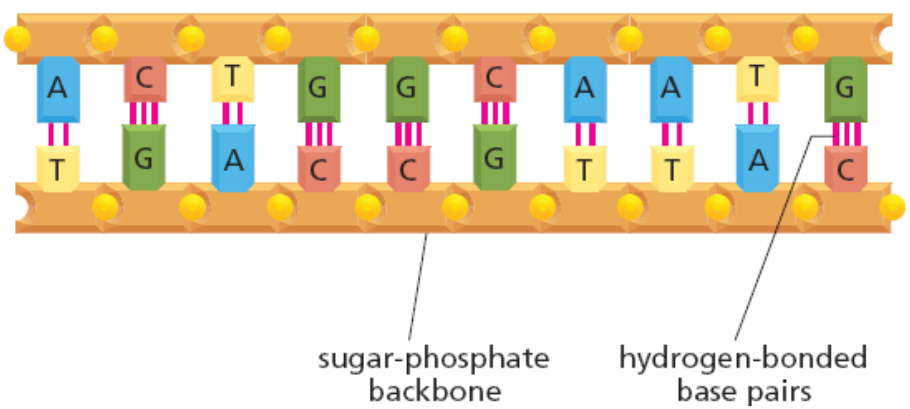
Building blocks of DNA: DNA Strands

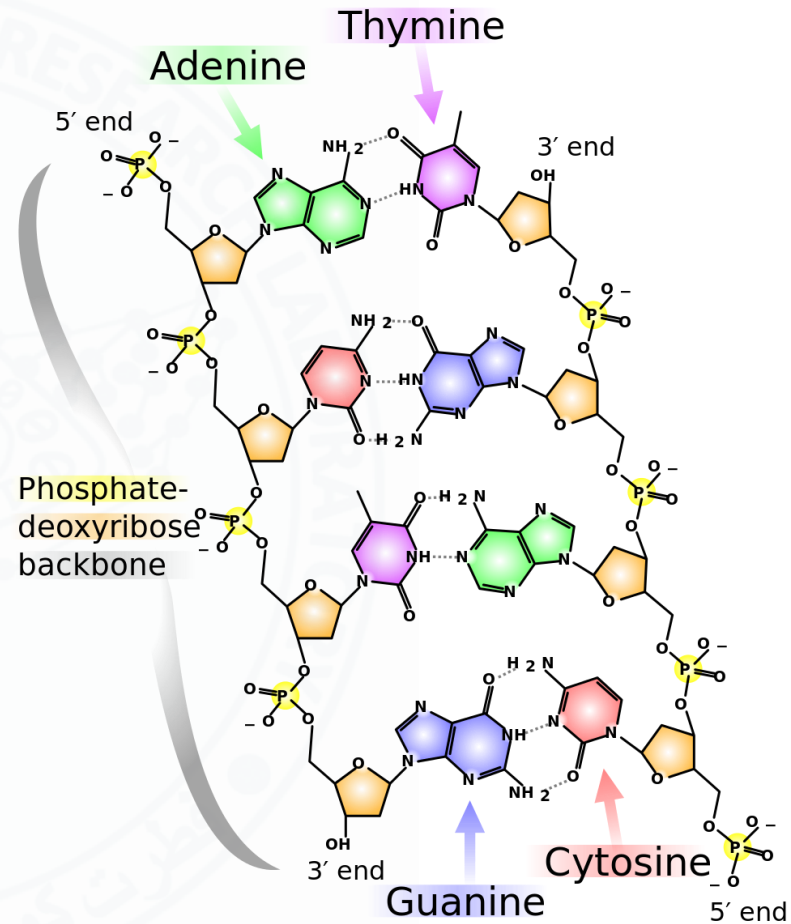
- Why Complimentarity?
  - Because it allows for efficient Hydrogen bonding



Building blocks of DNA: DNA Strands

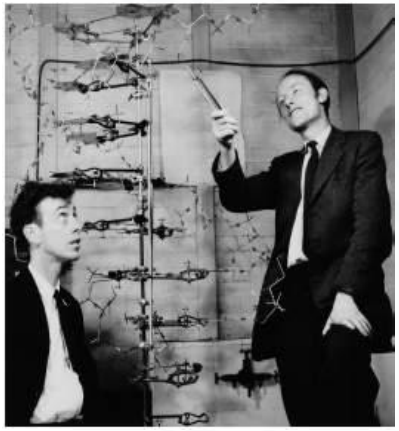
- Backbone
  - Formed by alternating sugar and phosphates





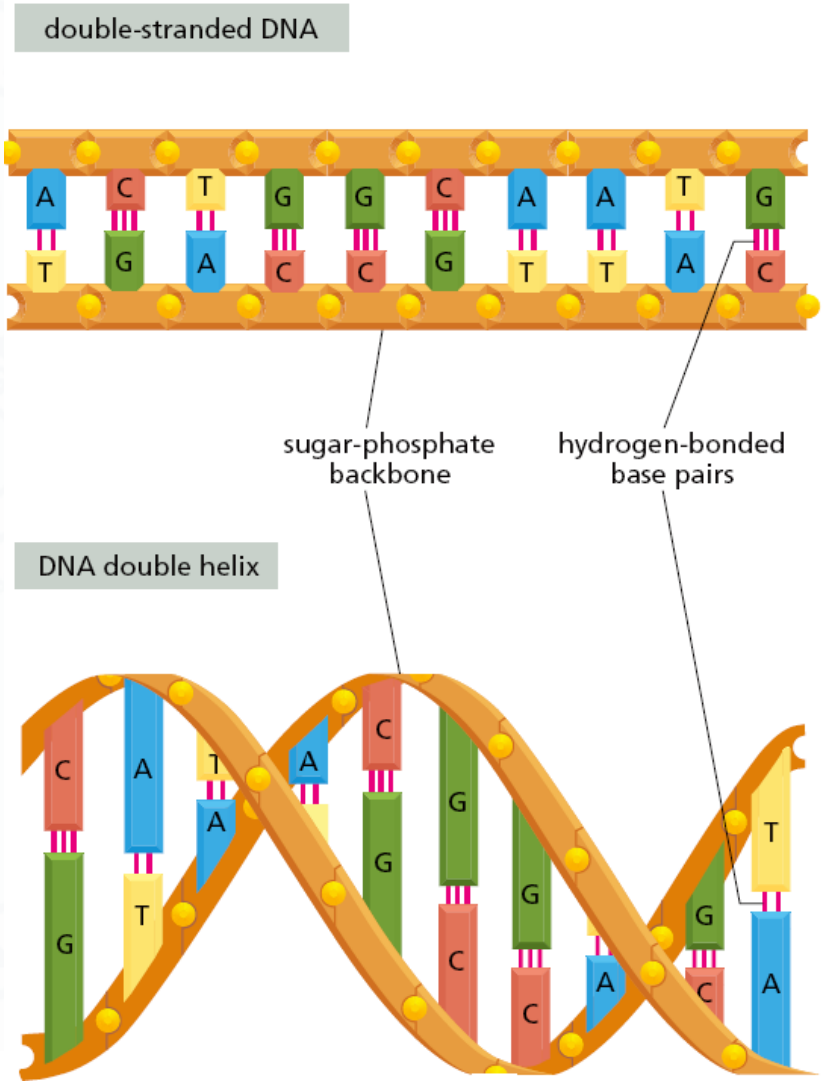
Building blocks of DNA: DNA Double Helix

