EE550
Computational Biology

Week 2 Course Notes

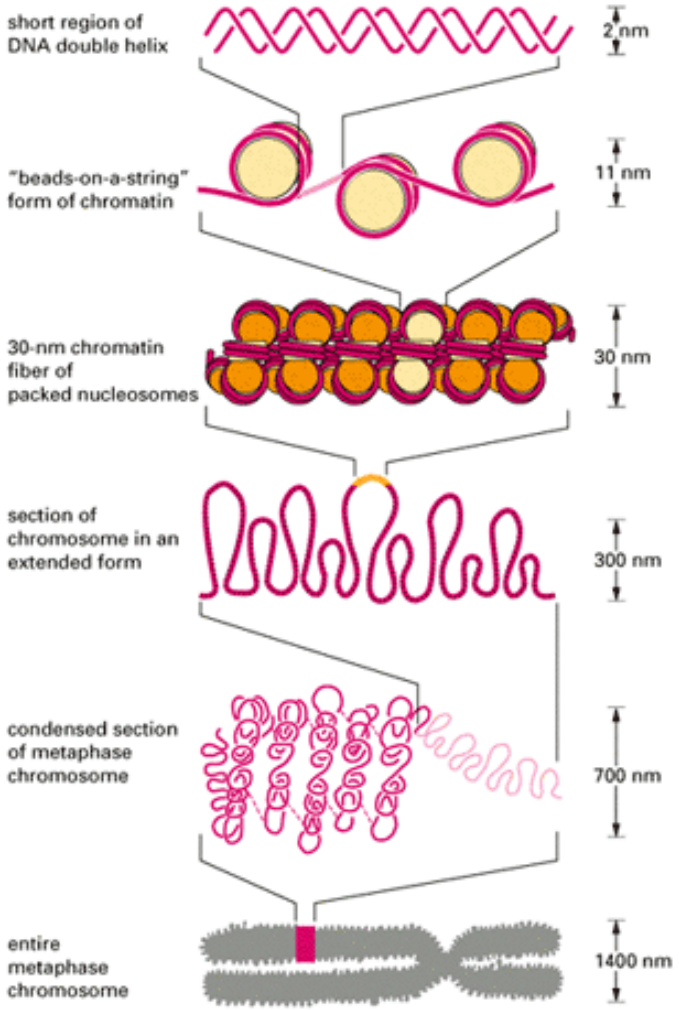
Instructor: Bilge Karaçalı, PhD

Topics

- Nucleic acid and protein structure
  - Nucleic acids
    - DNA
    - RNA
  - Proteins
    - Amino acids
    - Polypeptides
  - Biological information flow

DNA

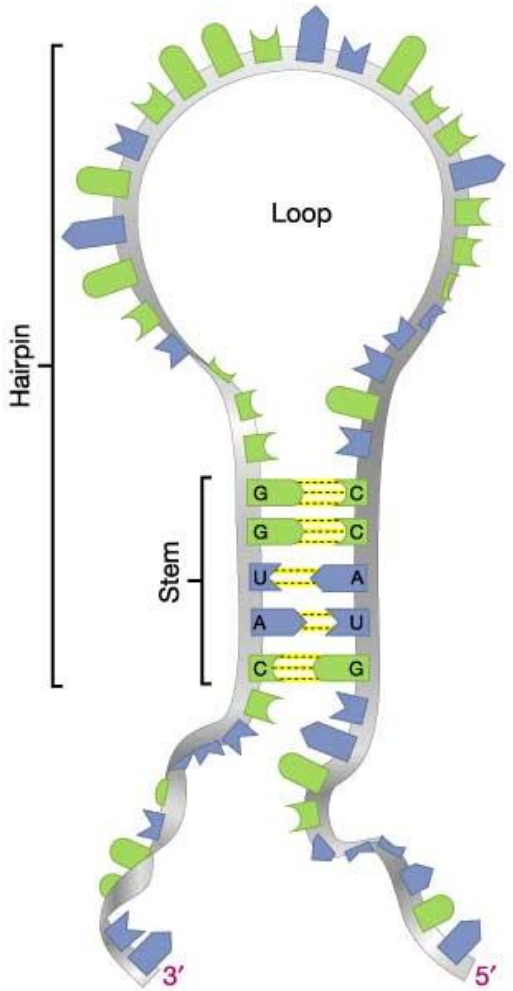
- Deoxyribonucleic acid encodes the genetic code
- It resides in the nucleus in the form of a paired linear sequence of four nucleotides
  - Adenine (A)
  - Guanine (G)
  - Cytosine (C)
  - Thymine (T)
- The linear sequence is formed by covalent bonds between successive nucleotides
- The DNA sequence has directionality
  - The carbon atoms in nucleotide bases are numbered
  - The covalent bond forms between the 3rd carbon of one base and the 5th carbon of another
  - The flow of the sequence is denoted from the 5´ end (upstream) to the 3´ end (downstream)



Source: http://library.thinkquest.org/C004535/media/chromosome_packing.gif

RNA

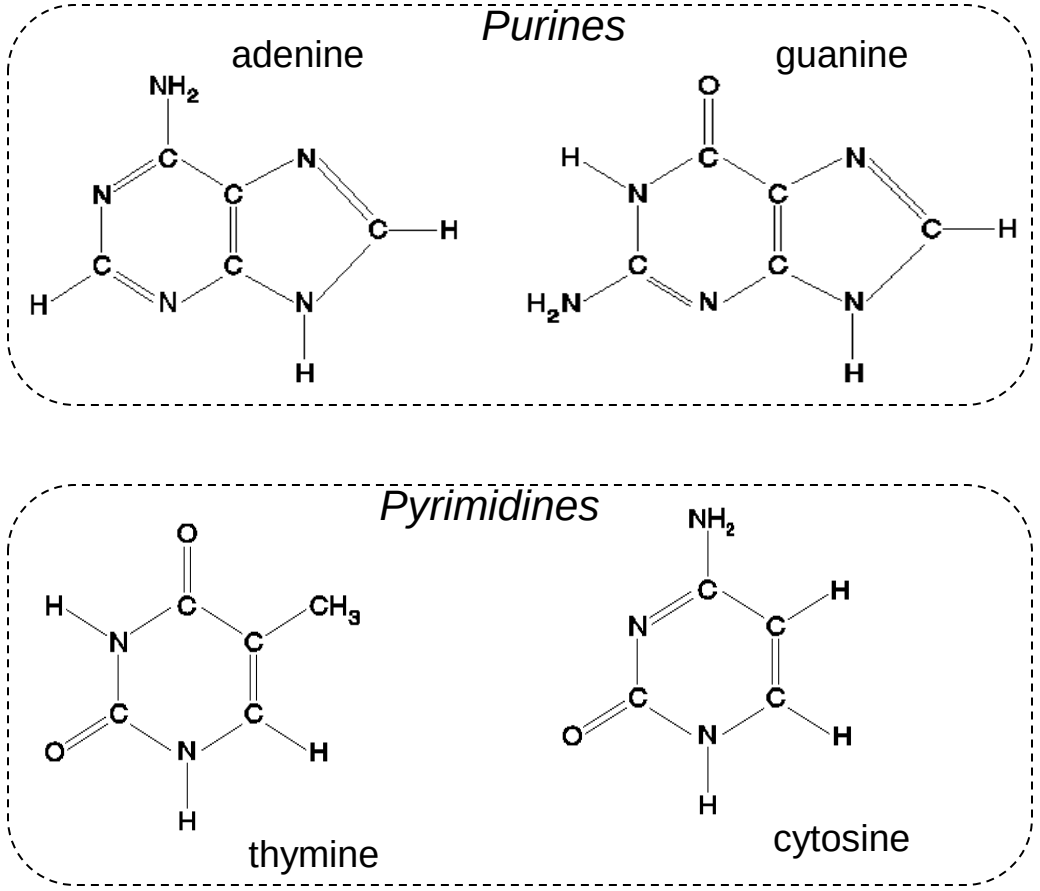
- Ribonucleic acid carries the genetic information from the nucleus to the cytoplasm
- It consists of a single strand of nucleotides
  - Thymine (T) is replaced by Uracil (U)
- It is synthesized by the transcription process that generates a complementary copy of a gene
  - Transcription produces the messenger RNA (mRNA) that encodes the proteins
- There are also non-coding types of RNA
  - Transfer RNA (tRNA)
  - Ribosomal RNA (rRNA)
  - MicroRNA (miRNA)
  - …



