Molecular Biology

Molecular Biology

The biochemical underpinnings of information transfer

ALEXEY MERZ; TIMOTHY CHERRY; AND KULLBERM

Contents

Syllabus for sessions Day 1 Hours 2, 3, 4; Day 2 Hours 1, 2, 3, 4; Day 5 Hours 1, 2; Day 21 Hours 1, 2.

Prepared by A.J. Merz, Ph.D. and T. Cherry, M.D., with assistance from Y. Kwon, Ph.D. Autumn, 2020, and Max Kullberg, Ph.D. and Pamela Langer, Ph.D. 2023

1. Protein, Nucleic Acid Building Blocks

Session Level Objectives (SLOs): after completing the session, students will be able to:

SLO 1. Summarize the key properties of amino acids (charge, polarity/hydrophobicity, aromatic character, reactivity) and the structure of the peptide bond.

SLO 2. Explain the "Central Dogma" of molecular biology.

SLO 3. Describe the general structure of nucleotides: sugar, phosphate, and base; explain the polarity of DNA and RNA chains.

SLO 4. Describe the structure of the DNA double helix and RNA secondary structure. Explain the forces that stabilize the double helix, in particular the role of water.

SLO 5. Explain how DNA and RNA polymers are synthesized. Explain the pivotal role of phosphotransfer reactions in DNA and RNA biology and in the cellular energy economy.

SLO1. Summarize the key properties of amino acid side chains (charge, polarity/hydrophobicity, aromatic character, and reactivity) and the structure of the amino acid backbone and peptide bond.

Proteins:

Proteins are the major components (by mass) of the body. Proteins have a multitude of functions: they provide structure (tensile strength; elasticity); they do mechanical work (moving chromosomes; contracting muscle); they sense the internal and external environment; they process and transmit signals in response to this information; and they carry out most of the enzymatic functions required for metabolism, for DNA replication and repair, and for gene expression. To understand the functions of proteins, we must think about their synthesis and structure.

Amino Acids & Polypeptides

A protein is made from one or more linear polypeptide chains — strings of covalently linked amino acids. The general structure of an amino acid is shown in Fig. 1. There are twenty major amino acids in eukaryotes, including humans. They differ by the side chain [R]. For the purposes of this course you **do not need to memorize** the structures of the amino acids. However, we will need to consider their chemical and physical properties.

Some amino acids we can make ourselves from other chemical precursors. Others, we cannot synthesize; these must be obtained through dietary intake. We'll consider amino acid metabolism in some detail later in the course.

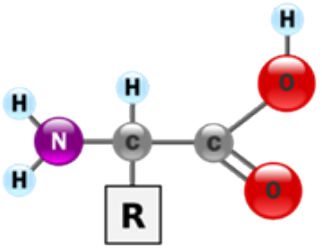


Fig. 1. A generic amino acid. Source: Wikimedia

Amino acid side chains are *chemically diverse*. Consequently, a polypeptide's properties result from its linear sequence of amino acid residues.

- Amino acids can be **hydrophilic** (*polar* or *charged*), or **hydrophobic** (*apolar*; "greasy").
- In a folded protein, hydrophilic side chains are usually exposed to the aqueous solvent. Hydrophobic side chains tend to be buried within the folded protein, so that they are shielded from

the aqueous solvent.

- Amino acid side chains have diverse **chemical reactivities**. For example, serine threonine, and tyrosine all have terminal hydroxyl groups that can form ester bonds. Arginine and lysine contain positively charged amine groups. Cysteine contains a redox-active sulfhydryl group.
- In a **polypeptide**, amino acids are linked in a linear chain, head-to-tail. Linkages between amino acids are called **peptide bonds** (Fig. 2).
- A polypeptide **backbone** has a polarity: At one end, there is primary amine group. This is the amino- or N-terminus (Figs 1,2). At the other end of the backbone is a carboxylic acid group. This is the carboxy- or C-terminus (Figs 1,2). A peptide bond can be severed by hydrolysis, liberating the amine on one residue and the carboxyl group on another. The enzymes that catalyze this reaction are called **proteases** or **peptidases**.
- Almost all polypeptides synthesized by cells are synthesized one amino acid residue at a time. New residues always added to the C-terminal end of the growing chain. For this reason, **we write down amino acid sequences from N-to-C**.
- The N-to-C **primary sequence** of a polypeptide, along with any additional covalent modifications to the polypeptide, controls how the polypeptide **folds** into a three- dimensional structure. This is a key point: **sequence controls structure, and therefore function.**

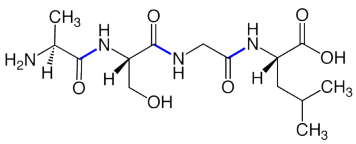


Fig. 2. Peptide bonds (blue) in a short (tetra)peptide. Source: Wikimedia

SLO 2. Explain the "Central Dogma" of molecular biology.

Sequence Information and the "Central Dogma" of Molecular Biology

The linear sequence of almost all polypeptides (there are a few exceptions) is stored, in encoded form, in the DNA of our genome. The flow of sequence information occurs whenever DNA, RNA, or protein polymers are synthesized:

DNA —*replication*—> **DNA** —*transcription*—> **RNA** —*translation*—> **polypeptide**

A critical point: each of these processes entails a series of chemical reactions. Each reaction is catalyzed by specific enzymes and is controlled by the laws of statistical thermodynamics. Consequently, *biological information transfer processes are never error-free*. They *cannot* be. Consequently, cells spend enormous energy, materials, and time to reduce and cope with errors in nucleic acid and protein synthesis. When these tactics fail, the consequences can be devastating.

A significant portion of this course will focus on errors in biological information transfer.

On the flip side, sequence changes (through DNA replication errors and other mutational processes) are responsible for all the richness and splendor of human genetic diversity. Understanding this diversity is essential to understanding and treating human disease and will only become more important as we are deluged with human DNA sequence data.

Fig. 3 shows that the cost of DNA sequencing has dropped faster than the price of electronic integrated circuits dropped from the 1970s to the present (Moore's Law). In 2016 it cost about $1000 to sequence a human genome.

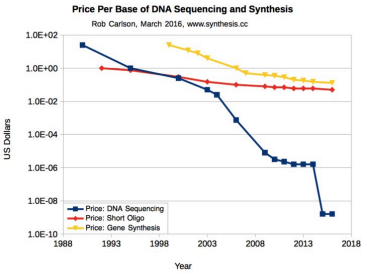


Fig. 3. Cost of DNA sequencing and synthesis, per base. The cost scale (y-axis) is logarithmic. Source: Rob Carlson www.synthesis.cc

At present, you won't see much DNA sequence data used in most clinical settings. But by the time you complete your medical training, the cost will have fallen to only a fraction of the current $1000 per genome — comparable to many standard lab tests.

Thus, it is essential that you should obtain a working understanding of human genetic variation and its consequences for health and disease.

SLO 3. Describe the general structure of nucleotides: sugar, phosphate, and base; explain the polarity of DNA and RNA chains.

The protomer of a DNA or RNA polymer is the **nucleotide** (Fig. 4). A nucleotide contains a **base**, a 5- carbon (**pentose**) sugar, and one or more **phosphate groups**. If the sugar doesn't have a phosphate group on it, the pentose-base unit is called a **nucleoside**.

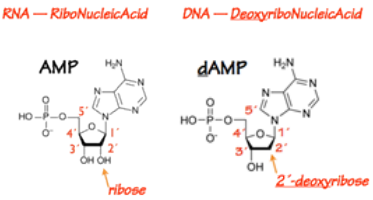


Fig. 4. DNA and RNA nucleotides. Note that positions on the pentose sugar are numbered 1′ (1-prime), 2′, etc.

In RNA the pentose is **ribose**. In DNA the pentose is **2′-deoxyribose** — ribose lacking a hydroxyl at its 2′position.

To each sugar is attached a **base** (Fig. 5). The base is always attached at the 1′position of the pentose sugar through a **glycosidic bond**.

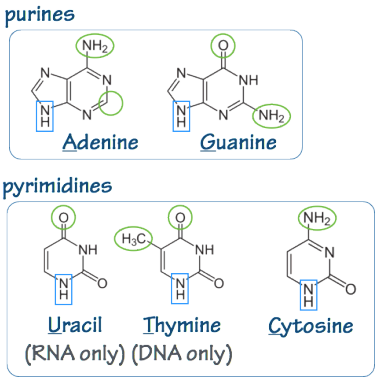


Fig. 5. Bases found in DNA and RNA chains. You do not need to memorize these structures but you should spend some time becoming acquainted with their features.

In DNA the bases are: adenine (A), thymine (T) guanine (G), and cytosine (C). In RNA, thymine (T) is replaced by uracil (U).

The **nucleotides** (base-sugar-phosphate) are called: adenosine (A), guanosine (G), thymidine (T), cytidine (C), and uridine (U). Often the names are written to indicate the phosphorylation state: Adenosine diphosphate (ADP), etc.

DNA and RNA chains are strings of linked nucleotides. Each chain (Fig. 6) consists of a **backbone** made out of alternating sugar and phosphate groups. Put a bit differently, the pentose sugars are linked by **phosphodiester bonds**.

As with proteins, DNA and RNA chains have **polarity**. This is defined by the orientation of the pentose sugar: one linking phosphate is attached at the 3′ position on the pentose, and one is attached at the 5′ position (see Fig. 4 and compare to the backbone in Fig. 6).

In biological polymerization reactions, nucleotides are always added at the 3′ end of an elongating chain. That is, the chain is polymerized 5′-to-3′. This leads to a convention: **we write DNA and RNA sequences 5′-to-3′**— unless explicitly specified otherwise.

SLO 4. Describe the structure of the DNA double helix and RNA secondary structure. Explain the forces that stabilize the double helix, in particular the role of water.

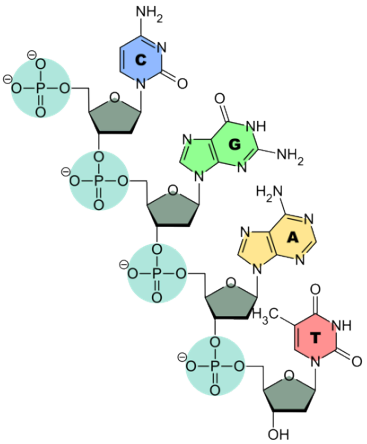


Fig. 6. DNA polymer. Source: Wikimedia

Strands of DNA or RNA can **hybridize** (anneal) to form **double-stranded structures** like the familiar DNA double helix.

- Specificity in hybridization is provided by **base-pairing**. A pairs with T (or U), G with C. Accuracy in pairing is promoted by favorable, **non-covalent hydrogen bonds** and by **shape complementarity**.
- The two pentose-phosphate backbones are at the exterior of the double helix. This makes sense: the sugars are very polar, and every phosphate group carries a negative charge. Both sugar and phosphate are hydrophilic — they like to interact with water.
- The two sugar-phosphate backbones run **antiparallel**, like the traffic on a two-way street: one strand runs 5′-3′, and the other runs 3′-5′.
- Consecutive base-pairs **stack** like plates at the center of the helix, so close together that water is excluded. This also makes sense: the bases are flat and relatively hydrophobic. Their flat surfaces favorably interact (stack) with one another and are shielded from aqueous solvent. *This also protects the bases from certain kinds of chemical attacks.*

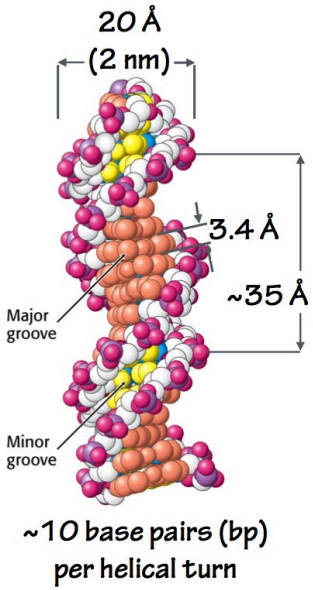


Fig. 7. DNA double helix.

- **To summarize**: the stability of a double helix is controlled by base-pairing and *also* by other forces: separation of the negatively- charged phosphates, favorable stacking interactions between the flat bases, and the resulting shielding of the hydrophobic bases from the aqueous solvent.

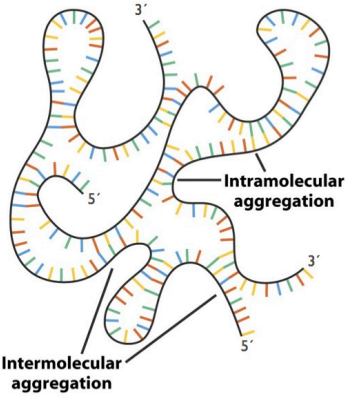


Fig. 8. Hybridization within and between RNA strands. Note that the backbones in hybridized regions are antiparallel.

- The double helix has a **minor groove** and **a major groove** (Fig. 7). The major groove is critical: it allows proteins to touch the bases and "read" the DNA sequence, as if by braille. Thus, regulatory proteins can identify and bind to specific short

DNA sequences, *without pulling the two strands apart*.

- RNA can hybridize with RNA, or with DNA. In RNA biology, hybridization between complementary sequences *on a single strand* is of special importance (Fig. 8). Hybridization allows the formation of *hairpin* structures with short regions of double helix. These *secondary structure* elements can combine to generate complex *tertiary structures* including tRNAs and ribosomes, which are critical in protein synthesis.
- DNA and RNA diagnostics including microarrays and in situ hybridization are based on sequence-specific hybridization of short **oligonucleotide probes** to DNA or RNA analytes (samples).

SLO 5. Explain how DNA and RNA polymers are synthesized. Explain the pivotal role of phosphotransfer reactions in DNA and RNA biology and in the cellular energy economy.

DNA replication and RNA transcription are chemical reactions that involve information transfer.

- To provide a **template** for DNA replication or RNA transcription, the DNA double helix must be locally pulled apart. Separation of the two strands is called **melting** or **denaturation**.
- A highly specialized **DNA or RNA polymerase enzyme** moves along a **template strand.** The polymerase elongates the **nascent chain** by testing the base-pairing of incoming nucleotides with the template.
- If the base-pairing is correct, the enzyme triggers the chemistry**: the incoming nucleotide is added to the 3′end of the nascent chain**. (Fig. 9
- To power polymerization, the incoming nucleotides are **NTPs** (nucleotide triphosphates, for RNA) or **dNTPs** (deoxynucleotide triphosphates, for DNA):

...pNpN-3′ + NTP
—polymerase—> **...pNpNpN-3′ + PPi**

Each lower-case "p" in the above scheme is one phosphodiester bond. The product of the reaction is a nascent chain with one additional nucleotide residue. A molecule of inorganic pyrophosphate (PPi = $(P_2O_7)^{4-}$) is evolved.

- The polymerization reaction is potentially reversible. To make the reaction irreversible, the enzyme **inorganic pyrophosphatase** destroys the evolved pyrophosphate in a highly favorable, effectively irreversible, reaction:

PPi + H2O
—pyrophosphatase—
> 2 Pi + heat

In later sessions, we will see the destruction of pyrophosphate used to make additional metabolic reactions, such as protein and lipid synthesis, irreversible.

- In most cases the polymerase enzyme remains tightly bound to the template and nascent strands, and the elongation cycle begins again. The ability of an enzyme to catalyze many polymerization cycles without falling off (dissociating) from a template is called **processivity**. Some DNA and RNA polymerases have processivities of a million bases or more.

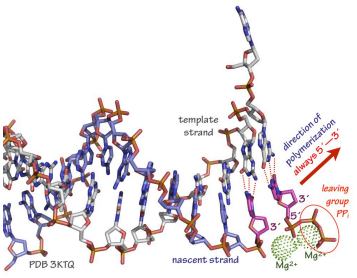


Fig. 9. Arrangement of template strand, incoming dNTP (in this case, dCTP), and stabilizing Mg2+ ions in a typical DNA polymerase active site. The polymerase itself is not shown in this rendering. Source: Merz, based on PDB 3KTQ

The main polymerases we'll think about in this course are the enzymes that catalyze DNA replication and RNA transcription. However, other DNA and RNA polymerase enzymes exist. There are RNA polymerases that use RNA as a template, and DNA polymerases that use RNA as a template ("**reverse transcriptases**"). Our cells use these alternative polymerases for specialized housekeeping functions such as telomere maintenance. RNA viruses and retroviruses use these classes of enzymes in their infection and replication cycles, as you'll see in the Infection and Immunity block.

2. Transcription, Translation

Session Level Objectives (SLOs): after completing the session, students will be able to:

SLO 1. Explain how cells overcome three major challenges to replicate their genomes.

SLO 2. Explain why different genes need to be expressed at different *rates*, in different *cells*, at different *times*.

SLO 3. Describe the functions of the three mammalian RNA polymerases.

SLO 4. Explain the structure of a mammalian gene, including regulatory and coding elements.

SLO 5. Describe how RNA polymerase II (RNAP II) is directed to gene promoters and the roles of general transcription factors.

SLO 6. Explain how transcription factors control which proteins are made in different cell types.

SLO 7. Outline the processing reactions that precede export of an mRNA from nucleus to cytoplasm.

SLO 8. Outline the major steps of protein synthesis, including the roles of mRNA, tRNA, tRNA synthetases and ribosomes.

SLO 9. Describe the basic ways in which microRNA (miRNA) molecules control gene expression.

Common Themes of Information Transfer

DNA —*replication*—> **DNA** —*transcription*—> **RNA** —*translation*—> **polypeptide**

There are common themes in these reactions.

- Biological polymerization reactions always have an intrinsic directionality (polarity): In DNA replication and RNA transcription, the **template strand is always read 3′-to– 5′**, and the **nascent chain is always synthesized 5′-to-3′**. In protein synthesis (translation), the **mRNA template is always read 5′-to-3′**, and the **nascent polypeptide is always synthesized N-to-C**.
- The polymerase must be accurately positioned at a start site on the template. In each case this process is called **initiation**, and in each case initiation entails several regulated steps.
- The polymerase has an **elongation** cycle, which is what it sounds like. This is the major biosynthetic stage.
- A signal on the template signals **termination** of polymerization. Termination involves disassembly of the elongation machinery and the release of templates and products.

This is a general framework. We will not focus on every stage for each process, but rather on key stages that illustrate important concepts.

SLO 1. Explain how cells overcome three major challenges to replicate their genome.

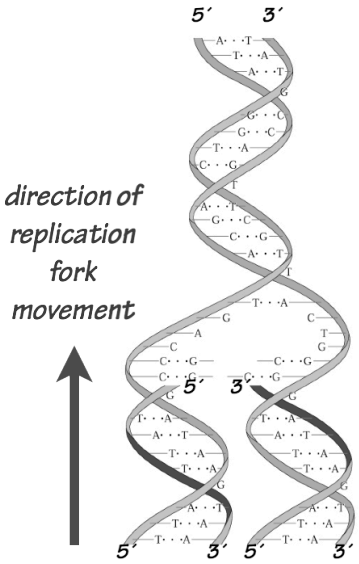


Fig. 1. DNA replication fork. Two nascent chains are synthesized from two template strands.

To replicate its genome, the cell must overcome several challenges:

- There are two DNA strands to be replicated.
- The two strands run in opposite directions (they're antiparallel).
- Replication must be accurate.
- Enormous amounts of DNA are replicated: 6 Gbp/cell (= 6×109 base pairs per cell).
- One and only one copy of each of the 46 chromosomes must be segregated into each daughter cell.

How is this done? Here we provide a cursory outline of DNA replication. Later in the course we will look more closely at replication, mitosis, and meiosis in the contexts of the cell division cycle, genetic inheritance, and cancer.

For now, there are **three core concepts** about DNA replication that you need to know:

- **DNA replication is semiconservative** (Fig. 1). This means that when a cell divides, the DNA duplexes in each daughter cell contain one of the parent cell's original DNA strands (which is used as a **template** for polymerization, and one newly synthesized or **nascent** strand.

- **DNA replication is semi-discontinuous**(Fig. 2). This means that the nascent strand associated with one of the two strands is synthesized in short segments called **Okazaki fragments**that are then knitted together.

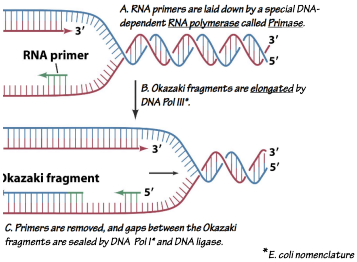


Fig. 2. Semiconservative, discontinuous replication. Each replication fork has a leading strand which is synthesized continuously, and a lagging strand which is synthesized discontinuously as a series of Okazaki Fragments. As each Okazaki fragment is completed, the preceding RNA primer is removed, and the fragments are linked, or ligated, by an enzyme called DNA ligase.

Discontinuous replication on one strand is necessary because the DNA polymerase can only add nucleotides to the 3′end of a nascent chain, but the template strands are **antiparallel** (Fig. 1).

An important **difference** between DNA and RNA polymerases is that DNA polymerases can only extend pre-existing nascent chains, while RNA polymerases can begin new chains from a single nucleotide. Thus, DNA polymerases invariably require a short RNA primer. The primer is made by a special RNA polymerase called primase (Fig. 2).

- **DNA replication is bidirectional** (Fig. 3). This means that at the DNA replication **origin** — the site where polymerization is initiated — two **replication forks**diverge from the

origin. The replication machinery at each fork synthesizes one **leading strand** and one **lagging strand**(which is assembled from Okazaki fragments).

Because human chromosomes can be hundreds of millions of base pairs in length, replication is done in parallel starting at many **replication origins** on each chromosome.

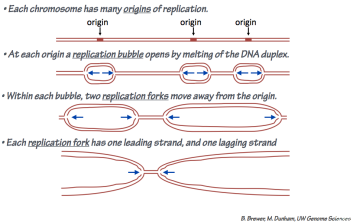


Fig. 3. Bidirectional replication of mammalian chromosomes. The DNA on each chromosome has many origins of replication. A replication bubble is opened at each origin. Each replication bubble has two replication forks that propagate away from the origin.

Each chromosome begins as *one* piece of double-stranded DNA. Replicating the very ends of the chromosomes, the **telomeres**, presents special problems on the lagging strand. Telomere DNA is therefore maintained by a special enzyme called **telomerase.**

SLO 2. Explain why different genes need to be transcribed at different rates, in different cells, at different times.

Most of the mammalian genome consists of non-coding DNA. A minority (~2%) of the human genome actually serves to encode mRNAs that are used as templates for protein synthesis. A similarly small fraction of the genome encodes other varieties of biologically important RNA molecules.

The fundamental question about gene expression is this: each of us has many different cell types, but only one genome. The different cell types are different because they make different proteins:

muscle cells make contractile proteins; nerve cells have the enzymes needed to manufacture neurotransmitters; osteoblasts have the protein machinery needed to manufacture bone, and so on.

Moreover, even a single cell type needs to make different proteins at different times: we synthesize and secrete insulin (a peptide hormone) when we eat. We remodel entire tissues and organs throughout growth, in response to injury, and during pregnancy. *How does this happen?*

DNA —*transcription*—> **RNA**
—*translation*—> **polypeptide**

The level of any given protein is a function of competing processes: synthesis and destruction. Protein synthesis requires an mRNA template, and the abundance of the mRNA encoding any given protein is also regulated by a balance of synthesis and destruction.

SLO 3. Describe the functions of the three mammalian RNA polymerases.

RNA Transcription

The cell makes RNA molecules that do different things:

- **mRNA** is the template for protein synthesis (translation).
- **tRNA** and **rRNA** are core parts of the protein synthesis machinery.
- Diverse RNA molecules are involved in regulating gene expression and other processes. Examples include micro RNA (**miRNA**) and long non-coding (**lncRNA**). Many other examples are emerging.

The various RNAs are made by three RNA polymerase (RNAP) enzymes:

- **RNAP I** makes most of the ribosomal **rRNA** — the most important part of the ribosome, the enzyme that synthesizes polypeptides.
- **RNAP II** makes all of the messenger **mRNA** — the templates for polypeptide synthesis.
- **RNAP III** makes transfer **tRNA** — the carrier of activated amino acids for polypeptide synthesis.

