CS481: Bioinformatics Algorithms

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

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs481/

CS481

- Class hours:
  - Mon 10:40 - 12:30; Thu 9:40 - 10:30
- Class room: EE517
- Office hour: Tue + Thu 11:00-12:00
- TA: Enver Kayaaslan (ekayaaslan@gmail.com)
- Grading:
  - 1 midterm: 30%
  - 1 final: 35%
  - Homeworks (theoretical & programming): 15%
  - Quizzes: 20%

CS481

- Textbook: An Introduction to Bioinformatics Algorithms (Computational Molecular Biology), Neil Jones and Pavel Pevzner, MIT Press, 2004
- Recommended Material
  - Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchison, Cambridge University Press
  - Bioinformatics: The Machine Learning Approach, Second Edition, Pierre Baldi, Soren Brunak, MIT Press
  - Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Dan Gusfield, Cambridge University Press
- (Most) of the course material is publicly available at: www.bioalgorithms.info

CS481

- This course is about **algorithms** in the field of bioinformatics:
  - What are the problems?
  - What algorithms are developed for what problem?
  - Algorithm design techniques
- This course is **not** about how to analyze biological data using available tools:
  - Recommended course: MBG 326: Introduction to Bioinformatics

CS481: Assumptions

- You are assumed to know/understand
  - Computer science basics (CS101/102 or CS111/112)
    - CS201/202 would be better
    - CS473 would be even better
  - Data structures (trees, linked lists, queues, etc.)
  - Elementary algorithms (sorting, hashing, etc.)
  - Programming: C, C++, Java, Python, etc.
- You don't *have to* be a "biology expert" but MBG 101 or 110 would be beneficial
- For the students from non-CS departments, the TA will hold a few recitation sessions
  - Email your schedules to ekayaaslan@gmail.com

Bioinformatics

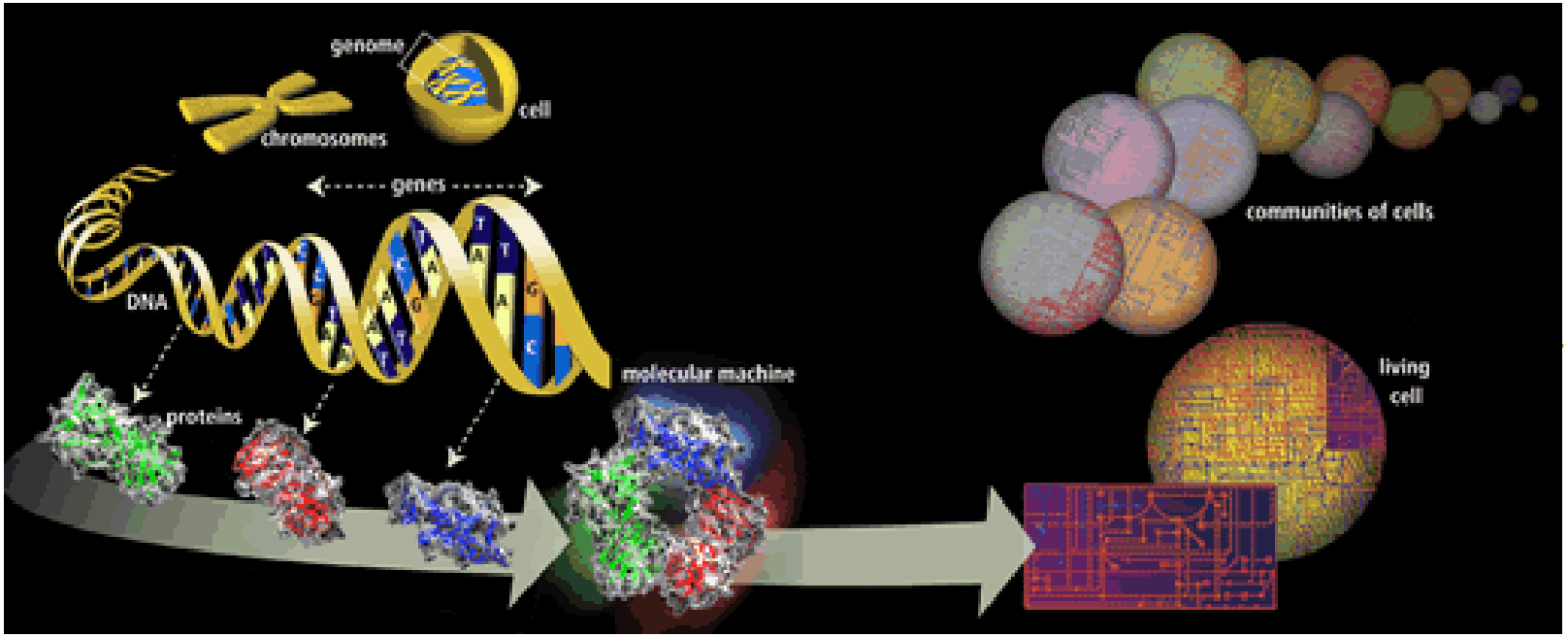
- Development of methods based on computer science for problems in biology and medicine
  - Sequence analysis (combinatorial and statistical/probabilistic methods)
  - Graph theory
  - Data mining
  - Database
  - Statistics
  - Image processing
  - Visualization
  - …..

CS 481

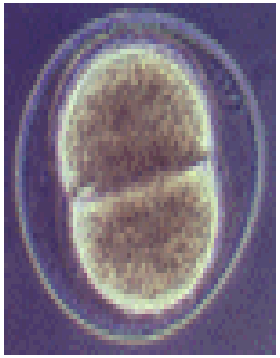
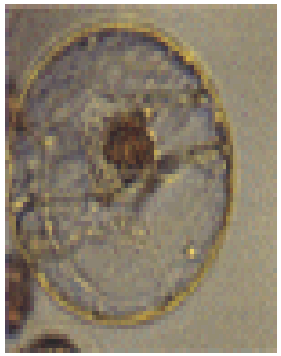
Bioinformatics: Applications

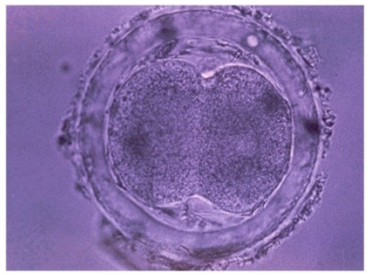
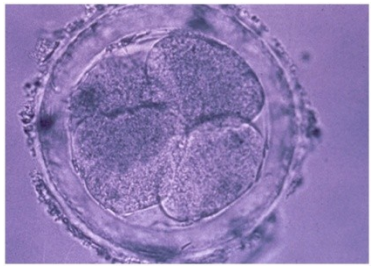
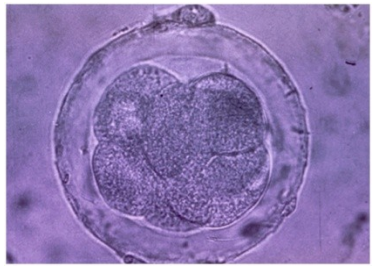
- Biology, molecular biology
- Human disease
- Genomics: Genome analysis, gene discovery, regulatory elements, etc.
- Population genomics
- Evolutionary biology
- Proteomics: analysis of proteins, protein pathways, interactions
- Transcriptomics: analysis of the transcriptome (RNA sequences)
- …

Molecular Biology Primer



What is Life made of?