- The two DNA strands twist around each other to form a double Helix



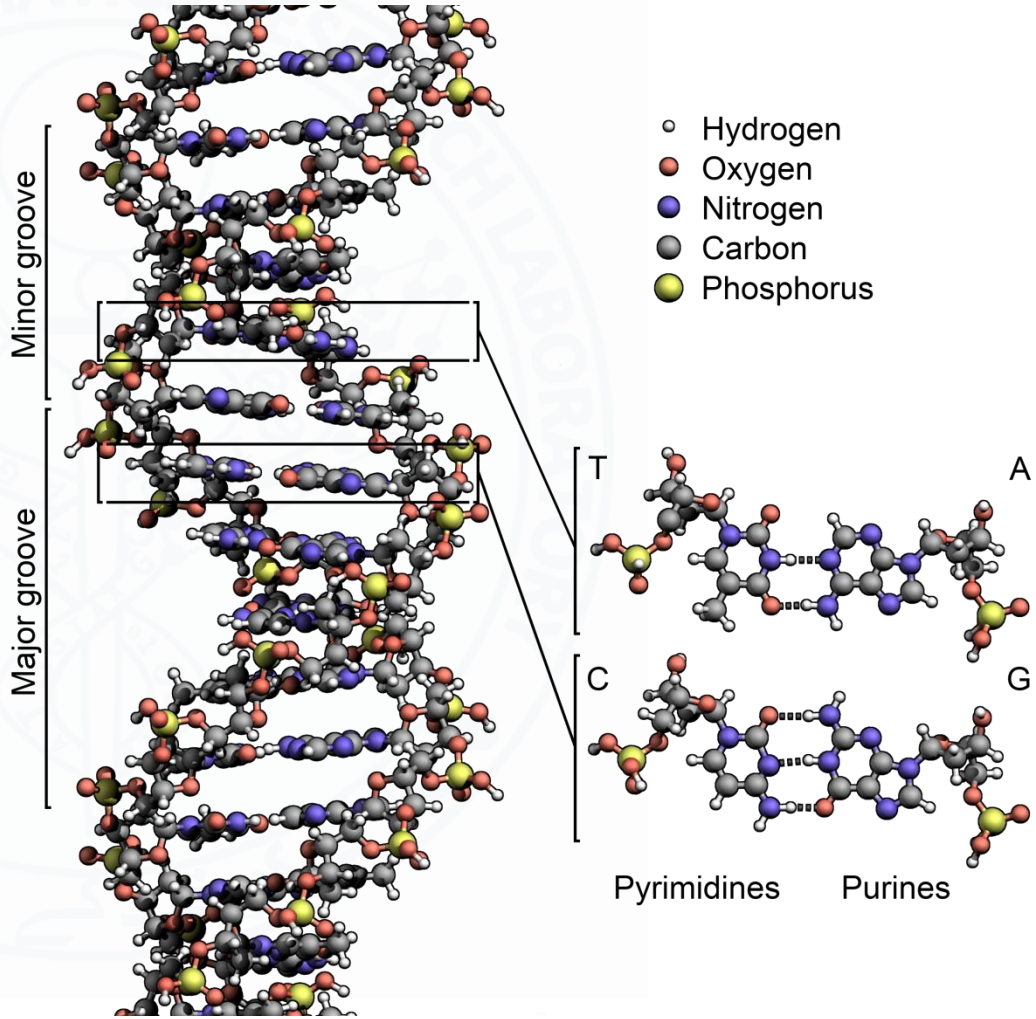
Watson & Crick with DNA model

Rosalind Franklin with X-ray image of DNA

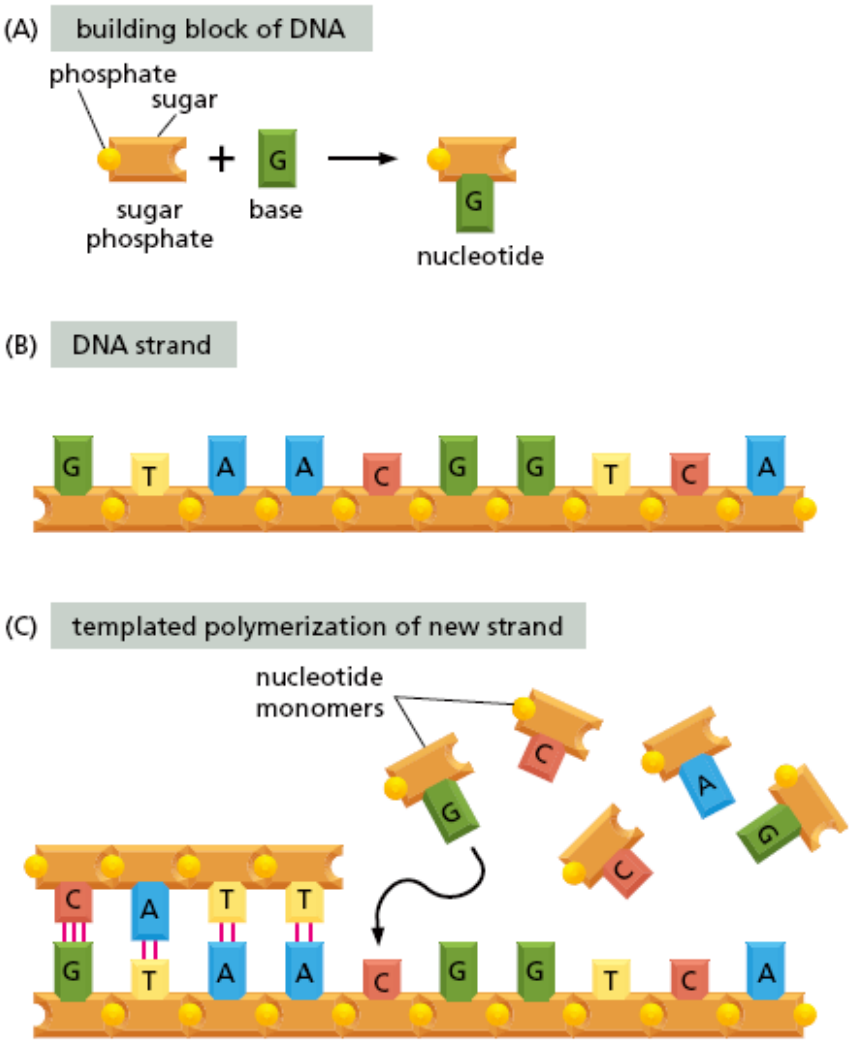


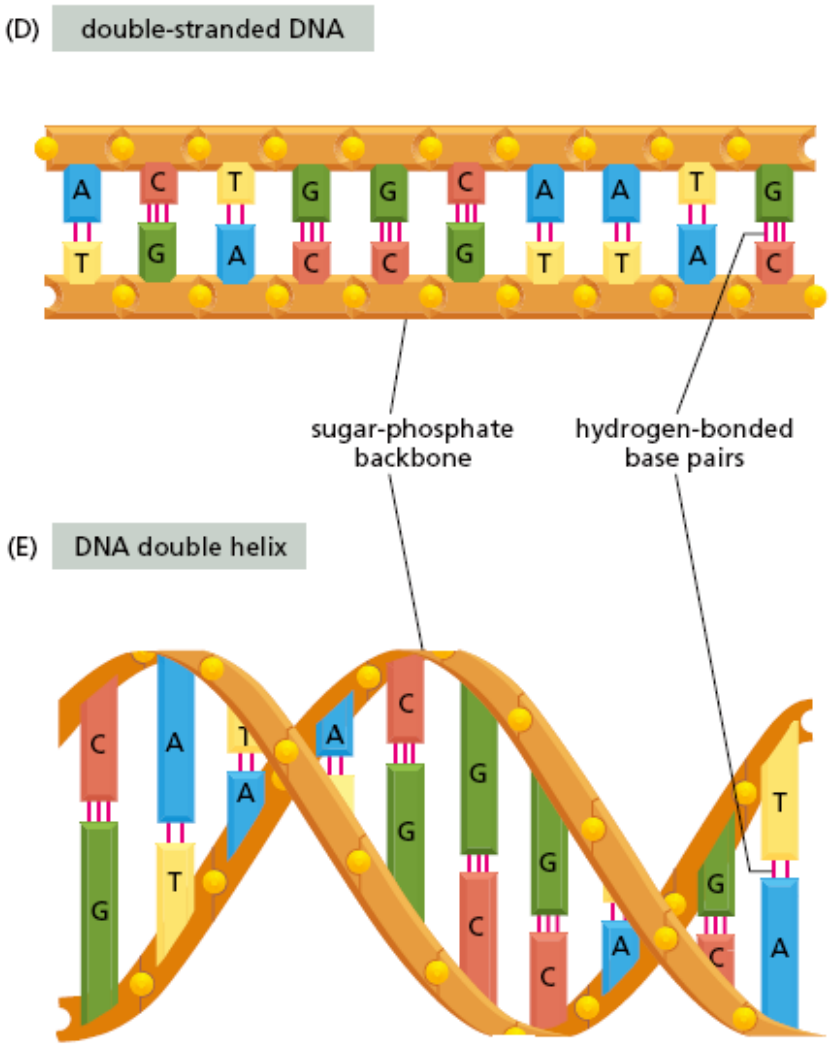
Building blocks of DNA: DNA Double Helix

- The DNA double helix is stabilized by the Hydrogen bonds between the strands and stacking of the aromatic bases