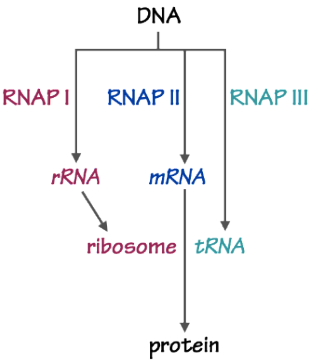


Fig. 4. Division of labor in transcription.

We will focus on transcription by RNAPII, because its activity controls the levels of mRNA templates for protein synthesis. The underlying principles by which RNAP I and III operate are similar.

SLO 4. Explain the structure of a mammalian gene, including regulatory and coding elements.

A gene contains two kinds of sequences:

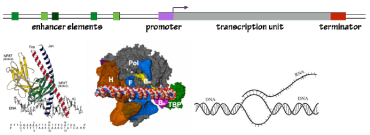


Fig. 5. Top: *Arrangement of a typical transcription unit and its regulatory DNA sequences (not drawn to scale).* ***Bottom Left:*** *Activating transcription factor proteins bound to a DNA enhancer sequence. Note that the transcription factors are "reading" the sequence by probing the major groove of the DNA double helix.* ***Middle:*** *RNAP II ("Pol") and the GTFs bound to a DNA promoter sequence.* ***Right:*** *Diagram of a transcription bubble with a nascent RNA chain. (In this case, the bubble would be moving from right-to-left. You should be able to identify the 5′ and 3′ ends of each DNA and RNA strand.) Sources: Nature 392:42; Kornberg, 2006 Nobel Lecture; Calladine, Understanding DNA.*

1. The **transcription unit** is the DNA sequence used as a **template** to synthesize RNA.
2. **Regulatory sequences** tell RNAP where to initiate and terminate transcription. They allow cells to control which genes are actively transcribed ("expressed"), and which are silent. These sequences can be further sub-divided:
    - The **promoter** sequence directs RNAP II and associated general transcription factors to the transcriptional **start site**.
    - **Enhancer** sequences, usually 10-30 bp in length, bind **transcription factors**, or **activator proteins**, that instruct RNA polymerase to become active at the promoter. Each gene is controlled by different enhancer elements.
    - Different cells contain specific sets of **transcription factors**. This is the main basis for cell-type-specific gene regulation! Enhancers can sit right next to the promoter, or tens of thousands of base pairs distant. Enhancers are usually upstream of the promoter but they can also be embedded within the transcription unit or even

downstream of it.

- The **terminator** tells RNAP that it has reached the end of the transcription unit.
- There are other regulatory elements in the genome as well. For example, **silencer** sequences bind factors that, as the name suggests, decrease transcription. **Insulator elements**ensure that different genes are subject to independent regulation.

SLO 5. Describe how RNA polymerase II (RNAP II) is directed to gene promoters and the roles of general transcription factors.

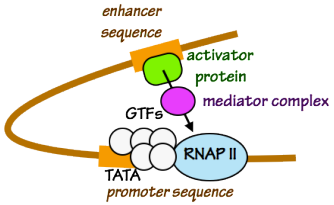


Fig. 6. Activation of RNAP by an activating transcription factor and a coactivator (in this example, "mediator complex"). The sequence TATA is often found in promoter regions. Enhancer sequences can be hundreds or thousands of base pairs distant from a promoter sequence.

Sequence of Events in Transcription:

1. Genomic DNA is packaged into **chromatin**. Transcription is regulated in part by how densely packaged a given gene is, and hence, how accessible its regulatory sequence elements are. Later, we will discuss how DNA packaging is controlled.
2. To initiate transcription of a gene, RNAP II must be directed to the promoter. This is done by the **General Transcription Factors** (Figs. 5 and 6). The GTFs recognize and bind to promoter sequences. They place RNAP II at the start site. GTFs then locally melt the DNA at the promoter, separating the two DNA strands to form a **transcription bubble**.
3. The GTFs are "general" transcription factors because they are **always needed** for initiation of transcription. However, the GTFs **do not** have the ability to regulate **when** RNAP actually initiates RNA synthesis.

4. In other words, GTFs are necessary but not sufficient for initiation of transcription.
5. **Transcription factors** that bind to regulatory promoter, so intervening DNA must loop out in order for transcription factors, coactivators, and the initiation complex (GTFs + RNAP II) to touch one another. **enhancer sequences** are needed to **activate** transcription by RNAP II and the GTFs (Figs. 5 and 6)
6. Activating transcription factors "talk" to RNAP II and the GTFs by binding to **coactivators** that touch RNAP II and the GTFs (Fig. 6). Together, these events cause the pre-initiation complex — RNAP II and the GTFs — to initiate RNA polymerization. As we will see in a later session, transcription factors can also control chromatin structure.
7. The chemistry of RNA polymerization is similar to the chemistry of DNA polymerization. The most important *differences* are:

    - No primer is needed for RNA synthesis.
    - **NTP**s are used for RNA synthesis, not 2-deoxy **dNTP**s.

8. As RNAP II elongates the nascent mRNA chain, it moves along the template strand of the transcription unit (Fig. 5). The replication bubble moves as RNAP II "crawls" along the DNA template strand. In other words, the DNA double helix melts in front of RNAP II and re- hybridizes (anneals) behind it.
9. When RNAP II reaches a **terminator** sequence (Fig. 5), the newly-synthesized RNA chain is released, RNAP is removed from the template strand, and the transcription bubble collapses.

SLO 6. Explain how transcription factors control which proteins are made in different cell types.

The reason we care so much about the mechanics of RNAP II transcription is that this process controls which mRNA transcripts are produced, and in what abundance. This in turn controls the specific repertoire of proteins that can be made by each cell.

Fig. 7 shows the regulatory sequences of genes that encode some key proteins made only in specific kinds of cells: skeletal muscle cells, heart muscle cells, and cells in the lens of the eye.

Each type of DNA enhancer element sequence is recognized and bound by specific activating transcription factors — shown here by colored shapes. Humans have about 2,000 different transcription factors.

By producing specific combinations of transcription factors, each cell specifies which subsets of genes are actively transcribed, and in what quantities.

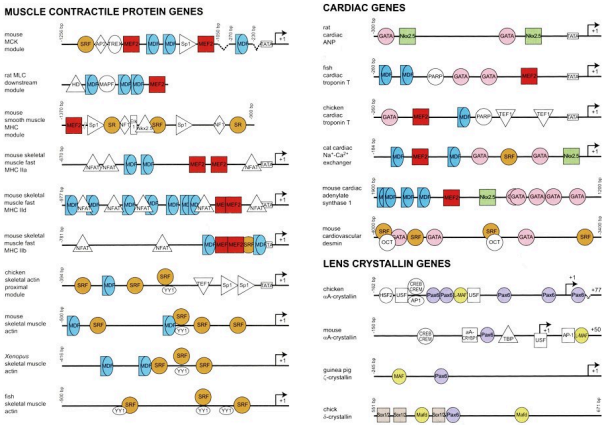


Fig. 7. Regulatory enhancer elements of protein-coding genes that are selectively transcribed in three different cell types. The arrow at the right end of each diagram identifies the promoter and the first base (+1) of the transcription unit. Source: Davidson, The Regulatory Genome.

This concept is so important that it bears repeating:

The specific array of transcription factors, present in a given cell, shapes that cell's pattern of gene expression — and thus, that cell's overall protein complement, the cell's identity (muscle, fibroblast, neuron, etc.) and its functional characteristics.

Another important point is that transcription factors are proteins. Consequently, **the genes that encode transcription factors are**

themselves subject to transcriptional regulation. By transcribing and translating specific transcription factors the cell can execute temporary or stable programs of gene expression in response to developmental cues and other signals, such as food or infection.

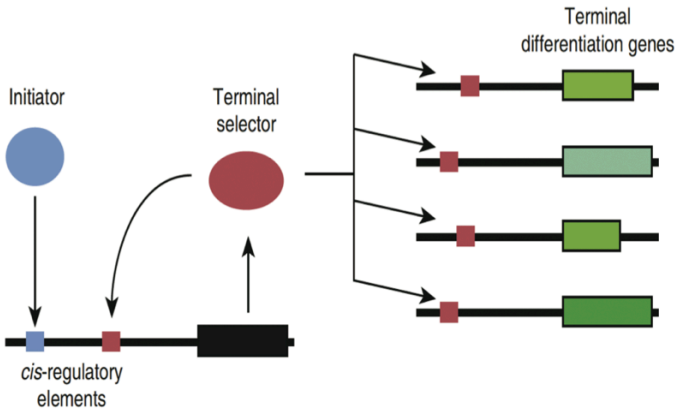


Fig. 8. Positive feedback loops maintain gene expression programs. Here, an "initiator" transcription factor binds an enhancer on the gene encoding a "terminal selector" transcription factor. When this gene is transcribed and the resulting mRNA is translated, the resulting protein binds to another enhancer in its own gene, and ensures that the terminal selector gene continues to be transcribed. The terminal selector also stimulates transcription of other genes needed for specific functions. Source: PNAS 110:7101

SLO 7. Outline the processing reactions that precede export of an mRNA from nucleus to cytoplasm.

mRNA Processing and Export

DNA replication and RNA transcription both occur in the cell's nucleus. However, proteins are synthesized in the cytoplasm. To

serve as templates for protein synthesis, mRNA molecules must be exported from the nucleus to the cytoplasm. This occurs at a special portal in the nuclear membrane, the **nuclear pore** (Fig. 9). The nuclear pore is an immense molecular assemblage that precisely controls the passage of mRNA and other macromolecules into, and out of, the nucleus. This is another theme that we will encounter again and again: biosynthetic products made in one cellular organelle are shuttled to another location — as with stations on an assembly line.

In the nucleus, the initial "raw" RNA transcript made by RNAP II must be **processed** (Fig. 10). Only then is the mature mRNA exported from the nucleus to the cytoplasm, where it will be used as a template for protein synthesis.

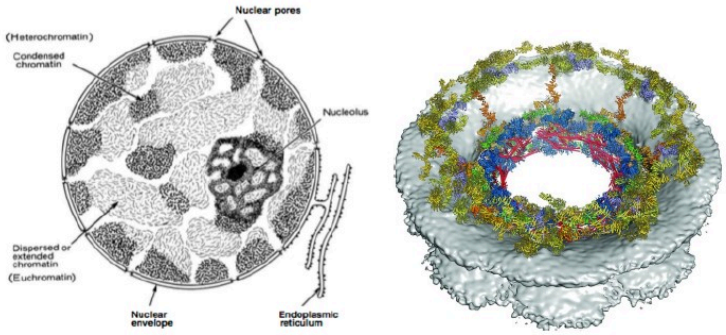


Fig. 9 Nuclear pores control the movement of macromolecules in and out of the nucleus. On the left, an entire nucleus is shown, with pores spanning the nucleus's double membrane. On the right, a partial structure of a single nuclear pore is depicted. The nuclear membrane is shown in gray. Protein components of the pore are shown in color, with each protein shown in a different hue. Sources: Fawcett, The Cell; Science, DOI: 10.1126/science.aaf1015

1. At the 5′end of the mRNA (the "beginning" of the transcript), a special nucleotide, **7– methylguanosine**, is attached through a covalent bond. **This is called the 5′ cap**.

- Later, in the cytoplasm, the 5′ cap will tell the protein synthesis machinery that the RNA bearing the cap is a messenger mRNA — a template for protein synthesis — and not some other type of RNA.

2. At the 3′ end of the transcript, a string of adenosine (A) nucleotides is added. This is the **poly-A tail** of the mRNA. The poly-A tail is constructed by a special enzyme (poly-A polymerase) through a **non-templated** polymerization process.

   - The poly-A tail will signal **export** of the mRNA from nucleus to cytoplasm. It will also control the **stability** (the half-life) of the mRNA once it's in the cytoplasm.

3. The mRNA is **spliced** to remove **introns** and ligate (join) **exons** together. This step is somewhat involved and *extremely* important, so we'll examine it in a bit more detail.

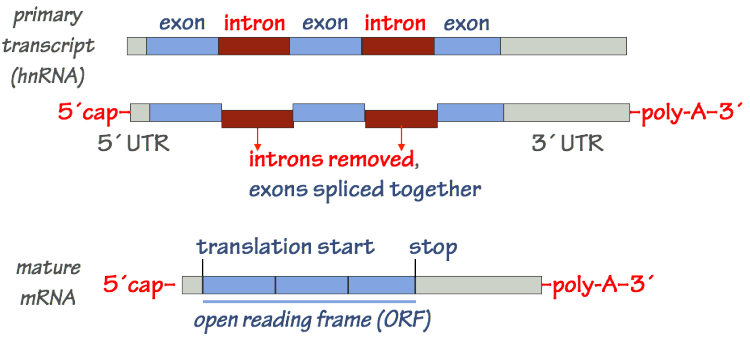


Fig. 10. Processing of primary transcripts in the nucleus yields mature mRNA molecules. Mature mRNAs are then exported to the cytoplasm to serve as templates in translation. The poly-A tail is usually 150-250 nt long. The three kinds of RNA processing machinery (the capping and poly-A addition enzymes, and the splicing machinery) are placed onto the growing transcript by binding to a flexible "tail" on the RNAP II enzyme itself. In other words, mRNA elongation by RNA polymerse II, and mRNA processing, are coupled reactions – all of which occur in the nucleus.

Differential mRNA splicing

The maturation of a large mRNA molecule may entail *dozens* of splicing reactions. In different cell types, these splicing reactions may be regulated so that not every exon ends up in each final, mature mRNA molecule. Consequently, **a single transcription unit may encode more than one mRNA variant**, with each derived from a different combination of exons (Fig. 11).

Differential splicing allows the ~20,000 protein-coding *genes* in the human genome to encode substantially more than 20,000 distinct mRNA templates and, thus, a much greater diversity of proteins.

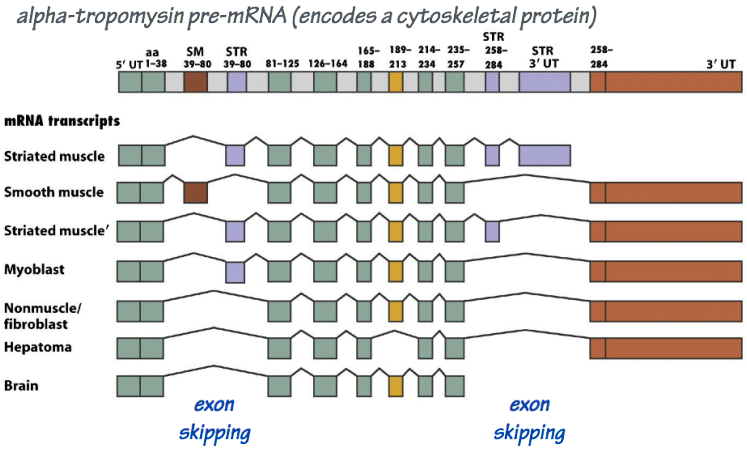


Fig. 11. Differential splicing of pre-mRNA's transcribed from a single gene can yield several different mature mRNA transcripts, each encoding a different polypeptide variant. The exons are shown as boxes, and the introns that are spliced out, by connecting lines. Note that the splicing pattern is different in different cell types, and that one cell type may produce more than one mRNA splice variant. Introns in the pre-mRNA are shown in gray. UT indicates un-translated regions at the 5′and 3′ends of the transcript. The 5′ cap and 3′poly-A tail are not depicted in this diagram, but would be present on each mature mRNA.

To summarize:

- Transcription initiation controls how many mRNA transcripts get made.
- Different cells have different activating transcription factors.
- Different genes have different enhancers that bind different transcription factors.
- mRNA cap addition signals that the mRNA will be a template for protein synthesis.
- Differential splicing controls which exons are in the mature mRNA template, and thus the sequence of the resulting polypeptide.
- The poly-A tail (along with other features of the mRNA) controls nuclear export and the stability of the mRNA — how long it persists in the cytoplasm.

SLO 8. Outline the major steps of protein synthesis, including the roles of mRNA, tRNA, tRNA synthetases and ribosomes.

Here we summarize how polypeptide chains are synthesized and how they fold into their correct three-dimensional configurations.

The notes for this section begin with two charts: first, the assignments of RNA codons to amino acids (the genetic code); second, the chemical and physical properties of the 20 regular amino acids listed by features. This table also includes the rare amino acid selenocysteine, sometimes considered as the "21st amino acid," although it is a modification of cysteine.

You do NOT need to memorize Tables 1 and 2! You **do** need to be able to apply their content.

Table 1

The Standard Genetic Code

First position (5´end)	Second position: U	Second position: C	Second position: A	Second position: G	Third position (3´end)
U	UUU ψ Phe	UCU Ser	UAU ψ Tyr	UGU Cys	U
	UUC ψ Phe	UCC Ser	UAC ψ Tyr	UGC Cys	C
	UUA ψ Leu	UCA Ser	UAA *Stop*	UGA *Stop*	A
	UUG ψ Leu	UCG Ser	UAG *Stop*	UGG ψ Trp	G
C	CUU ψ Leu	CCU Pro	CAU His	CGU Arg	U
	CUC ψ Leu	CCC Pro	CAC His	CGC Arg	C
	CUA ψ Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG ψ Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU ψ Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC ψ Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA ψ Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG ψ Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU ψ Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC ψ Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA ψ Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG ψ Val	GCG Ala	GAG Glu	GGG Gly	G

ψ = relatively hydrophobic residues

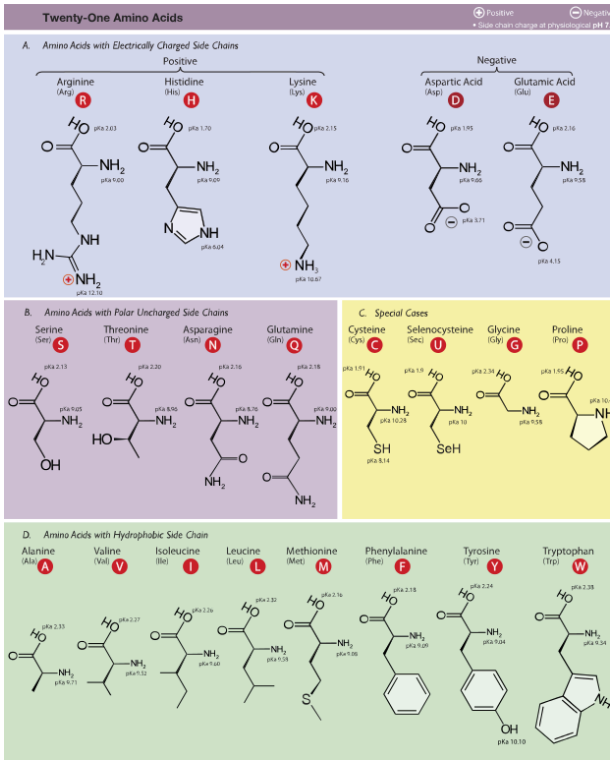


Table 2

Source: *Wikimedia*

Protein Synthesis: the Genetic Code

Like nucleic acid synthesis, protein synthesis is a template-directed process. However, in protein synthesis, the process is a bit more complicated because the template does not directly interact with the polypeptide product. How this works has never been explained more plainly than by Francis Crick, in an astonishing proposal that he made in 1955:

> *...Each amino acid would combine chemically, at a special enzyme, with a small molecule which, having a specific hydrogen-bonding surface, could combine specifically with the nucleic acid template. This combination would also supply the energy necessary for polymerization... there would be 20 different kinds of adaptor molecule, one for each amino acid, and 20 different enzymes to join the amino acid to their adaptors...*

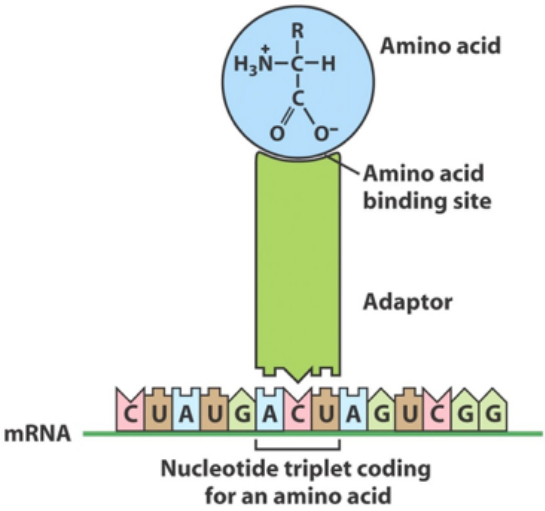


Fig. 12. The adapter hypothesis.

We now know that the "adapter" is **tRNA**. Crick's proposal was correct in every detail save one: there are 61 possible combinations of 3-base mRNA **codons** (see Table 1), so there are more than 20 tRNA "adapters." This system provides the biochemical basis of the **genetic code — the rules through which each 3-base codon in an mRNA template specifies one specific amino acid.**

tRNA and aminoacyl tRNA synthetase enzymes

1. Each tRNA has 2 "business ends":

   - The **anticodon** pairs with the 3-base **codon** on the mRNA template. Look closely at the diagram (Fig. 13). Note that as in other nucleic acid hybrids, the two strands are **antiparallel**.
   - The **aminoacyl acceptor site** is a terminal adenosine (A) nucleotide, where the carboxyl group of the amino acid is esterified to the 3′-OH of the adenine. **Note that this is a relatively unstable, high–energy bond**. It will make polypeptide elongation thermodynamically downhill, and hence favorable — exactly as predicted by Crick.

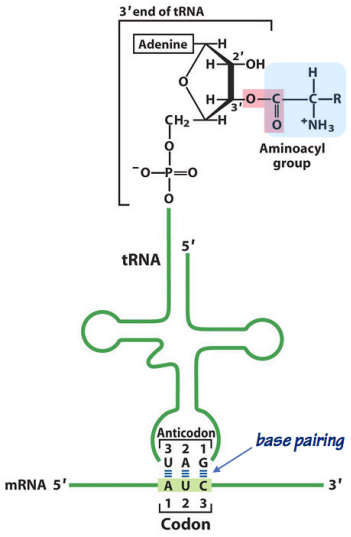


Fig. 13. An aminoacylated tRNA "reading" the genetic code.

2. *Any* given tRNA can (in principle) be esterified to *any* amino acid. However, **the fidelity of translation depends absolutely on the accurate matching of a tRNA bearing a specific**

anticodon to the corresponding amino acid.

3. The job of coupling each amino acid to its corresponding tRNAs is done by 20 different enzymes: the **aminoacyl tRNA synthetases** (also called aaRS enzymes). This is an absolutely key point: **the specificity of the genetic code is controlled by the tRNA synthetases.**
4. The tRNA synthetases produce aminoacylated tRNA molecules in a two-step reaction (Fig. 14).

    - First, an ATP is coupled to the amino acid to form an **amino adenylate**. That is, the amino acid is coupled to AMP. This is a high-energy **activated intermediate**.
    - In the first step, pyrophosphate (PPi) is also released. As we saw in DNA and RNA polymerization, pyrophosphate is immediately destroyed by pyrophosphatase, making the first sub-reaction an irreversible *committed step*.

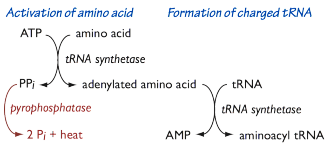


Fig. 14. Formation of aminoacyl tRNA.

    - Second, the aminoacyl group is transferred to the tRNA. The products are an aminoacyl tRNA, and AMP. Because we start with ATP, and end up with AMP and two inorganic phosphates, coupling of an amino acid to tRNA has an **energetic cost** of 2 ATP equivalents.

The Ribosome

The "polypeptide polymerase" is the **ribosome**, an enormous **ribonucleoprotein** complex. **The ribosome has two subunits**. Each subunit contains both RNA and many different polypeptides.