Source: http://www.uic.edu/classes/bios/bios100/summer2002/rna-loop.jpg

Nucleic Acid Structure

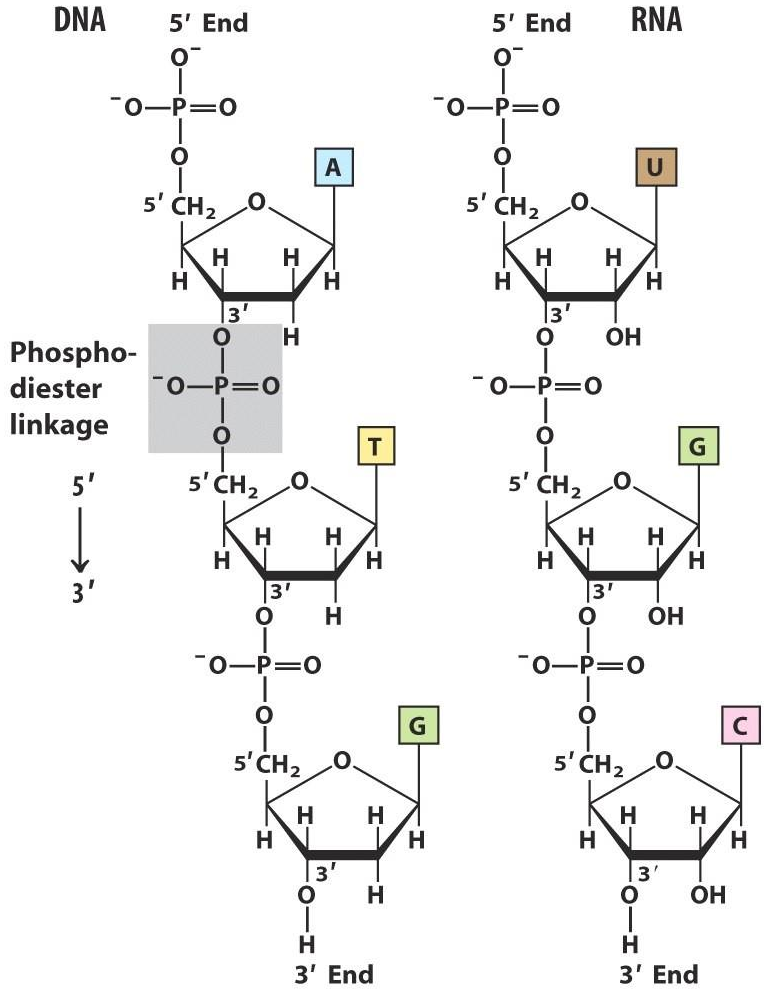
- Nucleotides are formed by sugar, phosphate and base
  - In the DNA, the sugar is the deoxyribose
  - In the RNA, the sugar is the ribose
  - The base determines the type of the nucleotide
- In the RNA, thymine is replaced by uracil that lacks the methyl group



Source: http://hal.wzw.tum.de/genglos/asp/genreq.asp?nr=155

Nucleic Acid Chains

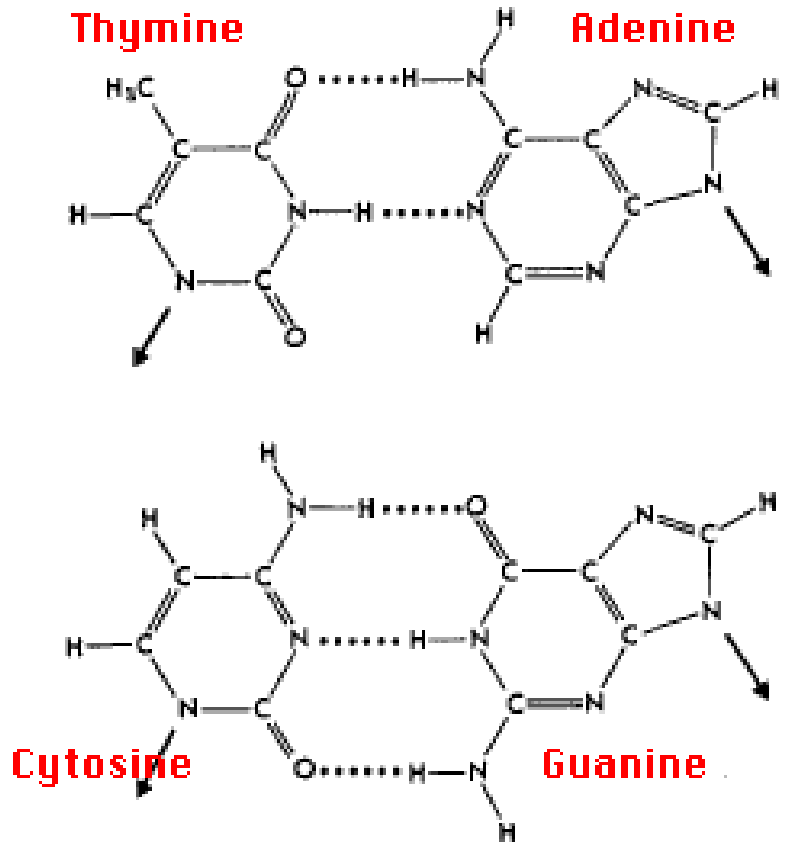
- Nucleotides in DNA and RNA are joined by phosphate ester bonds
  - between the phosphate component of one nucleotide and the sugar component of the next nucleotide (C-O)
- The phosphate binds the carbon no 3 in one sugar and the carbon no 5 in the other
- The directionality of the DNA (and the RNA) is therefore indicated by the carbon numbers left unbound at either end
  - The sequence is written from the 5th end to the 3rd end



Source: https://biochemix.wordpress.com/2014/04/17/nucleotides-and-nucleic-acids/