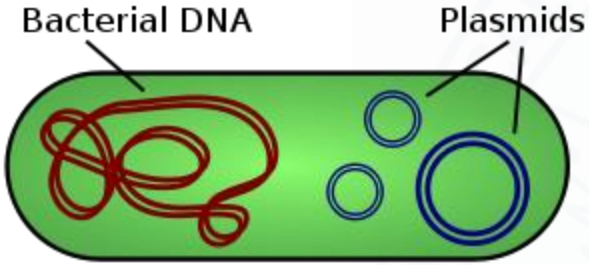


DNA's building blocks: Summary

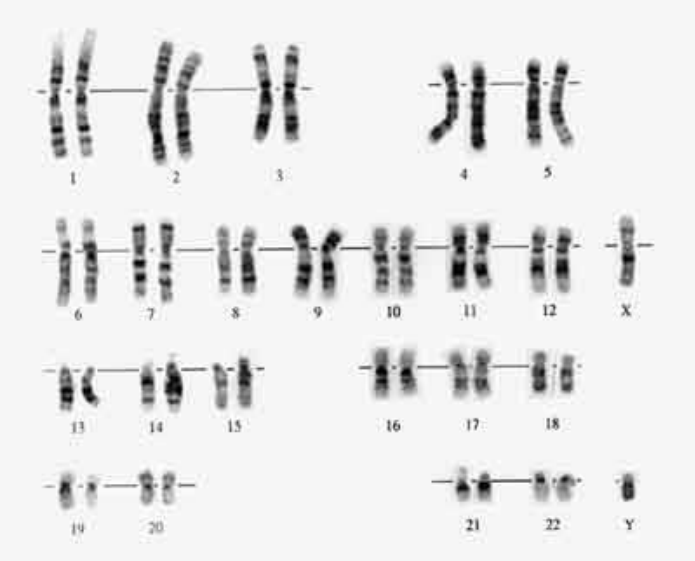




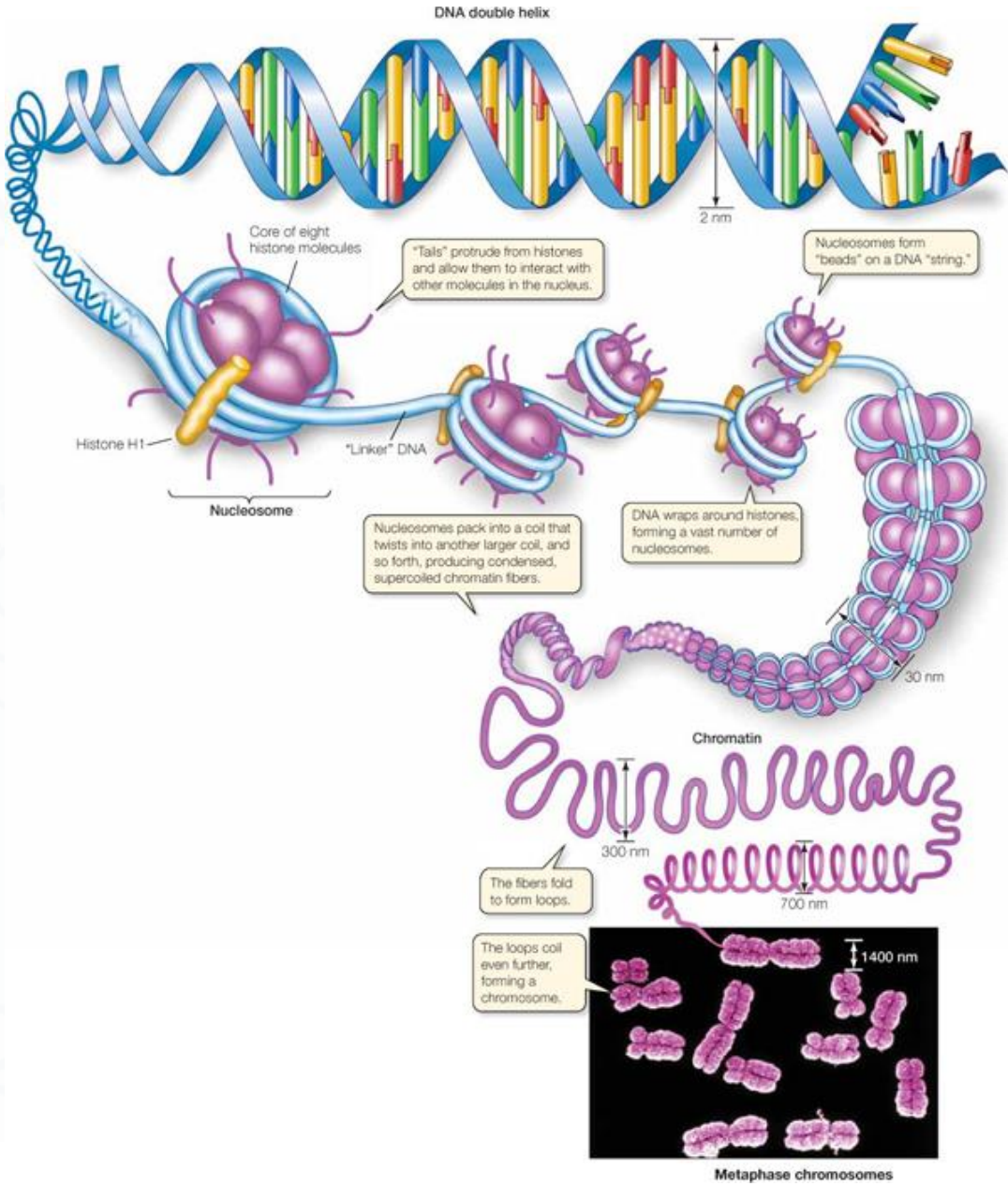
How is the program stored?



Circular DNA in Prokaryotes like *E. Coli*

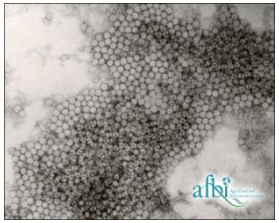


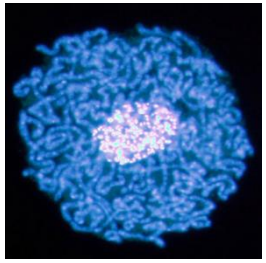
Chromosomal DNA in Eukaryotes (*H. Sapiens*)



Program size: DNA base pairs (bp)

Organism	# of base pairs	# of Chromosomes
Virus		
HIV	9193	1
SARS	29751	1
Porcine circovirus	1759	1
Prokayotic		
Haemophilus influenzae	1.8×10^6	1
Escherichia coli (bacterium)	4.6×10^6	1
Carsonella ruddii	159, 662 (0.16M)	1
Eukaryotic		
S. cerevisiae (yeast)	1.35×10^7	17
Drosophila melanogaster (fly)	1.65×10^8	4
Homo sapiens (human)	2.9×10^9	23
Paris japonica	150×10^9	-

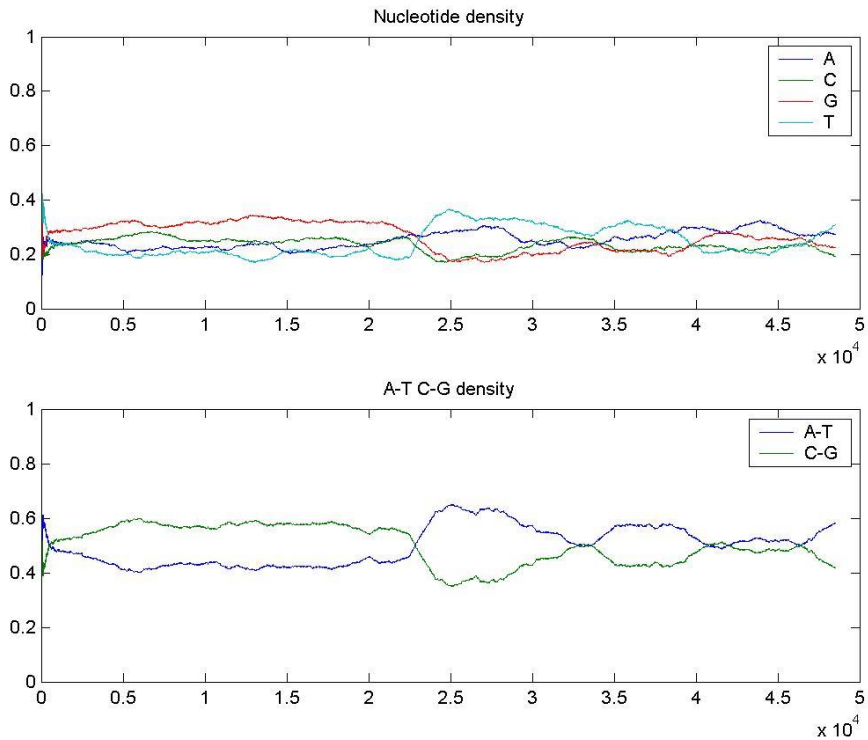




http://en.wikipedia.org/wiki/Genome
http://www.nature.com/news/2006/061009/full/news061009-10.html

