Taxonomy of Life

• Three domains: Eukaryotes, Bacteria (Eubacteria), Archaea.

• Two types of cellular architecture: Prokaryotic and Eukaryotic. Prokaryotes first appeared around 3.5 billion years ago, Eukaryotes first appeared around 1.8 billion years ago.

• Bacteria + Archaea = Prokaryotes.

• Prokaryotes are characterized by a simple cellular structure. They lack a membrane-bound nucleus. Eukaryotic cells have a nucleus, and generally have other membrane-bound compartments as well.

• Archea have the cellular architecture and metabolic machinery of bacteria, but their informational machinery is more similar to that of the Eukaryotes.

• Bacteria and Archaea are unicellular, Eukaryotes are either unicellular (e.g. yeast) or multicellular (e.g. mammals).

• Archea often live in extreme environments (high temperature, salinity, pH). Enzymes from Archea are often useful in biotechnology applications because they can function under unusual conditions (e.g. near the boiling point of water).

• Reference: www.ucmp.berkeley.edu/alllife/threedomains.html.
Structure of Cells

- Cells are bound by the cell membrane, also called the plasma membrane, which is a lipid bilayer. Beyond the cell membrane may be a cell wall (present in most bacteria and plant cells but not animal cells). The cell membrane is selectively permeable to ions and small molecules, and can actively transport such molecules into and out of the interior of the cell.

- Eukaryotic cells are generally highly compartmentalized. The interior can be partitioned into nucleus + cytoplasm. The nucleus contains the genome (DNA). The cytoplasm contains various complex subcellular structures called organelles (e.g. mitochondria, chloroplasts, lysosomes), and an aqueous compartment called cytosol (about 50% of the cellular volume). The cytosol contains a cytoskeleton (actin filaments and microtubules) that gives mechanical support and facilitates the movement of organelles within the cell.

- Larger cells have smaller surface/volume ratio, hence less diffusion of material into and out of the cell. This limits the size of a bacteria to about 1µm. Bacteria are often pill-shaped or rod-shaped to ensure that most of the cytoplasm is close to the cell membrane. The membranes and channels of a Eukaryotic cell allow more efficient transport within the cell. This allows Eukaryotic cells to have sizes around 30 – 50µm.

- Eukaryotes are more complex than Prokaryotes, but Prokaryotes are more evolved (e.g. have undergone more generations) hence may be said to be more perfectly adapted to their environment.
Some Cellular Functions

- Reproduction. Copy the DNA, replicate all organelles, apportion the contents to the daughter cells, mechanically divide the cell.

- Metabolism. Cellular energy (e.g. for driving kinetically unfavorable reactions) is largely derived by the hydrolysis of ATP (adenosine triphosphate) to ADP, then AMP. The reverse reactions (AMP to ADP to ATP) are necessary to restore the cell’s energy reserves. When no oxygen is available, this is primarily achieved by glycolysis, which breaks down glucose to pyruvate to drive the formation of ATP. In the presence of oxygen, many cells can further break down the pyruvate to obtain more ATP (citric acid cycle / oxidative phosphorylation). In Eukaryotes, these reactions take place in the mitochondria.

- Biosynthesis. Production of many types of biological molecules. Energy provided by ATP, reducing power by NADH, enzymes and coenzymes drive the reactions. Includes DNA replication, copying DNA to RNA, synthesizing proteins from RNA.

- Membrane transport. Pumping ions or small molecules into or out of the cell.

- Many specialized function of cells in multicellular organisms in addition to the above “housekeeping” processes, e.g. cell-cell communication and adhesion.
Biological Molecules

- Four atoms (C, H, N, O) make up > 98% of the mass of a cell.
- The mass of a cell is around 70% water. The remainder is protein (≈ 15%), lipids (≈ 3%), inorganic ions (≈ 1%), nucleic acids (≈ 3%), polysaccharides (≈ 2%), and various small metabolites (≈ 3%).

- **Macromolecule**: A large molecule consisting of tens of thousands to millions of atoms. Key biological macromolecules are proteins (polypeptides), nucleic acids (DNA), and polysaccharides.