Base Pairing of Nucleotides in the DNA

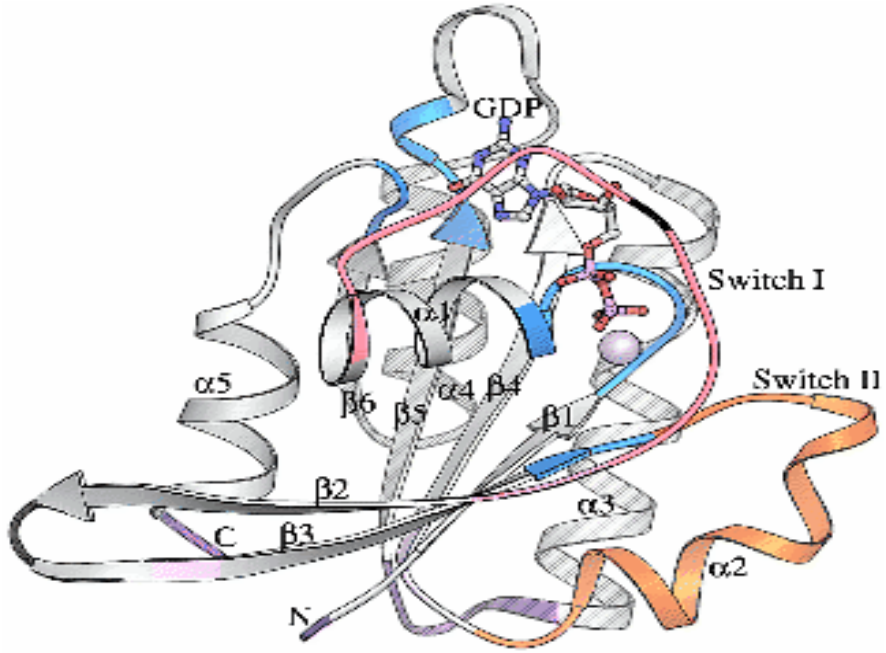
- The DNA sequence is paired with its complementary DNA sequence via hydrogen bonds to form the double helix structure (Watson and Crick, 1953)
  - Adenine – Thymine
  - Guanine – Cytosine
- This helix structure provides the DNA with the necessary stability
  - The helical structure is due to
    - the sugar-phosphate bond angle,
    - the stacking of the hydrophobic bases, and
    - the interaction between the complementary strands
  - The genetic code must be maintained in a stable molecule (Q: Why?)



Source: http://fig.cox.miami.edu/~cmallery/150/gene/chargaff.htm

Proteins

- Proteins are building blocks of life
  - function as enzymes that catalyze or inhibit biochemical reactions
  - carry out signaling and molecular transport
  - construct supporting structures
- Proteins are constructed as a sequence of amino acid residues
  - Out of a total of 20 naturally occurring amino acids
  - Polypeptides, oligopeptides
- The shapes of the proteins as well as the underlying amino acid structure determines the function of the protein
  - The amino acid sequence determines the primary structure
  - The presence of structural motifs determines the secondary structure
    - α helix, β sheet
  - The spatial structure that the protein folds into determines the tertiary structure

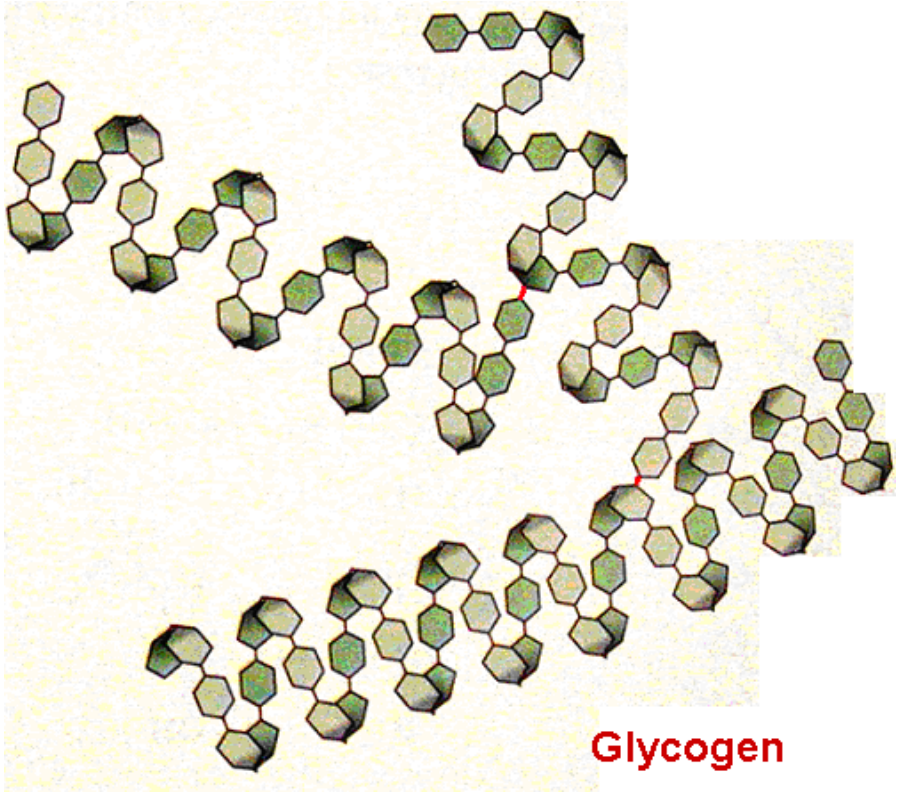


Rab6 GTPase with GDP (part of the Ras superfamily)

Source:
http://www.cs.stedwards.edu/chem/Chemistry/CHEM43/CHEM43/GTP/Index.htm

Carbohydrates

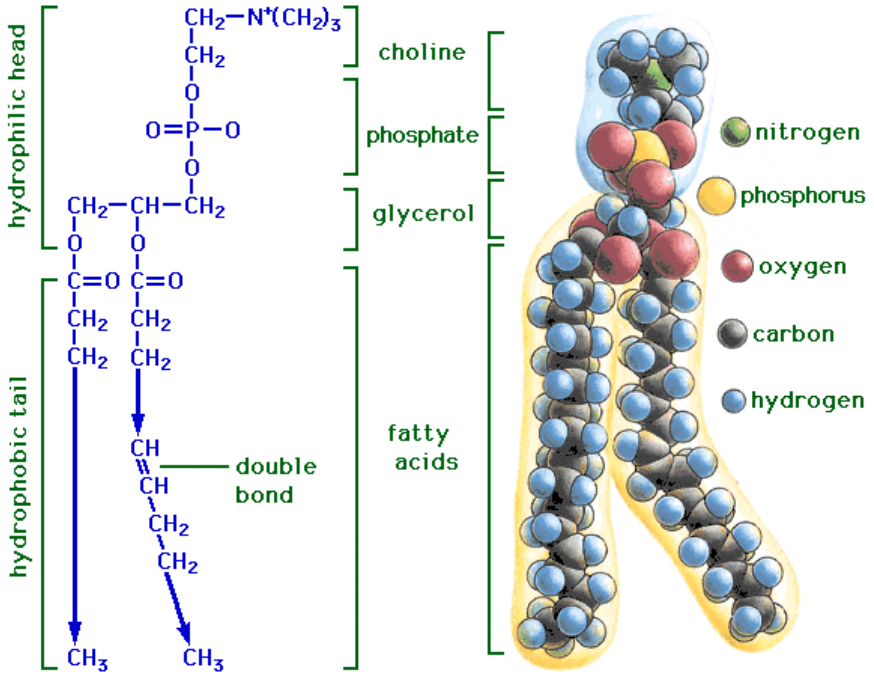
- Carbohydrates are complex sugar molecules
  - Monosaccharides, disaccharides, oligosaccharides, polysaccharides
- Polysaccharides are composed of a large number of monosaccharides
- In contrast to the sequential structure of proteins in terms of amino acids, **polysaccharides are not sequential**
  - The sequence of monosaccharides branch off frequently
  - This results in an exponentially increasing number of possible forms for a molecule with a given number of monosaccharides
- Polysaccharides play critical roles in almost all cellular processes from cell signaling to protein folding and stability