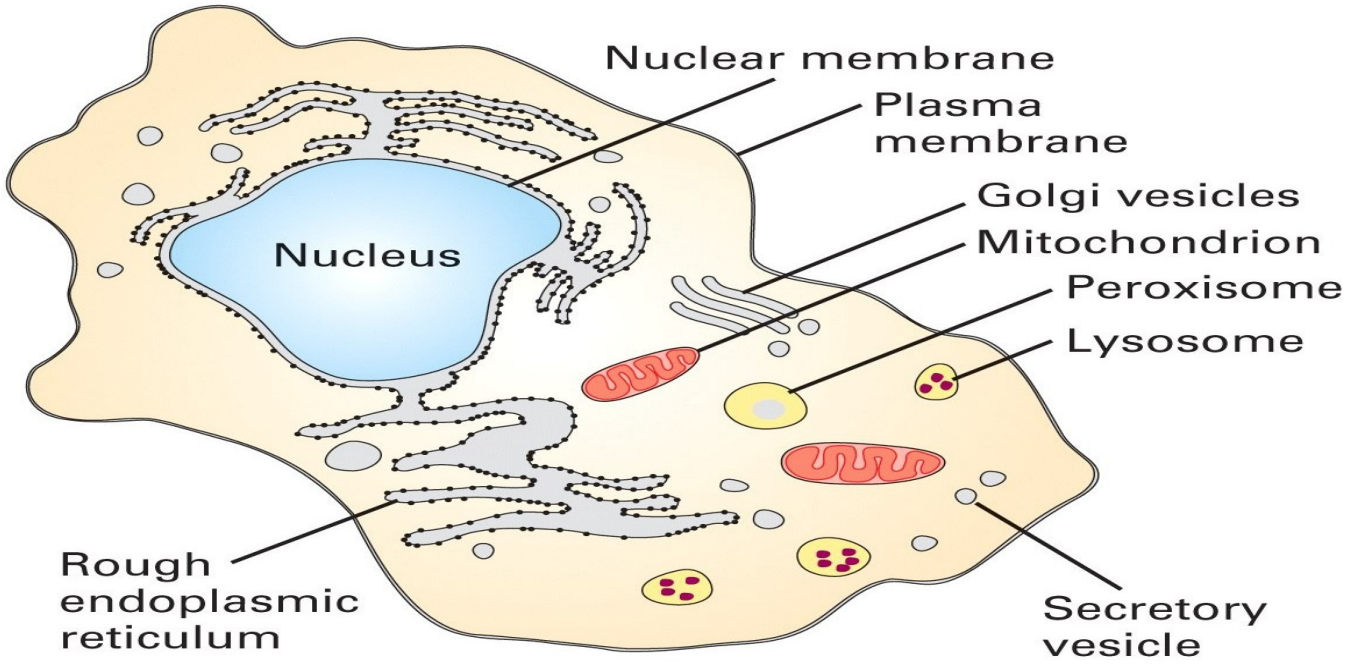




Cells

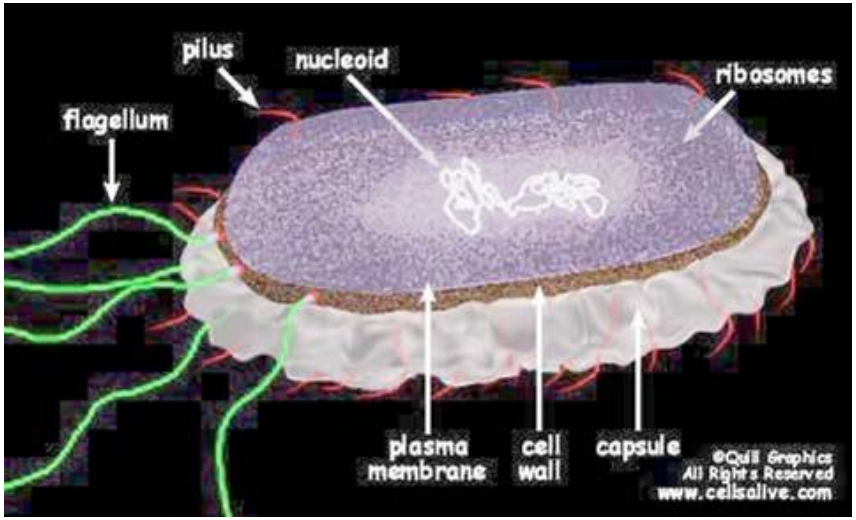
- **Fundamental working units** of every living system.
- Every organism is composed of one of two radically different types of cells:
  - **prokaryotic** cells
  - **eukaryotic** cells
- **Prokaryotes** and **Eukaryotes** are descended from the same primitive cell.
  - All extant prokaryotic and eukaryotic cells are the result of a total of 3.5 billion years of evolution.

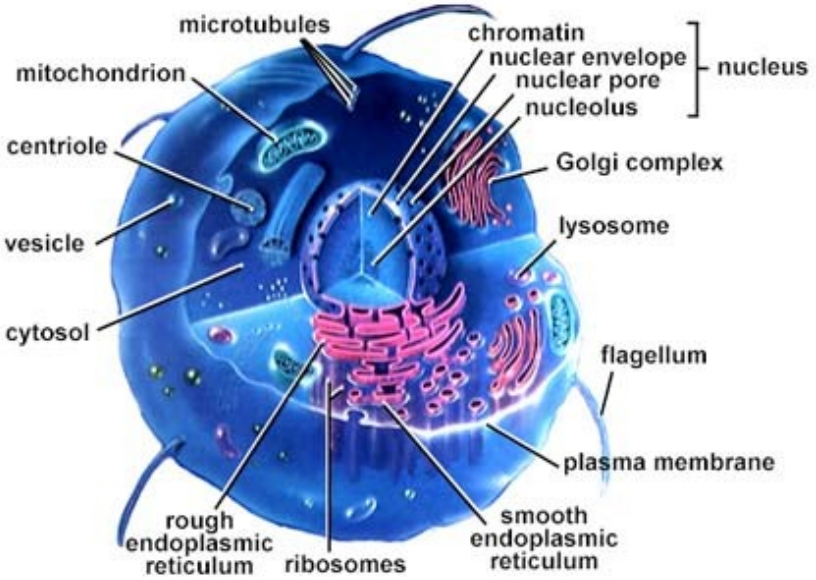
Life begins with Cell



- A cell is a smallest structural unit of an organism that is capable of independent functioning
- All cells have some common features

Prokaryotes *vs.* Eukaryotes





Prokaryotes and Eukaryotes

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

Cells Information and Machinery

- Cells store all information to replicate themselves
  - Human genome is around 3 billions base pair long
  - Almost every cell in human body contains same set of genes
  - But not all genes are used or expressed by those cells
- Machinery:
  - Collect and manufacture components
  - Carry out replication
  - Kick-start its new offspring

Some Terminology

- **Genome**: an organism's genetic material
- **Gene**: discrete units of hereditary information located on the chromosomes and consisting of DNA.
- **Genotype**: The genetic makeup of an organism
- **Phenotype**: the physical expressed traits of an organism
- **Nucleic acid**: Biological molecules(RNA and DNA)

More Terminology

- The **genome** is an organism's complete set of DNA.
  - a bacteria contains about 600,000 base pairs
  - human and mouse genomes have some 3 billion.
- Human genome has 23 pairs of chromosomes
  - 22 pairs of *autosomal* chromosomes (chr1 to chr22)
  - 1 pair of *sex* chromosomes (chrX+chrX or chrX+chrY)
  - Each chromosome contains many **genes**
- **Gene**
  - basic physical and functional units of heredity.
  - specific sequences of DNA that encode instructions on how to make **proteins**.
- **Proteins**
  - Make up the cellular structure
  - large, complex molecules made up of smaller subunits called **amino acids**.

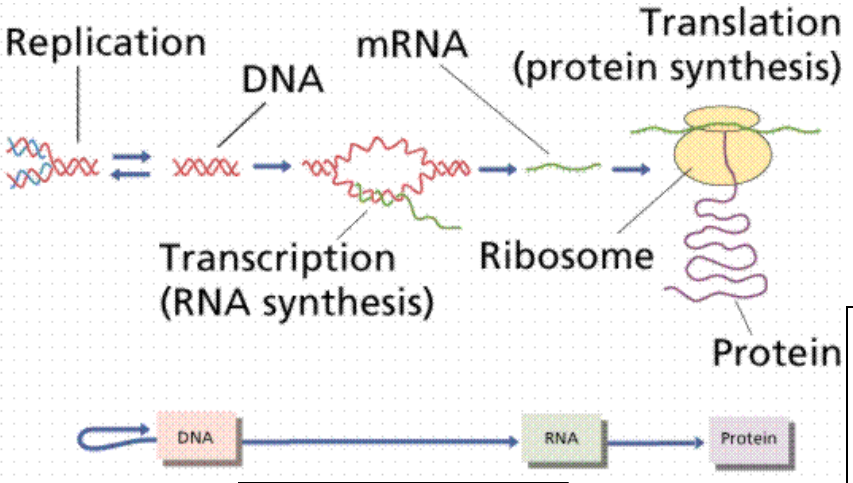
All life depends on 3 critical molecules

- DNAs
  - Hold information on how cell works
- RNAs
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body's major components (e.g. hair, skin, etc.)

Central Dogma of Biology

The information for making proteins is stored in DNA. There is a process (transcription and translation) by which DNA is converted to protein. By understanding this process and how it is regulated we can make predictions and models of cells.