Nucleotide composition of a genome

- Genomes often have a characteristic GC content which is more or less constant across a genome (value ranges from 24-72%)

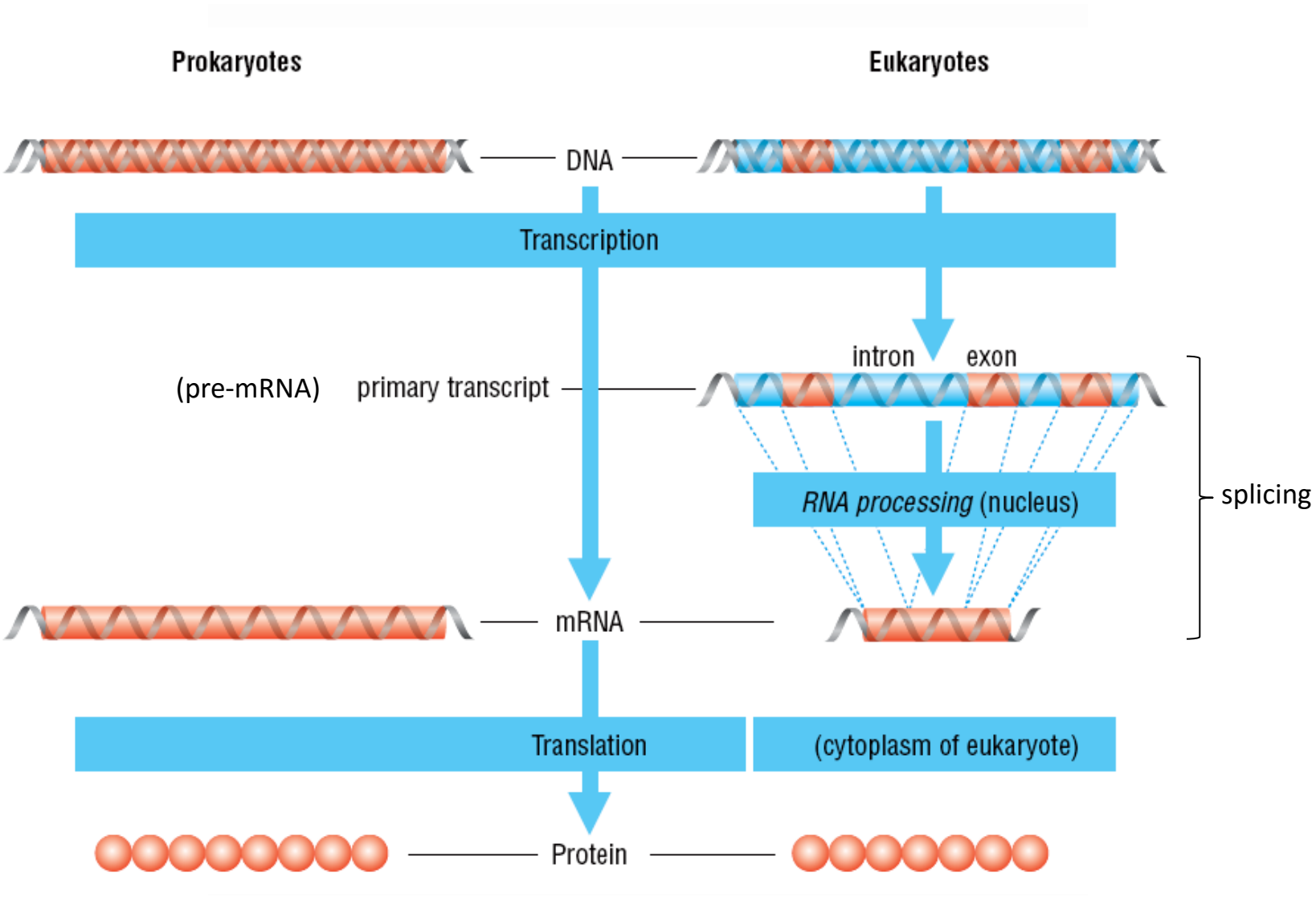


Nucleotide composition along the lambda phage genome

Data density in the genome

- Typical cell nucleus size:
  - 6 microns: same as dust particles
  - Stores all genetic material and a whole lot more
- DNA storage
  - 5.5 petabits / mm^3 of DNA
  - 90 PB / 41g of DNA (All of CERN's storage)
  - Can store information for thousands of years

Execution of the program: Central dogma of molecular biology



Central dogma of molecular biology

- It describes how genes are ‘expressed’: Gene Expression
- Transcription
  - DNA → mRNA: (A,T,G,C) to (A,U,G,C)
    - DNA bases to RNA bases
    - Output of transcription is called a `transcript` (remember transcriptomics?)
- Translation
  - mRNA→ Protein
    - RNA bases to amino acids
    - (A,U,G,C) to (A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Z)
- Splicing
  - pre-mRNA → mRNA
    - RNA to RNA
  - Exon
  - Intron

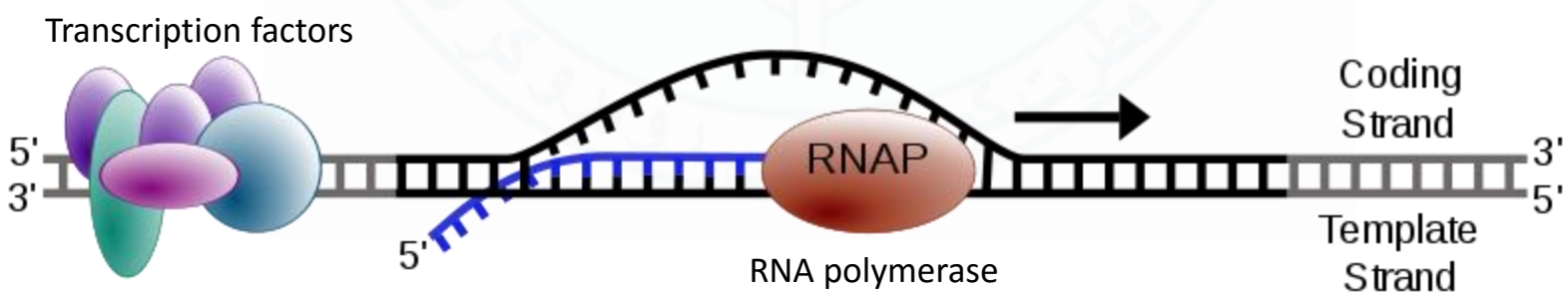
Transcription

- Performed by RNA polymerase after its binding to transcription factors
- Template strand of the DNA acts as the template for the production of RNA
- Coding strand contains the DNA version of the transcript sequence
- Either strand may be serving as the template: i.e., some genes run one way and some the other way and in some cases the same segment of double helix contains genetic information on both strands

TAATGCTAGACGTGTTCTAGGA

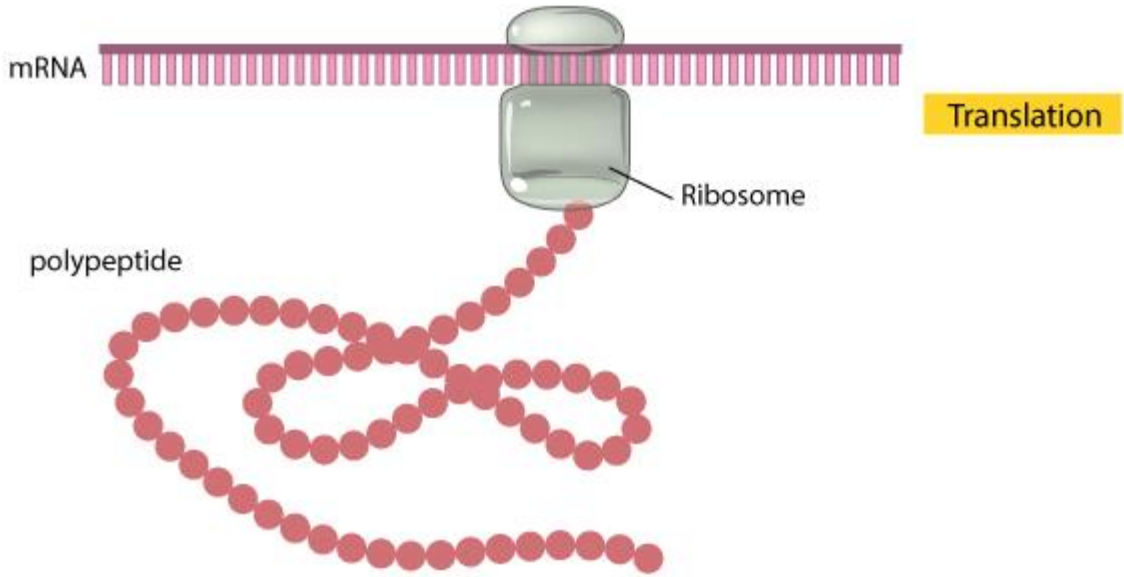
UAAUGCUAGACGUGUUCUAGGA

ATTACGATCTGCACAAGATCCT



Translation

- In translation the mRNA is converted into a sequence of amino acids
- Performed by Ribosome