- Chemistry concept: *electronegativity*. The degree to which an atom will take on additional electrons to become a negatively charged ion. Oxygen is more electronegative than nitrogen, which is in turn more electronegative than carbon.

- Chemistry concept: *polar molecule*. A molecule that has a partial negative charge near its more electronegative atoms, and a compensating partial positive charge near its less electronegative atoms. For example, water (partial negative at the oxygen, partial positive at the hydrogen).

- Key types of chemical bonds and interactions:
  - Covalent (90 kcal/mole): Shared electrons. Under biological conditions, each element prefers to form a fixed number of covalent bonds (phosphorus forms 5, carbon forms 4, nitrogen forms 3, oxygen forms 2, hydrogen forms 1). A missing bond gives the atom a negative charge. An extra bond gives the atom a positive charge. A pair of atoms can share more than one electron, leading to double and triple covalent
bonds. Single bonds rotate freely around the bond, double and triple bonds do not.

– Ionic (3 kcal/mole): Attraction of atoms with full opposite charges.

– Hydrogen (1 kcal/mole): Attraction of polar molecules with partial opposite charges. In the cell this generally takes the form of a H/O or H/N attraction.

– Hydrophobic forces: Nonpolar regions of large molecules are said to be hydrophobic, because they don’t form hydrogen bonds with water. Thus the water pushes Hydrophobic regions of a molecule together, to maximize the exposure of water to the polar part of the molecule.

– van der Waals (.1 kcal/mole): Charge fluctuations attract two molecules up to a point. Beyond this point there is a strong repulsion. At a moderate distance there is a very weak net attraction. This attraction is strongest at the van der Waal’s radius, which is a constant property of each element.

• Average thermal collisions have about .8 kcal/mole energy, so non-covalent bonds will often break and re-form in biological systems.

• Important carbon-based compounds:

  \[C_mH_n\]  
  \(-CH_2 - OH\)  
  \(-C - CO - C-\)  
  \(-COH\)  
  \(-COOH\)  
  \(-COOC-\)  
  \(-CNH_2\)  
  \(-CONHC-\)  
  hydrocarbons and aromatic hydrocarbons  
  alcohol  
  ketone  
  aldehyde  
  carboxylic acid, generally loses H and becomes negatively charged  
  ester, formed by condensation of an acid and an alcohol  
  amine, generally gains H and becomes positively charged  
  amide, formed by condensation of an acid and an amine
• Chemistry concept: \( \text{pH} \). A measure of the concentration of hydrogen ions (protons) in a solution. Specifically, it is the negative base 10 logarithm of this concentration: \( \text{pH} = -\log_{10}[H^+] \).

• Chemistry concept: Acid/Base. An acid is a molecule that releases a proton in solution. A base is a molecule that gains a proton in solution.

• Key chemical properties of biological molecules:
  – Three-dimensional shape (conformation): sheets, helices, ...
  – Polarity/Hydrophobicity: Relates to the molecule's ability to dissolve in water. Hydrophilic (polar) molecules form hydrogen bonds with water. Hydrophobic (nonpolar) molecules do not form hydrogen bonds with water. A large molecule can have various polar and nonpolar regions, which partially determine its 3D shape in an aqueous environment.
  – Charge: An excess or lack of electrons for certain atoms in the molecule, producing a net charge at that atom. A large molecule can contain several polar atoms with different charges. The net charge is a function of the pH.

• Molecular recognition: property of a certain molecule to bind or interact very specifically with a different molecule. These interactions are often transient, and have an equilibrium rate of association / dissociation.

• Recognition between macromolecules is a crucial component of the complexity of life. In particular it is essential for the storage, transmission, and processing of biological information.

• Dynamics of molecular recognition: Recognition requires proximity, proximity is limited by diffusion. In \( T \) time units, a
molecule diffuses an average distance that is proportional to \( \sqrt{T} \). Diffusion is rapid over short distances but slow over long distances. Small molecules diffuse more rapidly than large molecules.