1. The **small subunit** (it is **not** small, just not as big as the large subunit!) is the "**decoding center**." The small subunit's job is to match each codon on the template to a corresponding aminoacyl-tRNA. This is no easy task. There are 61 codons that

specify 20 amino acids (*Table 1*), so a large majority of the aminoacyl-tRNA molecules that enter the ribosome must be rejected.
2. When a correct codon-anticodon interaction is detected by the small subunit, the **large** subunit catalyzes the **peptidyltransfer reaction** — the chemistry of polypeptide elongation.

Remarkably, although each ribosome subunit contains both peptides and rRNA, both the decoding center within the small subunit, and the peptidyltransfer center in the large subunit, are **made of ribosomal rRNA**. **The enzymatic core of the ribosome is a "ribozyme"**. This is probably a relic of the ancient origin of ribosomes at the dawn of life, in the so-called RNA world.

Polypeptide synthesis

As in nucleic acid polymerization, there are three phases of polypeptide synthesis.

1. **Initiation**: the polymerase — the ribosome — must be placed precisely over the **start codon** on the mRNA template.
2. **Elongation**: this is where template-mediated polymerization of the polypeptide occurs.
3. **Termination**: A **stop codon** is identified, triggering release of the polypeptide and removal of the ribosome from the mRNA template.

Initiation of protein synthesis

As with DNA replication and transcription, initiation of protein synthesis is pretty complicated. And since the frequency of initiation controls the rate of protein synthesis, this step is also highly **regulated**.

Regulation occurs at both a global level (the cell asks how much protein synthesis it can support overall, given the available energy and resources), and for specific mRNA transcripts, which are translated with different efficiencies.

Steps in translation initiation

1. The ribosome small subunit (Fig. 15, small brown oval) assembles with **initiation factors**.

    - Together, they bind to the **5′cap of the mRNA**.
    - The small subunit and initiation factors crawl along the mRNA from 5′-to3′, **scanning** the mRNA for a **start codon**.
    - In most cases the start codon is **AUG**. If you look at **Table 1**, you'll see that **AUG** encodes the amino acid methionine (Met, M). Thus, the first amino acid in a polypeptide is usually Met.

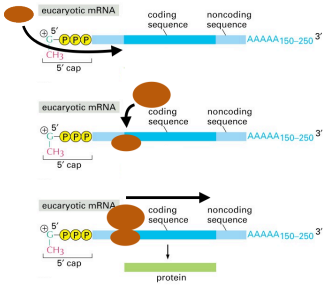


Fig. 15. Initiation of polypeptide synthesis by the ribosome. In this cartoon, the ribosome large and small subunits are the brown ovals. Translation initiation factors are not shown. Source: Alberts, Molecular Biol. of the Cell

2. Once the decoding center is accurately placed over the AUG start codon, the large subunit (in Fig. 15, the larger oval) docks onto the small subunit and mRNA, and the initiation factors fall off (dissociate). Now the **elongation cycle** can begin.

Two additional points about translation initiation must be emphasized.

First, as mentioned above, this step is highly regulated. Initiation factors are largely responsible for this regulation.

Second, the **accuracy** with which the ribosome's small subunit is placed over the start codon is **critical**: a positional error of ± 1 or 2 nucleotides will put the mRNA transcript **out–of–frame**, and result in a totally different, and incorrect, polypeptide sequence. Similarly, if a mutation in the genome inserts or deletes one or two DNA bases in the **coding region** of a gene, the resulting mRNAs will contain **frameshift errors**. When this happens, every codon following the

frameshift will be incorrectly decoded during translation. Most often this also results in early termination of synthesis.

Note: *There are two major, important differences in mRNA structure and translation initiation between eukaryotes (humans included) and bacteria.*

1. In bacteria, mRNA molecules do not have a 5′ cap or 3′poly-A tail (Fig. 8). *In addition, bacterial mRNA molecules often have a series of translation start sites, and a series of coding regions. Such an mRNA is referred to as "polycistronic."* ***In eukaryotes****, a mature mRNA template usually has only a single st*

   art site and encodes only a single polypeptide.
2. ***In bacteria, the first amino acid*** *is usually a Met deriviative,* ***fMet*** *(formylmethionine). Peptides beginning with fMet are recognized by our innate immune system as a* ***danger signal****, because they can indicate an active bacterial infection. You will learn more about bacteria, danger signals, and innate immunity in the Infections and Immunity Block.*

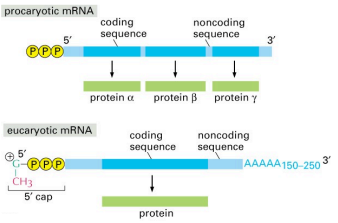


Fig. 16. mRNA structure in prokaryotes (bacteria) versus eukaryotes (including humans). Source: Alberts, Molecular Biology of the Cell

Polypeptide elongation

The assembled ribosome (large and small subunits) has three sites that can accommodate tRNA molecules: the A, P, and E sites. These names are shorthand for **a**minoacyl-tRNA, **p**eptidyl-tRNA, and **e**xit sites. It will become clearer in a moment why these names are used. The three sites are used in sequence. Here's how it works (Fig. 17).

1. At the A site, the ribosome samples incoming aminoacyl-tRNA (aa-tRNA) molecules. **The ribosome is looking for an aa–tRNA**

with an anticodon correctly pairs with the mRNA codon positioned under the A site. *Dozens of incorrect aa-tRNAs are rejected for each correct match.*

2. When a correct codon-anticodon match is identified, the ribosome initiates the **peptidyltransfer reaction** — the chemistry.

    - The growing polypeptide chain, still esterified at its C (carboxyl)-terminus to a tRNA, resides in the P site.
    - In the **peptidyltransfer reaction**, the nasent polypeptide is transferred from the peptidyl-tRNA sitting in the P site, to the aminoacyl residue sitting in the A site. *This is counterintuitive.* Inspect fig. 9 to see how it works.

3. Now the ribosome undergoes a ratchet-like twisting motion. This motion causes **translocation** of the mRNA, the peptidyl-tRNA, and the deacylated (discharged) tRNA. Now the peptidyl-tRNA is one residue longer, and in the P site. The deacylated tRNA is in the E site, where it is ejected from the ribosome. And the A site is unoccupied, ready to begin the cycle again for the *next codon* on the mRNA template.

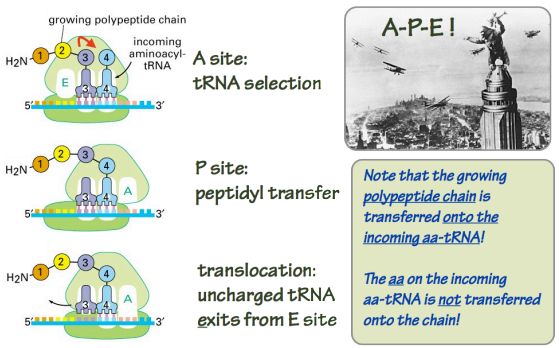


Fig. 17. Polypeptide elongation. Source: Alberts, Molecular Biology of the Cell

A final point about energy. Two ATP equivalents used to charge

each aa-tRNA. This powers the peptidyltransfer reaction. However, additional ATP equivalents are consumed during the tRNA selection and translocation portions of the elongation cycle. In terms of both energy and materials, protein synthesis is very expensive. Many cell types use most of their energy on protein synthesis.

SLO9. Describe the basic ways in which microRNA (miRNA) molecules control gene expression.

MicroRNAs (**miRNAs**) are post-transcriptional regulators of gene expression. More than 2,000 miRNAs have been annotated in human genome. 60% of all human genes are estimated to be regulated by one or more miRNA. miRNAs are **short noncoding RNAs**, usually about 22 nucleotides long. miRNA molecules are formed through **two major pathways**:

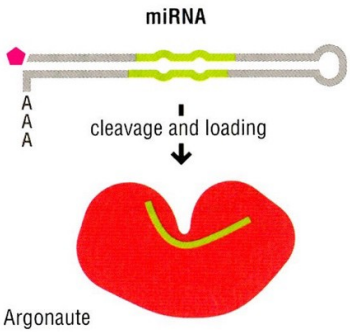


Fig. 18. Origin of micro RNA (miRNA) involved in gene regulation and protein synthesis. The miRNA is usually derived from a short "hairpin" region within a larger mRNA molecule. The miRNA is excised from the hairpin by an RNA endonuclease enzyme.

- Many miRNA precursors are coded as stand-alone genes, which can be transcribed by RNA polymerase II. Note that in the figure above, the miRNA is derived from a Pol II transcript with a 5′cap and 3′poly-A tail.
- As they are not coding sequences, miRNA precursors can also be derived from intron sequences that are embedded within other mRNA precursor transcripts. In these cases, splicing excises the intron and its miRNA from the coding exons of the

mature mRNA.

miRNA does not function alone. miRNAs bind to the **Argonaute** family of proteins in the cytoplasm. miRNA-Argonaute complexes bind to specific mRNA transcripts via complementary hybridization between the miRNA seed region, and target sequences in the 3'UTR of the targeted mRNA transcript. In most cases, miRNAs function to **repress** (decrease) the production of specific sets of proteins. miRNA-Argonaute-mRNA complex can repress protein expression in different ways:

- destabilization of the mRNA via shortening poly (A) tail;
- inhibition of translation initiation;
- cleavage and degradation of the target mRNA.

Note that these mechanisms are *post-transcriptional*, meaning that they operate on mature mRNA molecules in the cytoplasm. In contrast, transcription factors act in the nucleus to control the rate at which different mRNA molecules are synthesized (transcribed) by RNA polymerase II. The first miRNA was discovered in the nematode *C. elegans*. Studies in humans have revealed that mutations in miRNAs can cause or contribute to various human diseases. For instance, mutation in miRNA-96 is linked to hereditary progressive hearing loss, and deletion of the miR-17~92 cluster causes skeletal abnormality and growth defects. Dysregulation of miRNA function is also implicated as a causative factor in several cancers.

3. Protein Targeting, Vesicular Transport

Session Level Objectives (SLOs): after completing the session, students will be able to:

SLO 1. Describe the major organelles of the secretory pathway.

SLO 2. Understand the difference between targeting of integral m embrane proteins and secretory proteins. SLO 3. Outline how pro teins are folded and modified in the endoplasmic reticulum and G olgi organelles.

SLO 3. Outline how proteins are folded and modified in the endoplasmic reticulum and Golgi organelles.

SLO 4. Explain how proteins and lipids are packaged into carrie r vesicles, and how carrier vesicles fuse with target membranes i ncluding the plasma membrane.

SLO 5. Explain key mechanisms that underlie endocytosis, rece ptor recycling, and traffic to the lysosome.

SLO 6. Outline the synaptic vesicle cycle, and explain how clost ridial neurotoxins selectively block neurotransmission.

Nearly all proteins have distinctive subcellular (or, for secreted proteins, extracellular) localizations. Many proteins are directed to specific locations by "molecular zip codes" — structures within the protein that contain targeting information. Membrane and secreted proteins are encoded by 30-40% of our genes.

SLO 1. Describe the major organelles of the secretory pathway.

1. Secreted proteins, and most integral membrane proteins, are synthesized by ribosomes at the rough **endoplasmic reticulum (ER)**. These proteins can be co-translationally translocated *into* the ER membrane (for integral membrane proteins like ion channels), or they can be transferred *across* the ER membrane (for soluble secreted proteins like antibodies).
2. At the ER, proteins are **folded** with the aid of **chaperone**proteins. In many but not all cases secreted and membrane proteins are **glycosylated**: complex carbohydrates are covalently attached to the proteins by ER-resident enzymes.
3. Following folding, the proteins are packaged into **carrier vesicles** and transported to the **Golgi** apparatus. In the Golgi further post-translational modifications are performed and additional quality control checks are made. The proteins enter the Golgi complex at the ***cis* Golgi**, traverse the **medial Golgi**, and end up in the ***trans* Golgi network (TGN)**.
4. The TGN is the Grand Central Station of the secretory pathway: here, proteins are sorted into different carrier vesicles that can go to many different locations in the cell, including **endosomes**, **lysosomes**, and **secretory vesicles**.
5. There are two broad categories of secretory vesicles.

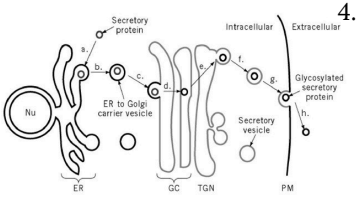


Fig. 1. General organization of the secretory pathway. Nu, nucleus. ER, endoplasmic reticulum. GC, Golgi complex. TGN, trans-Golgi network. PM, plasma membrane.

   - **Constitutive secretory vesicles** fuse with the plasma membrane "automatically," by default. This is a

"housekeeping" pathway.

- **Regulated secretory vesicles** fuse with the plasma membrane only in response to a signal. There are many different types of regulated secretory vesicles in different cell types. They may contain hormones (*e.g.*, insulin), neurotransmitters, blood clotting factors, etc. The commonality is that regulated secretory vesicles are held in reserve in the cytoplasm, until a signal indicates that their contents are required at the cell surface.

6. Fusion of a secretory vesicle at the plasma membrane is called **exocytosis**.

SLO 2. Understand the difference between targeting of integral membrane proteins and secretory proteins.

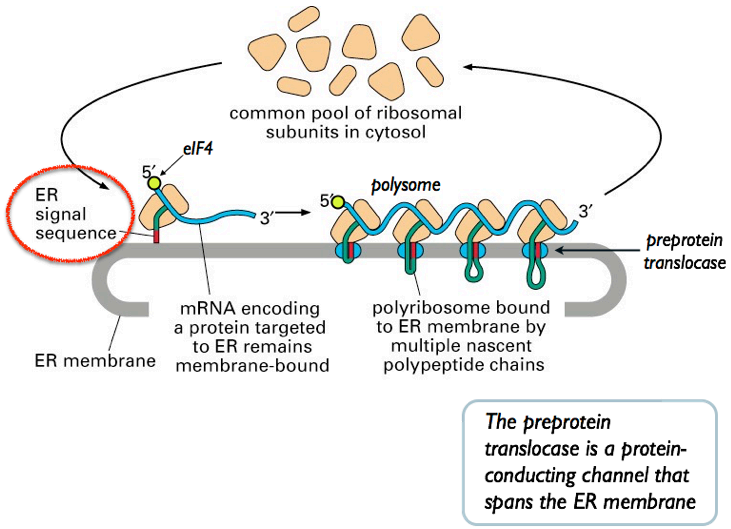


Fig. 2. Targeting of secreted and membrane proteins to the endoplasmic reticulum.

1. Integral membrane proteins, including receptors and ion channels, have one or more membrane-spanning domains that traverse the lipid bilayer.
2. Secreted proteins generally do not have transmembrane domains, and are soluble in the aqueous lumens of the ER, Golgi, and secretory vesicles.
3. Both integral membrane proteins, and secreted proteins, are usually **translocated into the ER** as they are synthesized. That is to say, **translocation is co-translational**.
4. A common protein machinery is used to target both secreted and integral membrane proteins at the ER: the **Preprotein Translocase**. This complex is also called the **Sec61 complex**.
5. Recall that the N- terminus is the first part of a protein to emerge from the ribosome. Secreted and integral membrane proteins emerging from the ribosome are marked by a "zip code" — an N- terminal **signal sequence**
6. The signal sequence is identified by the **Signal Recognition Particle (SRP)**. The SRP directs the mRNA, ribosome, and nascent polypeptide to the Preprotein Translocase.
7. As the polypeptide elongates, it is threaded through the preprotein translocase. For secreted proteins, the signal sequence is removed by a site-specific protease. This allows the protein to diffuse in the aqueous phase.

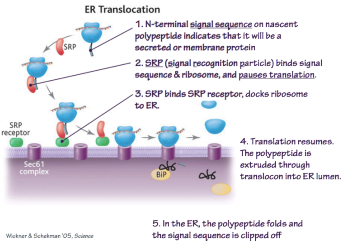


Fig. 3. SRP recognizes N-terminal signal sequences to target nascent polypeptides to the ER. BiP is a chaperone protein that helps secreted proteins to fold correctly.

SLO 3. Outline how proteins are folded and modified in the endoplasmic reticulum and Golgi organelles.

1. Chaperones in the ER lumen, such as BiP, intercept the nascent protein and attempt to help it fold correctly. In many cases, folding requires the assembly of multi-protein complexes (oligomerization).
2. Transmembrane proteins are assembled by allowing transmembrane segments to escape from the preprotein translocase laterally, into the membrane bilayer (Fig. 4). The transmembrane segments are hydrophobic. Thus integral membrane proteins are soluble (t

   hat is, they can diffuse) *within the two-dimensional plane of the membrane bilayer*, but they are not freely soluble in the aqueous phase.

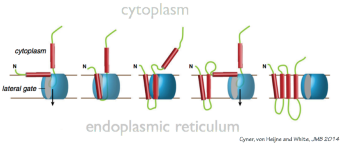


Fig. 4. Transmembrane domains are inserted into the ER membrane by lateral escape from the preprotein translocase.

3. 20% of the time or more, the folding of membrane proteins fails! The protein mis-folds: it must then be identified as a "dud" and degraded. The polypeptide is extracted back into the cytoplasm, and degraded by the proteasome (which you will learn about later in the course).

    - Mutations, or errors in transcription or translation, can cause mis-folding.
    - Chemical and physical stresses (oxidation, elevated temperature, etc.) can cause mis- folding.
    - Some amount of mis-folding happens through stochastic processes, even without unusual mutations or stresses.
    - Note that quality control occurs both in the cytoplasm and in the secretory system.

4. The protein folding and quality control machinery can be overwhelmed. This causes **ER stress**. The **ER stress response** is a **signaling pathway** that does thee major things:

    - It senses the amount of unfolded protein in the ER lumen;
    - it slows down translation (protein synthesis); and
    - it causes transcription of genes encoding ER chaperones.
    - Thus the ER stress response is a **homeostatic mechanism** that allows the efficiency of ER folding to respond appropriately to demand.

5. The ER stress pathway can be overwhelmed. At this point, the cell may choose self-destruction (apoptosis).
6. Proteins exported to the Golgi are subjected to additional quality control checks for proper folding and post-translational modification. If they fail these checks, they are often **retrogradely** transported back to the ER. At the ER they get an additional chance to fold or to be degraded.

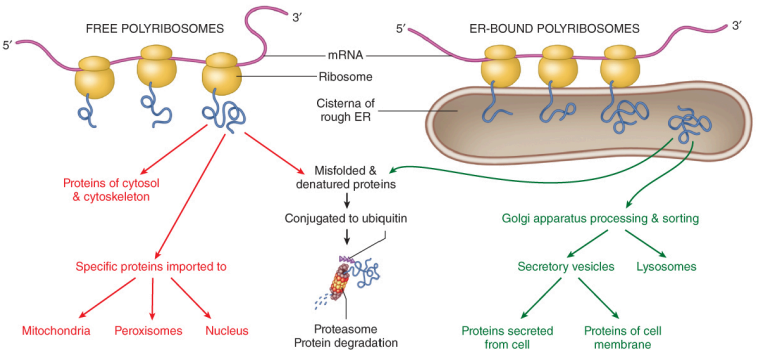


Fig. 5. Overview of protein targeting and protein quality control mechanisms.

SLO 4. Explain how proteins and lipids are packaged into carrier vesicles, and how carrier

vesicles fuse with target membranes including the plasma membrane.

Transport vesicles are membrane-delimited containers that carry lipids, integral membrane proteins, and solutes (including proteins) between organelles, and between organelles and the plasma membrane. We can outline in general how transport vesicles operate. Fig. 6 outlines the general themes in vesicular transport.

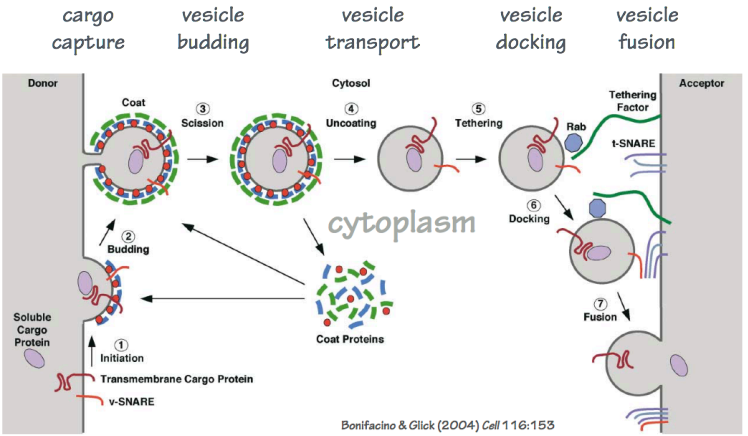


Fig. 6. General priniciples of transport vesicle budding and fusion.

1. Cargo lipids and proteins are identified by cytoplasmic **coat complexes**. Note that soluble secretory proteins must be identified by transmembrane receptors that traverse the membrane to engage with the coat complex. ***Note: different coat complexes assemble transport vesicles with different contents.***
2. The coat complex forms a shell that both captures cargo (lipids, proteins), and deforms the membrane (**budding**).
3. The coated transport vesicle detaches from the donor membrane.
4. The coat dissociates.

5. Proteins and lipids on the transport vesicle are recognized by **tethering factors** that attach the transport vesicle to an **acceptor membrane**. We can think of these as molecular "Velcro" – but this "Velcro" has specificity: only appropriate vesicles will tether to a given acceptor membrane. ***Note: Different vesicles and acceptor membranes use different tethering factors.***
6. Tethering leads to a tighter association: docking. During docking, the fusion complex is assembled. The fusion complex contains **SNARE proteins**.
7. Zippering of SNARE proteins is energetically favorable, and leads to **fusion** of the vesicle and acceptor membranes. **Different combinations of SNARE proteins can assemble. Only some of these actually trigger fusion — an additional layer of specificity.**

Here is an example of two specific pathways, operating in concert. If you examine this diagram closely, you'll see specific examples of all the general mechanisms described on the previous page.

1. The forward (**anterograde**) pathway uses the **COPII** coat to make vesicles that carry secretory cargo molecules from the ER to the *cis*-Golgi.
2. The backward (**retrograde**) pathway uses the **COPI** coat to retrieve SNARE proteins and cargo receptors such as K/HDEL back to the ER. The retrograde pathway also retrieves mis-folded proteins that have failed quality control checks in the Golgi.

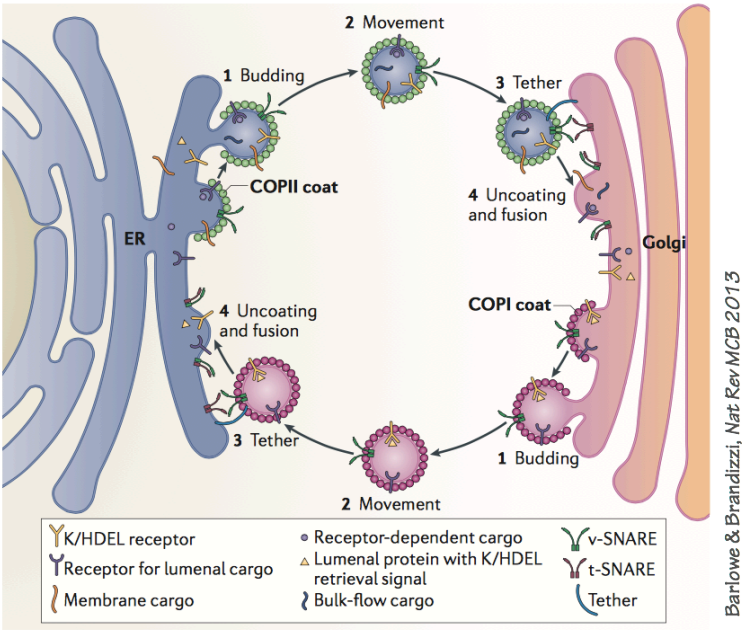


Fig. 7. Distinct coats control anterograde and retrograde traffic between the ER and Golgi.

SLO 5. Explain key mechanisms that underlie endocytosis, receptor recycling, and traffic to the lysosome.