Source: http://www.chm.bris.ac.uk/motm/glucose/glucosejm.htm

Lipids

- Lipids comprise a vast group of molecules that are the esters that the fatty acids form with glycerole
- Glycolipids and phospholipids form the cell membrane
  - Phospholipids form the bi-lipid membrane
  - Glycolipids, much fewer in number, carry out essential molecular recognition tasks
- The phospholipids also operate as substrates to signaling reactions catalyzed by activated receptor proteins embedded in the cell membrane
  - The phospholipids are decomposed into their constituents
  - These constituents travel across the cytoplasm and trigger a variety of reactions
- Lipids also serve as energy reserves
  - Excess glucose is stored in fat tissue
    - Provide 6 times more energy than glucose
  - Fat storage is regulated by the liver

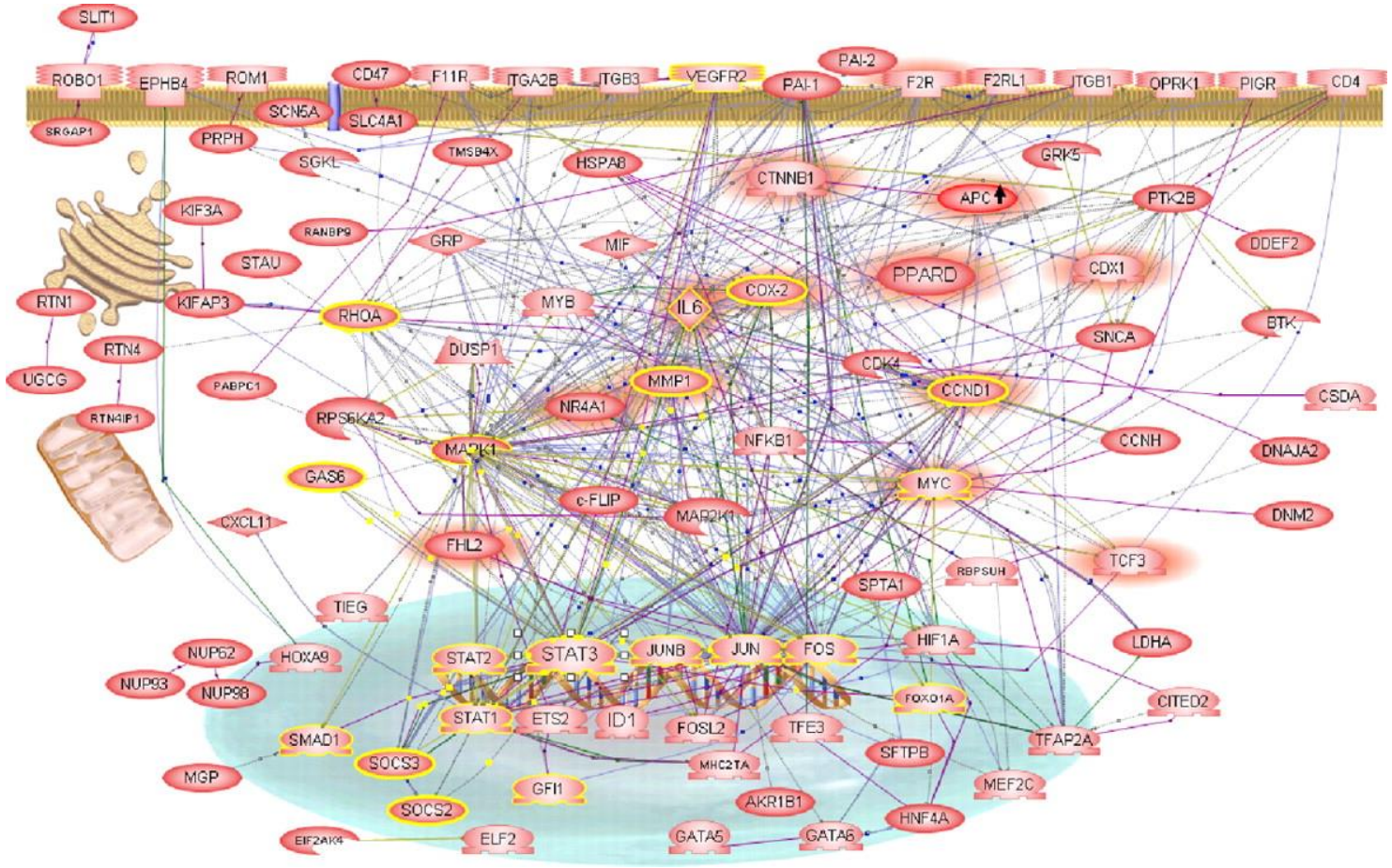


Phospholipid molecule

Source:
http://www.agen.ufl.edu/~chyn/age2062/lect/lect_06/4_18.GIF

Information Flow from Genes to Proteins

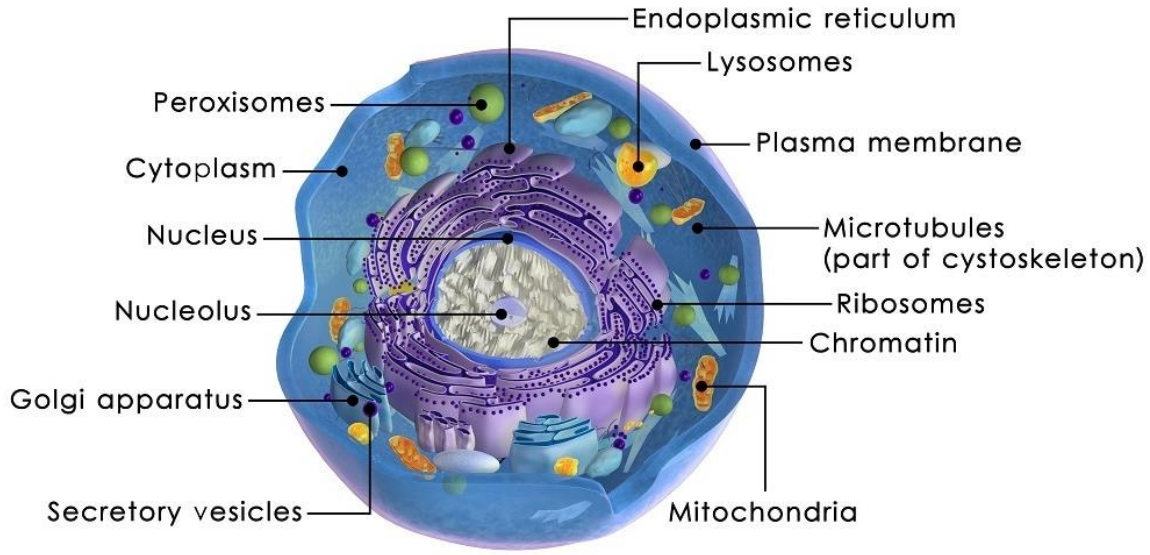
- Cells respond to their environment by initiating the synthesis of certain proteins and shutting the synthesis of others
  - Signaling molecules are picked up by receptor proteins embedded in the lipid membrane
  - These receptor proteins then create a cascade of reactions called the signaling pathway through phosphorylation and/or de-phosphorylation reactions
  - The signal eventually reaches the nucleus, triggering the cell's response by changing its protein composition → synthesis and degradation