Assembly



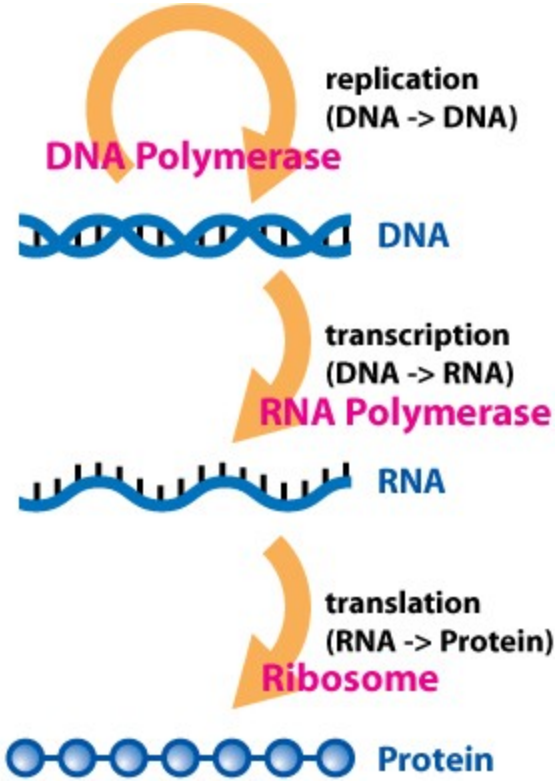
Sequence analysis

Gene Finding

Protein Sequence Analysis

Central dogma

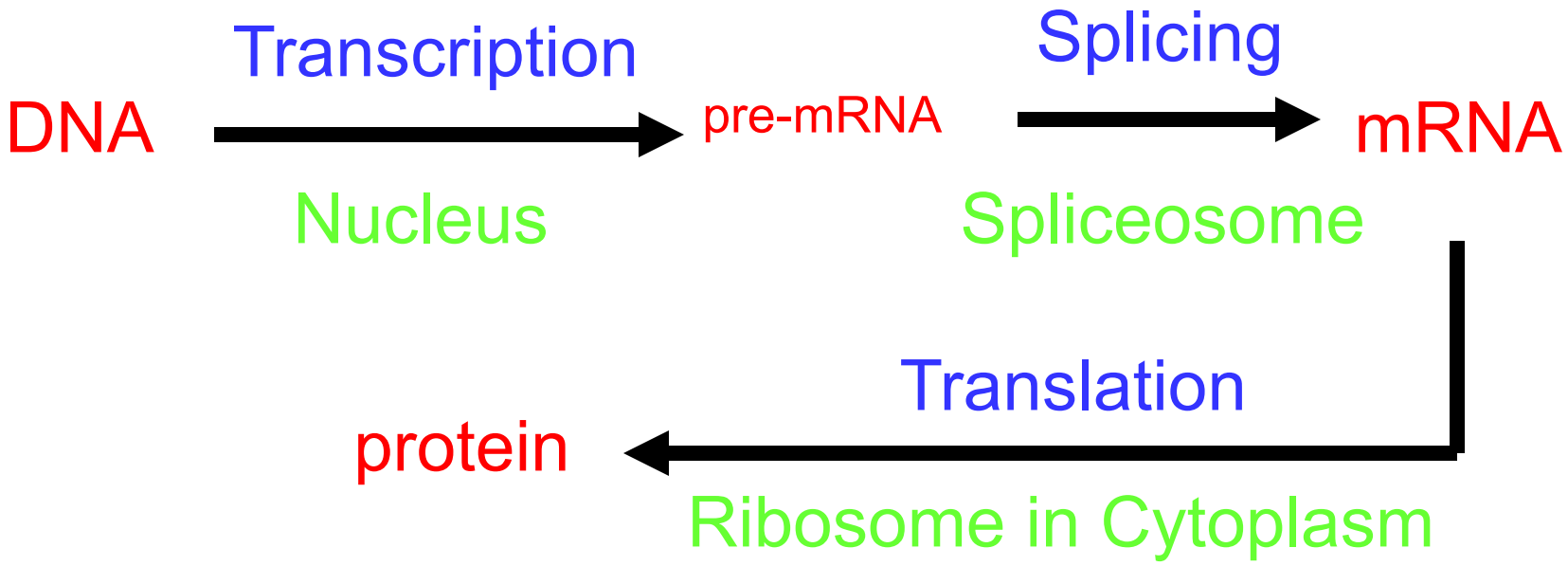
1970 F. Crick



Transcription: RNA synthesis

Translation: Protein synthesis

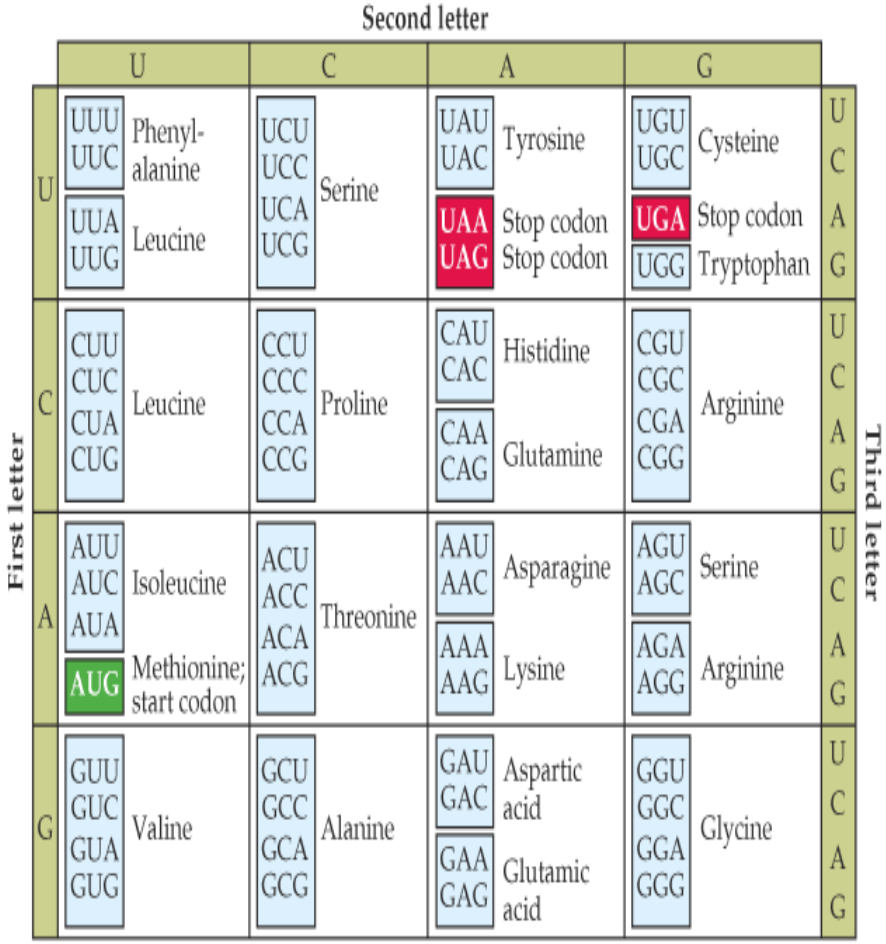
Central dogma



- Base Pairing Rule: A and T or U is held together by 2 hydrogen bonds and G and C is held together by 3 hydrogen bonds.
- Note: Some RNA stays as RNA (ie tRNA,rRNA, miRNA, snoRNA, etc.).

Cell Information: Instruction book of Life

- DNA, RNA, and Proteins are examples of strings written in either the four-letter nucleotide of DNA and RNA (A C G T/U)
- or the twenty-letter amino acid of proteins. Each amino acid is coded by 3 nucleotides called **codon**. (Leu, Arg, Met, etc.)



First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phenylalanine	UCU, UCC, UCA, UCG: Serine	UAU, UAC: Tyrosine	UGU, UGC: Cysteine	U C A G
	UUA, UUG: Leucine		UAA: Stop codon; UAG: Stop codon	UGA: Stop codon; UGG: Tryptophan	
C	CUU, CUC, CUA, CUG: Leucine	CCU, CCC, CCA, CCG: Proline	CAU, CAC: Histidine; CAA, CAG: Glutamine	CGU, CGC, CGA, CGG: Arginine	U C A G
A	AUU, AUC, AUA: Isoleucine	ACU, ACC, ACA, ACG: Threonine	AAU, AAC: Asparagine	AGU, AGC: Serine	U C A G
	AUG: Methionine; start codon		AAA, AAG: Lysine	AGA, AGG: Arginine	
G	GUU, GUC, GUA, GUG: Valine	GCU, GCC, GCA, GCG: Alanine	GAU, GAC: Aspartic acid; GAA, GAG: Glutamic acid	GGU, GGC, GGA, GGG: Glycine	U C A G

Alphabets

DNA:

$\sum$ = {A, C, G, T}

A pairs with T; G pairs with C

RNA:

$\sum$ = {A, C, G, U}

A pairs with U; G pairs with C

Protein:

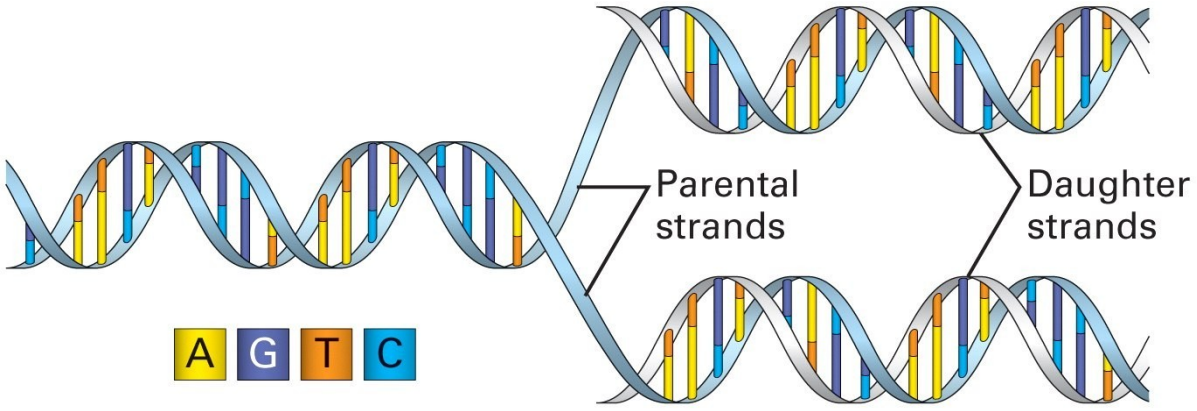
$\sum$ = {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y} *and*