- **Polymer**: A molecule constructed by sequentially binding members of a small set of *subunits* into a linear chain. Many important biological polymers are *linear, unbranched, oriented* polymers.
  - Nucleic acids (DNA and RNA) (polymer of nucleotides).
  - Proteins (polypeptides) (polymer of amino acids).
  - Lipids (polymer of fatty acids).
Structure of Nucleic Acids

- Linear, unbranched, oriented polymer of nucleotides.
- Two classes: DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). The difference is a single oxygen atom.
- A nucleotide is a base + a five carbon sugar + a phosphate.
- A nucleoside is a base + a five carbon sugar. The base is always linked to the same carbon (C1) of the sugar.
- The base is one of adenine and guanine (the purines, having two rings), or cytosine, thymine, and uracil (the pyrimidines, having one ring).
- The corresponding nucleosides are adenosine, guanosine, cytidine, uridine, and thymidine.
- If a single phosphate group is added monophosphate is appended to the name (e.g. adenosine monophosphate). diphosphate and triphosphate forms exist and are biologically important, but in DNA and RNA only the monophosphate forms occur, which are commonly denoted A, T, G, C, and U.
- Uracil only occurs in RNA, thymine only occurs in DNA.
- To form the DNA or RNA polymer, nucleotides are linked by a phosphodiester linkage. The phosphate group is bound to the 5’ carbon in the sugar of one nucleotide and to the 3’ carbon of the sugar in the next nucleotide.
- The phosphodiester bonds form the covalent backbone of a DNA or RNA molecule.
• Nucleic acids are oriented $3' \rightarrow 5'$ (unbound hydroxyl to unbound methyl).
Hybridization of Nucleic Acids

• Two strands (molecules) of DNA can pair via hydrogen bonds if they are *complementary*, meaning that every adenine on one molecule corresponds to a thymine on the other molecule, and every guanine on one molecule corresponds to cytosine on the other molecule.

• Each pair of complementary bases is held together with hydrogen bonds (two for A-T, three for G-C). The bonds are not flat, forming a *major groove* and a *minor groove*.

• The resulting complex is called a double-stranded DNA molecule, and often forms a double helix. The $3' \rightarrow 5'$ orientation on one strand is the reverse of the $3' \rightarrow 5'$ orientation on the complementary strand.

• Nucleic acid hybridization is an example of molecular recognition.

• RNA can form short double-stranded molecules with itself or with DNA, but these complexes are not as stable as the DNA/DNA complex.

• Hybridization plays an important role in biotechnology.
Chemical Structure of Proteins

- The subunits of a protein molecule are *amino acids* (sometimes called *residues* in this context). They consist of an α-carbon atom whose four bonds link to (i) a carboxyl group (*COOH*), (ii) an amino group (*NH₂*), (iii) one of 20 *side chains* (denoted *R*), and (iv) a hydrogen atom.

- The 20 side chains give rise to 20 amino acids (more than 20 amino acids are chemically stable, but only 20 occur in life).

- The chemical properties of the side chains determine the chemical properties and structure of the protein macromolecule. The side chains can be put into four groups: acidic, basic, uncharged polar, nonpolar.

- Amino acid polymers are formed in a condensation reaction (involving the loss of a water molecule). The carboxyl group of one molecule forms an *amide bond* with the amino group of the next molecule. One end of the polymer will have an unbound carboxyl group (the *carboxyl terminal end*), while the other end will have an unbound amino group (the *amino terminal end*). This gives the protein molecule its orientation.

- Structure organized as: primary (amino acid sequence), secondary (local 3D structure, e.g. helical or sheeted), tertiary (global 3D structure), quartenary (3D structure of a complex of several proteins).

- The secondary and tertiary structure are determined by the primary structure (mostly). But it is not known how to compute the 3D structure from the primary structure.
• Three-dimensional structure is a major determinant of function, along with charge distribution and location of polar and nonpolar regions.

• Extracellular proteins are synthesized in the cell, then secreted outside of the cell (transported through the cell membrane). These proteins can be stabilized by disulfide linkages (strong bonds that would normally be hydrolyzed in the cytosol).

• Proteins are heavily chemically modified beyond the basic polypeptide structure. A common modification is that addition of a small group, e.g., phosphorylation, methylation, adenylation, acetylation, carboxylation, glycosylation. These modifications can substantially change the structure and function of the protein. They can also serve to mark the protein for a destination (e.g. nucleus, mitochondria, extracellular space).
Genes, Chromosomes, and the Genome

- The DNA molecules in a cell comprise its genome. In prokaryotes the genome takes the form of a DNA tangle in the cytoplasm.