Angiogenic signaling network.
Source: http://www.pnas.org/content/pnas/104/31/12890/F2.large.jpg

Information Flow from Genes to Proteins

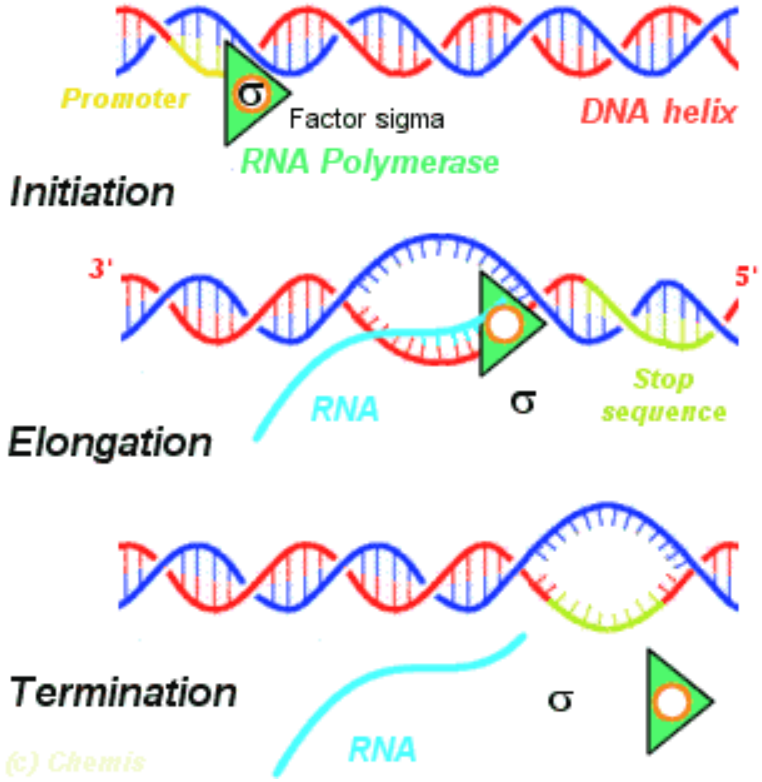
- Protein life cycle involves 6 processes carried out in succession
  - Transcription
  - Splicing
  - Translation
  - Post-translational modifications
  - Translocation
  - Degradation



Source:
https://www.news-medical.net/life-sciences/Eukaryotic-and-Prokaryotic-Cells-Similarities-and-Differences.aspx

Transcription

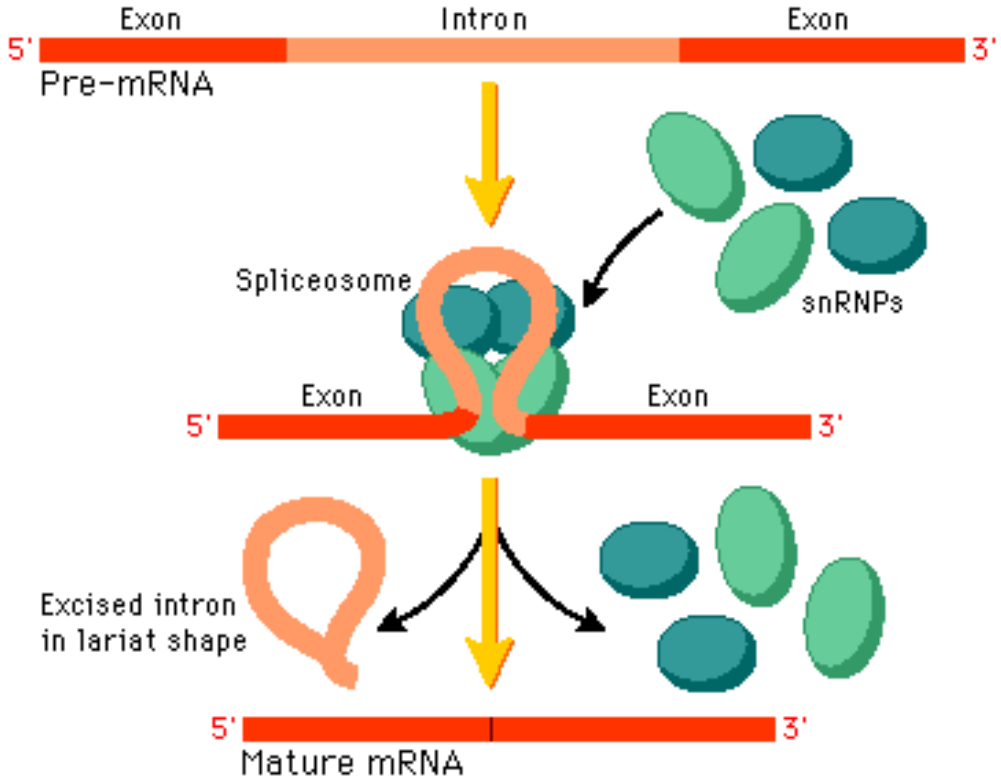
- Transcription is the process by which the genetic code is copied into a complementary mRNA sequence
  - Genetic code is represented by genes along the DNA
  - Each gene has a beginning and an ending, as well as a promoter region
    - Transcription factors bind to the promoter region to initiate the expression of the gene
    - Repressor molecules may also bind to the promoter to block the expression of the gene
  - Between the beginning and the ending sites, some parts of the code are not intended to go into protein coding
    - Introns\Exons
- Transcription is carried out by the RNA polymerase enzyme
  - The enzyme binds to the DNA with the help of the transcription factors and unwinds the double helix for access to a single strand
  - It then travels along the DNA downstream synthesizing the RNA
  - It stops and leaves the DNA when it encounters the nucleotide pattern of a Stop signal