B = N | D

Z = Q | E

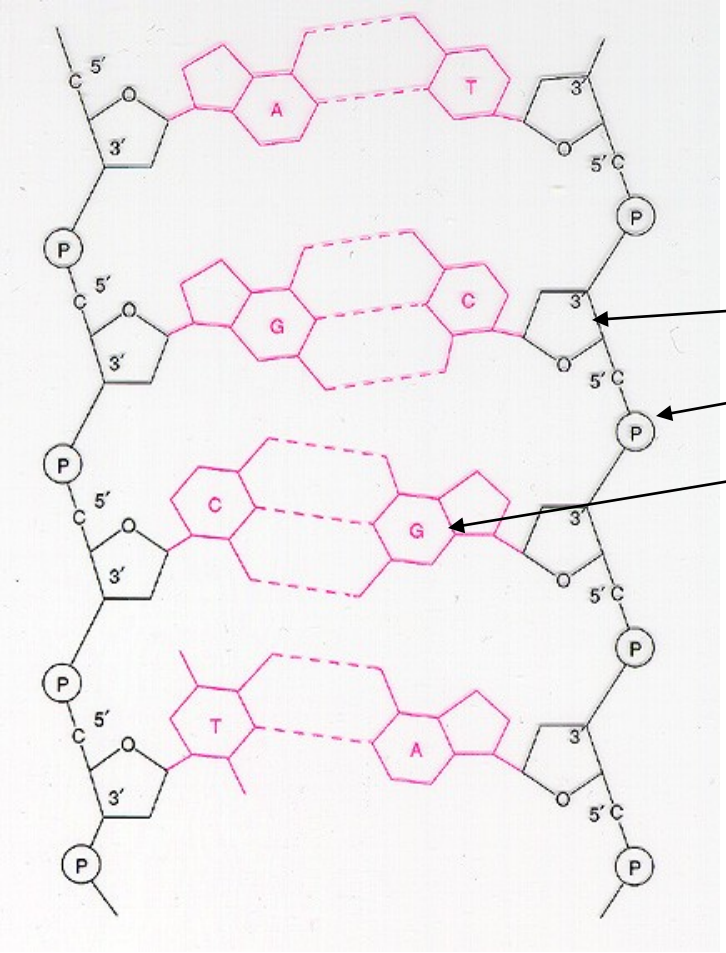
X = *any*

DNA: The Code of Life



- The structure and the four genomic letters code for all living organisms
- Adenine, Guanine, Thymine, and Cytosine which pair A-T and C-G on complimentary strands.

DNA, continued



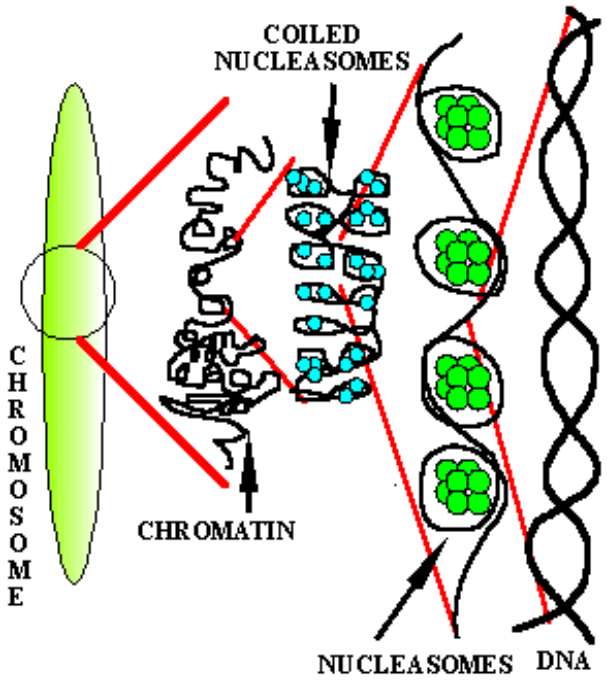
- DNA has a double helix structure which composed of
  - sugar molecule
  - phosphate group
  - and a base (A,C,G,T)
- DNA always reads from 5’ end to 3’ end for transcription replication

  5’ ATTTAGGCC 3’

  3’ TAAATCCGG 5’

DNA: The Basis of Life

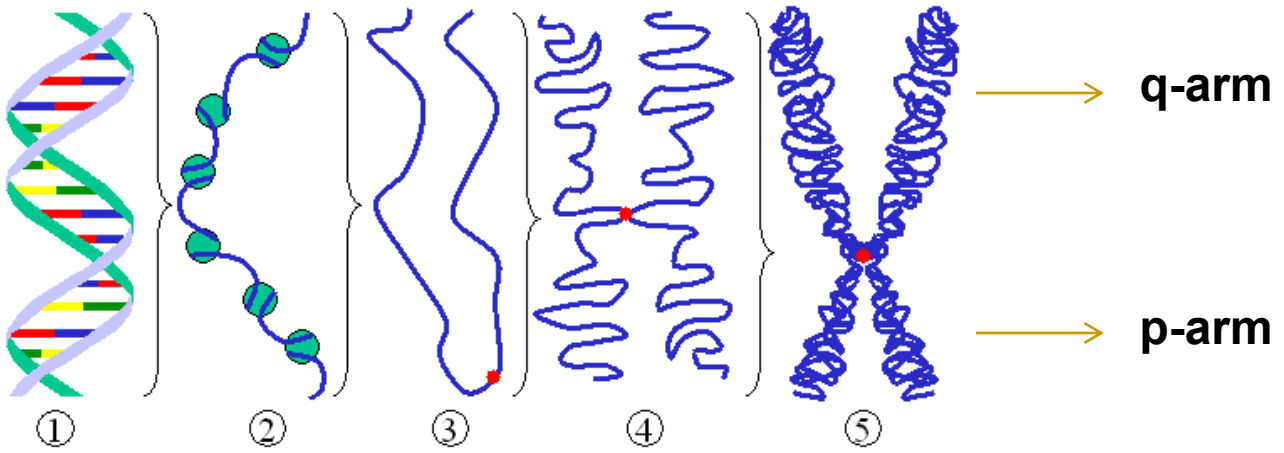
- Humans have about 3 billion base pairs.
  - How do you package it into a cell?
  - How does the cell know where in the highly packed DNA where to start transcription?
    - Special regulatory sequences
  - DNA size does not mean more complex
- Complexity of DNA
  - Eukaryotic genomes consist of variable amounts of DNA
    - Single Copy or Unique DNA
    - Highly Repetitive DNA



DNA is organized into Chromosomes

- Chromosomes:
  - Found in the nucleus of the cell which is made from a long strand of DNA, "packaged" by proteins called *histones*. Different organisms have a different number of chromosomes in their cells.
  - Human genome has 23 pairs of chromosomes
    - 22 pairs of *autosomal* chromosomes (chr1 to chr22)
    - 1 pair of *sex* chromosomes (chrX+chrX or chrX+chrY)
- Ploidy: number of sets of chromosomes
  - Haploid (n): one of each chromosome
    - Sperm & egg cells; hydatidiform mole
  - Diploid (2n): two of each chromosome
    - All other cells in mammals (human, chimp, cat, dog, etc.)
  - Triploid (3n), Tetraploid (4n), etc.
    - Tetraploidy is common in plants

Genetic Information: Chromosomes



- (1) Double helix DNA strand.
- (2) Chromatin strand (**DNA** with **histones**)
- (3) Condensed chromatin during interphase with **centromere**.
- (4) Condensed chromatin during prophase
- (5) Chromosome during metaphase

Chromosomes

Organism	Number of base pairs	number of chromosomes (n)
Prokayotic		
Escherichia coli (bacterium)	$4x10^6$	1
Eukaryotic		
Saccharomyces cerevisiae (yeast)	$1.35x10^7$	17
Drosophila melanogaster (fruit fly)	$1.65x10^8$	4
Homo sapiens(human)	$2.9x10^9$	23
Zea mays(corn)	$5.0x10^9$	10

Genome "table of contents"

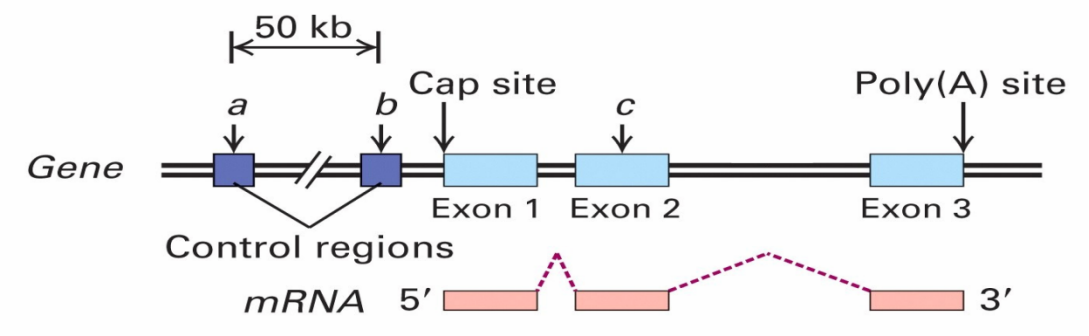
- Genes (~35%; but only 1% are coding exons)
  - Protein coding
  - Non-coding (ncRNA only)
- Pseudogenes: genes that lost their expression ability:
  - Evolutionary loss
  - Processed pseudogenes
- Repeats (~50%)
  - Transposable elements: sequence that can copy/paste themselves. Typically of virus origin.
  - Satellites (short tandem repeats [STR]; variable number of tandem repeats [VNTR])
  - Segmental duplications (5%)
    - Include genes and other repeat elements within

Genes

- What are genes?
  - Mendel definition: physical and functional traits that are passed on from one generation to the next.
- Genes were discovered by Gregor Mendel in the 1860s while he was experimenting with the pea plant. He asked the question:

Do traits come from a blend of both parent's traits or from only one parent?