- Eukaryotes have two genomes:
  - The nuclear genome is contained in the nucleus and contains the overwhelming majority of information. It consists of a number of linear DNA molecules bound with many types of proteins comprising a chromosome (most human cells have 23 homologous chromosome pairs of varying sizes). The size of the human nuclear genome is \( \approx 3.3 \text{GB} \) \((3.3 \times 10^9 \text{nucleotides})\). Sizes of this type always refer to the total number of nucleotides on one strand from each DNA double helix.
  - The mitochondrial genome is contained in the mitochondria and contains a subset of the information used in the mitochondria (most of the information used by the mitochondria is contained in the nuclear genome). The mitochondrial genome consists of a single circular DNA molecule that is largely free of bound proteins. The human mitochondrial genome consists of 16,569 nucleotides. A eukaryotic cell contains thousands of mitochondria (hence thousands of copies of the mitochondrial genome) while only one copy of the nuclear genome is present (during certain points of the cell cycle up to two copies may be present).

- The genome contains at least four kinds of information that are essential for the functioning of the organism:
1. Templates for the construction of proteins (protein coding genes).
2. Templates for the construction of non-coding RNA’s (ncRNA’s) (RNA genes).
3. Regions that control the expression of proteins and ncRNA’s (regulatory regions).
4. Regions associated with DNA replication (centromere, telomere, replication origins).

- In humans, around 93% of the mitochondrial genome is used for one of these four roles. This is true of only around 50% of the nuclear genome. The other half of the human nuclear genome is filled mostly with repeat-rich satellite sequences that arose from the activity of transposable elements. It is not clear what useful role, if any, these regions play for the host organism. Most vertebrates have a high fraction of satellite sequence, but other Eukaryotes such as yeast have much lower fractions.

- Base composition in the genome can characterized as

\[ \text{GC content} + \text{AT content} = 1. \]

Most organisms have GC content < .5. The human genome is estimated to have GC content of .42. However local GC content varies greatly.

- The “dogma of molecular biology” is DNA makes RNA makes Protein. That is, DNA is copied into RNA, and RNA is used as a template for constructing proteins. Exceptions to this rule include points 2, 3, and 4 above.
• A (protein coding) gene corresponds to a region on the chromosome that codes for a single protein, along with the regulatory sequences that control when the gene is expressed. Most (but not all) genes are contained on one chromosome. Other nonstandard arrangements are overlapping and interleaved genes.

• An RNA gene is the region on the chromosome that codes for one ncRNA.

• A gene that is expressed is being actively used for the construction of protein. Some genes are expressed only under certain conditions. In multicellular organisms some genes are only expressed in certain tissues.

• The genome evolves over many generations through amplifications, insertions, deletions, point mutations, segmental duplications, and whole genome duplications. In the lifespan of an individual these events are extremely rare, except in cancer.

• The Eukaryotic genome contains many pseudogenes, which are partially-mutated copies of coding fragments. They are the product of amplifications and insertions. Over time through the accumulation of point mutations they become unrecognizable relative to the coding fragment from which they originated.

• A CpG dinucleotide is an adjacent C and G base occurring in the 5′ → 3′ direction. In the genome overall, CpG’s appear far less frequently than would be expected by chance. If the GC content in humans is .42, then the individual G and C contents are .21, and the expected fraction of CpG is .21² = .044. The overall CpG frequency is in fact around .009. There is a biochemical basis for this, as CpG can undergo methylation followed by deam-
ination to give TpG. Certain regions of the genome called \textit{CpG islands} contain the expected CpG fraction of \( \approx .044 \). These regions tend to be rich in genes.

- Eukaryotic DNA is wrapped around histone proteins (small spherical protein balls). Between cell divisions the DNA is further packaged into loops of \textit{chromatin fiber} called \textit{euchromatin}. During cell division the DNA becomes even more highly condensed. Certain regions of the DNA called \textit{heterochromatin} are permanently condensed.