Source:
http://www.geneticengineering.org/chemis/Chemis-NucleicAcid/RNA.htm

Splicing

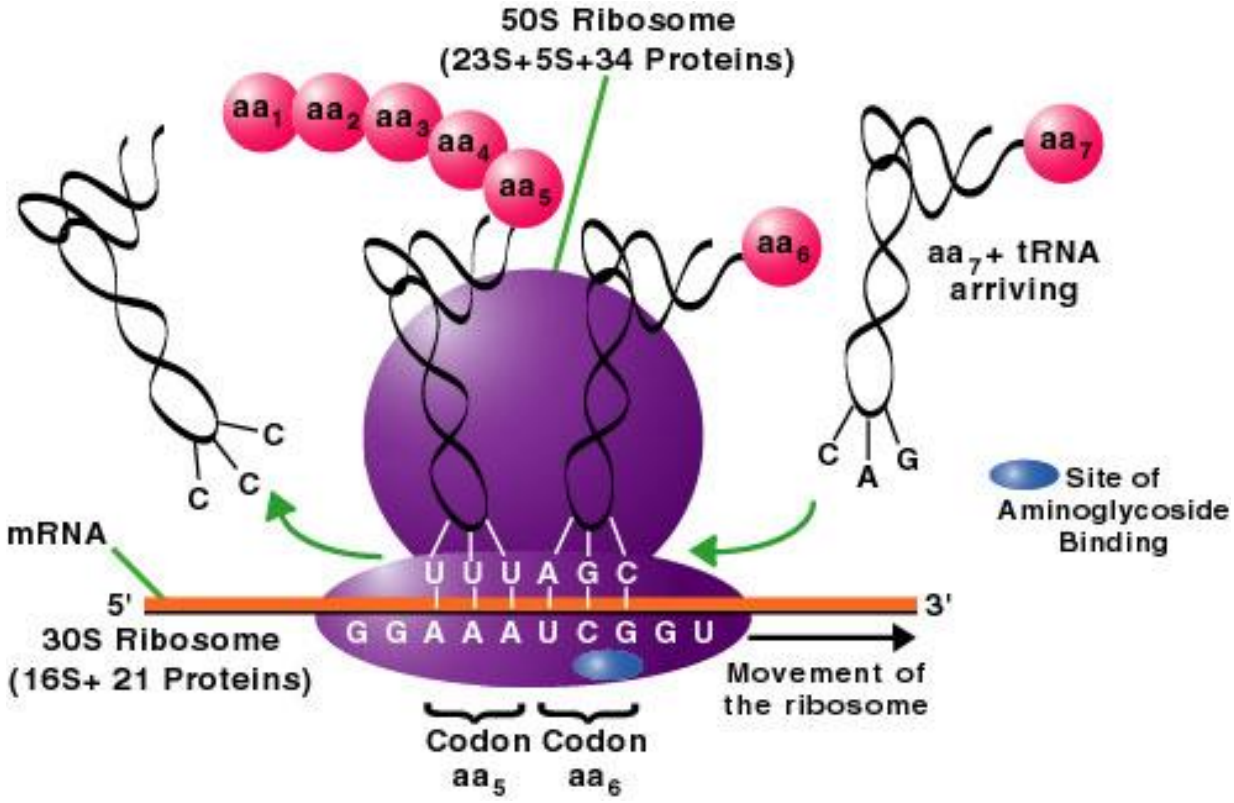
- The RNA polymerase synthesizes a precursor mRNA molecule
  - The pre-mRNA molecule has both introns and exons
- The process that removes the introns from the pre-mRNA molecule is called splicing
  - carried out by a complex of small ribonucleoproteins called the spliceosome
- As the introns are removed from the sequence of pre-RNA, the exons are stitched together to form the mature RNA
- The mature RNA then travels from the nucleus to the cytoplasm to carry out the protein synthesis message
- **Alternative splicing** refers to alternative ways in which a pre-mRNA molecule can be spliced into a different mRNA molecule
  - from ~40K genes to millions of proteins!!



Source:
http://www.cbs.dtu.dk/staff/dave/roanoke/genetics980408f.htm

Translation

- Translation refers to the synthesis of proteins according to the corresponding mRNA molecules
- Each nucleotide triplet along the mRNA sequence defines an amino acid
  - Each nucleotide triplet is termed a codon
  - Each of 64 possible codons encode for one of 20 different amino acids
    - The relationship is many to one
- The translation process is carried out in the cytoplasm by the ribosomes
  - The mRNA is grabbed by the ribosome
  - The tRNA collects the next amino acid encoded for by the next codon in the mRNA sequence
  - The polypeptide chain grows by the appending of successive amino acids
  - When a stop codon is encountered, the ribosome releases the mRNA and the polypeptide chain

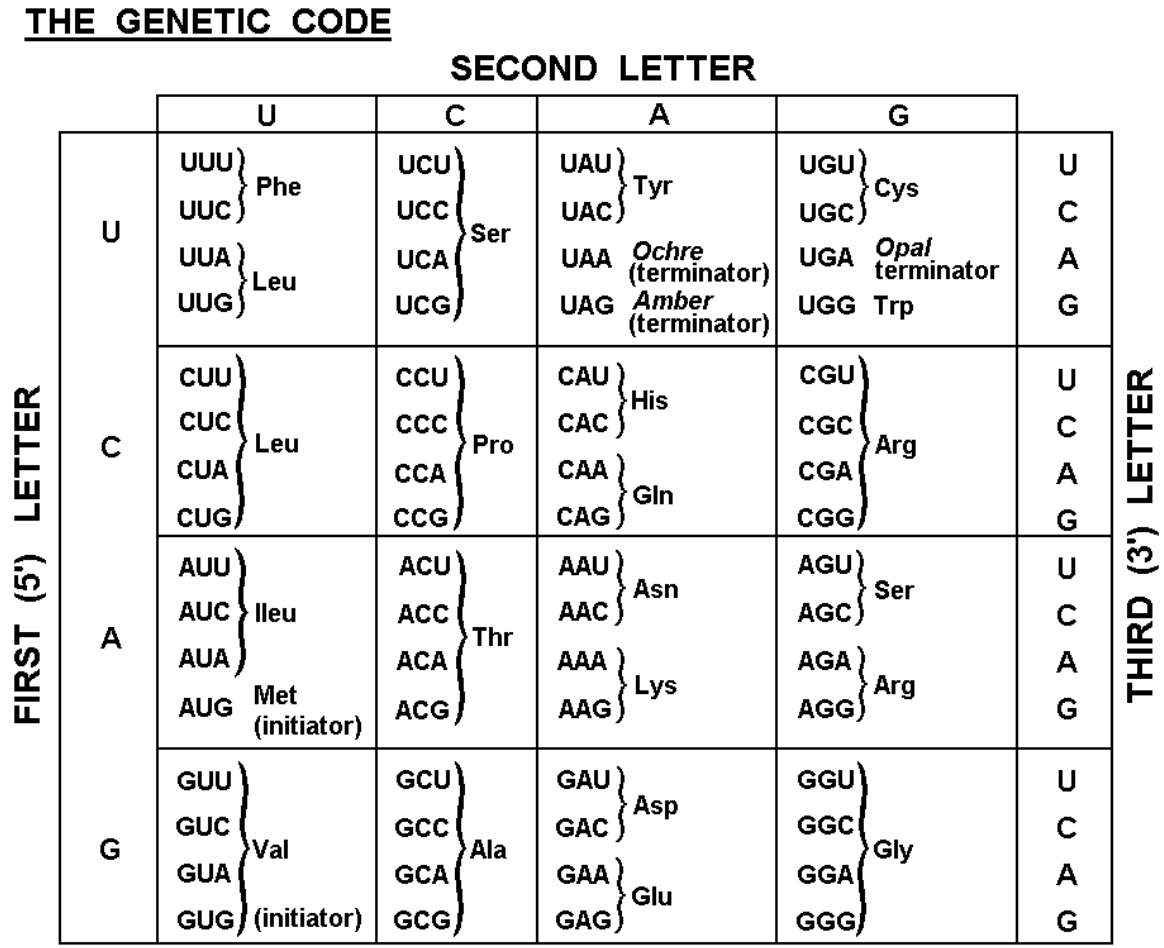


Source: http://www.scripps.edu/chem/wong/rna.html

Codons and Amino Acids

Amino Acid	Three Letter Code	Single Letter Code	Codons
Isoleucine	Ile	I	ATT, ATC, ATA
Leucine	Leu	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	Val	V	GTT, GTC, GTA, GTG
Phenylalanine	Phe	F	TTT, TTC
Methionine	Met	M	ATG
Cysteine	Cys	C	TGT, TGC
Alanine	Ala	A	GCT, GCC, GCA, GCG
Glycine	Gly	G	GGT, GGC, GGA, GGG
Proline	Pro	P	CCT, CCC, CCA, CCG
Threonine	Thr	T	ACT, ACC, ACA, ACG
Serine	Ser	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Tyr	Y	TAT, TAC
Tryptophan	Try	W	TGG
Glutamine	Gln	Q	CAA, CAG
Asparagine	Asn	N	AAT, AAC
Histidine	His	H	CAT, CAC
Glutamic acid	Glu	E	GAA, GAG
Aspartic acid	Asp	D	GAT, GAC
Lysine	Lys	K	AAA, AAG
Arginine	Arg	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop	---	-	TAA, TAG, TGA