Genes



- Regulatory regions: up to 50 kb upstream of +1 site
- Exons: protein coding and untranslated regions (UTR)
  1 to 178 exons per gene (mean 8.8)
  8 bp to 17 kb per exon (mean 145 bp)
- Introns: splice acceptor and donor sites, junk DNA
  average 1 kb – 50 kb per intron
- Gene size: Largest – 2.4 Mb (Dystrophin). Mean – 27 kb.

Genes can be switched on/off

- In an adult multicellular organism, there is a wide variety of cell types seen in the adult. eg, muscle, nerve and blood cells.
- The different cell types contain the *same* DNA.
- This differentiation arises because different cell types express different genes.
- Type of gene regulation mechanisms:
  - Promoters, enhancers, methylation, RNAi, etc.

Pseudogenes

- “Dead” genes that lost their coding ability
- Evolutionary process:
  - Mutations cause:
    - Early stop codons
    - Loss of promoter / enhancer sequence
- Processed pseudogenes:
  - A real gene is transcribed to mRNA, introns are spliced out, then reverse transcribed into cDNA
  - This cDNA is then reintegrated into the nuclear genome

Repeats

- Transposons (mobile elements): generally of viral origin, integrated into genomes millions of years ago
- Can copy/paste; most are fixed, some are still active
  - Retrotransposon: intermediate step that involves transcription (RNA)
  - DNA transposon: no intermediate step

Retrotransposons

- LTR: long terminal repeat
- Non-LTR:
  - LINEs: Long Interspersed Nucleotide Elements
    - L1 (~6 kbp full length, ~900 bp trimmed version): Approximately 17% of human genome
      - They encode genes to copy themselves
  - SINEs: Short Interspersed Nucleotide Elements
    - *Alu* repeats (~300 bp full length): Approximately 1 million copies = ~10% of the genome
      - They use cell's machinery to replicate
      - Many subfamilies; AluY being the most active, AluJ most ancient

Satellites

- Microsatellites (STR=short tandem repeats) 1-10 bp
  - Used in population genetics, paternity tests and forensics
- Minisatellites (VNTR=variable number of tandem repeats): 10-60 bp
- Other satellites
  - Alpha satellites: centromeric/pericentromeric, 171bp in humans
  - Beta satellites: centromeric (some), 68 bp in humans
  - Satellite I (25-68 bp), II (5bp), III (5 bp)

Segmental duplications

- Low-copy repeats, >1 kbp & > 90% sequence identity between copies
- Covers ~5% of the human genome
  - Both tandem and interspersed in humans, about half inter chromosomal duplications
  - Tandem in mice, no inter chromosomal duplications
- Gene rich
- Provides elasticity to the genome:
  - More prone to rearrangements (and causal)
  - Gene innovation through duplication: Ohno, 1970

Human Genome Composition

TABLE 10-1 Major Classes of Eukaryotic DNA and Their Representation in the Human Genome

Class	Length	Copy Number in Human Genome	Fraction of Human Genome, %
Protein-coding genes			
Solitary genes	Variable	1	≈15* (0.8)†
Duplicated or diverged genes in gene families	Variable	2–≈1000	≈15* (0.8)†
Tandemly repeated genes encoding rRNAs, tRNAs, snRNAs, and histones	Variable	20–300	0.3
Repetitious DNA			
Simple-sequence DNA	1–500 bp	Variable	3
Interspersed repeats			
DNA transposons	2–3 kb	300,000	3
LTR retrotransposons	6–11 kb	440,000	8
Non-LTR retrotransposons			
LINEs	6–8 kb	860,000	21
SINEs	100–300 bp	1,600,000	13
Processed pseudogenes	Variable	1–≈100	≈0.4
Unclassified spacer DNA	Variable	n.a.‡	≈25

*Complete transcription units, including introns.
†Protein-coding exons. The total number of human protein-coding genes is estimated to be 30,000–35,000, but this number is based on current methods for identifying genes in the human genome sequence and may be an underestimate.
‡Not applicable.

SOURCE: E. S. Lander et al., 2001, *Nature* **409**:860.

Genetic variation

- Changes in DNA sequence
  - Many types of variation
    - SNPs: single nucleotide polymorphism
    - Indels (1 – 50 bp)
    - Structural variation (>50 bp)
    - Chromosomal changes
      - Monosomy, uniparental disomy, trisomy, etc.

If a mutation occurs in a codon:

- Synonymous mutations: Coded amino acid doesn’t change
- Nonsynonymous mutations: Coded amino acid changes

The Good, the Bad, and the Silent

- Mutations can serve the organism in three ways:

A mutation can cause a trait that enhances the organism's function:

- The Good : Mutation in the sickle cell gene provides resistance to malaria.

- The Bad : A mutation can cause a trait that is harmful, sometimes fatal to the organism:

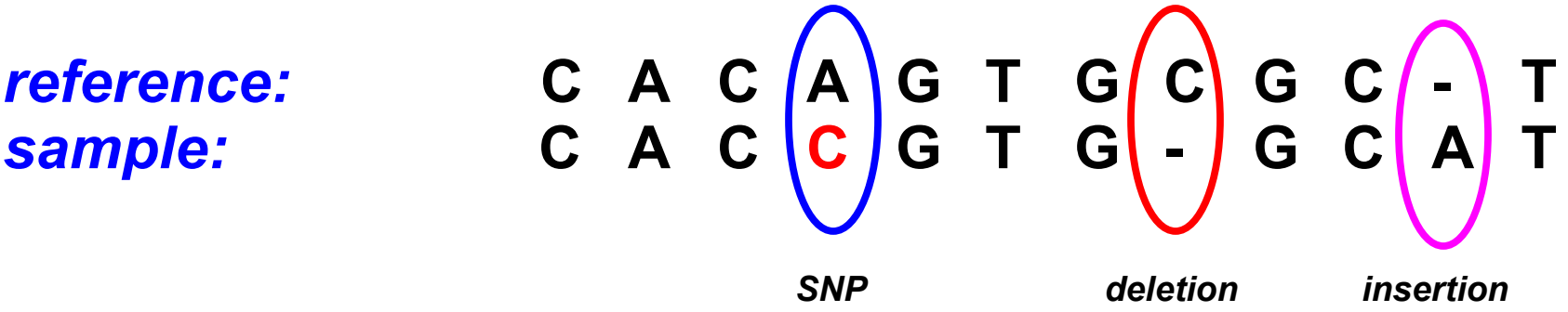
Huntington's disease, a symptom of a gene mutation, is a degenerative disease of the nervous system.

- The Silent: A mutation can simply cause no difference in the function of the organism.

Campbell, Biology, 5th edition, p. 255

SNPs & indels

SNP: Single nucleotide polymorphism (substitutions)
Short indel: Insertions and deletions of sequence of length 1 to 50 basepairs



- Nonsense mutations: create a stop signal in a gene before its natural stop (disease: thalassemia).
- Missense mutations: changes the gene sequence, produces a different protein (disease: ALS).
- Frameshift: caused by indels, shifts basepairs that changes codon order (disease: hypercholesterol).

Short tandem repeats

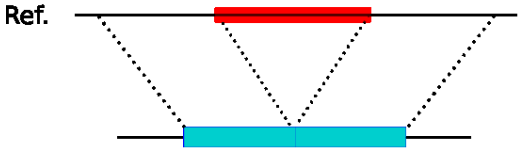
reference: **C A G C A G C A G C A G**

sample: **C A G C A G C A G *C A G* C A G**

- Microsatellites (STR=short tandem repeats) 1-10 bp
  - Used in population genetics, paternity tests and forensics
- Minisatellites (VNTR=variable number of tandem repeats): 10-60 bp
- Other satellites
  - Alpha satellites: centromeric/pericentromeric, 171bp in humans
  - Beta satellites: centromeric (some), 68 bp in humans
  - Satellite I (25-68 bp), II (5bp), III (5 bp)
- Disease relevance:
  - Fragile X Syndrome
  - Huntington's disease