Codon Table

nonpolar	polar	basic	acidic	(stop codon)

Standard genetic code

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine	UCA		UAA	Stop (*Ochre*) STOP	UGA	Stop (*Opal*) STOP	A
	UUG		UCG		UAG	Stop (*Amber*) STOP	UGG	(Trp/W) Tryptophan	G
C	CUU		CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	AUG[A]	(Met/M) Methionine GO	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

Translation

- Example below
- The end of the protein corresponding to the 5' end of the mRNA is called N-terminus whereas the other one is called C-terminus (we'll see why when we will talk about proteins more)

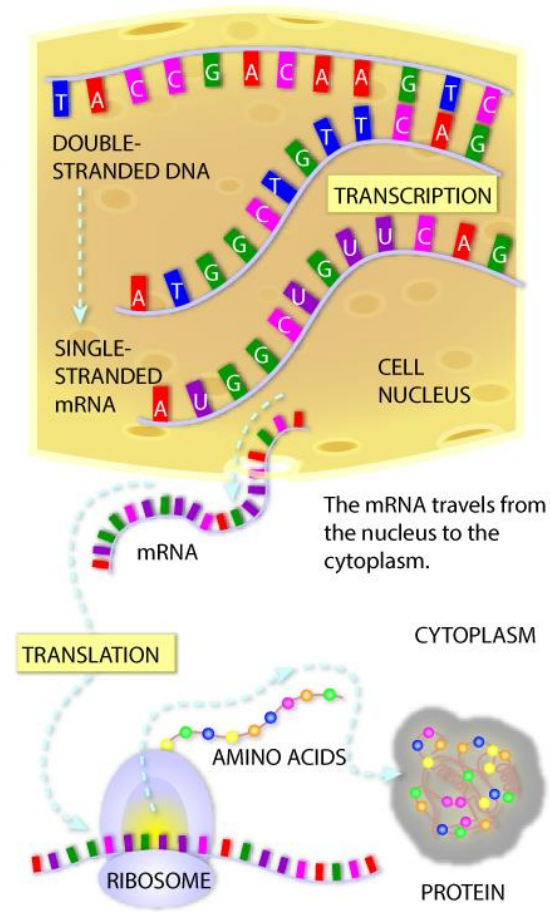
5' ⟶ 3'