- In a particular mammalian cell type, only about 3\% of the DNA is used as a template for constructing protein. For example, the human genome is large enough to encode about \( 2 \times 10^6 \) average-sized proteins (\( \approx 500 \) amino acids in the protein or \( \approx 1500 \) coding nucleotides in the gene). Based on mutational load only \( 1 \times 10^5 \) proteins can be essential. The best recent figure is that about \( 20,000 - 30,000 \) proteins are encoded in the genome. Most likely fewer than \( 5,000 \) proteins are ever synthesized in a particular cell.

- Gene density in the human nuclear genome averages \( \approx 10 \) genes per Mb. For individual chromosomes the range is from \(< 7 \) genes per Mb (chromosome 18) to about 20 genes per Mb (chromosome 19). The average intergenic distance is \( \approx 30 \)kb.

- Gene density in \textit{E. coli} is 1/1kb, in \textit{S. cerevisiae} it is 1/2kb, and in \textit{C. elegans} it is 1/5kb.

- The human mitochondrial genome encodes 37 peptides, with a gene density of 1/.45kb.
• The human genome contains at least $1.4 \times 10^6$ SNP’s (*single nucleotide polymorphisms*), which are individual nucleotides with variability in the current human population.

• Many mammalian genes can be organized into *gene families* consisting of a number of genes with similar sequences. The best-known example is the family of 90 *HOX* (*homeobox genes*).

• The correspondence between a protein-coding gene and its protein product is given by the *genetic code*, which is a 3:1 or *triplet code*, meaning that three consecutive nucleotides in the gene represent a single amino acid in the protein. A triplet of consecutive nucleotides is called a *codon*. The nuclear genetic code is universal, but the mitochondrial genetic code is not.

• For each gene, one of the two strands in the DNA double helix contains the code for the resulting peptide. This is called the *sense strand*. The other strand is called the *antisense* or *template* strand. The identity of the sense strand varies for different genes on the same chromosome.
Overview of protein synthesis

- The basic steps are:

  1. Transcription: A *messenger RNA* (mRNA) copy is transcribed from a region of DNA.
  2. The mRNA attaches to an RNA/protein complex in the cytoplasm called the *ribosome*.
  3. Translation: *Transfer RNA* (tRNA) molecules carry free amino acids to the ribosome, where they are linked in sequence to form a protein. The protein is assembled one amino acid at a time, from the amino-terminal end to the carboxyl terminal end.
  4. Peptide bonds between adjacent amino acids subunits are formed.
  5. The protein detaches from the ribosome. It may subsequently be heavily modified by other enzymes, and transported to a specific region of the cell.

- The steps of protein synthesis that involve mRNA synthesis are generically called *transcription*. Transcription occurs in the nucleus in Eukaryotes, and in the cytoplasm in bacteria.

- The steps of protein synthesis occurring at the ribosome that involve assembly of amino acids into the protein product are generically called *translation*.

- RNA genes are translated but not transcribed. The RNA transcript is directly functional in some way other than as a messenger.
• Amplification: a single gene may produce many mRNA. Each mRNA will be repeatedly translated into protein. In this way, a single gene can produce more than 10,000 protein products in a single cell cycle.

## Transcription

• mRNA is transcribed from DNA by an enzyme called RNA polymerase.

• Eukaryotes have 3 distinct types of RNA polymerase. RNA polymerase II transcribes all mRNA that are translated into proteins. RNA polymerase I and II transcribe genes coding for ncRNA’s like ribosomal RNA and tRNA. Bacteria have a single RNA polymerase, and a single type of RNA. RNA polymerases are large, having molecular weight around 500,000 Daltons.

• RNA polymerase in Eukaryotes is compartmentalized, meaning that it is only found inside the nucleus.

• Steps of transcription:

  1. Free RNA polymerase floats in the nucleus, randomly colliding with the DNA strand. When the RNA polymerase hits a special DNA sequence called a promoter, it sticks (this is the case in bacteria, in Eukaryotes the situation is more complex).

  2. *Initiation:* The RNA polymerase unwinds a short segment of DNA, and catalyzes the formation of a phosphodiester linkage between the first two nucleotides in the sequence.