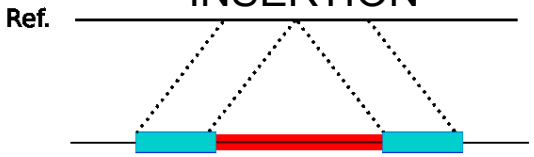
Structural Variation

DELETION

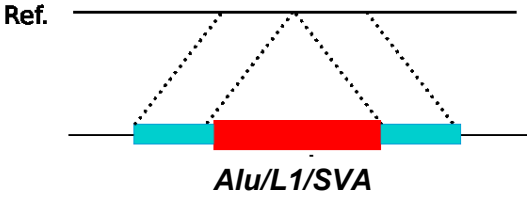


Autism, mental retardation, Crohn's

NOVEL SEQUENCE INSERTION

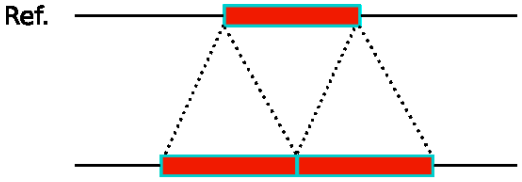


MOBILE ELEMENT INSERTION

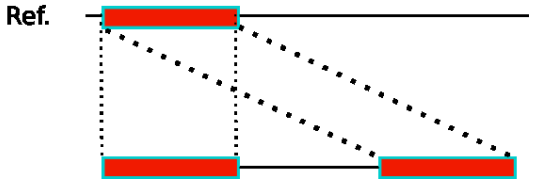


Haemophilia

TANDEM DUPLICATION

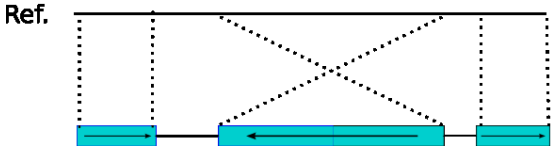


INTERSPERSED DUPLICATION

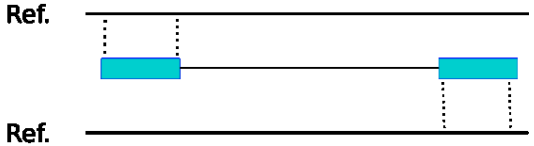


Schizophrenia, psoriasis

INVERSION



TRANSLOCATION

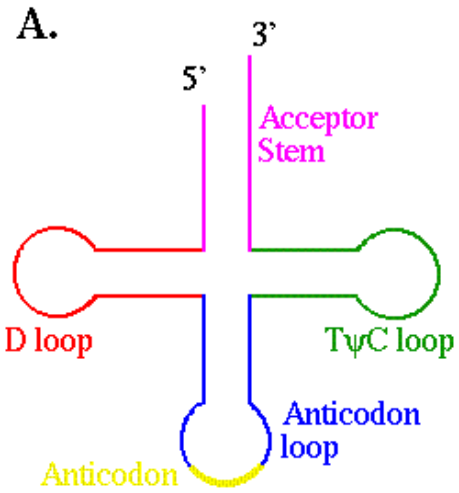


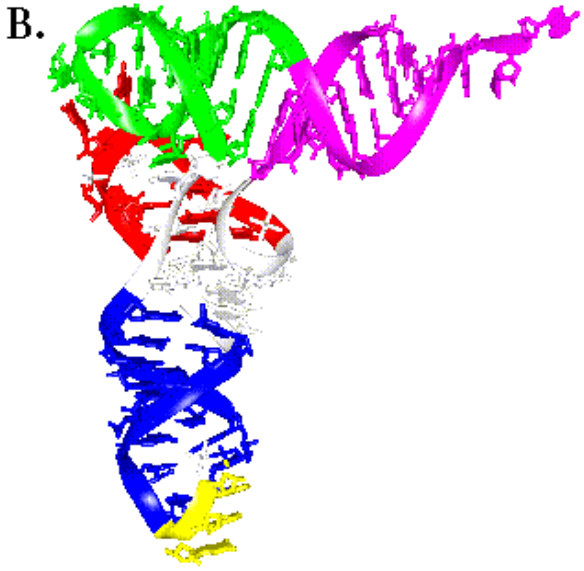
Chronic myelogenous leukemia

RNA & PROTEIN

RNA

- RNA is similar to DNA chemically. It is usually only a single strand. T(hymine) is replaced by U(racil)
- Some forms of RNA can form secondary structures by "pairing up" with itself. This can have change its properties





DNA and RNA can pair with each other.

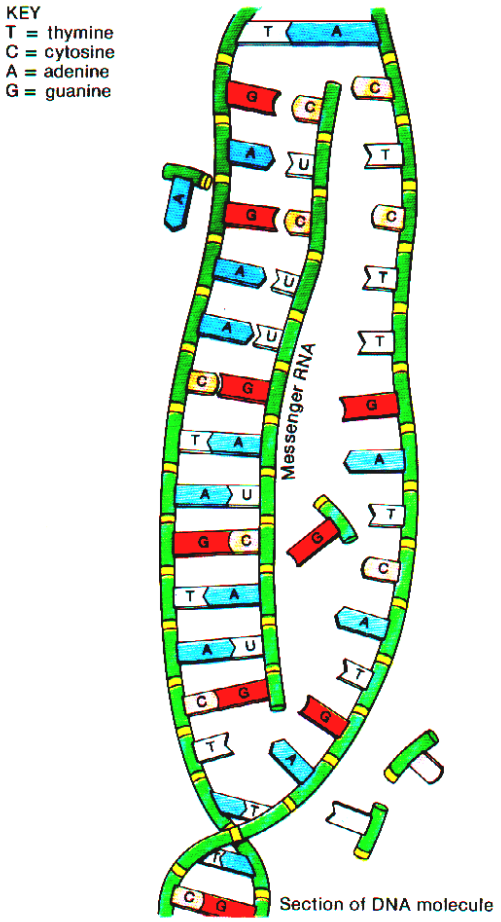
tRNA linear and 3D view: http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif

RNA, continued

- Several types exist, classified by function
  - mRNA – this is what is usually being referred to when a Bioinformatician says “RNA”. This is used to carry a gene’s *m*essage out of the nucleus.
  - tRNA – *t*ransfers genetic information from mRNA to an amino acid sequence
  - rRNA – *r*ibosomal RNA. Part of the ribosome which is involved in translation.
  - Non-coding RNAs (ncRNA): not translated into proteins, but they can regulate translation
    - miRNA, siRNA, snoRNA, piRNA, lncRNA

Transcription

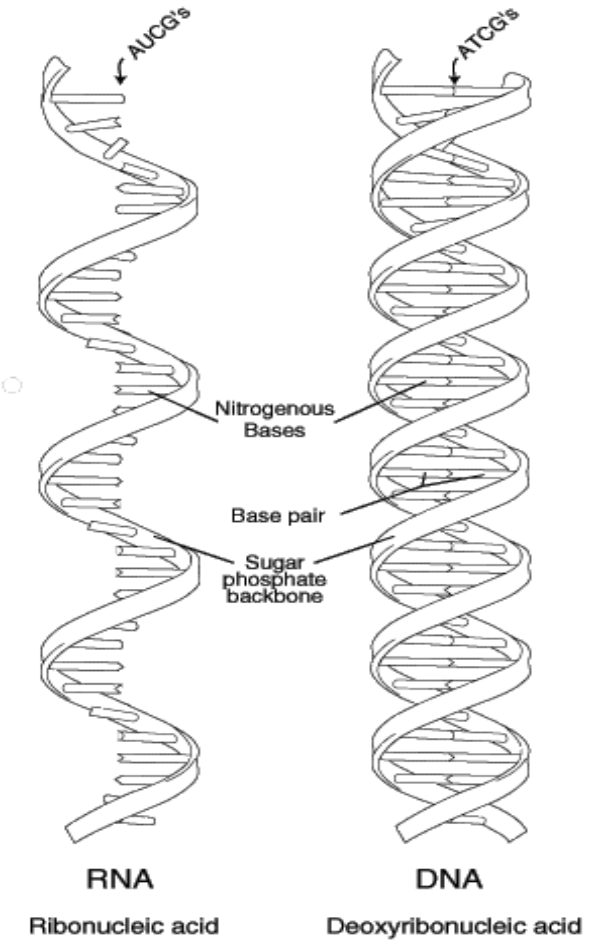
- The process of making RNA from DNA
- Catalyzed by “transcriptase” enzyme
- Needs a promoter region to begin transcription.
- ~50 base pairs/second in bacteria, but multiple transcriptions can occur simultaneously



http://ghs.gresham.k12.or.us/science/ps/sci/ibbio/chem/nucleic/chpt15/transcription.gif

DNA → RNA: Transcription

- DNA gets transcribed by a protein known as *RNA-polymerase*
- This process builds a chain of bases that will become mRNA
- RNA and DNA are similar, except that RNA is single stranded and thus less stable than DNA
  - Also, in RNA, the base uracil (U) is used instead of thymine (T), the DNA counterpart