UAAUGCUAGACGUGUUCUAGGA

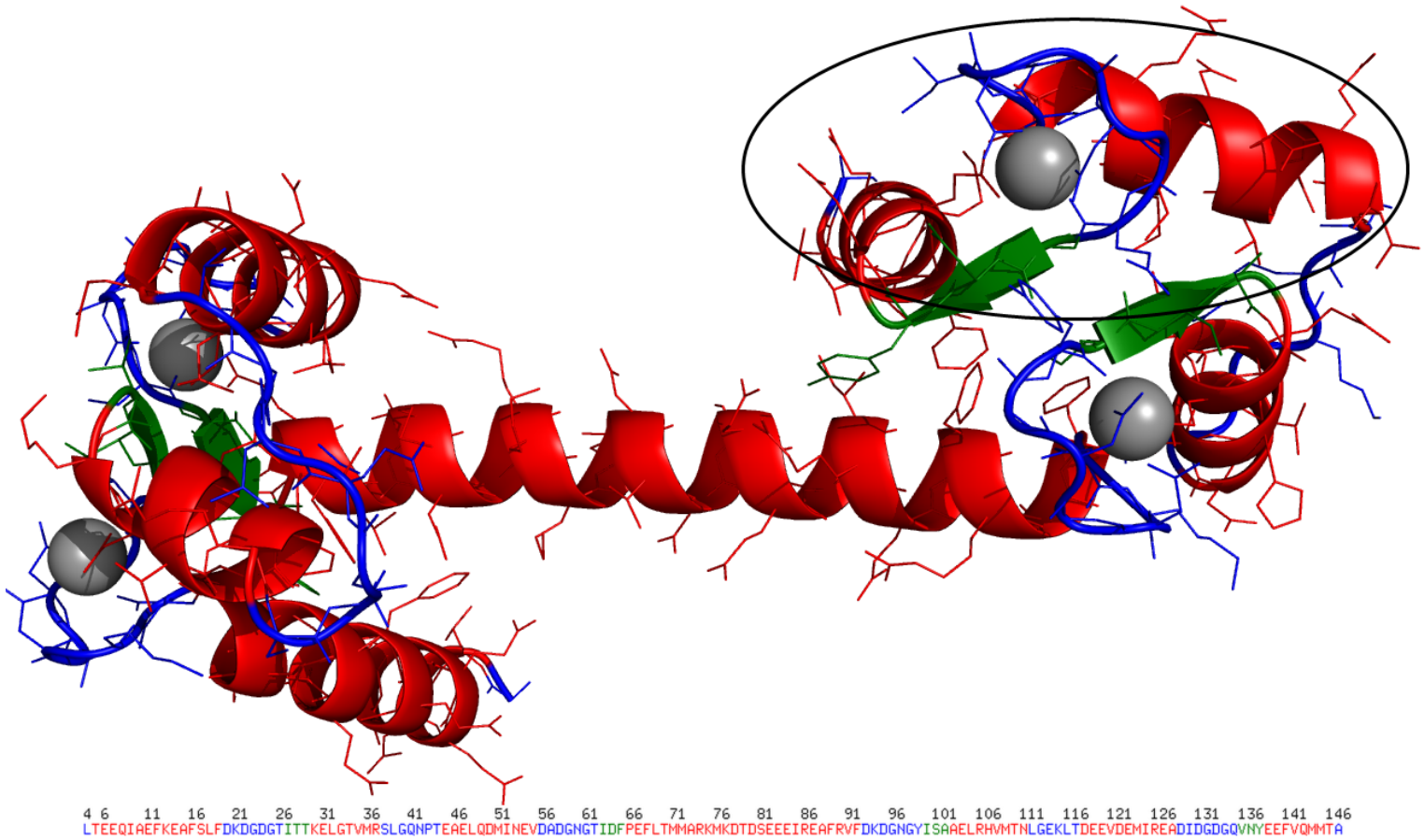
UA**AUG**CUAGACGUGUUC**UAG**GA

MLAVF

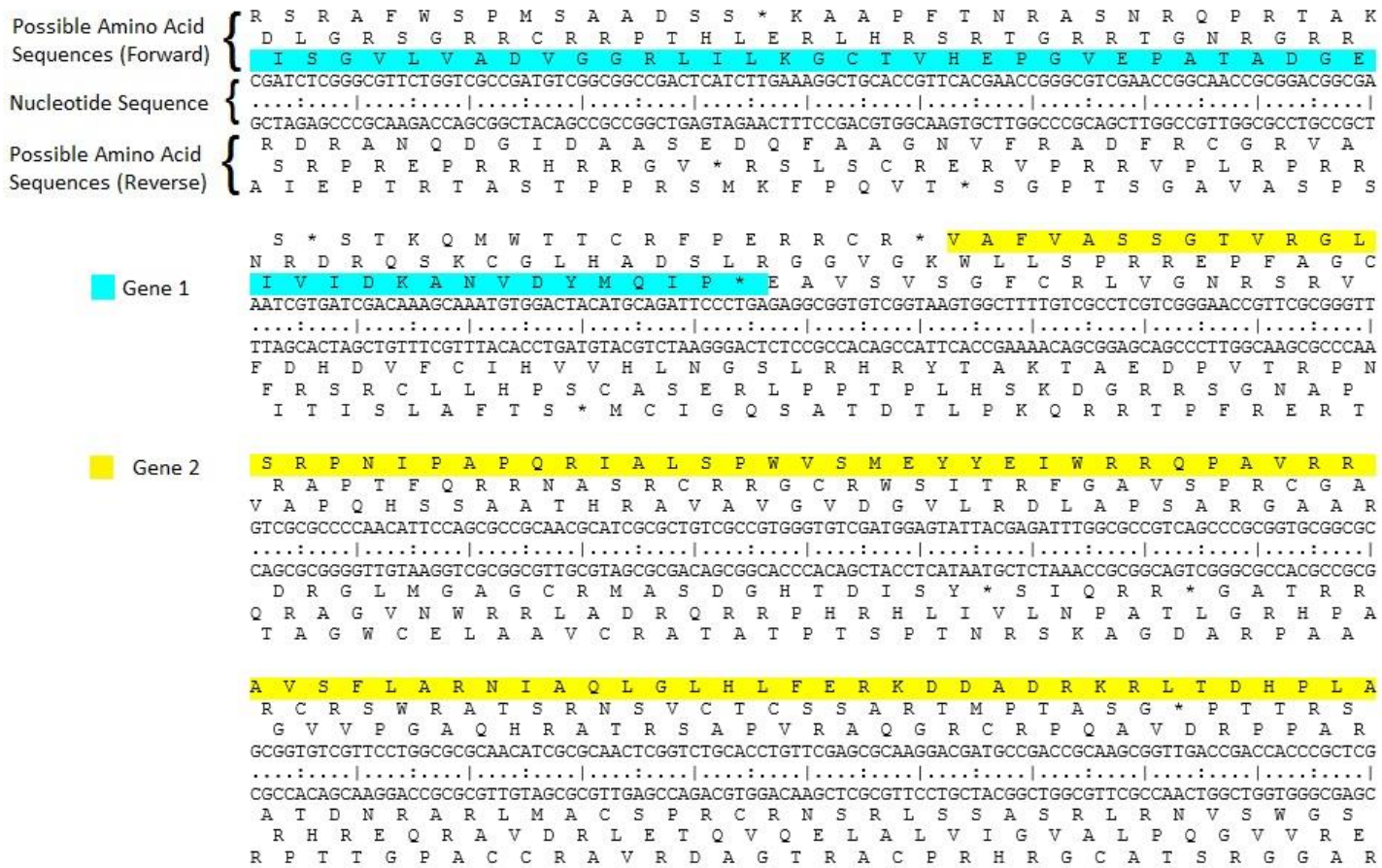
N-terminus ⟶ C-terminus



Proteins


4 6 11 16 21 26 31 36 41 46 51 56 61 66 71 76 81 86 91 96 101 106 111 116 121 126 131 136 141 146
LTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTA

Open reading frames



1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A

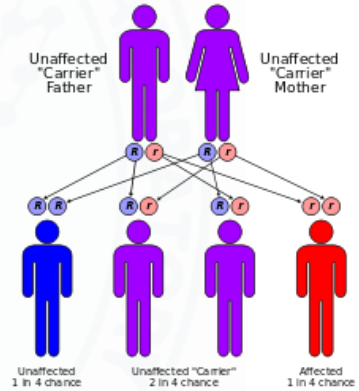
Non-Sense Mutation

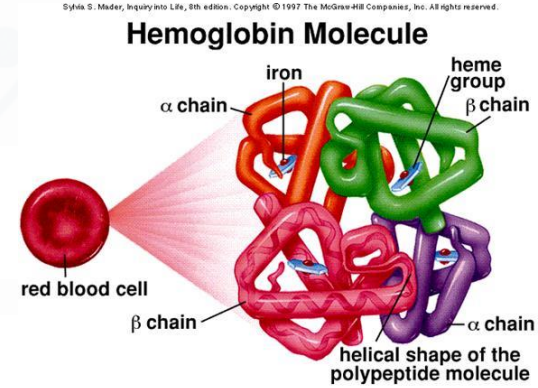
- A point mutation in a sequence of DNA that results in a premature stop codon
  - Protein product is incomplete or non-functional
- Beta-Thalassemia
  - Results from a single point mutation
    - HBB gene on chromosome 11
    - Reduction in production of hemoglobin
      - HBB blockage over time leads to decreased Beta-chain synthesis
    - Having a single gene for thalassemia may protect against malaria
    - One of the most commonly inherited disorders in Pakistan
    - With a prevalence rate of 6 % in the Pakistani population
    - 5000-9000 children every year

```
    DNA: 5' - ATG ACT CAC CGA GCG CGA AGC TGA - 3'
         3' - TAC TGA GTG GCT CGC GCT TCG ACT - 5'
   mRNA: 5' - AUG ACU CAC CGA GCG CGA AGC UGA - 3'
Protein:      Met Thr His Arg Ala Arg Ser Stop
```

↓

```
    DNA: 5' - ATG ACT CAC TGA GCG CGA AGC TGA - 3'
         3' - TAC TGA GTG ACT CGC GCT TCG ACT - 5'
   mRNA: 5' - AUG ACU CAC UGA GCG CGU AGC UGA - 3'
Protein:      Met Thr His Stop
```