If you imagine how the secretory pathway operates in isolation, there is a problem: membrane is continuously deposited on the plasma membrane, but not retrieved. The plasma membrane grows and grows. Thus, there is a second system that brings membrane back into the cell: the **endocytic pathway**.

The endocytic system also has a **degra**

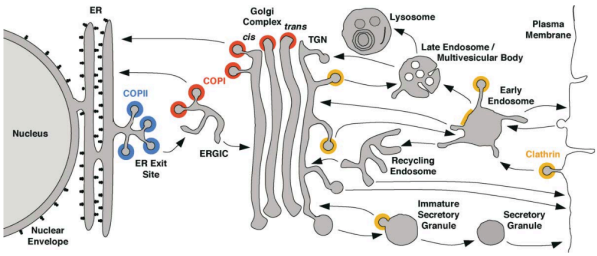


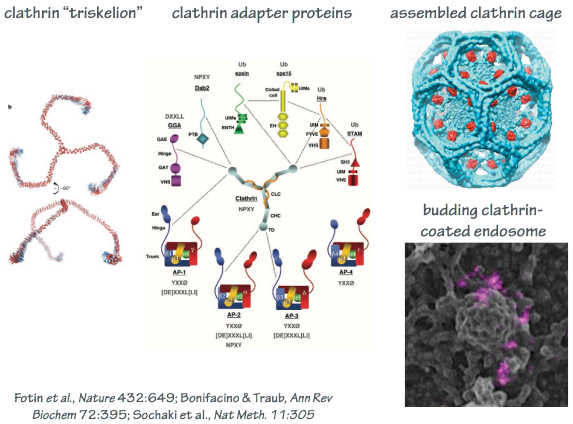
Fig. 8. Overview of secretory and endocytic systems, and coat complexes that mediate transport within them. Note: ERGIC is an ER-Golgi intermediate compartment. You don't need to know about it.

dative role, bringing membrane lipids, proteins, and surface-adsorbed particles into the cell, where they can be routed back into the secretory pathway or, alternatively, sent to the hydrolytic **l**

ysosome for destruction and recycling. The endocytic system also has a **recycling pathway**, as we'll see below.

While the **secretory pathway** begins at the **ER** and ends with **exocytosis** (exit or ejection from the cell), the **endocytic pathwa**y begins at the plasma membrane with **endocytosis** (taking into the cell).

In endocytosis, the coat consists of an inner shell that recognizes cargo and an outer shell that organizes membrane deformation. The outer shell complex, **clathrin**, also operates at other steps of intracellular transport (Fig. 9). The inner shell adaptor system is diverse, with molecules specialized for capturing different cargo at different locations in the cell.



Fotin *et al.*, *Nature* 432:649; Bonifacino & Traub, *Ann Rev Biochem* 72:395; Sochaki *et al.*, *Nat Meth.* 11:305

Fig. 9. The clathrin vesicle coat system. Note: You do not need to know the specific names of the different clathrin adapter proteins. Just know that different adaptors exist.

SLO 6. Outline the synaptic vesicle cycle, and explain how clostridial neurotoxins selectively block neurotransmission.

At chemical synapses, vesicles loaded with neurotransmitter fuse with the plasma membrane of the presynaptic cell. This is a **regulated exocytosis** event. This is followed by endocytosis.

Together, these events constitute the **synaptic vesicle cycle**. This is an example of an **endocytic recycling pathway**: the donor membrane from which the vesicle is derived, and the acceptor membrane with which the vesicle fuses, is the same membrane — the plasma membrane.

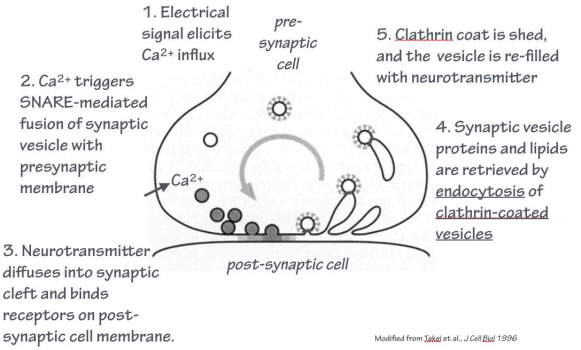


Fig. 10. The synaptic vesicle cycle.

Clostridial neurotoxins are secreted by bacteria of the genus *Clostridium*, including the species that cause botulism and tetanus.

The clostridial neurotoxins bind to receptors on the plasma membrane, are endocytosed, and then the active toxin "payload" is translocated across the membrane into the cytoplasm. The active toxin is a **protease** enzyme that requires a bound Zn2+ metal ion for activity (a zinc protease).

These proteases have *exquisite* selectivity: **they recognize only SNARE proteins involved in exocytosis**, and cleave them. This potently blocks membrane fusion and neurotransmitter release — and, therefore, blocks neurotransmission (see Fig. 6).

4. Protein Structure and Function

Session Level Objectives (SLOs): after completing the session, students will be able to:

SLO 1. Know the elements of protein secondary, tertiary, and quaternary structure.

SLO 2. Explain the roles of hydrophilic vs. hydrophobic aminoacyl residues in protein folding.

SLO 3. Explain the importance of correct protein folding, chaperone proteins, and how misfolding can lead to pathology.

SLO 4. Understand common post-translational modifications of proteins (phosphorylation; disulfide bond formation; glycosylation) and know why specific modifications occur predominantly on proteins within the cytoplasm or in extra-cytoplasmic environments.

SLO 5. Know that different proteins are targeted to specific locations inside and outside of cells.

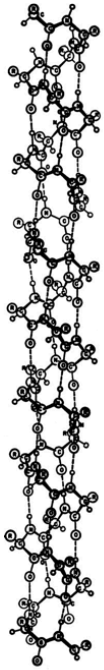
SLO 1. Know the elements of protein secondary, tertiary, and quaternary structure.

If you want to understand function, study structure. — F. Crick

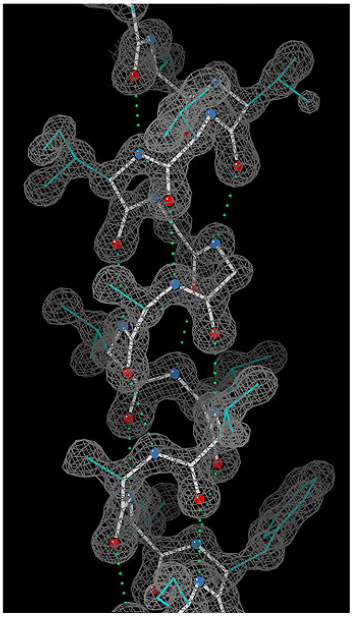
As a nascent polypeptide emerges from the ribosome, it must **fold** into a specific, functional, three- dimensional structure. **The functional "native" fold of a protein is determined by the linear sequence of amino acids in the polypeptide**.

We think about folding as a hierarchical process:

- A polypeptide's **primary structure** its **linear sequence of amino acids**. This sequence, as you have just seen, is specified by the sequence of codons in an mRNA template.
- **Secondary structure** elements are "folding motifs" that form through local interactions between residues within the polypeptide chain (H-bonding, salt bridges, van der Waals interactions, *etc.*). The most common and important secondary structure motifs are the **α–helix** (Fig. 1) and the **β– sheet** (Fig. 2).



Pauling, Corey, and Branson., 1951



Wikipedia

Fig. 1. α-helix motif. The polypeptide backbone forms a right-handed helix. The helix is stabilized by hydrogen bonds formed between backbone amino and carbonyl groups on successive turns of the helix. The amino acid side chains (blue) project outward from the helix.

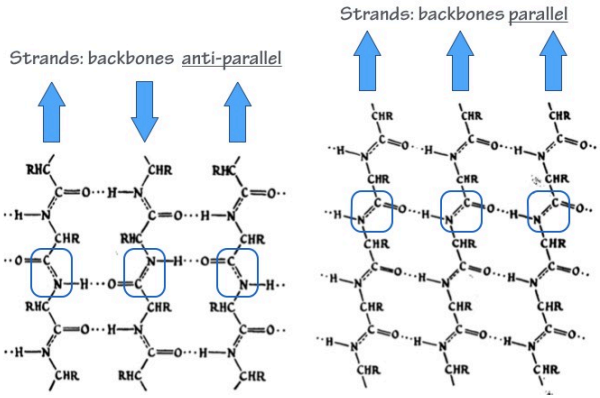


Fig. 2. β-sheet motif. Hydrogen bonds between backbone amine and carbonyl groups connect adjacent segments of the polypeptide. The side chains project above and below the sheet. The strands can run parallel (then N- and C-termini are on the same side) or anti-parallel. Source: Pauling & Corey, 1951.

- **Tertiary structure** describes the overall arrangement of a polypeptide's secondary structure elements. This is the overall 3-dimensional fold of the polypeptide. Many proteins contain mainly α-helix or β-sheet folds. Others use both kinds of folds; an example is shown in Fig. 3.
- Many, many proteins operate as larger **complexes**. The assembly of more than one polypeptide into a protein complex is the **quaternary structure**. This can mean as few as two small polypeptides, or an assemblage as big as the nuclear pore or — even bigger — silk or human hair.

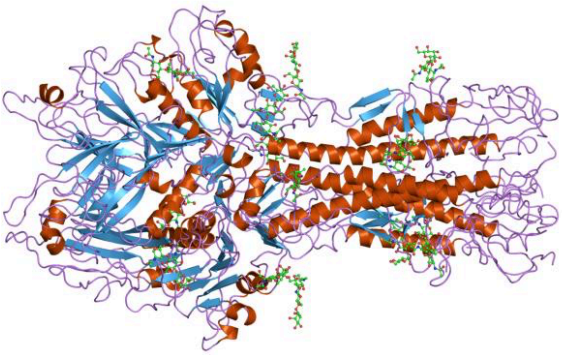


Fig. 3. Influenza virus HA protein. This protein is a homotrimer, meaning that the quaternary structure is a complex containing three identical copies of a single type of polypeptide. In this rendering, each of the three chains has both α-helix (red corkscrews) and β-sheet (blue arrows) secondary structure folds, connected by "loop" segments (purple). The polypeptides are post-translationally glycosylated with carbohydrate molecules (green).

SLO 2. Explain the roles of hydrophilic vs. hydrophobic aminoacyl residues in protein

folding.

The principles that control protein folding are the *exactly* same ones that we have already seen with RNA and DNA: The hydrophobic effect, charge interaction and repulsion, Van der Waaals contacts, etc.

1. Water molecules form many hydrogen bonds with one another, and they have high entropy (they can diffuse freely, translate, and rotate).
2. Hydrophobic (greasy) amino acid side chains are surfaces where water cannot hydrogen bond (this is an enthalpic penalty). Near these surfaces the water has reduced *entropy*, as well. Because water "hates" hydrophobic side chains, these chains "want" to be shielded from the aqueous solvent. Thus, **hydrophobic amino acid residues tend to be buried within folded portions of the protein.**
3. Hydrophilic (polar or charged) amino acid side chains can form energetically favorable hydrogen bonds with water. They are often exposed to the aqueous solvent. If they cannot interact with the solvent (if they are buried), they generally interact with other portions of the polypeptide through hydrogen bonds or salt bridges.
4. Additional inter-chain interactions that contribute to protein stability include van der Waals contacts, aromatic stacking interactions (analogous to the base stacking that we saw with DNA and RNA), and electrostatic repulsion between similarly charged (-/- or +/+) groups on the polypeptide.

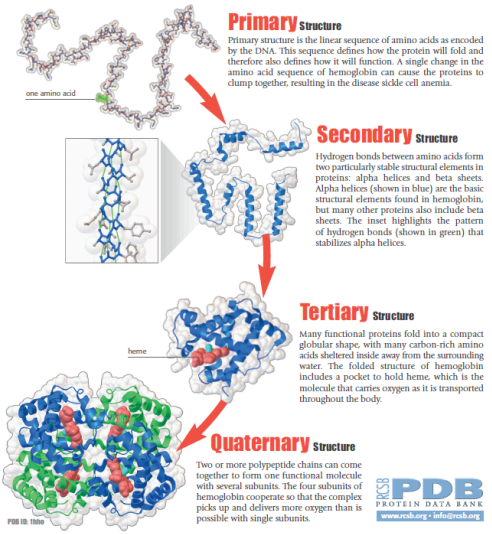


Fig. 4. Summary of the hierarchy of protein folding. Notice that, in contrast to the flu virus protein in Fig. 3, hemoglobin is an exclusively α-helical protein.

SLO 3. Explain the importance of correct protein folding, chaperone proteins, and how misfolding can lead to pathology.

Protein folding, chaperone proteins, and how misfolding can lead to pathology

Mutations that cause protein sequence changes, or errors in transcription or translation, can change the balance of forces that we have just described, causing **misfolding** of a protein loss of its function. Other mutations may still allow a protein to fold more or less correctly, but change the protein's activity. For example, some mutations result in ion channels that open more easily than they otherwise would, leading to neurological disorders.

Both within our cells, and in extracellular spaces (cartilage, blood, cerebrospinal fluid, *etc.*), proteins are present **at extremely high**

overall concentrations. This dense proximity means that proteins will touch other proteins both during and after folding, with enormous potential for inappropriate interactions that can lead to non-specific aggregation.

You're probably familiar with one protein aggregation process: making Jell-O™. We start with a clear aqueous solution of soluble proteins at high concentration. We then heat the solution so that the proteins unfold. That is, they are **denatured**. As the unfolded proteins cool, they aggregate into a single disordered gel.

In cells, protein aggregates are major sources of cytotoxicity, and — as we will see — they contribute to pathologies ranging from Alzheimer's disease to type II diabetes.

Mutations can cause proteins to misfold at elevated rates, but even non-mutant proteins sometimes misfold, especially in the presence of stresses such as heat or oxidation. To mitigate inappropriate contact between un-folded or partially-folded proteins, cells use special proteins called **chaperones**. There are many different chaperones. Some passively shield proteins from inappropriate contacts.

Others use energy from ATP hydrolysis to mechanically pull apart proteins that have formed inappropriate contacts, giving the proteins a "second chance" to fold correctly.

When a protein cannot fold correctly even with the assistance of chaperones, the cell may recognize the misfolded polypeptide as hopeless, and mark it for destruction. This cellular surveillance process operates in almost every cell and is called **protein quality control**. The quality control system has at least two branches: the ubiquitin–proteasome system, and the autophagy–lysosome system. We'll discuss these systems later in the block.

SLO 4. Understand common post-translational modifications of and know why specific modifications occur predominantly on proteins

within the cytoplasm or in extra- cytoplasmic environments.

Once synthesized, most polypeptides undergo **covalent post-translational modifications**. These fall into several different categories. The following list is not comprehensive! It's illustrative, showing some important examples.

1. **Proteolysis**. Many proteins are precisely clipped before they are fully functional. For example, many digestive enzymes are made as inactive **proenzymes** — a form safe for transport through sensitive cellular compartments. Upon secretion into the digestive tract an inhibitory portion of the polypeptide is clipped off, and the enzyme is activated.

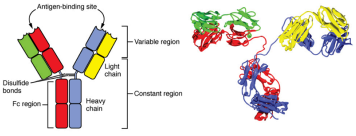


Fig. 5. An antibody (IgG) molecule. Each IgG is a heterotetramer containing four polypeptides: two identical light chains and two identical heavy chains. IgG is an entirely β-sheet protein. The quaternary structure of the complex is stabilized by non-covalent interactions between the chains, and also by disulfide bonds that covalently cross-link the two heavy chains together.

2. **Disulfide bonding**. The terminal sulfhydryl group on the amino acid **cysteine** can be oxidized to form a **cysteine–cysteine disulfide bond**.

    - Disulfide bonds are most often used to form mechanically stabilizing cross-links within a polypeptide chain or to cross-link two chains together in a protein complex.
    - In general, the cytoplasm and nucleus of a cell have a chemically reducing potential, while the extracellular environment has a relatively oxidizing potential. What this means: we *very* seldom see proteins with disulfide bonds in the cytoplasm or nucleus, but *lots*of secreted proteins such as antibodies (Fig. 5), and many cell-surface proteins, have

disulfide bonds.

3. **Glycosylation**. Sugars are attached to most, but not all, **secreted and cell surface proteins**. Usually these are short-chain, branched carbohydrates. These sugars are used in cell–cell recognition and in cell signaling processes, and they can stabilize and protect proteins that are exposed to harsh extracellular environments. On the down side, viruses such as HIV and SARS use glycosylation to shield their surface proteins from recognition and attack by our immune systems (Fig. 6). Very few cytoplasmic or nuclear proteins are glycosylated (though the ones that are may be of great importance).

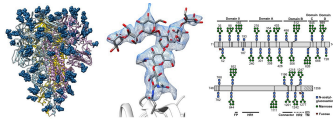


Fig. 6. The glycan "shield" of a Coronavirus spike protein. Like the Influenza HA protein shown in Fig. 3, the Spike protein is used by Coronaviruses to gain entry into host cells during infection. The spike consists of a protein homotrimer, anchored to the viral envelope (membrane) at the base. Attached to each monomer are over twenty complex carbohydrate molecules (blue). Each sphere represents a hexose. The structure of the carbohydrates, ascertained using mass spectrometry, is shown in schematic form on the right.The carbohydrates both stabilize the spike protein and shield it from proteases, antibodies, and other host defenses.Source: D. Veesler, UW Biochemistry. Nat Struct Mol Biol. (2016) 23(10):899-905

4. **Phosphorylation**. The covalent transfer of phosphate groups from ATP to polypeptides (protein phosphorylation) is a critical regulatory mechanism that controls almost every aspect of cell physiology.

    - The phosphotransfer reaction is mediated by **protein kinase enzymes**. The recipient is always an amino acid residue with a **terminal hydroxyl group** on its side chain: serine, threonine, or tyrosine. The product is a **phosphoester**.
    - **Phosphorylation is reversible**. Hydrolysis of the phosphoester removes the phospphoryl group from the

protein. **Dephosphorylation** is catalyzed by **protein phosphatase** enzymes.
    - Several hundred kinases and phosphatases are encoded in the human genome.
    - Protein kinases and phosphatases were discovered here in the UW School of Medicine, by Professors Ed Krebs and Eddie Fischer. For their discoveries, these lifelong friends and collaborators shared the Nobel Prize.

SLO 5. Know that proteins are targeted to specific locations inside and outside of cells.

Different proteins have different functions — and the proteins must carry out their functions in different locations. Inside a cell, for example, some proteins operate in the cytoplasm, some in the nucleus, some within organelles such as mitochondria, and some within the plane of the plasma membrane. Many proteins are secreted from cells. Examples include the collagen that holds our tissues together (Fig. 7), antibodies and other proteins in blood and serum, digestive enzymes in the gut, and polypeptide hormones such as insulin.

Each protein must have a way to get to its site of action. **Protein targeting** typically involves a specific amino acid sequence within the polypeptide that serves as a "molecular zip code" used to direct the protein to its destination.

For example, the RNA polymerase II complex consists of several polypeptides. It is synthesized, folded, and assembled into a complex in the cytoplasm, but it has a "nuclear localization sequence" that directs the folded complex through the nuclear pore and into the nucleus, where it will transcribe mRNA molecules. As we'll see, some mutations cause proteins to go to incorrect locations, resulting in disease.

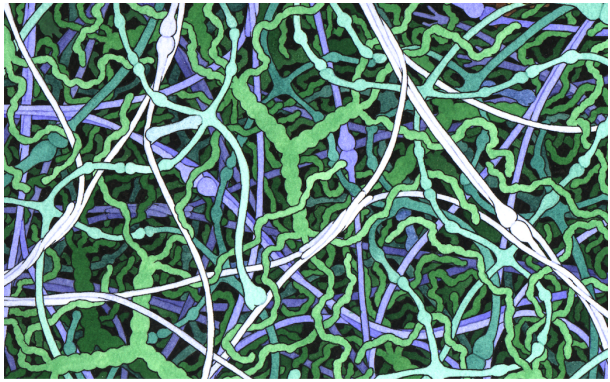


Fig. 7. Extracellular matrix. Many big and small proteins including collagen (long white and violet fibers) assemble into complex structural webs that link cells together within tissues. Each of these proteins carries a signal sequence that directs its secretion after it is synthesized inside the cell. Many secreted proteins are heavily post-translationally modified. Source: David Goodsell

5. Hemoglobin Disorders

Session Level Objectives (SLOs): after completing the session, students will be able to:

SLO 1. Explain the diversity of protein–protein and protein–ligand interactions.

SLO 2. Understand how the quaternary structure of hemoglobin and explain the function of the heme prosthetic group.

SLO 3. For a general ligand-receptor pair, be able to explain and calculate the relationship between kon, koff, and KD.

SLO 4. Describe the mechanistic bases of hemoglobinopathies.

SLO 5. Explain the key properties of enzymes. Explain why many enzymes contain bound prosthetic groups.

SLO 6. Explain and calculate the relationships between kcat, Km, and Vmax. Michaelis-Menten enzyme kinetics

SLO 1. Explain the diversity of protein–protein and protein–ligand interactions.

In this session, we begin to think seriously about how proteins do their jobs. Proteins come in a dazzling variety of shapes, sizes, abundances, and tissue distributions.

Nearly without exception, all proteins are functionally similar: **what proteins do is recognize specific chemical entities, and bind — stick — to them.**

That's it. That's almost (though not quite) the whole deal. **Proteins stick to specific things.** Antibodies stick to antigens to mediate immune responses during infection. Cell adhesion molecules allow cells to stick to each other and to extracellular matrix proteins.

Extracellular matrix proteins stick to each other, and to cells. Transcription factors recognize and stick to specific enhancer sequences in our genes. Odorant receptors stick to specific volatile molecules (from perfume to putrescine). Hormone receptors stick to insulin, estrogen, or other hormones.

Enzymes are *also* most simply understood through their ability to stick to things. They stick to their substrates, and they stick to transition states more tightly, favoring the formation of those otherwise unfavored states, and accelerating reaction rates. Enzymes often bind products less tightly, allowing their dissociation (release) from the enzyme.

Within membranes, ion channels, transporters, and pumps are again sticking to substrates and using a series of binding steps to move things from one side of a membrane to the other.

Molecular motors like myosin do the same thing again. Myosin sticks and un-sticks to the actin thin filament. The order of myosin-actin sticking-unsticking is **coupled** to myosin sticking (binding) to ATP, to the ATP hydrolysis transition state, and finally to the release of ADP and Pi. Here, coupling of *two* binding cycles allows energy derived from ATP hydrolysis to make myosin's binding and un-binding to actin *directional* — and that is the power stroke that makes our muscles contract.

With a general quantitative description of a protein's binding characteristics we can understand an enormous amount about what a protein does, how it does it, and how it can fail in its functions, leading to pathology. The same concepts, as we will see in the next session, allow us to think about how drugs interact with their molecular targets.

After all, most drugs are just chemical entities that particular proteins recognize, and stick to.

SLO 2. Understand the structure of hemoglobin and explain the function of heme in hemoglobin.

We begin this exploration with a protein we've already seen. Hemoglobin (**Hb**) is a protein present at

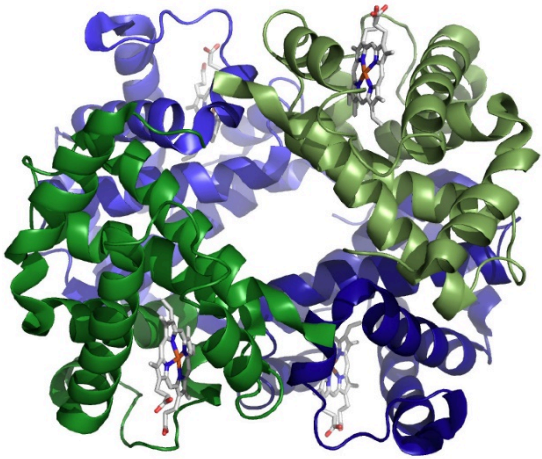


Fig. 1. Heterotetramer of α2β2 Hb. Note that there are four chains, each with a bound heme group. Rendered from PDB 2DHB dataset (M. Perutz, 1970).

enormously high concentration in the cytoplasm of red blood cells. The most important (but not only) thing hemoglobin sticks to is molecular oxygen (O2).