Transcription, continued

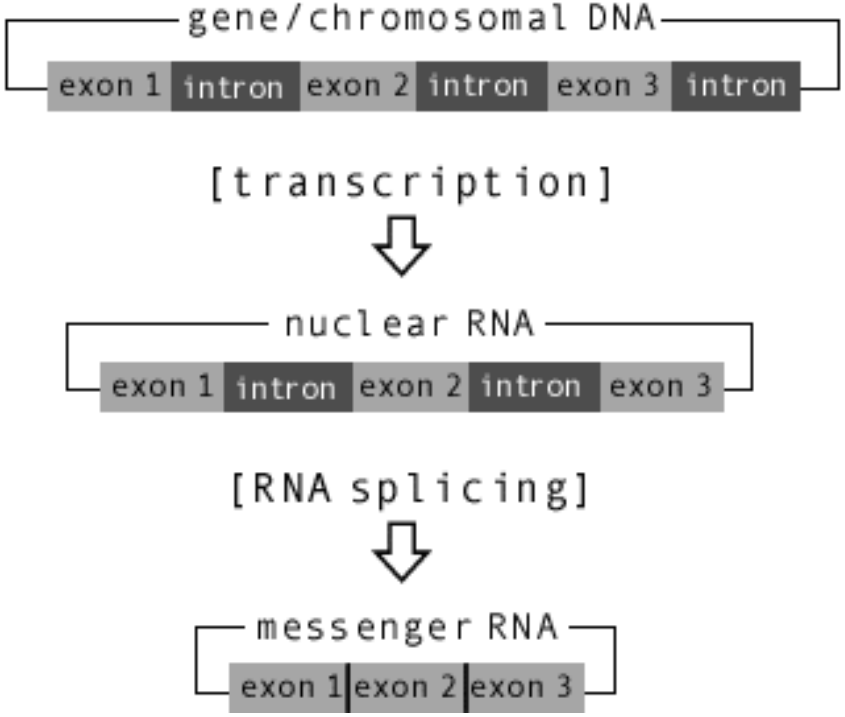
- Transcription is highly regulated. Most DNA is in a dense form where it cannot be transcribed.
- To begin transcription requires a promoter, a small specific sequence of DNA to which polymerase can bind (~40 base pairs “upstream” of gene)
- Finding these promoter regions is a partially solved problem that is related to motif finding.
- There can also be repressors and inhibitors acting in various ways to stop transcription. This makes regulation of gene transcription complex to understand.

Splicing and other RNA processing

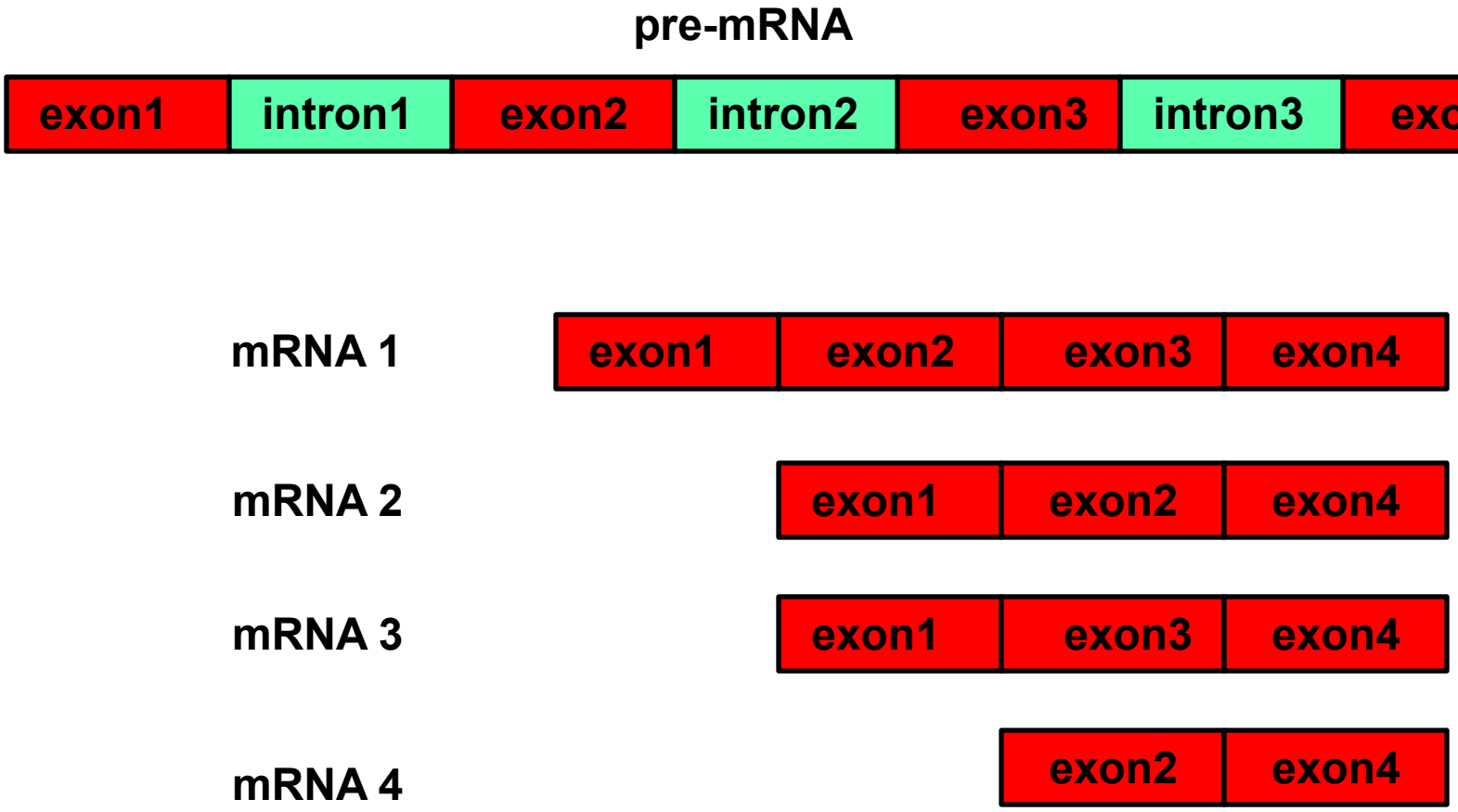
- In Eukaryotic cells, RNA is processed between transcription and translation.
- This complicates the relationship between a DNA gene and the protein it codes for.
- Sometimes alternate RNA processing can lead to an alternate protein as a result. This is true in the immune system.

Splicing (Eukaryotes)

- Unprocessed RNA is composed of Introns and Extrons. Introns are removed before the rest is expressed and converted to protein.
- Sometimes alternate splicings can create different valid proteins.
- A typical Eukaryotic gene has 4-20 introns. Locating them by analytical means is not easy.



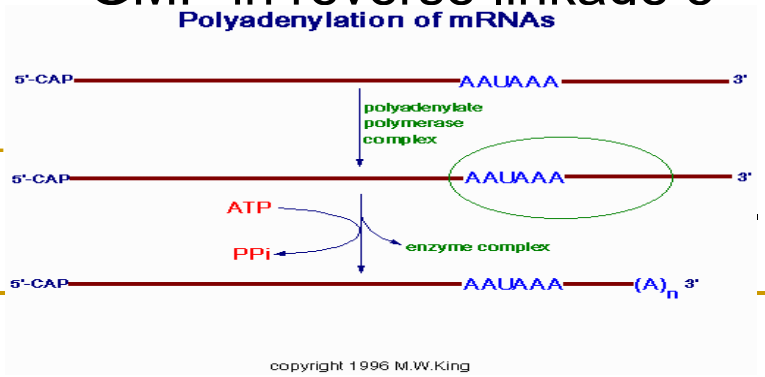
Alternative splicing



Posttranscriptional Processing: Capping and Poly(A) Tail

Capping

- Prevents 5’ exonucleolytic degradation.
- 3 reactions to cap:

1. Phosphatase removes 1 phosphate from 5’ end of pre-mRNA
2. Guanyl transferase adds a GMP in reverse linkage 5’
3. 

Poly(A) Tail

- Due to transcription termination process being imprecise.
- 2 reactions to append:

1. Transcript cleaved 15-25 past highly conserved AAUAAA sequence and less than 50 nucleotides before less conserved U rich or GU rich sequences.
2. Poly(A) tail generated from ATP by poly(A) polymerase which is activated by cleavage and polyadenylation specificity factor (CPSF) when CPSF recognizes AAUAAA. Once poly(A) tail has grown approximately 10 residues, CPSF disengages from the recognition site.

Proteins: Workhorses of the Cell

- 20 different **amino acids**
  - different chemical properties cause the protein chains to fold up into specific three-dimensional structures that define their particular functions in the cell.
- Proteins do all essential work for the cell
  - build cellular structures
  - digest nutrients
  - execute metabolic functions
  - Mediate information flow within a cell and among cellular communities.
- Proteins work together with other proteins or nucleic acids as "molecular machines"
  - structures that fit together and function in highly specific, lock-and-key ways.