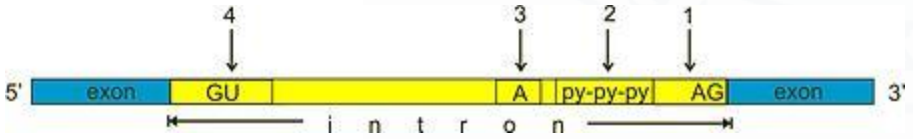


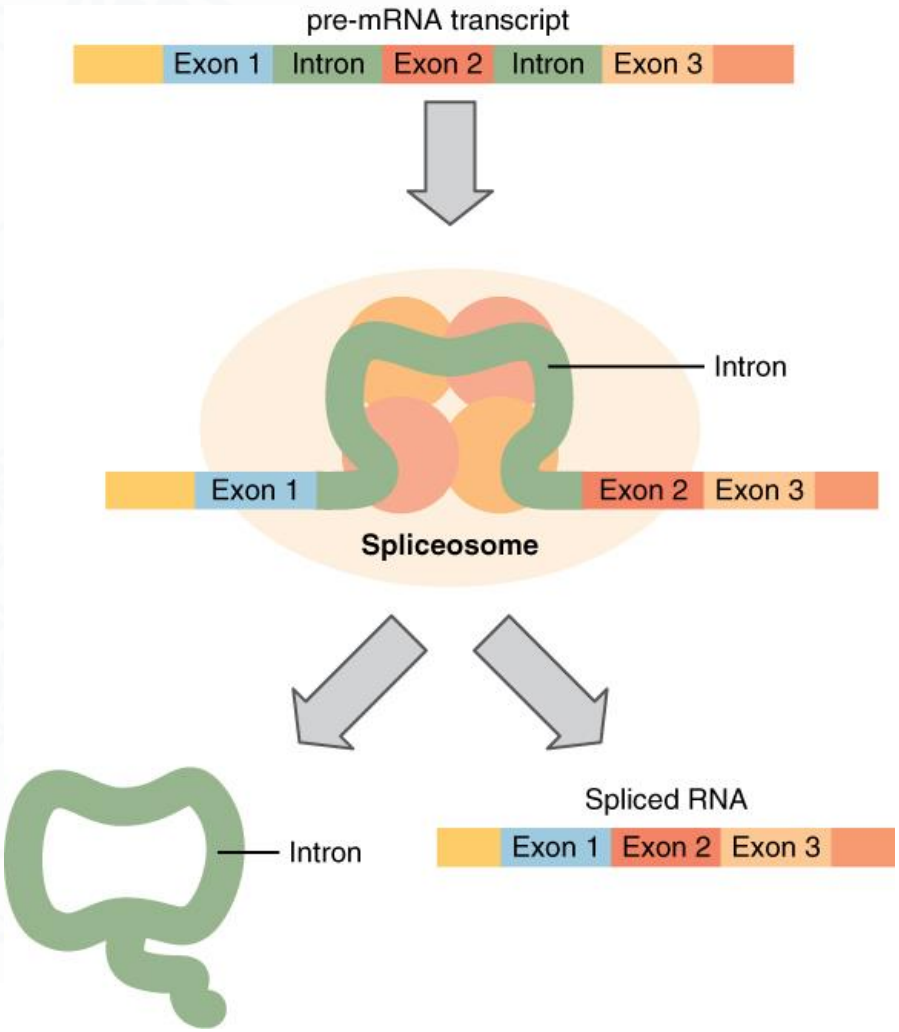


Thalassaemia Federation of Pakistan

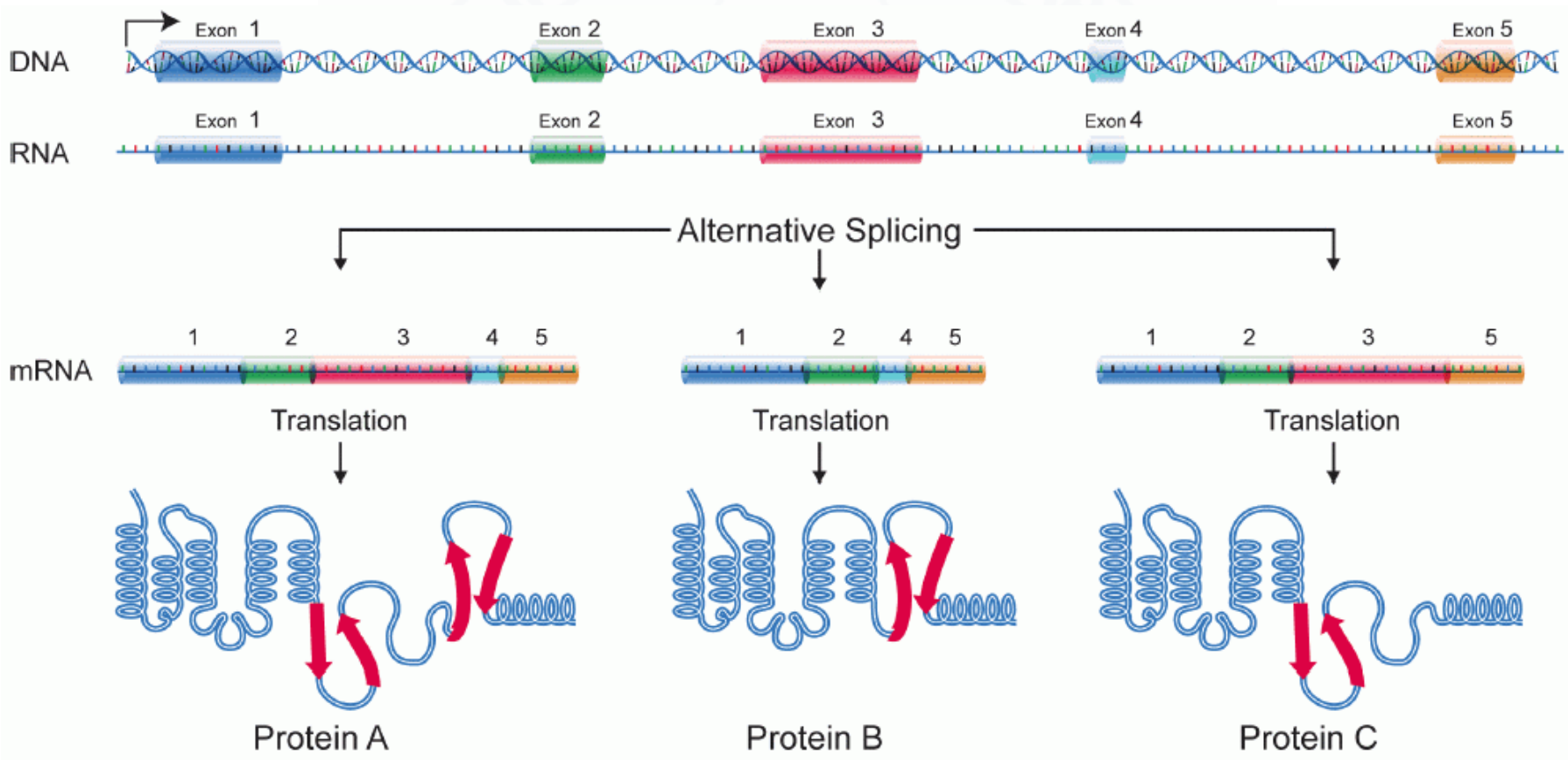
Splicing

- Modification of the nascent pre-messenger RNA (pre-mRNA) transcript in which introns are removed and exons are joined
- Human genome:
  - ~180,000 exons for ~21,000 genes
  - ~9 exons/gene





Alternative Splicing



Three different isoforms

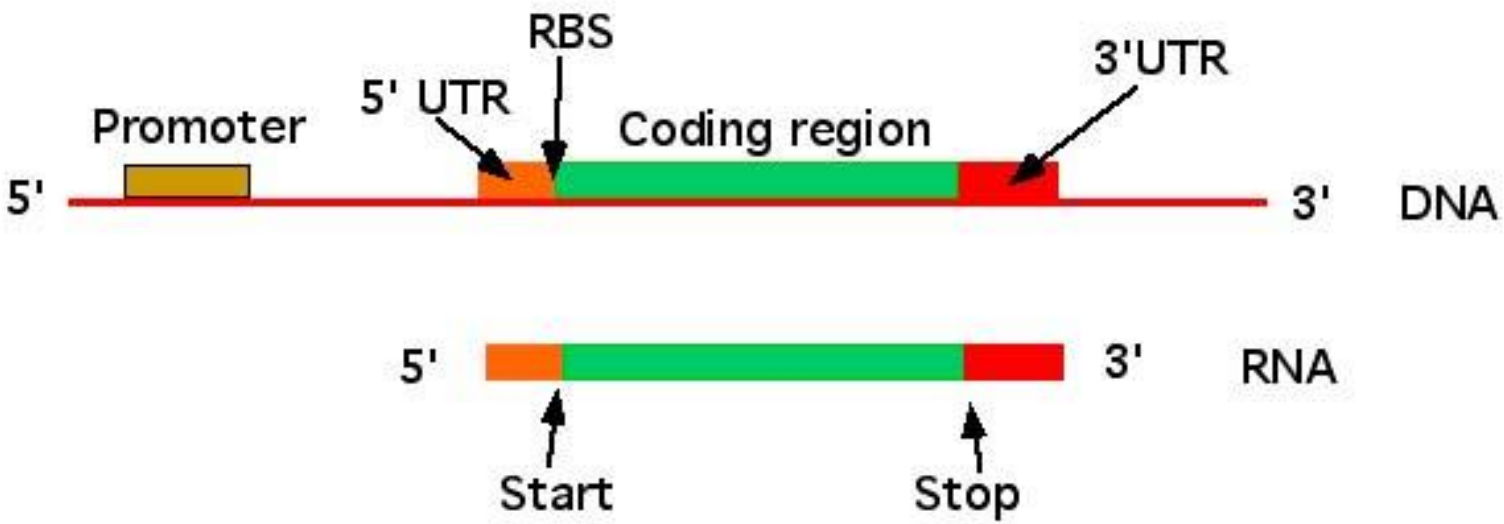
Alternative Splicing

- ~95% of multi-exonic genes in the human genome are alternatively spliced
- Explains how ~20K genes can produce a whole lot more proteins