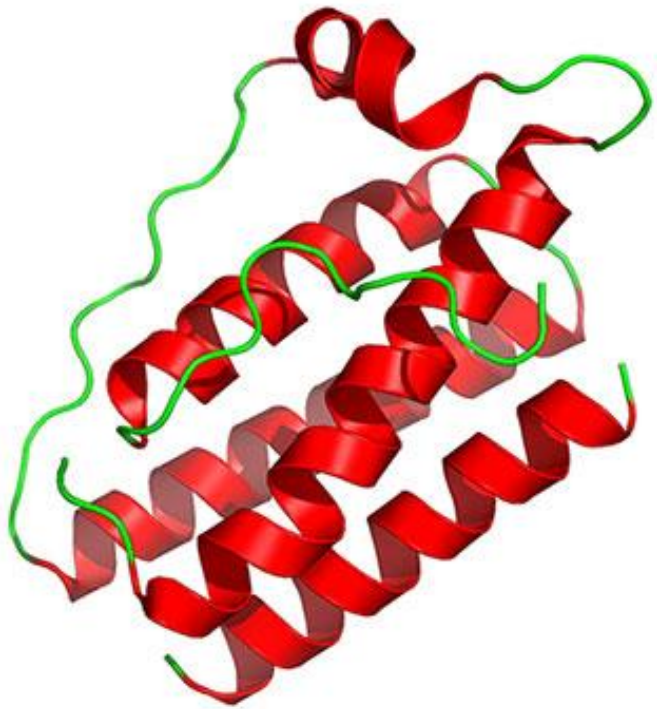
Codons and Amino Acids



Source: http://www.msstate.edu/dept/poultry/pics/gnscht.gif

Post-Translational Modifications

- Once a polypeptide chain is synthesized, it undergoes a variety of operations that turn it into functioning proteins
- The operations may induce
  - the addition of extra molecules like sugars (glycosylation) and acetyl groups (acetylation)
    - For proper folding
  - structural alterations in the form of establishment of di-sulfide bonds
    - Again, for proper folding
  - the chemical changes at the amino acid level like deamination (glutamine to glutamic acid or asparagine to aspartic acid) or citrullination (arginine to citrulline)
  - cleaving to generate functioning units from non-functioning peptide chains

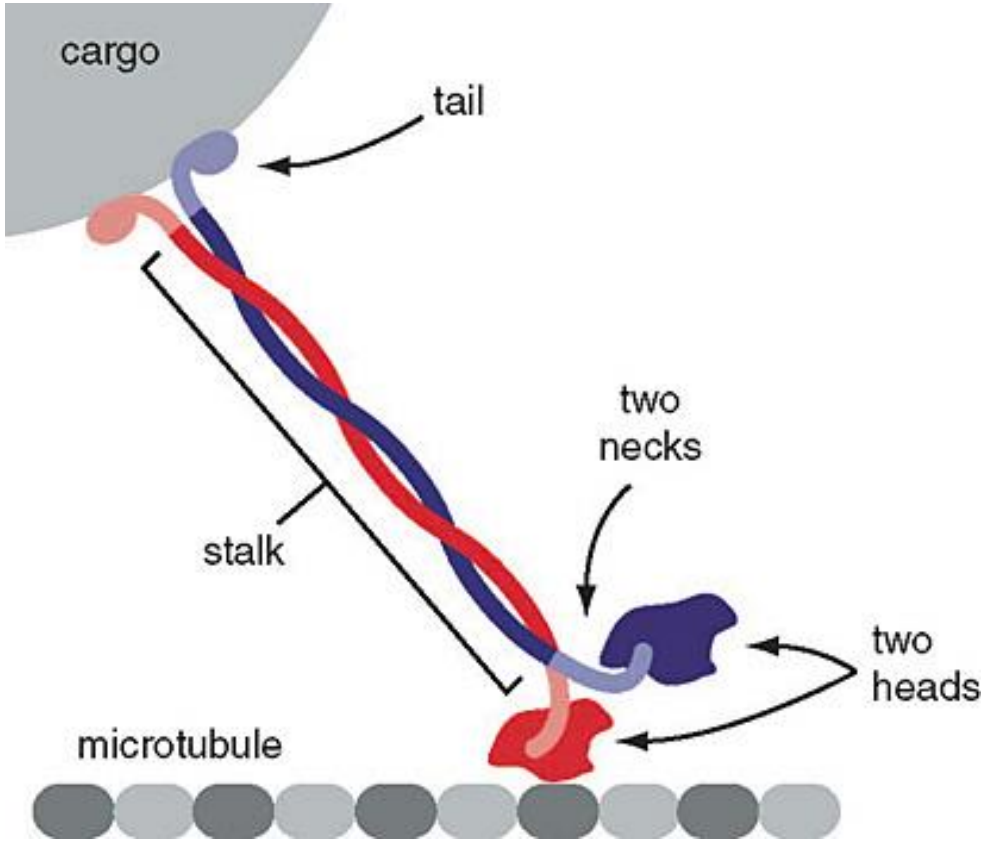


Leptin protein complex

Source: http://www.3dchem.com/molecules.asp?ID=154

Translocation

- The cell is composed of many compartments specialized to perform specific tasks
- The synthesized proteins are taken to the compartments for their functional specialization via molecular transport mechanisms
- Molecular transport inside cells are carried out by **cargo proteins**
  - Cargo proteins are equipped with a molecular sack suitable for fetching the protein to be transported
  - They also have a pair of extensions (head) that move the cargo proteins along microtubules all the way to their target locations



Source: http://news-service.stanford.edu/news/2003/december10/gifs/Kinesin_Proof2.jpg

Protein Degradation

- Proteins carry out highly specific functions in cells
- The amounts of different proteins are adjusted to match the transient needs of a cell's biomolecular mechanism
- This requires not only the synthesis but also the removal of the proteins that are no longer needed from the intracellular environment
- The process that eliminates proteins is termed protein degradation
- Protein degradation is carried out in subcellular organelles called **lysosomes**
  - The **excess protein amounts** are identified by the molecular mechanism via specific proteins that label the excess proteins for degradation (a.k.a. ubiquitination)
  - The proteins marked for degradation are taken to the lysosome by the molecular transport mechanisms
  - In the lysosome, polypeptide chains are hydrolyzed and decomposed into their constituent amino acids