Structure of Hb

1. Adult Hb (**HbA**) is a complex of four polypeptides: two α chains and two β chains (**α2β2**).
2. Fetuses and infants have fetal **HbF** containing two **γ (gamma) chains** instead of β chains (**α2γ2**). The α, β, and γ chains are not identical, but they have very similar primary sequences and a nearly identical, all α-helical, tertiary fold.
3. Terminology note: α -helices and β-sheets are **not** the same as the α and β chains of Hb.
4. Separate genes encode mRNA templates for the various Hb chains.
5. Each of the four chains cradles one heme molecule (Figs 1 and 2): a porphyrin ring that coordinates an ion of iron (Fe2+) at its cent

   er. The bound heme is an example of a **prosthetic group** — a non- amino acid bound to an enzyme and required for its activity. As we'll soon see, vitamins often serve as prosthetic groups in enzymes.
6. Within each subunit (or chain), amino acid side-chains and the bound heme operate together to bind one O2 molecule (Fig. 2). Thus, the binding capacity of a hemoglobin heterotetramer is 4 O2.

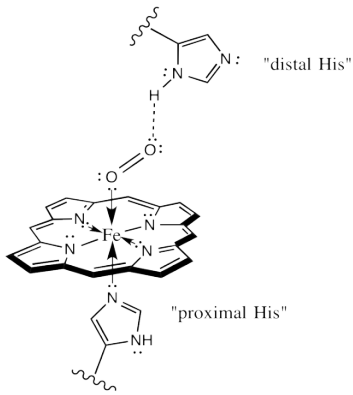


Fig. 2. Mechanism of O2 binding by Hb. The proximal His (histidine) is covalently bound to the heme iron. The distal His helps to coordinate the O2 molecule.

7. It follows that a *one Hb tetramer* can be ¼, ½, ¾, or completely

saturated with O2.

8. With huge numbers of Hb molecules, O2 saturation of the *population* can be anywhere from 0 to nearly 100%.

Function: what Hb needs to do

As erythrocytes pass through our lungs, their Hb binds O2. The erythrocytes flow with the blood to our peripheral tissues where they dump the O2. The tricky bit is that Hb needs to hold on to its precious cargo of O2 until it is in a part of the body that's most in need of O2.

In other words, Hb needs to bias its O2 binding characteristics to accelerate dissociation in relatively hypoxic environments, rather than dumping O2 randomly. This, along with control of vasoconstriction, allows us to maintain a relatively shallow pO2 gradient from our lungs to our fingers, toes, and brain in the periphery.

SLO 3. For a general ligand–receptor pair, be able to explain and calculate the relationship between kon, koff, and KD.

O2 is a **ligand**, and we can say that HbA is a **receptor** that binds it. As the partial pressure of oxygen, pO2, increases, the fractional saturation of HbA increases (Fig. 3).

The first thing to notice is this is an *ensemble measurement* of many molecules of HbA. We note that **there is a concentration of O2 where 50% the receptor (HbA) is 50% saturated by its ligand (O2).**

1. This concentration is the **dissociation constant (K_D)**, and it indicates how tightly a ligand binds its receptor. For this reason, **K_D** is also called the **affinity constant**.
2. **K_D** has units of concentration. For a gas, we may use partial

pressure (units: mm Hg, atmospheres, Pa, pounds per inch^2 (p.s.i.), *etc.*). For solutes in liquid we will more often use concentration per volume (M, g/L, *etc.*).

3. We say that an interaction with a *small* K_D (say, 10^{-9} M) is **high-affinity**. We say that an interaction with a *large* K_D (say, 10^{-3} M) is **low-affinity**.
4. For a simple ligand-receptor pair, L + R ⇌ LR, KD = ([L][R])/[LR] where [L], [R], and [LR] are the concentrations of the ligand, the receptor, and the complex, at equilibrium (when the concentrations of free and receptor-bound ligand are not changing).

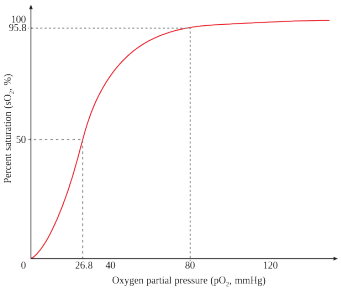


Fig. 3. Fractional saturation of HbA at different O2 partial pressures. Source: Wikimedia

5. **KD** is an *equilibrium* constant, which reflects two **rates**: the rate at which a particular ligand sticks to the receptor (**association**), and the rate at which that ligand falls off (**dissociation**). For the simple case, L + R ⇌ LR, $\mathbf{K_D = k_{off}/k_{on}}$

   where k_{on} is the rate constant for association and *koff* is the rate constant for dissociation. k_{on} has units of inverse concentration and time (M^{-1} s^{-1}), and K_{off} has units of inverse time (s-1).
6. Take the time to satisfy yourself that if the association rate constant increases, the receptor-ligand affinity *increases*. Conversely, if dissociation rate constant increases, the affinity *decreases*.

Now take another look at the *shape* of the saturation curve for HbA. What does the shape of the curve tell us? We will be thinking through this question in class.

SLO 4. Describe the mechanistic bases of hemoglobinopathies.

Diseases caused by mutations that alter hemoglobin function are probably the most prevalent and best-understood of all Mendelian genetic disorders. At least 1 in 15 people carry genetic variants that contribute to hemoglobin-related disorders. To understand the hemoglobinopathies we need to start by looking at the genes that encode the various Hb chains, and their expression patterns (Fig. 4).

1. First, note that during gestation, Hb is initially produced in the yolk sac and then in the liver. Toward the end of gestation, Hb production gradually switches to the bone marrow.
2. Concomitant with the changes in the *locations* of Hb production, *the chain types* produced are also changing. The major form of fetal and infant Hb, **HbF** (**α2γ2**), is gradually supplanted by adult **Hb A (α2β2).**

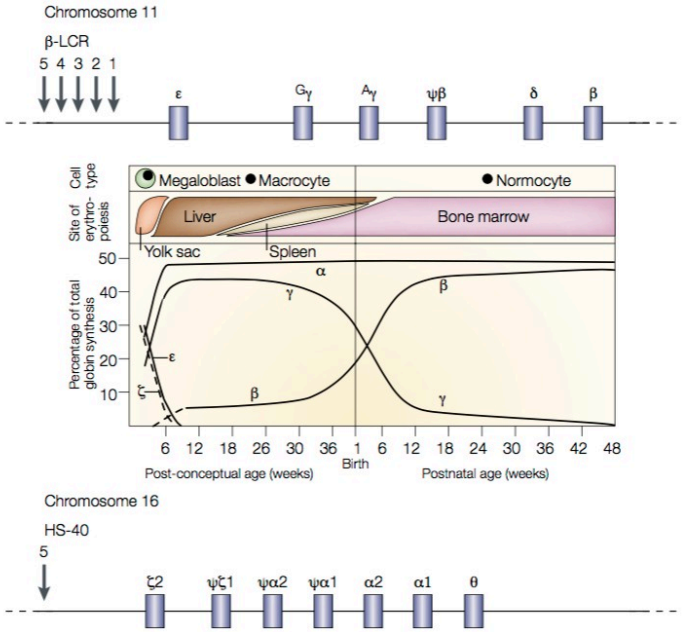


Fig. 4. Expression patterns of Hb α-like and β-like chains. On top is a map of the β-like globin loci on chromosome 11. The Locus Control Region (LCR) contains a series of enhancer sites required for expression of all β-like globin loci. Additional enhancer and repressor elements then control the expression of each specific globin gene. On the bottom is a map of the α-like globin loci on chromosome 16. ***Note that there are two*** *identical copies of the* ***α-globin gene.*** *Source: Weatherall, Nat Rev Genetics 2001.*

3. These changes are called **Hb switching**. They represent a classical example of **developmental regulation of gene transcription**. The mechanisms through Hb switching occurs include both the types of mechanisms we have already seen, and "epigenetic" changes in DNA packaging (chromatin regulation) that we'll see in more detail later.
4. When we look at the arrangement of the genes that encode Hb chains, we see that they are arranged in clusters: the β-like loci sit together on chromosome 11, and the α-like loci sit together

on chromosome 16.

5. Remarkably — and unusually — each set of genes is arranged *in its temporal order of expression.*

Sickle cell disease (hemoglobin S disease)

Sickle Cell Anemia is the most prevalent single hemoglobinopathy. It falls within the broader class of **hemolytic anemias.**

Note — Cyanosis vs. Pallor. *You cannot manifest red (or blue) color without blood. Pallor (being pale) reflects anemia. Cyanosis (being blue) indicates that blood is poorly oxygenated. As mentioned above, HbS carries O2 normally, but leads to clogging of capillaries and hemolysis, the latter resulting in anemia.*

1. Sickle disease results from a single nucleotide **missense substitution,** βS, that changes a codon for glutamate to a codon for valine at amino acid position 6 in the Hb β–chain (β–globin).
2. The β S mutation has **no effect on the ability of Hb to carry O2**.
3. Heterozygote carriers of the sickle allele, referred to as βS, generally present few or no symptoms. Extreme physical exertion can lead to rhabdomyolisis.
4. Heterozygote carriers exhibit partial resistance to malaria. The βS allele appears to have appeared *de novo* multiple times. Its prevalence is elevated in populations in the Mediterranean, Africa, and southern Asia, all areas with endemic malaria.
5. Heterozygous carriers of the

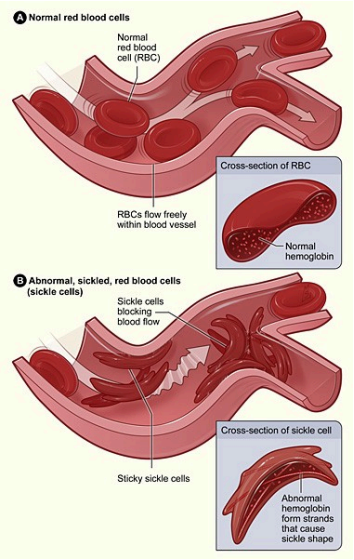


Fig. 5. Effect of HbS (α2Sβ2S) on erythrocyte morphology and function. Source: National Institutes of Health.

sickle trait produce a mix of Hb tetramers: α2A β2A (normal HbA), α2A β2S (**sickle HbS**), and α2A βA βS.

6. Homozygotes with sickle cell disease produce mainly α2A β2S (sickle) Hb tetramers.
7. The O2-free form of HbS, deoxyHbS (deoxy- HbS), is five times less soluble than deoxyHbA.
8. At the high concentrations of HbS that are present in homozygotes, **deoxy–HbS tetramers assemble into long, higher-order filaments**. These deoxy-HbS filaments distort the normal rounded shape of erythrocytes, leading to clogging of capillaries. The fibers can also puncture the cell's membrane, causing erythrocyte lysis (**hemolysis**).
9. Sickle crisis is an episode of extreme pain lasting hours to days. Generally, crisis is thought to result from sickled erythrocytes blocking blood flow, particularly to bones. A short (2 min.) video on sickle crisis is here:

https://www.nejm.org/do/10.1056/NEJMdo005311/full/

This is one of many examples you'll see of ***protein folding diseases.***

Other hemoglobinopathies

In addition to S disease there are a large number of mutations in the various Hb chains that can lead to disease. These fall into three categories:

1. **Structural hemoglobinopathies**. These disorders are generally caused by **missense mutations** that alter the **primary structure** of Hb chains. Sickle (HbS) disease is an example.
2. **Thalassemias**. These disroders are caused by **imbalances** in the amounts of α- and β-chain synthesis, degradation, or Hb tetramer assembly, leading to excess production of unassembled globin chains. *Note: some structural hemoglobinopathies can result in thalassemia.*
3. **Hereditary persistence of fetal hemoglobin (HSPS)**. These are regulatory disorders in which **the switch from HbF to HbA fails to occur** in early childhood. HSPS by itself does not lead to

major pathology, but it can be a strong genetic modifier of other hemoglobinopathies and thalassemias.

An absolutely key point is that **there are two identical copies of the Hb α-chain gene** on chromosome 16 (See fig. 4). This means that most people have *four copies* of the α-chain gene. Consequently, recessive disorders of Hb β–chains are far more frequent, and tend to be more severe, than disorders of the Hb α-chains.

Moreover, because of the *temporal order* of Hb chain expression (Fig. 3), disorders of Hb β–chains tend to manifest only in childhood (recall that HbA is α2β2), while disorders of Hb α-chains can begin manifesting prenatally (HbF is α2γ2).

Structural hemoglobinopathies

For the Hb β–chain (β-globin) alone, mutations leading to the synthesis of well over 700 different structural variants of the protein have been identified. Additional variants in other globins can also cause disease.

The Hb structural variants were originally named with letters (HbS, HbE, *etc*., and later, they were named by the location where the carriers of the mutations were identified (Hb Hammersmith, *etc*.). Structural mutations can lead to changes in the O2 saturation curve (and cyanosis, as in Hb Hammersmith), changes in Hb solubility, and hemolysis (as in HbS), to thalassemias, or combinations of these defects.

In addition, the iron in the heme center can be oxidized by bound O2, from ferrous (Fe^{2+}) to ferric (Fe^{3+}) iron. Hb containing an oxidized heme center **cannot bind O2, and** is called **methemoglobin.** An enzyme, **methemoglobin reductase** converts the heme iron to the ferrous (Fe^{2+}) state, and restores its ability to carry O2.

Mutations that impair methemoglobin reductase cause gradual conversion of Hb to methemoglobin, resulting in cyanosis. Moreover, some structural variants of Hb (such as Hb Hyde Park) cannot productively engage with methemoglobin reductase, and

these also gradually convert into methemoglobin (also leading to cyanosis).

Thalasseimias

A vast variety and number of mechanisms and mutations cause imbalances in globin chain abundance. The resulting diseases occur through a variety of mechanisms, with diverse presentations and severities. Collectively, these diseases are called **thalassemias**.

1. **α–thalassemia**. Recall that there are two α-globin genes on chromosome 16 (Fig. 4). α-thalassemias most commonly arise through deletion of entire α1 or α2 genes.

    - The silent *carrier* genotype is $-\alpha/\alpha\alpha$, resulting in 75% of normal α-globin production.
    - The α-thalassemia *trait* occurs in two forms: $--/\alpha\alpha$, and $-\alpha/-\alpha$. These genotypes result in 50% of normal α-globin production.
    - Simple α-thalassemia disease is associated with the $--/-\alpha$ genotype. Only 25% of the normal α-globin amount is produced, and an excess of Hb β4 tetramers are produced. This variant is also called HbH or Hb Bart's.
    - The $--/--$ genotype results in assembly of Hb γ4 tetramers. Because this genotype is embryonic lethal, the switch from γ to β expression never occurs.

2. **β–thalassemia**. A wide variety of mutations can cause reduced production of β globin chains, and β-thalassemia. We review these because they illuminate and summarize many of the ways that mutations can change protein production.

    - The **locus control region** (LCR; Fig. 4) is needed for transcription of the entire set of β-like globin chains. Deletions within the LCR can decrease or eliminate transcription of all of the β-like genes.
    - Similarly, mutations in the β-chain **promoter** can impair

transcription of the β-chain mRNA alone, but leave transcription of the other β-chain genes intact.

- Mutations in the β-chain transcription unit can prevent formation of a functional mRNA template:
  - Defects in **5′cap** addition;
  - mRNA **splicing** defects;
  - Failure of **poly–A** tail addition;
  - Point mutations that eliminate the **start codon**;
  - Insertions or deletions (indels) that cause **frameshifts**;
  - Introduction of premature **stop codons** (**nonsense** mutations).
  - Missense mutations (like those in the structural hemoglobinopathies) can also cause instability β-chain *protein* and its destruction by the **protein quality control** system.

3. As with sickle cell trait (HbS), heterozygotes carrying the α- and β-thalassemia traits are partially protected from malaria. These traits (and hence, the associated diseases) occur at elevated levels in populations affected by endemic malaria. Because most disease alleles are in carriers compared to affecteds for recessive disorders (2pq>>q2), a fitness advantage to carriers often outweighs the loss of fitness among those with disease. (As used here, fitness refers to survival to reproduction.)

SLO 5. Explain the key properties of enzymes. Explain why many enzymes contain bound prosthetic groups. Describe enzymatic co-factors and how deficiency of these factors leads to disease.

Enzymes

Enzymes are highly selective **catalysts**.

1. Enzymes bring together specific **substrate** molecules in specific geometries, and accelerate chemical reactions that convert the substrates into specific **products**.
2. As catalysts, enzymes **cannot change the thermodynamics — the overall free energy balance — of a reaction** (Fig. 6).
3. Instead, enzymes change the **kinetics** of the reaction. They do this by reducing **activation energy barriers** that would otherwise prevent the reaction from occurring at a biologically relevant rate.
4. By bringing specific substrates together in the appropriate geometry, and by shielding intermediates from reactive non-substrate molecules, enzymes **suppress the formation of off– pathway products**.
5. The vast majority of enzymes are proteins. (But, as we saw with the ribosome, a small critical subset of enzymes have catalytic centers made out of RNA.)
6. Many enzymes use covalently or non-covalently bound **prosthetic groups** to control the electronic environment within their active sites, promoting catalysis. Prosthetic groups are sometimes, but not always, vitamins. Thus, vitamin deficiency often leads to compromised function in specific enzymes, leading to metabolic or structural pathologies.
7. In summary, we can think of enzymes as receptors that bind ligands (substrates) in highly specific ways, to promote specific chemical reactions among those substrates.

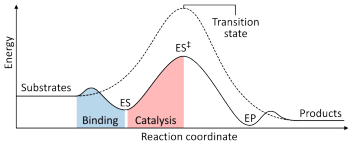


Fig. 6. Principle of enzyme-mediated catalysis. E, enzyme. S, substrate(s). P, Product(s). The enzyme does not alter the overall free energy balance of the reaction pathway. Rather, the enzyme lowers the activation energy for formation of the transition intermediate. Source: Wikimedia

SLO 6. Explain and calculate the relationships

between kcat, Km, and Vmax. Michaelis-Menten enzyme kinetics

As with ligand-receptor interactions, a small number of parameters provides a vivid description of an enzyme's critical properties.

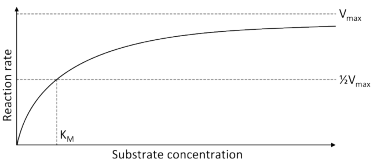


Fig. 4. Michaelis-Menten enzyme saturation curve. Source: Wikimedia

1. **V** is the reaction rate at which substrate is converted to product, under some specified set of conditions (substrate concentration, temperature, pH, *etc*.).
2. The $\mathbf{V_{max}}$ is the *maximum* rate at which the enzyme can convert substrate to product. $\mathbf{V_{max}}$ occurs when the substrate is at a sufficiently high concentration that its availability is not rate-limiting.
3. The $\mathbf{K_M}$, or **Michaelis–Menten constant**, is the concentration of substrate at which the rate of product formation per molecule of enzyme is half-maximal. Thus, **at a substrate concentration [S], where** $\mathbf{V = V_{max}/2}$, $\mathbf{K_M = [S]}$
4. $\mathbf{K_M}$ **is closely analogous to** $\mathbf{K_D}$, because each refers to the concentration of ligand or substrate where 50% of the receptor or enzyme is occupied.
5. $\boldsymbol{k_{cat}}$ is the rate constant for the enzyme-catalyzed conversion of the enzyme-bound substrate to product: For the scheme in Fig. 3, $\boldsymbol{k_{cat}}$ is the rate for ES ⇾ E+P. *kcat* is also called the enzyme's *turnover rate*.
6. The **catalytic specificity** of an enzyme for any given substrate can be defined as a ratio, $\boldsymbol{k_{cat}}$ **/** $\mathbf{K_M}$
7. By comparing k_{cat} /K_M ratios for one enzyme and various substrates, we can learn how **selective** an enzyme is. For this reason k_{cat} /K_M is often called the **specificity constant**.
8. *For example*: the active site pockets of DNA polymerase

enzymes have high affinity (small K_M) for dNTP (DNA) nucleotides, but extremely low affinity (very large KM) for NTP (RNA) nucleotides. This is because DNA polymerase active sites are usually shaped so that the 2′-OH group of dNTPs does not fit within the pocket.

9. Thus for DNA polymerases k_{cat} /K_M for dNTPs is relatively large, while k_{cat} /K_M for NTPs is very small. *We therefore say that DNA polymerases are selective for dNTP substrates versus NTP substrates.*
10. As we'll see in following sessions on Pharmacology, *competitive enzyme inhibitors* have high affinities for enzyme active sites (low K_M), but small — or zero — k_{cat}. Competitive inhibitors "clog" an enzyme's active site, preventing legitimate substrates from binding.

6. Genetics Introduction

7. Epigenetics

Session Level Objectives (SLOs): after completing the session, students will be able to:

SLO1.Understandthefundamentalsofchromatinstructureandremodeling.

SLO2.DescribethemechanismsbywhichcovalenthistonemodificationsandDNAmethylationresultinepigeneticregulationofgeneexpression

SLO3.Demonstratehowepigeneticmodificationsresultinimprintinganddistinguishhowimprintingleadsto Prader-Willior Angelman syndromes.

SLO4.Illustratehownon-codingRNAsandcovalentepigeneticmodificationscooperativelyregulatemammalianX-inactivation

EPIGENETICS

Epigenetics was first defined in the 1940's by Conrad Waddington as "the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being." Today, Epigenetics most commonly refers to heritable changes of gene expression (across cell divisions or across generations) that do not involve changes in DNA sequence.

Epigenetic regulation most commonly occurs as a result of covalent modification of either DNA or of the DNA-packaging histone proteins. We will begin with a discussion of the fundamentals of chromatin structure to provide context for these covalent modifications and their effects. Chromatin is an intrinsically repressive environment for gene expression; much of the DNA is masked by protein. Transcriptional activators can recruit proteins that regulate chromatin accessibility. Cellular mechanisms that control accessibility to DNA in chromatin include covalent DNA

modifications, covalent histone modifications, and ATP-dependent chromatin remodeling. These are frequently all referred to as epigenetic mechanisms, though formally only mechanisms that are potentially heritable through cell divisions are considered epigenetic. Epigenetic changes in humans are typically not heritable through the germline, as epigenetic modifications are erased during gamete formation.

SLO 1: Understand the fundamentals of chromatin structure and remodeling.

Chromatin

Chromatin is a complex of DNA, proteins and RNA that allows chromosomes to be packaged and organized within a cell's nucleus. Maintenance and remodeling of chromatin structure is essential for the proper regulation of gene expression, DNA replication, cell division and prevention of DNA damage. The fundamental unit of chromatin is the **nucleosome,** a structure comprised of 8 **histone** proteins around which DNA is wrapped.

Histones

Histones are highly conserved, positively charged (lysine- and arginine-rich) proteins. All histones share a related histone fold domain comprised of 3 alpha helices, and the core histones also have additional C- and/or N-terminal extensions known as the histone tails.

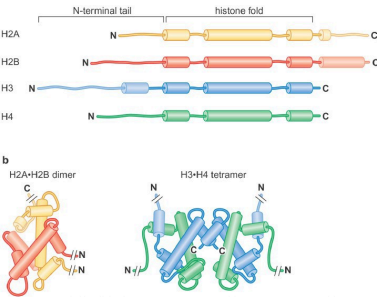


Figure 1. Schematic of histone monomers H2A, H2B, H3, and H4 (above). Schematized structure of histone dimers H2A-H2B (left) and tetramers H3-H4 (right).

There are four core his

tones: H2A, H2B, H3 and H4 (Figure 1). Two each of histones

H3 and H4 form the H3-H4 tetramer, and H2A and H2B form dimers. One H3-H4 tetramer and two H2A-H2B dimers form the histone octamer (also called the "histone core particle.".