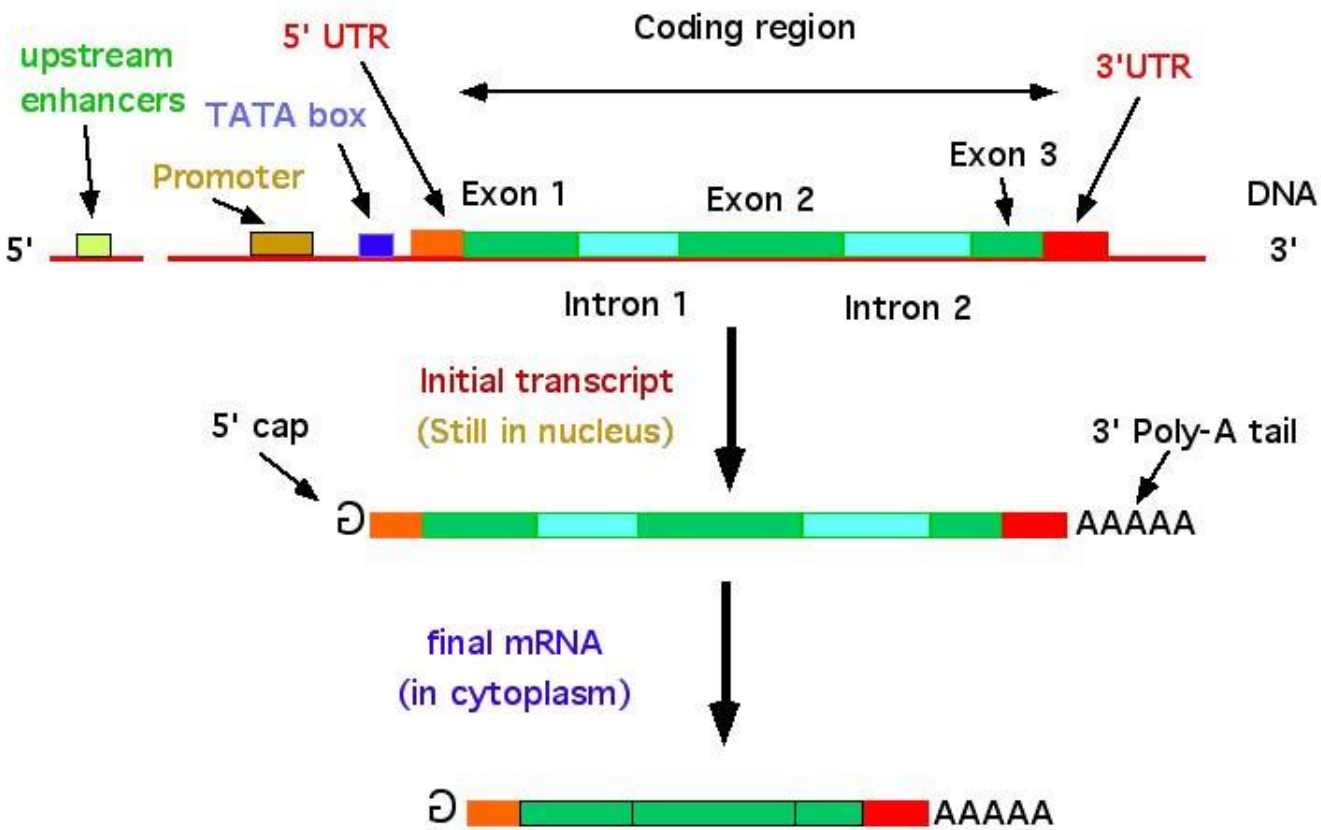
Post translational modifications

- Proteins can change even after they are created through PTMs

Gene structure: Prokaryotes



Gene structure: Eukaryotes



Human Genes

- Regulatory regions: up to 50 kb upstream of +1 site
- Exons: 1 to 178 exons per gene (mean 8.8)
  8 bp to 17 kb per exon (mean 145 bp)
- Introns: 1 kb – 50 kb per intron (mean ~2,000 bp)
- Gene size: Largest – 2.4 Mb (Dystrophin). Mean – 27 kb.

Conclusions

- Readings
  - Required: Pevzner chapter 3
  - Recommended:
    - Wikipedia!
    - J. Cohen, “Bioinformatics: an Introduction for Computer Scientists,” ACM Comput Surv, vol. 36, no. 2, pp. 122–158, Jun. 2004.
    - B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, “Cells and Genomes,” 2002.
- History
  - http://en.wikipedia.org/wiki/History_of_genetics

Problems in Bioinformatics

Genomics:
- Assembly
- Detection of variation
- GWAS

RNA:
- Gene expression
- Transcriptome assembly
- Pathway analysis

Protein:
- Mass spectrometry
- Structure prediction
- Protein-Protein interaction

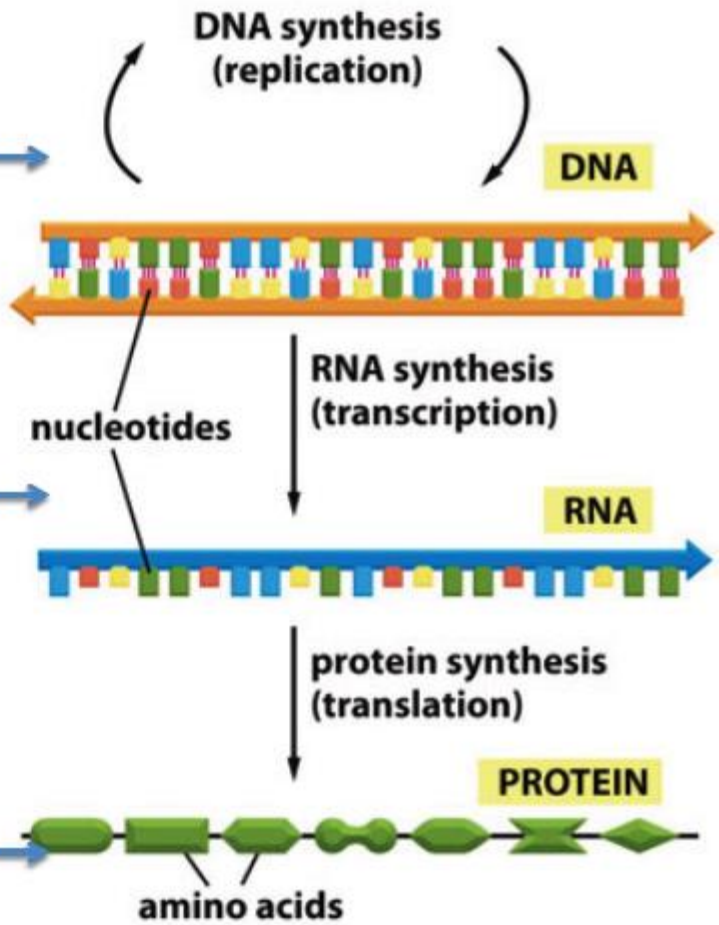


Figure 1-2 Essential Cell Biology 3/e (© Garland Science 2010)

Basic skills required

- Understanding of the scientific method
- Biology/Chemistry knowledge
- Programming
- Algorithms
- Databases
- AI & ML

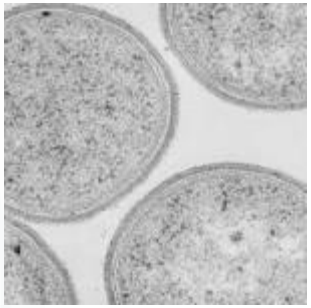
- Need to feel comfortable in interdisciplinary area
- Primary data from others
- Abstract thinking to address important biology & CS problems

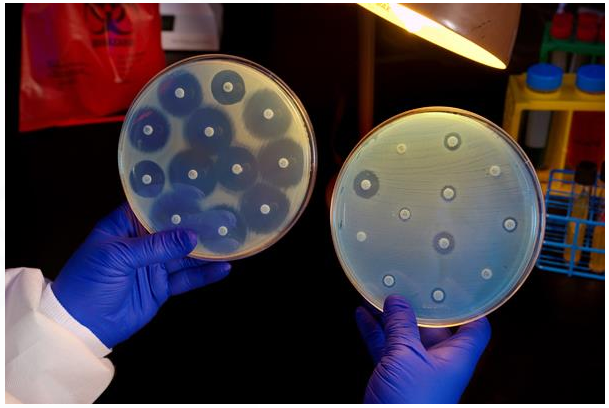
Biology vs. CS

Biology	CS
Everything is true or false in computer science	Either True or False
Strive to understand the complicated, messy natural world	Seek to build their own clean virtual worlds
Obsessed with being the first to discover something	Obsessed with being the first to invent or prove something
Comfortable with the idea that all data have errors	Not!
Typically have to complete one or more 5-year post-docs...	Get high-paid jobs after graduation

Assignment-1: Python Warmup

- Brevibacillus brevis
- Circular genome
  - 6,296,436bp (6.3M)
- Produces a antibiotic!





  - Download as fasta file from
    - http://bacteria.ensembl.org/brevibacillus_brevis_nbrc_100599/Info/Index

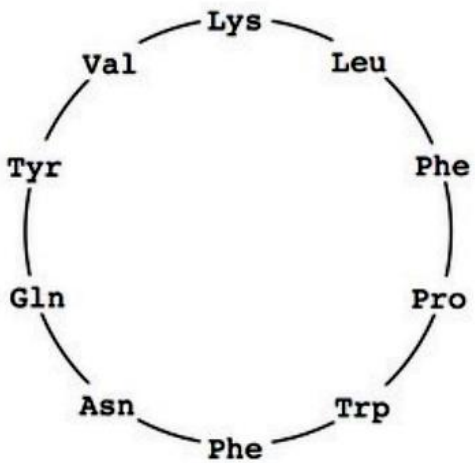
Brevibacillus_brevis_nbrc_100599.GCA_000010165.1.23.dna.chromosome.Chromosome.fa

Assignment

- Write program
  - Count the number of k-mers in overall genome
    - k=1: count A, T, G, C
      - What is the percentage of G+C in the Genome? 47.3%
    - K=2: count AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, CC
    - And so on
    - Plot the frequencies for k=1, k=2
  - Plot the frequencies for k=1 and k=2 in a window of an arbitrary size and then experiment with different window sizes (100, 1000, 10000)

Assignment

- Write a program which when given a protein sequence finds the start location of that sequence in the genome together with the `sense` or the directionality
  - 6 ORFs, Circular
- Find the sequence for the antibiotic circular mini protein (peptide) Tyrocidine B1



1 Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr
2 Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr-Val
3 Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr-Val-Lys
4 Phe-Pro-Trp-Phe-Asn-Gln-Tyr-Val-Lys-Leu
5 Pro-Trp-Phe-Asn-Gln-Tyr-Val-Lys-Leu-Phe
6 Trp-Phe-Asn-Gln-Tyr-Val-Lys-Leu-Phe-Pro
7 Phe-Asn-Gln-Tyr-Val-Lys-Leu-Phe-Pro-Trp
8 Asn-Gln-Tyr-Val-Lys-Leu-Phe-Pro-Trp-Phe
9 Gln-Tyr-Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn
10 Tyr-Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln

Python Help

- Reading File
- Plotting (matplotlib.pyplot)
- String Functions (count)
- Use Lists or Strings

End of Lecture-2