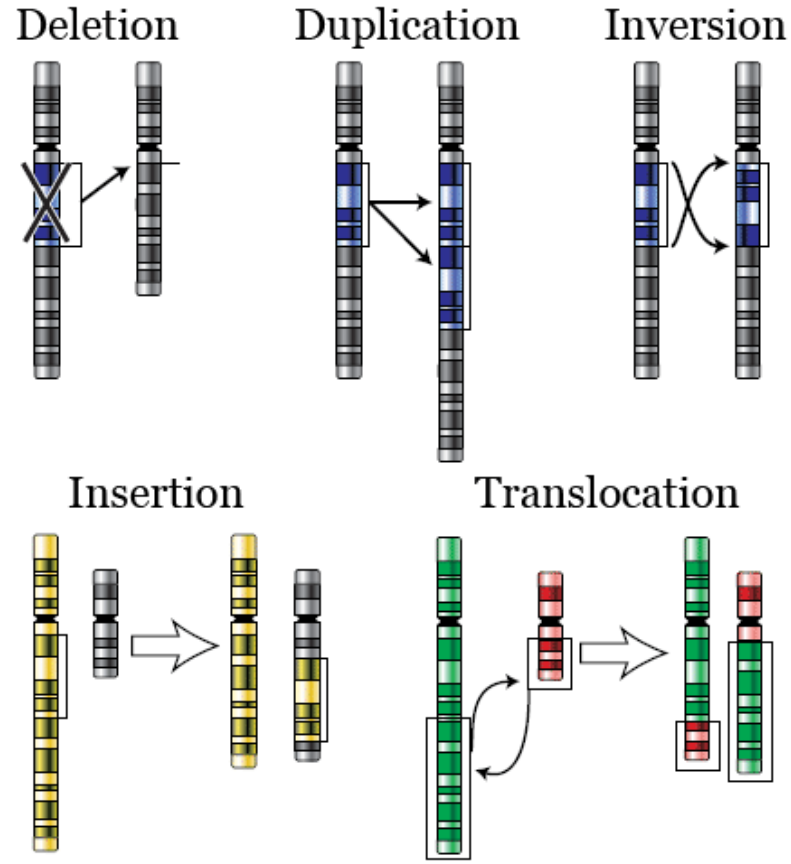
Illustration: The Inner Life of the Cell

Gene and Protein Expression

- **Protein expression** is the term used to cover the steps from transcription to post-translational modifications
- **Gene expression** refers to the synthesis of mRNA after splicing
  - Hence, gene expression is an intermediary to protein expression
- The expression levels of proteins in a cell characterize the reactionary effort of the cell to its environment
  - Intracellular environment
    - Cellular metabolism
  - Extracellular environment
    - Cell-to-cell signaling, hormonal signaling, etc.
- Measurement of protein expression levels are extremely problematic
  - Proteins are typically identified via mass spectroscopy techniques that identify the expression levels of a known set of proteins
  - The proteins that may be critical for the biological hypothesis in consideration may not be known a priori
  - Furthermore, the expression levels required for activity of certain proteins may be lower than the sensitivity of mass spectroscopy
- In contrast, measuring gene expression can be achieved in a high-throughput manner using DNA microarrays
  - The expression levels of hundreds of thousands of gene probes can be assessed in a single run
  - On the downside, issues with accuracy, comparability and noise are abound

DNA Sequence Analysis

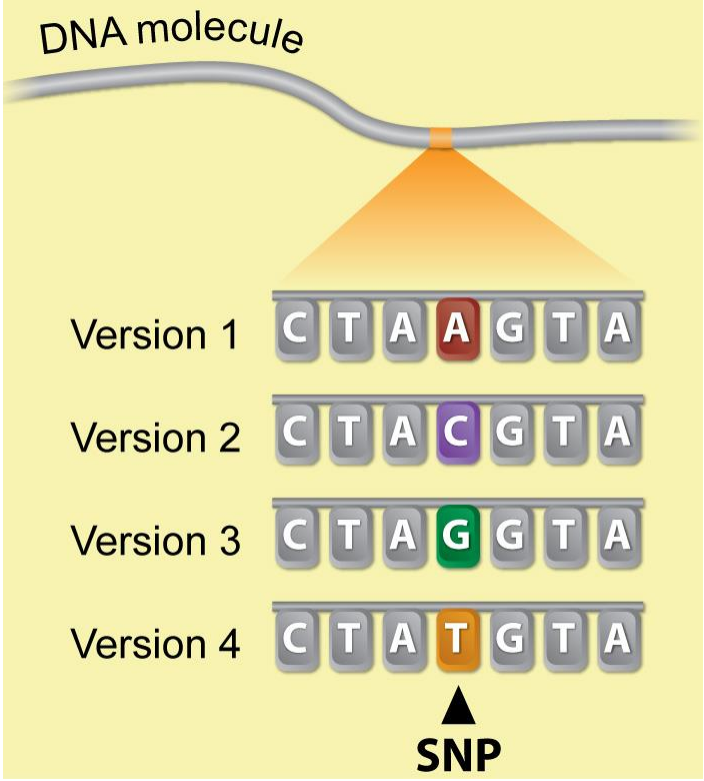
- Genetic variation
  - Genetic variation is essential for the ability of organisms to adapt to changing environmental conditions
    - Different genotypes achieve different genetic fitness resulting in different proliferation rates
    - Gradually more fit genotypes become to dominate the population
  - Genetic variation results from various stochastic processes
    - Genetic mutations
      - Insertion
      - Deletion
      - Inversion
      - Translocation
      - Duplication
    - Sexual reproduction



Source: https://www.ck12.org/biology/mutation/lesson/Mutation-Types-BIO/

Example: Single Nucleotide Polymorphisms

- A single nucleotide polymorphism is an alterations of just one base pair at a specific DNA position
  - Also termed point mutations
- SNPs are inherited in part by the descendants of the individual
  - Some from the mother, rest from the father
- In different individuals, observing the same set of SNPs indicates common family lineage
- The odds of observing several different SNPs in two unrelated individuals drops sharply with increasing number of SNPs
  - Hence the strength of DNA evidence in crime scene investigations



Source: https://learn.genetics.utah.edu/content/precision/snips/

Protein Sequence Comparisons

- Mutations in the DNA sequence implicate alterations in the proteins encoded by the corresponding genes
  - **Alteration** of a single nucleotide *can* replace an amino acid with another without affecting the remaining sequence
  - **Deletion** of a single nucleotide alters the corresponding amino acid as well as all those that follow
- Protein shape and function are governed by its amino acid sequence
  - Mutations that cause alterations in the sequence affect the protein function
    - Cancer is linked with hyper-activity of growth factors of inhibition of tumor suppressor proteins
    - Sickle cell disease is due to the replacement of a single nucleotide, causing a change in a single amino acid in the sequence of the components of the hemoglobin complex, in turn reducing its oxygen carrying capacity
  - Proteins with similar sequences are likely to carry out similar functions
    - Protein families possess preserved sequence motifs
    - Presence of these motifs in a newly sequenced protein signals its membership in the corresponding protein family

Example: Protein Sequence Alignment

- Similarity of the amino acid sequence indicates similarity of the protein function
  - Orthologs: Proteins whose sequence is largely conserved across different species
  - Paralogs: Proteins with similar sequences within the same genome, indicating common origin

common origin → homology

- In order to assess the sequence similarity of two different proteins, sequence alignment procedures are used
  - The alignment of two sequences requires inferring the proper amino acid replacements and deletions to arrive at a minimally extended common sequence
    - Replacement likelihood of individual amino acids can be assessed via 20 by 20 replacement matrices
    - Bits of amino acid sequence segments missing in the other protein are accounted for by unknown sequence stretches
  - Dynamic programming provides optimal alignments
    - Feasible alignments are achieved by suboptimal but fast algorithms

Summary

- The molecular machinery in living cells operates as a tightly regulated and finely tuned system of many components
- Analyzing this massively parallel system requires elucidating
  - Protein-protein interactions
  - Protein-DNA interactions
  - Enzymatic activity for glycosylation, acetylation, phosphorylation, …
- Numerically efficient biomedical signal processing algorithms are in need
  - Statistically viable predictions
  - Using incomplete and potentially misleading biological data