The Nucleosome: the building block of chromatin

Figure 2. The structure of the nucleosome composed of DNA wrapped around 8 histone proteins.

The nucleosome is formed when 146 bp of DNA wraps around the histone octamer (Figure 2), forming multiple ionic bonds between the DNA phosphate backbone and lysine and arginine residues in the histones. Linear DNA is wrapped around a succession of histone cores. DNA connecting adjacent nucleosomes is referred to as linker DNA. There is an average of ~50-60 bp of DNA in each linker region, though this can vary significantly at individual locations in the genome. Linker histones, such as Histone H1, interact with linker DNA to impact levels of chromatin folding.

Chromatin folding

The nucleosome represents the first level of DNA packing, but in the nucleus DNA is condensed 10,000-50,000 fold over its fully extended length. An array of nucleosomes is also called a **10- nm fiber** (its actual diameter is 11-nm). The 10-nm fiber is further folded via interactions with other proteins and association with an underlying proteinaceous scaffold that gives structure to both interphase chromatin (300nm) and the structure of metaphase chromosomes (700nm & 1400nm) (Figure 3).

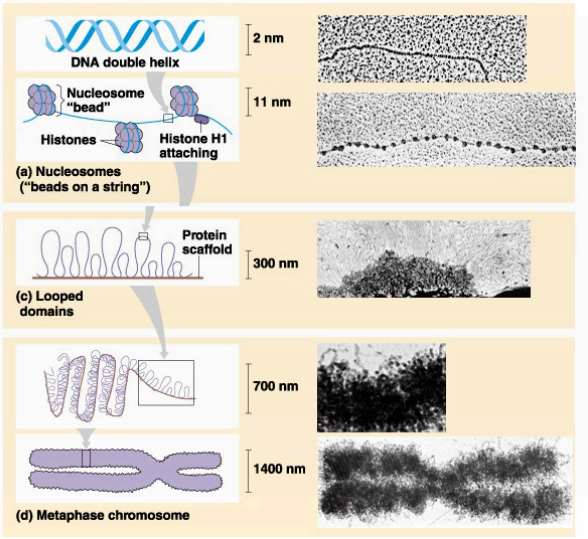


Figure 3. Levels and scales of chromatin organization.

ATP-dependent chromatin remodeling.

There is a large superfamily of ATP-dependent chromatin remodelers that use the energy of ATP hydrolysis to alter histone-DNA contacts, resulting in nucleosome sliding, histone or nucleosome

displacement from the DNA, and other alterations in nucleosome structure. The particular outcome depends on the remodeler involved and other protein factors acting at a particular site. Remodeling can lead to either accessible chromatin for gene activation (Figure 4), or can repress transcription (e.g. by moving nucleosomes to cover key promoter elements). There are numerous subfamilies within the superfamily of remodelers; the most important are the SWI/SNF, ISWI, CHD, and INO80 families. Most ATP-dependent remodelers exist in complexes of 2-20 subunits. Members of these complexes have been implicated in human diseases, including neurodevelopmental diseases and multiple cancers.

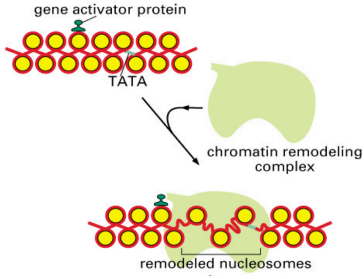


Figure 4. A schematic of chromatin remodeling. In this model a transcription factor (gene activator protein) is able to bind DNA in a region of tightly packed nucleosomes. This transcription factor then recruits a chromatin remodeling complex. This complex displaces nucleosomes to open up the DNA potentially for transcription and/or binding of additional transcription factors.

SLO 2: Describe the mechanisms by which covalent histone modifications and DNA methylation result in epigenetic regulation of gene expression.

Post-translational modifications of histones All histones have N-terminal tails, and some

also have C- terminal tails. Tails are flexible extensions that

protrude

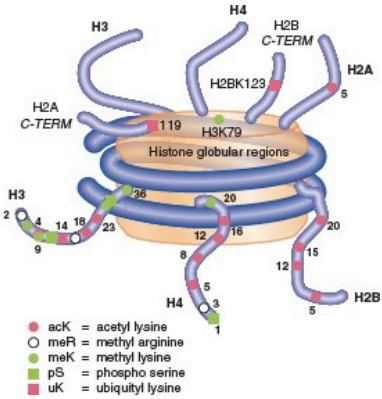


Figure 5. Diagram of histone tail modifications.

fromthenucleosome.Multiplepost-translational modifications occur in the tails (Figure 5), though some also occur in the cores. The number of known covalent modifications is extensive and range from very large moieties such as ubiquitin or SUMO, to the small molecule modifications shown here. These include the best-characterized modifications, such as lysine acetylation, lysine and arginine methylation, and serine or threonine phosphorylation. Lysine can be mono-, di- or tri-methylated, and arginine can be mono- or di- methylated. Note that both phosphorylation and acetylation change the net charge, while methylation, even trimethylation, retains the positive charge on lysine and arginine (Figure 6).

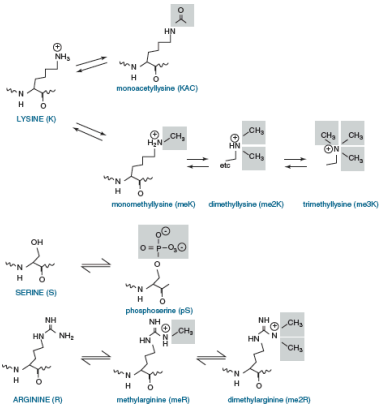


Figure 6. Examples of amino acid modifications found on histone tails.

Effects of histone modifications

Charge effects:

The positive charge on histone tails results in ionic interactions with the negative phosphate backbone of the DNA, both within a single nucleosome as well as between different nucleosomes in higher order structures. Acetylation of histone tails reduces these interactions and promotes unfolding of higher order structures and assists movement of the nucleosome along the DNA (Figure 7). ***Histone acetylation is associated with transcriptional activation.***

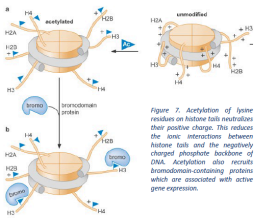


Figure 7. Acetylation of lysine residues on histone tails neutralizes their positive charge. This reduces the ionic interactions between histone tails and the negatively charged phosphate backbone of DNA. Acetylation also recruits bromodomain-containing proteins which are associated with active gene expression.

Changestobindingsites: Post-translational modification of histone proteins can also produce novel binding sites for other proteins, which can result in a variety of functional outcomes. Enzymes that add epigenetic marks to chromatin are called ***writers***. Proteins that recognize these marks are called ***readers***. Enzymes that remove these marks are ***erasers***. Note that these readers & writers are specialized: they recognize specific amino acid(s) within particular histone(s).

Histone methylation can be associated with either activation or repression, depending on the specific site of modification and thus the specific binding site generated (Figure 8). It is important to recognize that **all histone modifications are reversible**, but

patterns of modificationmay be stably maintained through multiple cell divisions.

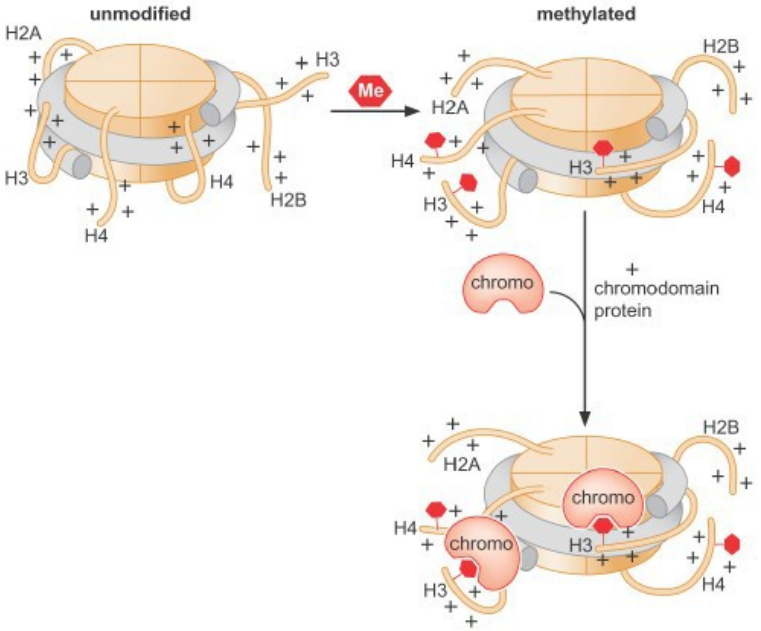


Figure 8. Methylation of amino acids on histone tails can be associated with transcriptional activation or repression depending on the amino acid and its position. Some methylated lysine residues recruit chromodomain-containing proteins which contribute to repressive chromatin complexes.

Some important classes of writers, readers and erasers:

Histone acetyltransferases (HATs) acetylate lysines (*writer*).

Histone methyltransferases (HMTs) methylate lysines or arginines (*writer*).

Bromodomains are common protein motifs that recognize acetylated lysines; they are commonly found in transcriptional activators, as well as in chromatin remodeling complexes (*reader*).

Chromodomains are one family of methyllysine binding domains, frequently found in repressive complexes (*reader*). Several other methyllysine and methylarginine binding motifs are known as well.

Histone deacetylases (HDACs) and **histone demethylases** remove acetyl and methyl marks, respectively (*erasers*). HDACs are generally found in repressor complexes. Histones can be extensively and simultaneously modified at numerous (different) sites within a single nucleosome. Modifications of the same residue are mutually exclusive (such as acetylation vs. methylation of the same lysine). The **histone code hypothesis** states that **combinations of modifications define chromatin function and accessibility** (Figure 9).

The "meanings" of many modifications have been defined, but the complexities of iinteractions of many modifications have yet to be clarified. Histone acetylation is almost exclusively associated with transcriptional activation; while histone methylation is more frequently associated with transcriptional repression (eg H3- K27me2,3), some specific histone methylation sites are associated with activation (eg H3K4me2,3). Note that while the focus here is on transcription, histone modifications are also critically involved in all DNA transactions, including replication, recombination, and repair.

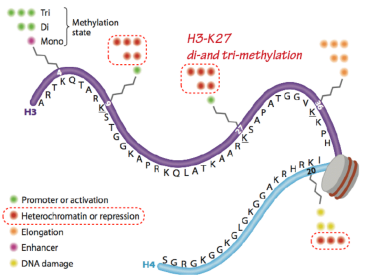


Figure 9. Cartoon representation of the histone code hypothesis.

Post-translational modification of DNA

The most extensive (and best-studied) covalent modification of DNA is **DNA methylation** (Figure 10). It occurs on the 5C position of cytosine within **CpG dinucleotides**. ("CpG" is just a shorthand that specifies a 5'-CG-3' sequence.) CpG dinucleotides are underrepresented (less frequent than expected by chance) in most of the genome, but enriched 100-fold in regions known as **CpG islands**.

DNA methylation is typically a *silencing* epigeneticmark that is very stable through cell divisions. DNA methylation often occurs at CpG islands within gene promoter regions to silence gene expression. Commonly genes are progressively methylated during cell differentiation as genes whose expression is not needed for a particular cell type are silenced.

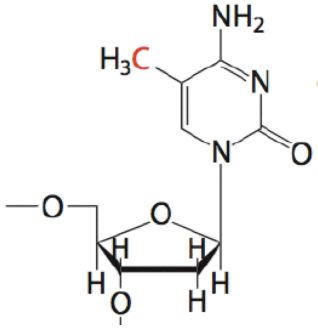


Figure 10. Methylated cytosine

DNA methyltransferases (DNMTs) are the writers that perform this function. **DNA demethylases** are the erasers. **Methyl-C binding domains (MBDs)** are a common motif in readers of DNA methylation.

Gene silencing. Proteins that read and write repressive marks interact with each other, creating positive feedback loops that lead to sustained, self-reinforcing silencing (Figure 11). In addition, DNA methylation is maintained through replication with nearly 100% efficiency, ensuring methylation patterns are preserved through cell divisions. Thus, a fully silenced state is one of the most stable epigenetic states.

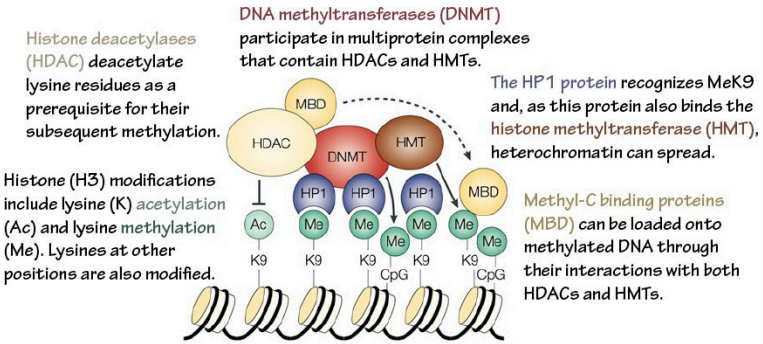


Figure 11. Readers and writers of repressive marks create positive feedback loops to maintain silenced chromatin.

SLO 3: Demonstrate how epigenetic modifications result in imprinting and distinguish how imprinting errors lead to Prader-Willi or Angelman syndromes.

Imprinting: differential expression of paternal vs. maternal alleles.

1. Imprinting is a specialized subtype of gene silencing. Patterns of allelic expression are predetermined based on the parent of origin of the allele. The allele from one parent is methylated and silenced (imprinted), while the allele from the other parent is active. Gender- specific patterns of imprinting occur during germ cell formation. First there is global demethylation followed by a sex-specific imprinting pattern that is laid down in primordial germ cells. These patterns are then maintained through fertilization and subsequent cell divisions (with the exception of

some tissue-specific demethylation of certain genes) (Figure 12).

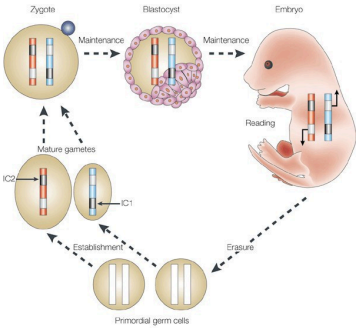
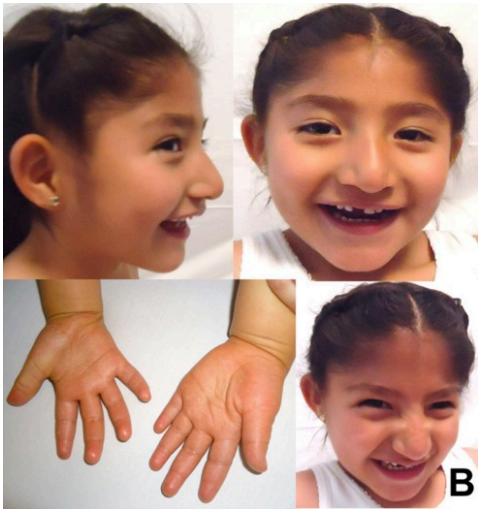


Figure 12. Schematic of parental imprinting. The epigenetic marks that drive parent-specific repression of certain genes (imprinting) are erased during the formation of primordial germ cells. Imprinting centers (ICs) then establish regions of repressed chromatin in the mature gametes that are unique to either the egg or sperm. These repressed regions are maintained in the fertilized zygote and throughout the lifetime of the resulting organism until those marks are once again erased in the new generation's germline.

Diseases of imprinted loci. Several human syndromes result from mutations that impact imprinted regions. Two very different syndromes, Prader-Willi Syndrome (PWS) and Angelman Syndrome (AS), can actually be caused by the *identical*deletion in patients. PWS is characterized by chronic feelings of hunger and is the leading genetic cause of marked obesity; AS features severe motor and cognitive impairment. ~70% of patients have a 4-5 Mb deletion of a region known as the PWS-AS critical region (Figure 13). **Which syndrome a patient exhibits is determined by *which parent contributed the deletion***. The typical deletion encompasses a large region with numerous imprinted genes, including two that are paternally imprinted (*UBE3A* and *ATP10C*) and multiple maternally imprinted genes.

A



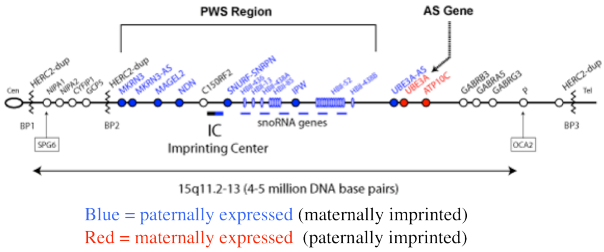


Figure 13. A) young boy with Prader-Willi Syndrome (PWS). B) a 5-year-old girl with Angleman Syndrome. C) Region of chromosome 15 affected in PWS and AS Syndrome. Common chromosomal breakpoints BP1, 2, 3 are diagrammed.

Prader-Willi syndrome results when the deletion comes from the father due to the loss of paternally expressed genes in the PWS critical region (i.e. none of the genes shown in blue in the figure are expressed, because they are imprinted in the intact maternal copy, and deleted in the paternal copy).

Angelman syndrome results when the deletion comes from the mother. In this case, it is the loss of the maternally expressed UBE3A gene (in the ubiquitin pathway) that results in the syndrome (mutations just in UBE3A alone also cause AS).

70% of the cases of both PWS and AS are due to deletion of the PWS-AS critical region. Both syndromes can also result from uniparental disomy, in which the individual inherited both copies of chromosome 15 from the mother (PWS, ~25% of cases) or the father (AS, ~7% of cases). In these cases, both chromosomes have the imprinting pattern of a single parent, and thus silence the same set of alleles on each copy. The remaining cases of PWS or AS result from mutations in genes that control imprinting, UBE3A mutations (AS), or other chromosomal rearrangements (Figure 14).

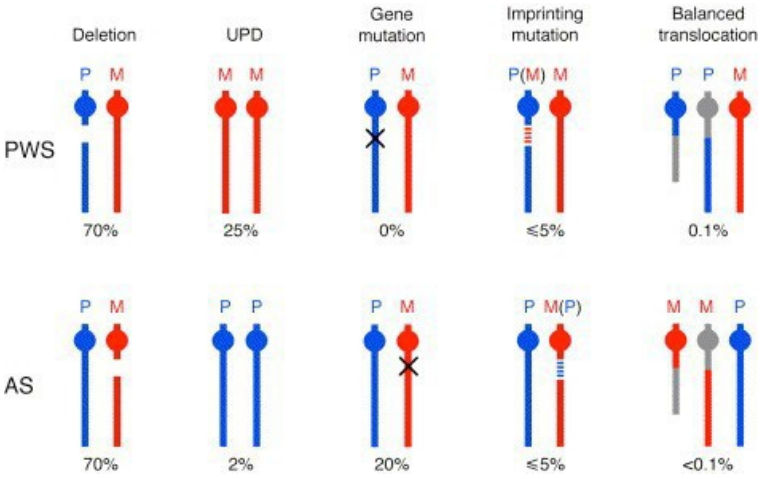


Figure 14. Structural chromosomal variants that can cause Prader-Willi Syndrome (PWS) or Angelman syndromes (AS). Paternally-derived chromosomes (P) are depicted in blue. Materinally-derived chromosomes (M) are depicted in red.

SLO 4: Illustrate how non-coding RNAs and covalent epigenetic modifications cooperatively regulate mammalian X inactivation.

X-inactivation results in silencing of an entire chromosome. As described earlier, histone and DNA modifications can act in concert to establish stably silenced regions of the genome. *Heterochromatin (Figure 15)* is defined as highly compacted throughout the cell cycle, late replicating, and

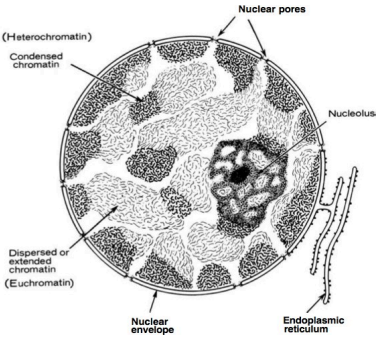


Figure 15. Chromatin organization within the nucleus of the cell.

transcriptionally silent domains of chromatin. (All other regions are collectively known as euchromatin.) Some heterochromatic regions are always found as heterochromatin, such as telomeres and centromeres: these are **constitutive heterochromatin**. Other regions maybe euchromatic or heterochromatic depending on developmental stage, cell type, or other features. These are regions of **facultative heterochromatin**. The most dramatic example of facultative heterochromatin is the

silent X chromosome in female mammals.

All female mammals are mosaic for expression of X chromosome genes, familiar to most in the coat color of calico cats (Figure 16). Because female mammals have two X chromosomes and males have one, the process of dosage compensation ensuresthat males and females receive the same levels of gene expression of X- linked genes. In mammals, dosage compensation is accomplished by [almost] entirely silencing one X chromosome in each cell in females. This silencing is random and occurs early in development; thus the adult is a patchwork of tissues in which cells were derived from an ancestor in which one of the two

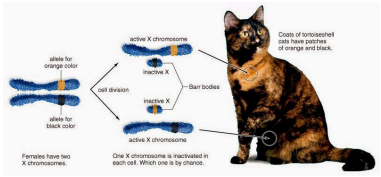


Figure 16. The phenomenon of X inactivation leads to the mosaic coat pattern of calico cats.

X chromosomes was silenced. The process was discovered by Mary Lyon and X chromosome inactivation is often referred to as "Lyonization." The process of X-inactivation includes *counting* (all but one X will be silenced in individuals with >1 X), a *choice* where one X is designated to become the active X (Xa) and other(s) to become inactive (Xi), and finally the process of triggering extensive heterochromatin formation on Xi. Critical to this process are **noncoding RNAs (ncRNAs)** known as *Xist*

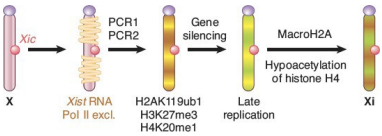


Figure 17. The stepwise silencing of one X chromosome during development.

and *Tsix*. **Xist** is a remarkable RNA that is transcribed stably from the *Xic* locus on the future Xi, and physically coats the chromosome to trigger the silencing process. Initially, both X chromosomes express *Xist*, which is destabilized by the presence of an antisense transcript from the same locus, *Tsix*. Selection of the future Xi involves inactivation of *Tsix* and stabilization of *Xist*. *Xist* accumulation on the Xi excludes RNA polymerase II, recruits silencing complexes, induces multiple histone modifications and inclusion of a variant of histone H2A found only on Xi, and finally DNA methylation of gene promoters on the Xi (Figure 17).

The role of X-inactivation in disease. Specific X-linked genes are mutated in many diseases such as Rett Syndrome, duchenne muscular dystrophy, color blindness and hemophila A. We typically think of X-linked conditions affecting males because they only have one X chromosome. Females, however, may be heterozygous for the mutation. In other words they may have one mutated and one normal allele, although any given cell expresses only one allele because of X-inactivation.

Females can be affected by X-linked mutations as a result of skewed Lyonization, but may have distinct clinical phenotypes because some of their cells are normal, while other cells express the mutated gene on the X chromosome. For example, young girls with pathological mutations in one copy of the X-linked Methyl CpG-binding Protein gene (MECP2) present with Rett Syndrome, an autism-like neurodevelopmental disorder with an onset between 6-18 months. In contrast, boys with MECP2 mutations, who survive to term, present with different phenotypes including generalized mental retardation. Aside from being located on the X-chromosome, the MECP2 gene produces a protein product that binds to methylated DNA throughout the genome to regulate transcription. Rett Syndrome, therefore, is not only influenced by the epigenetics of X-inactivation, but also affects the epigenetic state of the cell through the misregulation of expression of methylated DNA. The severity of an X-linked condition will depend on the pattern of X- inactivation. Just as calico cats can have coat

color spots of variable size and coverage, X- inactivation in humans can be skewed for a given body region or organ system.