  3. *Elongation:* Additional nucleotides are added in sequence, from the 5’ to the 3’ end (relative to the mRNA and sense
DNA strand, it is $3' \rightarrow 5'$ relative to the template DNA strand). A short RNA/DNA double helix is formed upstream of the enzyme, beyond which the DNA helix recoils.

4. **Termination:** A particular DNA sequence called a *termination signal* is reached. Elongation process stops, and the newly-formed RNA sequence and RNA polymerase detach from the DNA. The DNA recoils.

- The typical rate of RNA synthesis is $\approx 30$ nucleotides per second.

- Many genes have promoters that match the ideal promoter only partially. The degree of the match determines the transcription rate, ranging from the highest rate (perfect match to the ideal promoter), to the lower rates that occur when there is only a partial match to the ideal promoter.

- A Eukaryotic gene contains *introns* and *exons*. The introns do not correspond to amino acids in the product protein.

- The product of transcription is a *primary transcript* which comprises the introns, exons, and the $3'$ and $5'$ untranslated regions (UTR’s). While still inside the nucleus, RNA processing enzymes (the *spliceosome*) remove the introns, in a process called *RNA splicing*.

- Introns exist in one of three phases depending on whether the intron begins at position $3n$, $3n+1$ or $3n+2$ relative to the start codon. These are referred to as *phase 0*, *phase 1*, and *phase 2* introns respectively.

- The first two nucleotides of most introns are GT, and the final two are AG.
• In Eukaryotes, 5′ capping and a 3′ cleavage take place. The cleavage usually takes place 20-30 bp downstream of a sequence resembling AAUAAA. Subsequently a poly-A tail is added to the 3′ end of the transcript.

• Different proteins can be made from the same gene, by modifying the way that the primary transcript is spliced.

• Bacterial genes and mitochondrial genes have no introns – their genes correspond directly to proteins via the 3:1 code.

• Rough breakdown of the distribution of different mRNA species in a typical mammalian cell:

<table>
<thead>
<tr>
<th></th>
<th># transcripts</th>
<th># genes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abundant</td>
<td>12000</td>
<td>4</td>
<td>48,000</td>
</tr>
<tr>
<td>Intermediate</td>
<td>300</td>
<td>500</td>
<td>150,000</td>
</tr>
<tr>
<td>Scarce</td>
<td>15</td>
<td>11,000</td>
<td>165,000</td>
</tr>
</tbody>
</table>

**Translation**

• Translation is the process by which a messenger RNA is used to build a protein.

• The word length for the genetic code is 3, therefore, there are three *reading frames* on a DNA or RNA strand. Only one reading frame is read for each gene. The correct reading frame is located by *initiation factors* that bind to the promoter, and then locate the *start codon* (AUG). The start codon is usually embedded is a consensus sequence of the form GCCPuCCAUGG, where Pu can be either purine.

• Translation starts directly after the start codon, and continues to the *stop codon* (one of UAA, UGA, UAG).
The tRNA molecule functions as an adapter in the information transfer from RNA to protein. Specifically, the tRNA binds to three successive bases on the mRNA called a codon. Each tRNA matches a codon in the mRNA to the corresponding amino acid in the genetic code.

Each tRNA molecule is capable of carrying only a single type of amino acid. Each amino acid can be carried by at least one, but generally several types of tRNA molecules.

tRNA is an ncRNA molecule consisting of \( \approx 70 - 90 \) bases. The molecule is folded into a cloverleaf, with the 3’ and 5’ ends nearly adjacent. The carboxyl end of the amino acid is attached to the 3’ end of the tRNA. At the opposite end of the tRNA molecule is the anticodon, which matches the codon on the mRNA strand.

Enzymes called aminoacyl tRNA synthetases catalyze the reaction that covalently joins the amino acid to the tRNA molecule. There is a unique aminoacyl tRNA synthetase for each of the 20 amino acids. The resulting complex is known as an aminoacyl tRNA complex.

It is the tRNA molecule, and not the ribosome that catalyzes the formation of the peptide bonds in the growing amino acid sequence.