Uncovering the code

- Scientists conjectured that proteins came from DNA; but how did DNA code for proteins?
- If one nucleotide codes for one amino acid, then there would be 4^1 amino acids
- However, there are 20 amino acids, so at least 3 bases codes for one amino acid, since $4^2 = 16$ and $4^3 = 64$
  - This triplet of bases is called a “codon”
  - 64 different codons and only 20 amino acids means that the coding is degenerate: more than one codon sequence code for the same amino acid

RNA → Protein: Translation

- Ribosomes and *transfer-RNAs* (tRNA) run along the length of the newly synthesized mRNA, decoding one codon at a time to build a growing chain of amino acids ("peptide")
  - The tRNAs have anti-codons, which complimentarily match the codons of mRNA to know what protein gets added next
- But first, in eukaryotes, a phenomenon called splicing occurs
  - Introns are non-protein coding regions of the mRNA; exons are the coding regions
  - Introns are removed from the mRNA during splicing so that a functional, valid protein can form

Translation

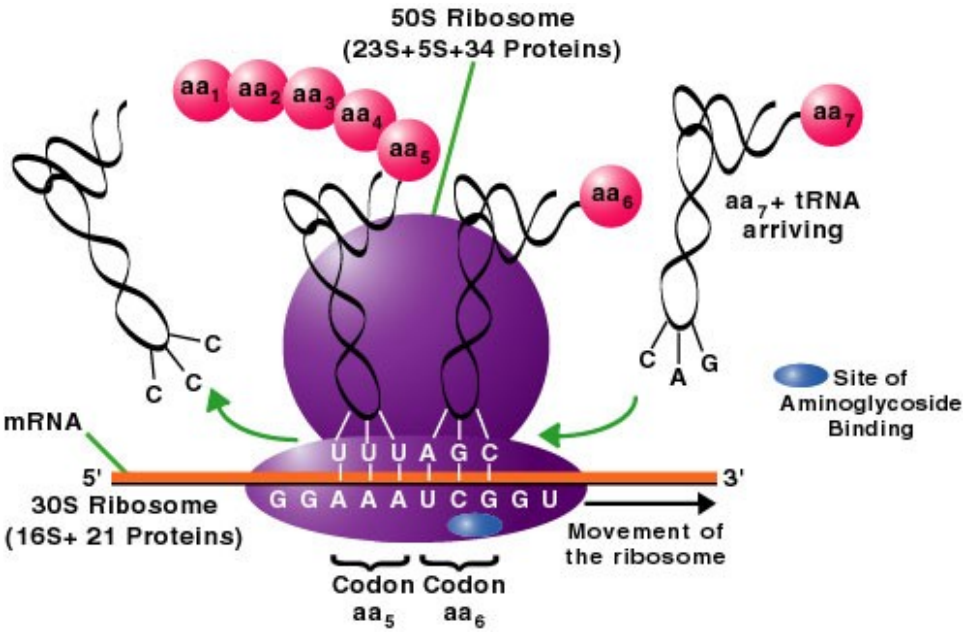
- The process of going from RNA to polypeptide.
- Three base pairs of RNA (called a codon) correspond to one amino acid based on a fixed table.
- Always starts with Methionine and ends with a stop codon

FIRST POSITION	SECOND POSITION: U	SECOND POSITION: C	SECOND POSITION: A	SECOND POSITION: G	THIRD POSITION
U	phenyl-alanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
	* methionine		lysine	arginine	A
					G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

* and start

Translation, continued

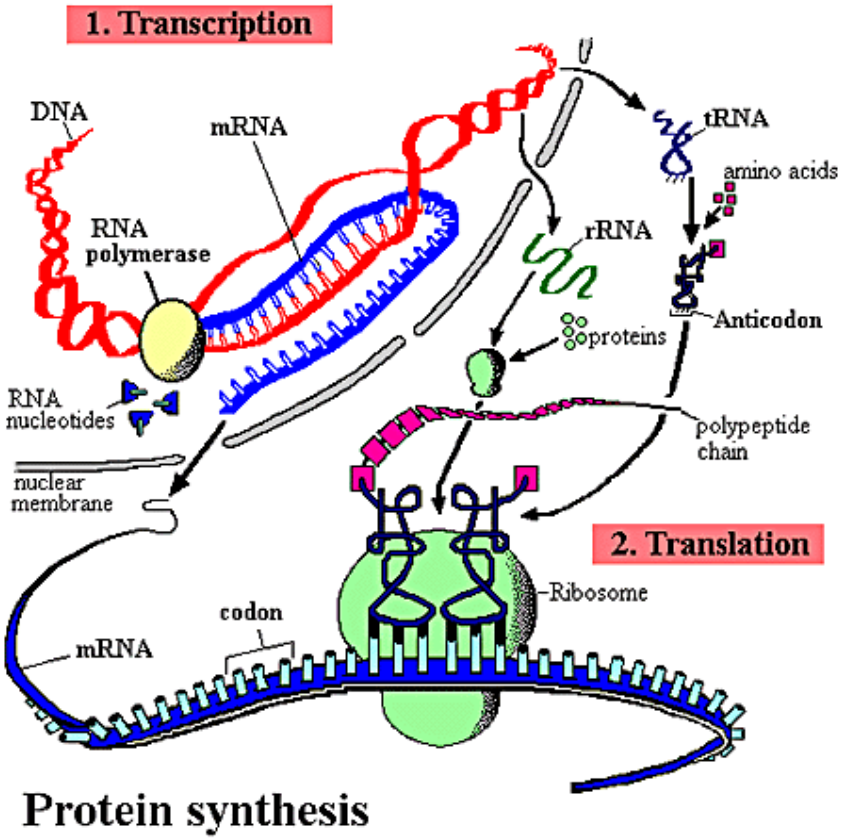
- Catalyzed by Ribosome
- Using two different sites, the Ribosome continually binds tRNA, joins the amino acids together and moves to the next location along the mRNA
- ~10 codons/second, but multiple translations can occur simultaneously



http://wong.scripps.edu/PIX/ribosome.jpg

Protein Synthesis: Summary

- There are twenty amino acids, each coded by three-base-sequences in DNA, called “codons”
  - This code is degenerate
- The **central dogma** describes how proteins derive from DNA
  - DNA → mRNA → (splicing?) → protein
- The protein adopts a 3D structure specific to it’s amino acid arrangement and function



Protein synthesis

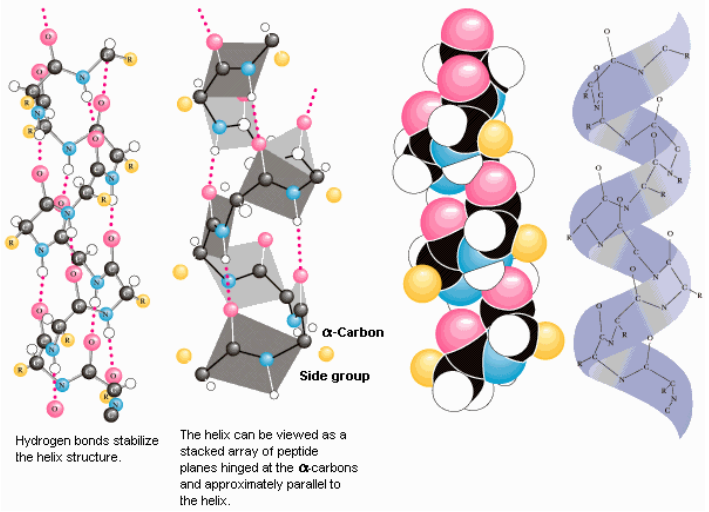
Proteins

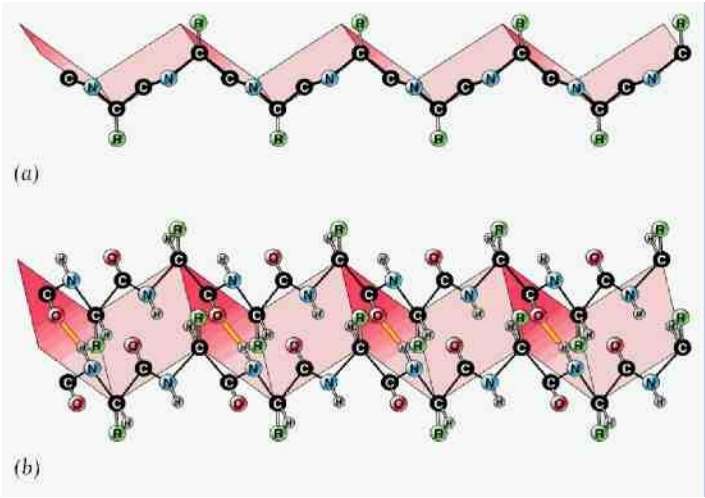
- Complex organic molecules made up of amino acid subunits
- 20* different kinds of amino acids. Each has a 1 and 3 letter abbreviation.
- http://www.indstate.edu/thcme/mwking/amino-acids.html for complete list of chemical structures and abbreviations.
- Proteins are often enzymes that catalyze reactions.
- Also called “poly-peptides”

*Some other amino acids exist but not in humans.

Protein Folding

- Proteins tend to fold into the lowest free energy conformation.
- Proteins begin to fold while the peptide is still being translated.
- Proteins bury most of its hydrophobic residues in an interior core to form an α helix.
- Most proteins take the form of secondary structures α helices and β sheets.
- Molecular chaperones, hsp60 and hsp 70, work with other proteins to help fold newly synthesized proteins.
- Much of the protein modifications and folding occurs in the endoplasmic reticulum and mitochondria.





Protein Folding

- Proteins are not linear structures, though they are built that way
- The amino acids have very different chemical properties; they interact with each other after the protein is built
  - This causes the protein to start fold and adopting it's functional structure
  - Proteins may fold in reaction to some ions, and several separate chains of peptides may join together through their hydrophobic and hydrophilic amino acids to form a polymer

Protein Folding (cont'd)

- The structure that a protein adopts is vital to it's chemistry
- Its structure determines which of its amino acids are exposed carry out the protein's function
- Its structure also determines what substrates it can react with