8. Sickle Cell Disorder Mutations, Lab Techniques, Integration

FUERSTPG

Mutations, Molecular Biology Techniques and Sickle Cell Disease

Session Level Objectives (SLOs): after completing the session, students will be able to:

1. Types of Mutations: Correlate gene mutations with effects on protein expression, structure and function.

SLO1 Quiz1 **SLO1 Quiz2**

2. Molecular Biology Techniques: Explain how different techniques are used to detect different types of biological molecules including northern, western and Southern blotting, ELISA, PCR, Sanger sequencing, RNAseq, (FACS) cell sorting and HPLC.

SLO2 Quiz1 **SLO2 Quiz2**

3. Sickle Cell Disease: Know the common mutations that give rise to sickle cell disease (C and S) and interpret these from fetal (F) and adult (A) hemoglobin on diagnostic tests for both carrier and disease states.

SLO3 Quiz1

Synopsis: In this session we will consider mutations and molecular biology techniques and apply this knowledge to learn about sickle cell disease. Sickle cell disease is a debilitating disorder

and sickle cell patients disproportionately suffer from health care disparities. The out of class material will focus more heavily on SLOs 1 and 2 while SLO3 is covered more thoroughly in class. All SLOs are relevant clinically and for testing purposes.

SLO1: Correlate gene mutations with effects on protein expression, structure and function.

There are two primary types of changes, or mutations, that occur in DNA. The first type of change is a **base substitution**. When a base substitution occurs, one letter of DNA is changed to another letter, for example A -> G. The second type of mutation is the insertion/deletion mutation, or **indel mutation**, in which the number of letters in a DNA sequence increases or decreases, for example GATC -> GAC (1 base deletion) or GATC -> GAATC (1 base insertion). A specific type of indel mutation is a base repeat expansion mutation, wherein a repeated number of bases is increased, for example CCG36 -> CCG42, which would be the addition of 6 new CCG repeats for a total of 18 new nucleotides.

Consequences of base substitution mutations

Base substitutions may occur within genes or in non-coding sequence outside of genes. Within genes, the base substitutions may occur within coding sequence or in non-coding sequence, for example within introns. The consequence of base substitutions varies widely depending on where the mutation is located.

Base substitutions in coding sequence can have four different effects on the coding sequence of the gene:

Synonymous mutation: The DNA changes, the mRNA changes, but the amino acid that is coded for does not change. An example would be GCA -> GCG, both codons of which code for alanine. If you consult the genetic code below you will notice that bases in the third position of the codon tolerate base substitutions.

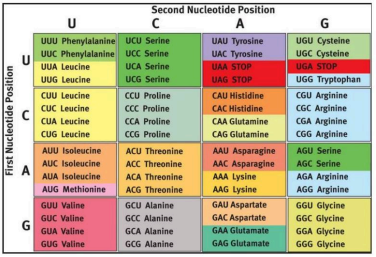


Figure 1 Genetic Code

You will notice that some base substitutions in the third position will change the amino acid that is coded for and that some changes in other positions will not (see Arginine, Serine and Leucine, for example). *In* ***rare*** *cases synonymous mutations can cause disease, for example if they disrupt normal splicing or regulatory sequences. Most synonymous mutations are silent mutations- that is there is no change in phenotype.*

Nonsense mutation: A nonsense mutation changes an amino acid coding codon to one of the three stop codons, for example UGG -> UGA. This will terminate the protein early, which likely will have a profound effect on the resulting protein, depending on how close the new stop codon is to the native (normal) stop codon. Sometimes this will trigger a process called nonsense mediated decay, wherein mRNA with a premature stop codon is degraded. A *premature stop codon introduced by an indel mutation is not a nonsense mutation!* It is a premature stop codon introduced by an indel mutation. A premature stop codon introduced by an indel mutation can however result in the process of nonsense-mediated decay of the mRNA transcript.

Missense mutation: A missense mutation is a base substitution in which the amino acid that is coded for changes, for example AGA -> AGC. Some missense mutations have little or no effect on protein structure and function, for example a mutation that changes leucine to isoleucine may not have a detectable impact on protein structure and function. These are conservative missense mutations. A missense mutation that changes the class of amino acid or changes an amino acid with special properties is more likely to be a non-conservative deleterious mutation, for example the mutation GAA (glutamate) -> CAA (glutamine) will change the charge of the amino acid that is coded for. *The mutation that results in sickle cell*

disease, GAG -> GTG (glutamate -> valine), is a non-conservative deleterious missense mutation.

Read through mutation: A rare type of base substitution changes a stop codon to an amino acid. An example would be TGA -> CGA (stop -> arginine). Read through mutations will result in longer proteins. Translation will continue until a stop codon is encountered in what would otherwise be the 3' UTR.

Base substitutions in non-coding sequence: Changing a nucleotide base outside of coding sequences is likely to have no detectable outcome but if the base substitution is in regulatory sequence the change could be very deleterious. Examples include intron regulatory sequences, promotors and miRNAs. The vast majority of non-coding sequence does not play a regulatory or other role.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://uw.pressbooks.pub/fmrmolbio/?p=202#h5p-1

Consequences of insertion deletion (indel) mutations

Indel mutations in coding sequence: Indel mutations add or remove bases, hence the name. The impact of indel mutations in coding sequences depends on whether the mutation is in frame and whether it preserves the reading frame.

If the indel mutation is divisible by three and is in frame it will preserve existing or remaining codons and insert or delete codons.

In frame and **preserves reading frame** insertion example:

GAA CGC AUG TTT CCA ACC TCC TCC Codons (wild type)

E R M F P T S S Coded Amino Acids

GAA CGC AUG ACA GAC TTT CCA ACC TCC TCC Codons (mutant)

E R M T D F P T S S Coded Amino Acids

Out of frame but **preserves reading** frame insertion example:

GAA CGC AUA CAG ACG TTT CCA ACC TCC TCC Codons (mutant)

E R I Q T F P T S S Coded Amino Acids

It should also be noted that an indel mutation that preserves the reading frame can introduce a stop codon, for example if the introduced nucleotides code for a stop codon.

If the indel mutation is not divisible by three it will not preserve the reading frame of the transcript. In this case there is little difference if the mutation is in or out of frame- both examples will be very deleterious.

In frame does not preserves reading frame insertion example:

GAA CGC AUG ACA ACT TTC CAA CCT CCT CC Codons (Mutant)

E R M T T F Q P P Coded Amino Acids

Out of frame does not preserves reading frame insertion example:

GAA CGC AUA CAA CGT TTC CAA CCT CCT CC Codons (Mutant)

E R I Q R F Q P P Coded Amino Acids

It is likely that this type of mutation will introduce an early stop codon, but in some cases the indel mutation might cause translation to miss the native (normal) stop codon and code for a longer protein. This can occur if the mutation is close to the stop codon.

Indel mutations in non-coding sequence: Similar to base changes outside of coding sequences, indel mutations in non-coding sequence are much less likely to have a detectable outcome. If the

indel is in regulatory sequence the change could be very deleterious however. Examples include intron/exon junctions and intron regulatory sequences, promotors and miRNAs. A specific example is the CGG repeat in Fragile X that results in methylation and inactivation of gene expression. Such nucleotide expansion repeat disorders affect the nervous system and will be covered in more detail in genetic and pathology sessions. The vast majority of non-coding sequence does not play a regulatory or other role however and the indel mutation in non-coding sequence is more likely to not have an impact because of this.

Gene duplication and deletion indel mutations: Sometimes indel mutations are very large and result in the duplication or deletion of an entire gene or genes.

Silent mutations: Any type of mutation can be a silent mutation (also referred to as a neutral mutation) – a mutation which does not result in a change to phenotype.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://uw.pressbooks.pub/fmrmolbio/?p=202#h5p-2

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://uw.pressbooks.pub/fmrmolbio/?p=202#h5p-3

An interactive H5P element has been excluded from this

version of the text. You can view it online here: https://uw.pressbooks.pub/fmrmolbio/?p=202#h5p-4

SLO 2. Explain how different techniques are used to detect different types of biological molecules including northern, western and Southern blotting, Histochemistry, ELISA, PCR, RT-PCR, Sanger sequencing, RNAseq, (FA) cell sorting and HPLC.

Many of the techniques we will briefly review utilize electrophoresis (Figure 2), the process of separating molecules based on size and or charge through a matrix (agarose, acrylamide etc.) by applying an electrical charge across the matrix. The molecules that migrate can be stained with a variety of dyes and other labels to visualize them.

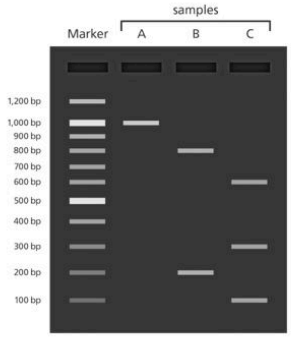


Figure 2. Example of electrophoresis. DNA samples are loaded in the sample wells and migrate towards the cathode because DNA has a negative charge. The DNA molecules migrate based on their size.

Figure 2 Example of Electrophoresis

Other techniques make use of antibodies that are specific for (typically) a protein antigen. Antibodies may be polyclonal (for multiple B-cells) or monoclonal (from a single B-cell lineage). Antibodies are made by injecting the antigen into an animal that will have an immune response to the antigen and then isolating the antibodies or B-cells from the animal.

What follows is a **very brief** synopsis of a variety of molecular biology techniques that you will encounter in the curriculum.

Southern Blot: Like Northern blot but for DNA. DNA molecules

are typically cut with a restriction enzyme before running them on a gel matrix by electrophoresis. Southern was a person, which is why Southern is capitalized.

Northern Blot: RNA molecules are separated by electrophoresis. Typically a radioactive complementary DNA or RNA molecules are used to detect and identify the presence, size and abundance of a specific type of RNA molecule. This technique is not frequently used anymore because of the advent of RNAseq.

Western Blot: Protein samples are separated by electrophoresis and specific protein molecules are detected with an antibody.

PCR (Polymerase Chain Reaction): Short DNA oligo primers are used to selectively amplify a DNA sequence with repeated cycles of strand denaturation and DNA synthesis*.

PCR: Short DNA oligo primers are used to selectively amplify a DNA sequence in polymerase chain reaction. The DNA may then be run or a gel and visualized, sequenced or used for a downstream application.

RT-PCR (reverse transcription-PCR): PCR but using cDNA as a template. cDNA is generated by reverse transcription of RNA. The COVID-19 PCR test is an RT-PCR test. Reminder- cDNA is generated by reverse transcribing RNA, typically mRNA. This differs from genomic DNA (gDNA) which would include intron sequences that are spliced out of mRNA molecules.

ELISA: Enzyme-linked immunosorbent assay. The ELISA test is a useful and common test to detect the presence of a specific antigen or antibodies to a specific antigen. ELISA tests are rapid tests and very frequently used clinically or over the counter for home use. Example: over the counter pregnancy tests are a type of ELISA test. ELISA tests involve incubating a small plastic well with solutions containing a mix of antigens and or antibodies. If the antibodies and antigens interact then they will remain in the well during washing steps. Otherwise, they will be flushed out of the well when the wells are washed.

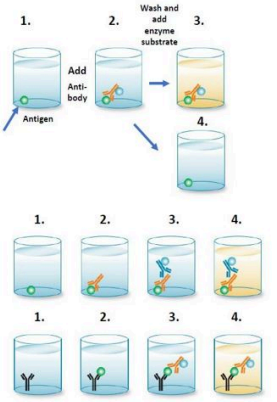
There are several types of ELISA test:

Direct ELISA: An antigen is immobilized on a well. Antibody

samples are labeled with an enzyme or fluorescent tag. The test measures if the antibody binds to the antigen. This can be used to measure if there was an immune response to an antigen.

Indirect ELISA: Like the direct ELISA an antigen is immobilized on a well. The well is incubated with an antibody (the "primary" antibody). A second "secondary" antibody that is labeled with an enzyme and binds to the primary antibody is then used to determine if the primary antibody bound to the antigen. The primary antibody binds to antigen the secondary antibody will bind to the first antibody and the result will be a positive reaction. If the primary antibody does not bind to the antigen then the secondary antibody will be washed out, resulting in a negative result.

Sandwich ELISA: A well is coated with an antibody specific for an antigen. If the antigen is present it will bind to this antibody and remain attached to the well during washes. A second labeled antibody that also binds the antigen is then added. If there is antigen "sandwiched" between the two antibodies a positive result will occur. If there is no antigen the labeled second antibody will not stick in the well after washes.



Direct ELISA
1. Antigen is fixed to the well. 2. An enzyme linked antibody is added. The enzyme linked antibody may or may not bind the antigen. The well is washed and a substrate for the enzyme that develops a color is added. 3. If the enzyme linked antibody binds the antigen it will remain after washes and give a positive result because the enzyme will cleave the substrate. 4. If the enzyme linked antibody does not bind to the antigen it will be washed out, resulting in a negative test result, because there will be no enzyme to cleave the substrate.

Indirect ELISA
1. Antigen is fixed to the well. 2. A primary antibody containing solution is added. The antibody may or may not bind the antigen. 3. The well is washed out and a second enzyme linked antibody that binds the primary antibody (if it is present; if it did not bind the antigen it will be washed out) is added. In this case the antigen that the enzyme linked secondary antibody binds to is the *primary antibody*! 4. The well is washed and a substrate for the enzyme is added. If the labeled secondary antibody is present, it will cleave the substrate and a colored product will develop.

Sandwich ELISA
1. An antigen binding antibody is fixed to the well. 2. A solution that may or man not contain an antigen that the bound antibody binds to is added. 3. The well is washed out and a second enzyme linked antibody that binds the antigen of interest (if it is still present; if it did not bind the fixed antibody it will be washed out) is added. 4. The well is washed and a substrate for the enzyme is added. If the labeled secondary antibody is still present it will cleave the substrate and a colored product will develop.

Competitive ELISA (not shown): ELISA tests can be modified to include a known amount of a competitor antibody or antigen. This can permit one to measure how much antigen or antibody there is in a solution.

Sanger Sequencing: Sanger sequencing, named after a scientist, is a very accurate "gold standard" method of sequencing a single purified population of DNA, which are typically generated by PCR.

Next generation sequencing: A lower accuracy (compared to Sanger sequencing) but high throughput method of sequencing an entire population of nucleic acid molecules. This technique has largely supplanted techniques like microarrays and northern blots.

FACS/flow cytometry: Fluorescent Activated Cell Sorting. A cell population is labeled, for example with a fluorescent antibody to CD4, and counted and or sorted on a cell sorter. This technique can be used to detect specific populations of immune cells based on their expression of specific receptors.

HPLC: High performance liquid chromatography: samples are separated based on their hydrophobicity and similar properties. The nature of some molecules can be detected based on their very specific sizes, or the sizes of their breakdown products, after separation by mass spectrometry.

Histochemistry/Cytochemistry: A technique by which tissues (histo) or cells (cyto) are probed and marked with specific labels, most often antibodies. When antibodies are used this is referred to as **immunohistochemistry** (IHC) or **immunocytochemistry** (ICC). The labels can be used to identify if a cell type expresses a specific protein, for example estrogen receptor/progesterone receptor testing of breast cancer biopsies.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://uw.pressbooks.pub/fmrmolbio/?p=202#h5p-5

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://uw.pressbooks.pub/fmrmolbio/?p=202#h5p-6

SLO 3. Know the common mutations that give rise to sickle cell disease (C and S) and interpret these from fetal (F) and adult (A) hemoglobin on diagnostic tests for both carrier and disease states.

Sickle cell disease is a type of hemoglobinopathy that is caused by a mutation in b-hemoglobin. b-hemoglobin combines with a-hemoglobin in a 2:2 ratio to form adult (A) hemoglobin. Note the Greek letters denote the individual protein chains while the Latin letters denote the hemoglobin molecules made up of four hemoglobin chains (Figure 3):

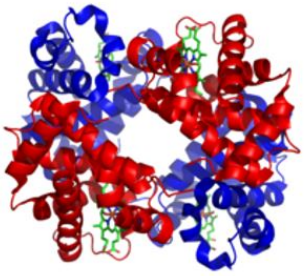


Figure 3. Structure of quaternary hemoglobin. α-chains are in red and β-chains are in blue.

alpha-hemoglobin: one of two molecules in hemoglobin. Combines with b or g hemoglobin to form hemoglobin tetramers.

beta-hemoglobin: the adult hemoglobin molecule that combines with a hemoglobin.

gamma-hemoglobin: the fetal hemoglobin molecule that combines with a hemoglobin.

A-hemoglobin: Adult hemoglobin composed of two a and two b hemoglobin subunits (a2b2).

F-hemoglobin: Fetal hemoglobin composed of two a and two g hemoglobin subunits (a2g2).

There is a developmental switch from A to F type hemoglobin early in postnatal life. In fact some therapies for sickle cell disease try to reactive production of fetal F type hemoglobin.

The primary disease causing mutation responsible for sickle cell disease is a missense mutation that results in a substitution of valine in place of glutamic acid at amino acid #6. Valine is a hydrophobic amino acid and the mutation causes hemoglobin molecules to form chains because of hydrophobic interactions between valine and other hydrophobic amino acids in other b-hemoglobin molecules. A similar mutation, glutamic acid at position 6 -> lysine is responsible for hemoglobin C disease. Note, in this nomenclature the first methionine of the protein, which is removed during translation of many but not all proteins, is not counted. Hemoglobin C disease is mild- the main complication of the mutant allele is that it can combine with the hemoglobin S mutation to cause sickle cell disease.

S-hemoglobin: Adult hemoglobin composed of two a and two b hemoglobin subunits (a2bE>V2).

C-hemoglobin: Adult hemoglobin composed of two a and two b hemoglobin subunits (a2bE>K2).

*Note: hybrid hemoglobin tetramers can form that are (a2bb*E>V). *These hybrid hemoglobin molecules actually inhibit sickling by blocking the chain formation that leads to sickling of red blood cells.*

Sickle cell and hemoglobin C diseases are recessive:

Alleles	**Outcome**
bE6->V/ bE6->V Hemoglobin S only	Sickle cell disease
bE6->K/ bE6->K Hemoglobin C only	Hemoglobin C-disease
bE6->V/ bE6->K Hemoglobin C, Hemoglobin S	Sickle cell disease
b/bE6->K Hemoglobin A, Hemoglobin C	Hemoglobin C trait
b/bE6->V Hemoglobin A, Hemoglobin S	Sickle cell trait
b/b Hemoglobin A only	Typical

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://uw.pressbooks.pub/fmrmolbio/?p=202#h5p-7

Please attend class or review the session recording for a more complete overview of sickle cell disease. This is important content. The clinical aspects of sickle cell disease will be more thoroughly covered in the blood and cancer block.

9. Cystic Fibrosis and Mutation-Specific Therapies

Session Level Objectives (SLOs): after completing the session, students will be able to:

SLO 1. Explain the molecular and cellular basis of cystic fibrosis.

SLO 2. Explain how a single mutation can cause different manifestations in a variety of tissues.

SLO 3. Explain how CF treatments can ameliorate symptoms and predict difficulties implementing them effectively in patients.

SLO 4. Describe the advantages and limitations of mutation-specific molecular therapies

CF is the most common life-shortening single-gene disorder in the U.S. and Northern Europe, affecting ~30,000 persons in U.S. alone. It is an autosomal recessive disorder. In the US, 1 in 28 white people are carriers; the incidence is much smaller in other ethnic groups (in people of African and Mediterranean origin, sickle cell anemia and G6PD deficiency are more common than CF; those diseases will also be covered in FMR). The median life expectancy has gradually increased from ~5 years before the 1950's to >40 years today in the U.S. with the development of better treatments. However, a 2017 study in the Annals of Internal Medicine reported that life expectancy with CF in Canada was 10 years longer than in the U.S., and the gap has been widening!

In infants, CF commonly presents first as failure to thrive. This is due to blockage of the pancreatic duct, leading to incomplete digestion of food. Malabsorption and caloric needs persist throughout a CF patient's life, but the most severe morbidities and mortality of CF in adults are due to the pulmonary diseases that develop.

Newborns are screened for CF in all states, initially using a blood test for trypsinogen (a pancreatic enzyme) that has a 90% false positive rate. Many states do routine DNA testing for the most common mutations. Diagnosis is usually confirmed with a sweat chloride test (CF patients have more salty sweat). Since treatments are now available that are specific to some of the mutations responsible for the disorder, a DNA test to determine the gene defect is always indicated if this test is positive.

CF is caused by mutations in both alleles of the *CFTR* gene on chromosome 7, resulting in a lack of functional Cystic Fibrosis Transmembrane conductance Regulator protein (CFTR), an epithelial cell plasma membrane chloride channel that also regulates other ion channels in the same cell. Ion channels allow ions to flow across the membrane in the direction that is down a concentration or potential gradient.

The severity of CF symptoms varies with the type of mutation; the most common mutation results in misfolding and degradation; it causes complete absence of CFTR from the plasma membrane. Therefore, homozygotes for this mutation have severe symptoms. Other mutations that still retain some CFTR functionality lead to milder forms of CF.

A life-threatening clinical manifestation of CF is buildup of thick, viscous mucus in the airways (CF is sometimes called mucoviscidosis in Europe), causing partial or total occlusion and allowing bacteria to colonize. Lack of chloride secretion, due to absent or defective CFTR, results in retention of Cl^- and Na^+ ions, and therefore retention of water in the lung epithelial cells. This causes the airway surface fluid layer to dry up. The viscous mucus can no longer be effectively cleared by the beating cilia, and this facilitates colonization of the airway by opportunistic biofilm-forming pathogens, such as *Pseudomonas aeruginosa*, *Staphylococcus aureus* and *Burkholderia cepacia*. Even though the bacteria can usually be controlled with antibiotics, they are never completely eliminated, and recurrences of lung infections are frequent. The bacteria induce an inflammatory immune response,

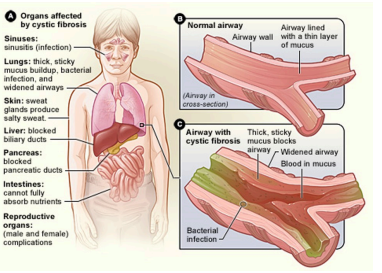
primarily of neutrophils, which produce reactive oxygen species to kill the bacteria. Consequently, collateral damage to the lung tissue accumulates over the years, resulting in inflexible scar tissue that diminishes breathing capacity. Ultimately, due to the inflexibility and mucus accumulation, the lung capacity becomes so diminished that a lung transplant becomes the only option. Donor lungs are in short supply, and the operation is risky (it is often a combined heart-lung transplant). It can result in dramatic and immediate improvement in breathing, but can also lead to rejection of the donor lung or to additional opportunistic infections because of the immunosuppressant drugs that are given to prevent rejection.

Similar osmotic effects cause the secretions of the pancreas and biliary system to become thick and viscous. As a result, digestive enzymes (proteases, lipase, amylase) and bicarbonate produced in the pancreas do not reach the duodenum. This results in maldigestion and nutrients being lost in the stool. In the liver, the same process blocks the bile duct and leads to a deficiency of bile salts, which are needed to solubilize fats in the lumen of the small bowel. This further worsens fat absorption, leading to fat in the stool and deficiency of fat soluble vitamins. Malnutrition ("failure to thrive" in infants) is a major problem in CF. Keeping CF patients properly nourished is important, but difficult. Besides having problems digesting and absorbing their calories, their caloric need is higher due to difficulty breathing. CF also results in decreased appetite which can lead to fewer calories taken in. When their nutritional needs are properly managed, CF patients do better in terms of lung function and overall health. In adults, accumulated damage to the pancreas can lead to lack of insulin or delayed insulin secretion from the pancreas; this results in CF-related diabetes.

CF affects other organs with epithelial cells, including the skin (salty sweat and reduced sweat volume) and reproductive organs (males are almost always infertile due to absence of the vas; females can also be infertile, but there are cases of well-treated women with CF successfully having children).