The ribosome manages the pairing of the anticodon of the tRNA molecule to the codon on the mRNA strand. The specificity of this reaction depends only on the tRNA, and not the amino acid that it carries (e.g., if the amino acid is chemically changed after the aminoacyl-tRNA complex is formed, the modified amino acid is inserted in the protein in the positions assigned to the
unmodified amino acid).

- RNA sequences, like proteins, can be post-translationally modified. tRNA molecules undergo an unusually large amount of post-translational modification.

- By mass, the ribosome is about half RNA and half protein. It is believed that the primary catalytic activity of the ribosome is effected by the RNA, not the protein.

- Protein synthesis makes about 1 error per 10,000 residues. Errors can occur at two points: i) the wrong amino acid may be attached to the tRNA molecule, and ii) the wrong tRNA molecule may be matched to the mRNA codon by the ribosome. Both points have a checking and error correcting mechanism.

- Prokaryotic protein synthesis is sufficiently different from Eukaryotic protein synthesis that certain chemicals will inhibit prokaryotic synthesis without affecting eukaryotic synthesis. These are the antibiotics.
Control of Gene Expression

• Points of expression control in a Eukaryote:

1. Transcription: mRNA for different genes are transcribed from the DNA at vastly different rates.
2. RNA processing: The mechanisms that control the splicing out of the introns may proceed at different rates. This may limit expression, since the introns must be spliced out before the mRNA can be transcribed.
3. RNA transport: mRNA’s are synthesized in the nucleus, while protein synthesis occurs on the ribosome, which is external to the nucleus. The nuclear membrane acts as a regulator by selectively permitting only certain mRNA’s to enter the cytoplasm.
4. Translation: the ribosome acts as a regulator, by translating different mRNA’s at different rates.
5. mRNA degradation: enzymes in the cytoplasm target certain mRNA types for degradation, lowering their half life and preventing them from being translated.
6. Protein activity: some proteins are synthesized in an inactivated form, and only become active when they are post-translationally modified by an enzyme. Other proteins are synthesized in an activated form, but may be deactivated soon after they are created.
Transcriptional control is generally effected by DNA binding proteins or gene regulatory proteins that enhance or diminish the ability of RNA polymerase to initiate transcription.

DNA binding proteins slide along the major groove (more rarely along the minor groove) and can recognize their binding sites without unzipping the helix.

The region of a DNA binding protein that contacts the DNA often belongs to one of a number of families such as zinc finger, helix-loop-helix, leucine zipper, helix-turn-helix.

A transcriptional repressor inhibits transcription by binding to the promoter and blocking access by RNA polymerase.

A transcriptional activator binds to an inefficient promoter and causes it to become more efficient at binding RNA polymerase.

Many proteins influence gene expression indirectly by binding as a ligand to a DNA binding protein and changing its activity. These proteins do not directly contact the DNA.

Many bacterial genes are grouped into clusters called operons that share a promoter and are transcribed as a unit.

In Eukaryotes RNA polymerase will generally not bind to the promoter in isolation. Other proteins called ubiquitous (general) transcription factors as well as tissue-specific transcription factors must be present as well. Among the ubiquitous transcription factors, the most common example is TFIID, which binds to the TATA box around 25bp upstream of the TSS. Other prominent examples are the CCAAT box (-75bp) and INR element (+1bp).
• A *cis-acting* regulatory element is a DNA sequence that controls the expression of an adjacent coding region. Generally they are upstream, but they may be downstream or located in introns.

• A *trans-acting* regulatory element is a translated DNA sequence that influences the expression of a distantly located gene indirectly through its protein or RNA product.

• *Enhancers* and *silencers* are cis-acting elements that may sit somewhat further from the TSS. Their mode of action may be through the DNA molecule folding back on itself bringing the element within close spatial proximity to the core promoter.

• *Response elements* are *cis-acting* elements that respond to environmental changes (often mediated through a signal transduction pathway, e.g. involving hormones and G-protein coupled cell surface receptors).

• Some proteins control gene expression by affecting the chromatin structure.

• DNA methylation is a means of permanently inhibiting transcription in certain genomic regions. This is often used to block the expression of certain genes in a tissue-specific fashion. Methylation patterns are inherited by the daughter cells giving rise to an *epigenetic* mode of transcriptional control.