Standard CF treatments are designed to ameliorate the

symptoms, and have led to dramatic increases in life expectancy over the past decades. They don't cure the disorder, so it is important to consistently apply them to prevent long-term deterioration. Adherence can be a challenge for children and adolescents, because the daily treatments do not provide visible short-term results. These treatments include: 1) Chest compressions to loosen mucus; 2) Inhaling hypertonic saline (using a nebulizer) to rehydrate mucus; Pulmozyme (DNase) can be added to reduce the viscosity of mucus by cutting up DNA released from lysed cells, and bronchodilators (*e.g.* albuterol, a β2 adrenergic receptor agonist) are used to relax airway smooth muscle. 3) Inhaled and oral antibiotics (*e.g.* Tobramycin or azithromycin, which inhibit bacterial protein synthesis). 4) Pills containing pancreatic digestive enzymes (amylase, lipase, proteases) taken with meals. 5) High calorie diet.



CFTR Structure and Function:

CFTR contains a bundle of transmembrane helices that form a chloride-permeable channel and three cytoplasmic domains: two nucleotide binding domains (NBD), and a regulatory (R) domain. Opening the channel requires phosphorylation of the R domain at multiple sites by cAMP-dependent protein kinase (PKA), as well as ATP binding and hydrolysis by the NBDs. CFTR is related to the ATP-driven pump P-glycoprotein, which drives many chemotherapy

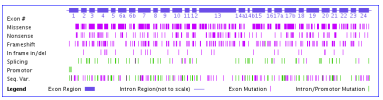
drugs out of cancer cells and leads to multi-drug resistance when overexpressed. However, CFTR is a *channel*, not a *pump*, which means it can only allow ions to flow down an electrochemical gradient.

Although the number of CFTR molecules in a cell membrane is small, they have a large effect on osmoregulation (it doesn't take a lot of open channels to let a lot of ions flow). CFTR also regulates another chloride channel and a sodium channel.

The most common mutation that causes CF is due to a deletion of 3 bases. It is called ΔF508, because the mutant protein is missing a Phe residue (F), amino acid #508 out of the 1480 in the sequence. It is in the NBD1 domain, but not involved in ATP binding.

CFTR Gene:

The *CFTR* gene contains 24 exons spanning almost 200,000 base pairs on chromosome 7. Since less than 4,500 bases are needed to code for the amino acids in the protein, most of the sequence is in the introns. More than 2,000 mutations have been discovered in the *CFTR* gene, although only about 200 of these have been clearly shown to cause CF. As seen in the diagram below, they are scattered all across the gene.



Missense mutations change a codon to specify a different amino acid, usually due to a single base change.

Nonsense mutations change a codon to a stop codon.

Frameshift mutations result when 1 or 2 bases are inserted or deleted in an exon (or any number not a multiple of 3). The remaining sequence downstream will code for the wrong amino acids and usually result in a truncated protein due to a stop codon in the other reading frame.

In frame insertions or deletions will result in one or more amino acids added or missing, but the rest of the sequence will be correct. The ΔF508 mutation is one of these.

Splicing mutations generally occur due to base changes at the exon/intron boundary and can result in an intron being retained

in the mRNA and translated (usually leading to an incorrect and truncated sequence) or result in an exon being skipped. This may also result in a frameshift in the next exon.

There are many possible ways a mutation can lead to CF, and there are examples for almost all of them. A mutation in the promoter may result in inadequate transcription. A frameshift or splice junction mutation can result in a protein that bears little resemblance to CFTR. Such dysfunctional polypeptides are typically rapidly degraded to prevent protein aggregation. CFTR with the ΔF508 mutation, even though it has only one missing residue out of 1480, causes the protein to misfold and then be degraded before it ever gets to the plasma membrane. Other mutant CFTR proteins that reach the plasma membrane may be dysfunctional, either because they don't open when they are supposed to (dysregulation) or don't conduct enough chloride ions when they are open. Some mutations result in partial function due to reduced synthesis, reduced proper splicing (activation of alternate splicing), or more rapid degradation (turnover).

The geographical distribution of CFTR mutations varies widely. In people of Northern European descent, 90% of the mutant alleles are ΔF508; in Ashkenazi Jews, 48% are W1282X. These are called "founder effects". In native populations of Sub-Saharan Africa and Asia, CF is relatively rare.

An understanding of the effects of a mutation can lead to drug therapies targeted to the effect of a particular mutation. Only 5% of normal CFTR function can help prevent lung and pancreatic disease symptoms, so even minor improvements in expression or function can have a big impact. Three examples that have led to new CF therapies are given below:

G551D:

This mutation has an Asp substituted for the Gly normally at position 551. It is found in 3-4% of CF patients, and results in a channel that does not open properly even though it is present in the membrane. The Cystic Fibrosis Foundation funded a project that first yielded a compound called a "potentiator", because it increases

chloride flow through G551D (as well as through normal CFTR and some other mutant forms). When it entered clinical trials it was named ivacaftor. A phase III study led from UW in 2011 showed significant improvements in G551D patients taking this drug in addition to standard therapy. Note that most of these patients have ΔF508 in their other CFTR allele, but only one of the genes needs to produce functional protein. When the FDA approved this as a new drug in 2012, it was given the brand name Kalydeco. In 2017, the FDA expanded the approval to treatment of 33 rare mutations, which doubled the number of patients for which Kalydeco was indicated.

ΔF508:

The most common mutation. Key observations were: 1) Individuals homozygous for the ΔF508 mutation transcribe and translate the gene, but have no CFTR in their plasma membranes. 2) When researchers expressed a human CFTR gene with the ΔF508 mutation in insect cells grown at 28C, the protein was found in the plasma membrane of these cells and it functioned as a chloride channel. These observations led to the hypothesis that the ΔF508 CFTR is absent in the plasma membranes of CF patients because it misfolds at 37C, but is stable enough to fold at 28C. Misfolded proteins are exported from the ER and degraded via the ubiquitin-proteasome system. This led to a search for small molecules (called "correctors") that could help stabilize ΔF508 CFTR a little and get it to fold properly. The first compound that did this successfully was lumacaftor. It did not help patients by itself, but in 2014 a Phase III trial (again led by UW) in combination with ivacaftor reported improvement in most ΔF508 homozygotes. The FDA approved this combination (called Orkambi) in 2015. It only improves lung function (FEV_1, the volume of air that can be exhaled in 1 min) by a few %, but every bit helps. Second generation correctors tezacaftor and elexacaftor in triple combinations with ivacaftor (called Trikafta) increased FEV_1 by 14% in clinical trials. Trikafta was approved in 2021 for use in patients with one or two ΔF508 alleles.

G542X, W1282X:

Nonsense mutations are present in about 5-8% of CF patients.

Ataluren was developed to cause read-through of stop codons due to nonsense mutations, of which G542X and W1282X are the most common in CF. A Phase III trial reported in 2012 showed modest improvement and slowed disease progression in CF patients with nonsense mutations. The drug was licensed in Europe, but is not approved in the US. Another Phase III trial in 2017 showed no significant improvement in FEV_1. The FDA also rejected it for use in muscular dystrophy that year. What problems might be caused by a drug that causes read-through?

Gene therapy for CF has been attempted by various means without success since the 1990's. One major obstacle has been targeting it to the airway stem cells that constantly regenerate the airway epithelia. In 2020 researchers demonstrated that they could take a patient's blood cells, reprogram them to generate induced pluripotent stem cells (iPSCs) and then coax them into becoming basal airway stem cells. This opens the possibility of correcting a CF mutation in these cells and returning them to the patient.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://uw.pressbooks.pub/fmrmolbio/?p=195#h5p-1

A few final comments:

Every genetic disorder is "rare", except in certain in-bred populations (*e.g.* the Faroe islands have the highest incidence for CFTR ΔF508); most disorders have incidences of 1 in 10,000 or less.

However, there are thousands of known genetic disorders, so the possibility that a patient has one of them should not be discounted.

Mutations occur almost everywhere and are found in everyone. We all have mutations in a number of important genes. Most are recessive, but we are all carriers of genetic disorders.

Missense mutations are the most common; they can be benign or severe depending on the substitution.

Nonsense mutations and frameshifts almost always result in translation of dysfunctional protein that could misfold and aggregate. Cells have mechanisms for degrading mRNAs with premature stop codons and degrading misfolded proteins to mitigate these effects.

Disease-causing mutations may be in introns or in promoter and enhancer regions thousands of bases from the coding sequences. Promoter and splicing mutations often result in reduced expression of normal protein, rather than complete absence.

If you sequence a gene (or a whole genome) and find variations from the "normal" DNA sequence of a gene, they may be harmless. Unless they have been found in other patients and characterized, you can't assume they cause a disorder.

Although methods have been developed to replace or repair defective genes in the laboratory, translating that to a cure can be extraordinarily difficult. Think about what cells you would need to fix in a CF patient and how to get the DNA into them.

Additional information about CF can be found at Dynamed (under "Top Resources" on the UW Health Sciences Library website), and at the Cystic Fibrosis Foundation.

10. DNA Replication and Repair

DNA Replication and Repair

Session Learning Objectives:

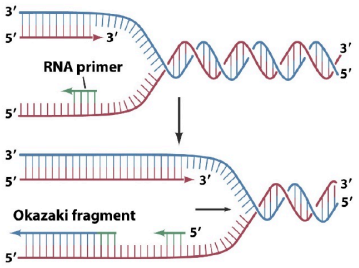
1. Illustrate DNA replication and identify proteins that are targets for inhibiting DNA replication.
2. Explain why telomere replication presents special problems and the disorders that could develop if defective, such as dyskeratosis congenita.
3. Describe the major sources of DNA damage and errors and the pathways used to recognize and correct these errors.
4. Analyze how defects in different DNA repair pathways lead to specific syndromes, including cancer-predisposition syndromes: Li-Fraumeni syndrome, Lynch syndrome, Xeroderma pigmentosum, Ataxia telangiectasia and hereditary breast and ovarian cancer (HBOC) syndromes.
5. Describe how DNA repeat expansion relates to the presence and severity of specific disorders: Fragile X syndrome/Fragile X-associated tremor/ataxia syndrome (FXTAS), Huntington disorder, myotonic dystrophy.
6. Describe how repeated DNA sequences and homologous recombination contribute to the appearance of interstitial deletion syndromes.

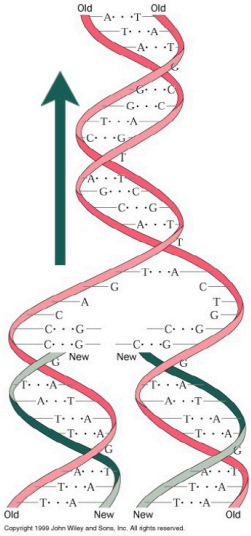
SLO 1. Illustrate DNA replication and identify proteins that are targets for inhibiting DNA replication.

DNA replication

DNA replication is semiconservative: Each daughter strand contains one of the original parent strands and one new strand.

The template strands are antiparallel, and **polymerase can only synthesize DNA in one direction (5′→ 3′).** The fork moves in one direction (i.e. direction of DNA unwinding; green arrow), while replication of each strand is in the opposite direction—yet still simultaneous. The strand that is being synthesized in the same direction as fork movement is the **leading strand**, and it is synthesized continuously as the fork unwinds. The **lagging strand** is synthesized as a series of short **Okazaki fragments**. Each Okazaki fragment requires a new primer. **Primase** is the complex of enzymes that makes these primers.



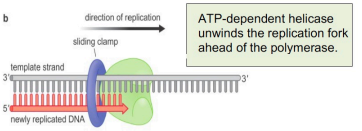


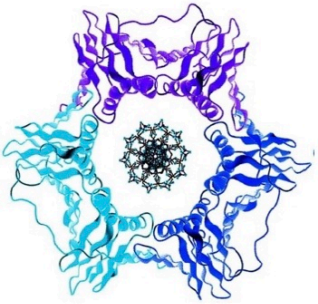
DNA Polymerases require an initiating primer made from either DNA or RNA. RNA polymerases do not. **Primase** is a special DNA-dependent RNA polymerase that synthesizes primers for the leading strand. A DNA polymerase extends these primers, then the main replicative DNA polymerases synthesize the leading and lagging strands. During ongoing lagging strand synthesis, the primers are removed, any gaps between the Okazaki fragments are **filled in by DNA polymerase**, and finally nicks are sealed by **DNA ligase.**

Other essential components of replication:

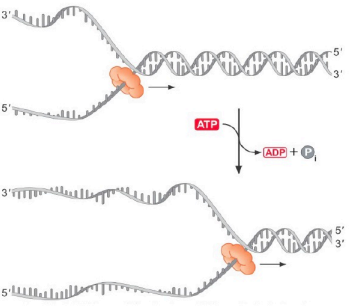
The **sliding clamp** is a processivity factor (keeps polymerase complex from falling off the DNA). The human sliding clamp is **PCNA**

(Proliferating Cell Nuclear Antigen), which is often used in histology as a marker of DNA synthesis and cell proliferation. It is a trimeric complex assembled around the DNA by an ATP-dependent clamp loader on the leading strand and at each new Okazaki fragment.





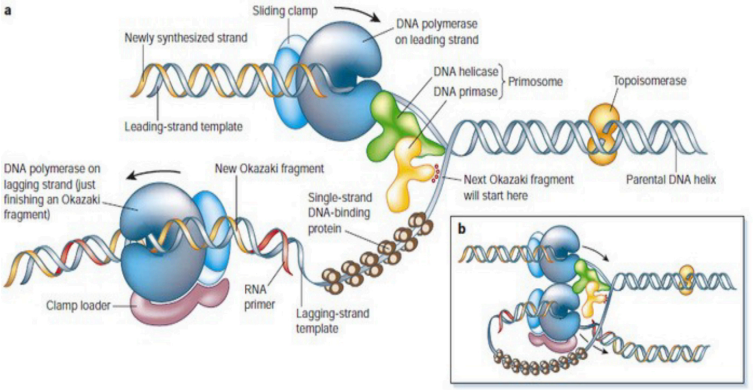
PCNA sliding clamp structure



Another marker for cell proliferation is Ki-67 (MKI67) (not shown here and not to be confused with PCNA). Ki-67 is associated with the transcription of ribosomal RNA in the nucleus and is present throughout an active

cell cycle. Ki-67 is increased during the S phase where DNA is being synthesized.

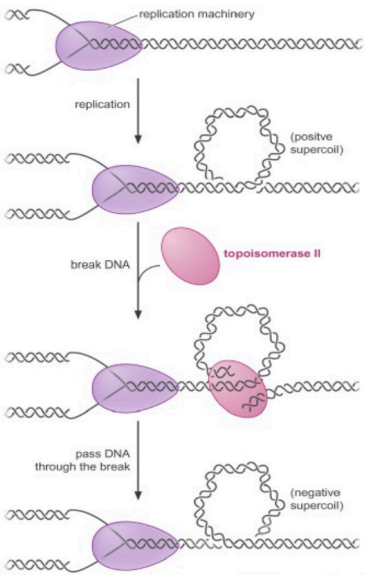
All of these key activities are shown in the figure below showing proteins at the replication fork; the inset shows how the leading and lagging strands are coordinated. As each Okazaki fragment is synthesized, the loop grows until the polymerase encounters the 5' end of the prior Okazaki fragment. Then the loop is released, a new clamp is loaded, and the next fragment begins.



SLO 2. Explain why telomere replication presents special problems and the disorders that could develop if defective, such as dyskeratosis congenita.

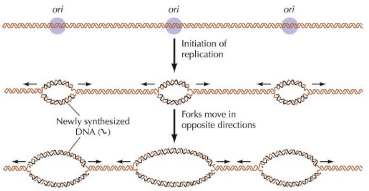
Topoisomerases relieve the superhelical tension (positive supercoiling) that occurs ahead of a moving replication fork, by breaking the DNA to relieve the torsion, then resealing the break.

Topoisomerase inhibitors are powerful anticancer drugs and antibiotics. For example, camptothecin-derived compounds (like Topotecan, used for ovarian & lung cancer) block replication by converting the topoisomerase reaction into a dead-end reaction with the topoisomerase covalently linked to broken DNA. **This results in genome fragmentation and cell death.**



Telomerase dysfunction results in premature death of tissues that rely heavily on stem cell divisions. Mutations in TERT, TERC and several other genes cause the genetic disorder **Dyskeratosis congenita** (DC), which results in irreversible degeneration of skin tissue, early hair graying and loss and bone marrow failure, among other clinical presentations.

Chromosomes are replicated by many polymerase complexes, operating in parallel. Each chromosome has many replication origins. At each origin a replication bubble opens by melting the DNA duplex. Within each bubble, two replication forks move away from the origin, each with a leading and lagging strand. Replication terminates when forks collide or reach the end of the chromosome. This creates two problems.

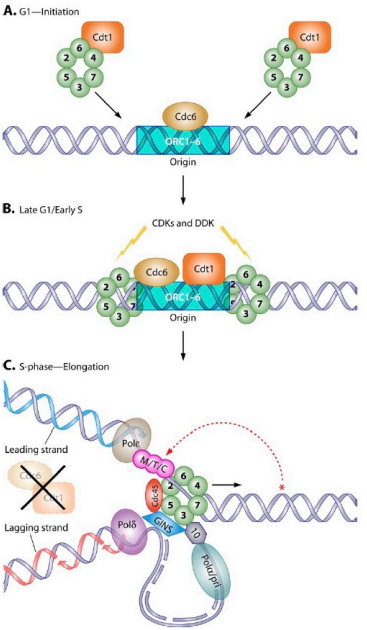


Problem one: How, with so many origins (that fire at different times during S phase), can you ensure any given origin initiates once and only once per cell division cycle?

The solution: Origin licensing.

ORC (Origin Recognition Complex) binds to origins throughout the cell cycle, marking each initiation site. ORC then brings in additional proteins that load the **MCM complex** (Panel A and B in figure). MCM is the replication helicase, and similar to the sliding clamp, encircles DNA and can't fall off.

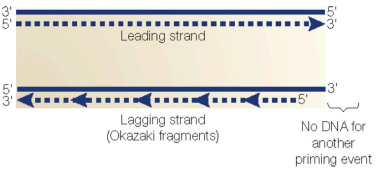
When the signal arrives to initiate replication (a cascade of phosphorylation events), the MCM helicase is activated to fire the origin, and **MCM loading proteins (Cdc6/Cdt1)** are destroyed to prevent re-licensing! (Panel C in figure). They will not return until the G1 phase of the cell cycle.



Problem 2: The "end replication problem." At each end of the chromosome, replication results in a 3' overhang because after removal of the endmost RNA primer on the lagging strand, there is no upstream sequence to

prime the DNA polymerase filling in. Thus, chromosomes shorten with every replication cycle!

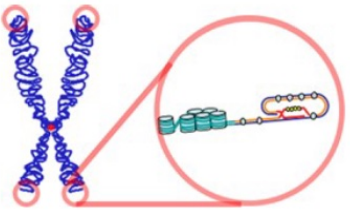
The solution: Telomeres and telomerase.



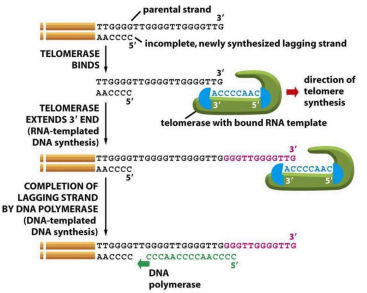
Telomerase is made of two components:

TERT, an RNA-dependent DNA polymerase (aka, reverse transcriptase)

TERC (Telomerase RNA component or hTR), an RNA molecule that provides the template for the telomeric repeat sequence.



Telomerase is active during embryogenesis and in stem cells, but not in more differentiated cells. Thus, telomeres shorten with each cell division in these cells. (Cancer cells frequently reactivate telomerase to achieve replicative immortality.) Telomerase synthesis involves binding to the 3' end of the G-rich telomeric parental strand and aligning with the complementary **TERC** RNA template. **TERT** then adds 6 deoxynucleotides using the RNA template and translocates to the new 3' end of the DNA to repeat the process.



An interactive H5P element has been excluded from this

version of the text. You can view it online here:
https://uw.pressbooks.pub/fmrmolbio/?p=461#h5p-3

SLO 3. Describe the major sources of DNA damage and errors and the pathways used to recognize and correct these errors.

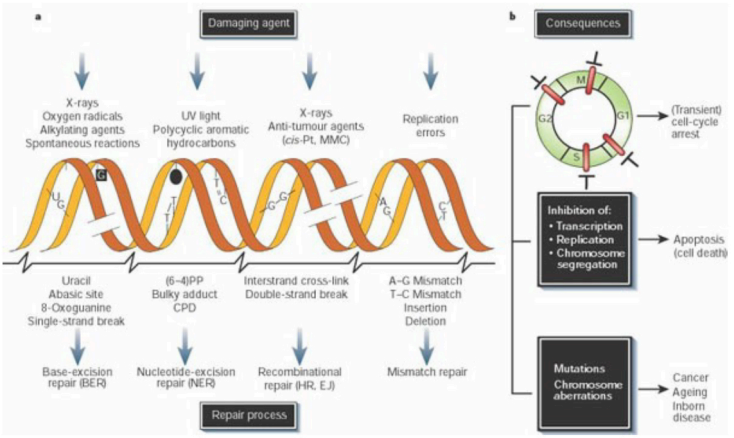
SLO4. Analyze how defects in different DNA repair pathways lead to specific syndromes, including cancer-predisposition syndromes: Li-Fraumeni syndrome, Lynch syndrome, Xeroderma pigmentosum, Ataxia telangiectasia and hereditary breast and ovarian cancer (HBOC) syndromes.

DNA repair

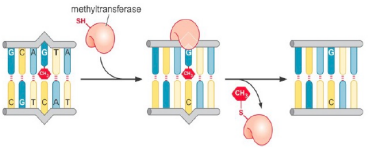
Cells are subjected to many sources of DNA damage, both extrinsic (such as UV or ionizing radiation) and intrinsic (such as errors of replication, endogenous reactive oxygen species, and spontaneous lesions). Different sources of damage create different types of DNA lesions, which in turn are recognized and repaired by specialized repair pathways. DNA damage triggers cell-cycle arrest (checkpoint activation), which will be maintained until either the damage is repaired or the persistence of damage triggers apoptosis. Unrepaired damage can result in fixation of a mutation, from single base pair changes to major chromosomal rearrangements.

DNA Repair pathways

DNA repair pathways fall into several broad categories: direct reversal, in which damage is directly reversed in situ with no cleavage of the DNA backbone; excision repair pathways (BER, NER, MMR), in which a single strand of DNA containing damage is excised (cleaved 5' and 3' of the damage and removed) and replaced by DNA polymerases using the undamaged complement as a template; and double strand break repair (DSBR) pathways that contend with damage resulting in breakage of both strands of DNA.



Only one type of direct reversal exists in human cells: direct removal of alkylation damage. Base alkylation, most commonly methylation or ethylation, usually occurs due to exposure to alkylating drugs such as those used in chemotherapy.

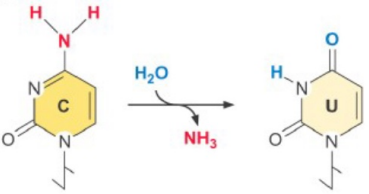


Repair methyltransferases transfer the alkyl group directly to a cysteine residue in the repair protein. This is a suicide repair mechanism, as the alkyl group is then permanently covalently attached to the protein.

Base excision repair (BER)

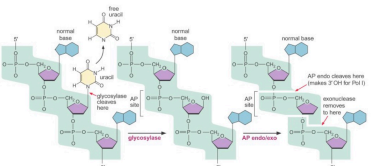
One common source of DNA damage in the normal cellular environment is spontaneous deamination of cytosine—it occurs at ~100 C's/cell/day! Cytosine deamination creates uracil, easily recognized as an error. However, deamination of 5-methyl cytosine yields thymine and causes a T-G mispairing. [This may be part of the reason CpGs are relatively rare outside of CpG islands—it is harder to repair accurately. To compensate, repair of T-G mispairs outside of replication are biased toward replacing the T.]

Spontaneous depurination (loss of the adenine or guanine base from the deoxyribose sugar) is even more frequent, at ~5000 events/cell/day.



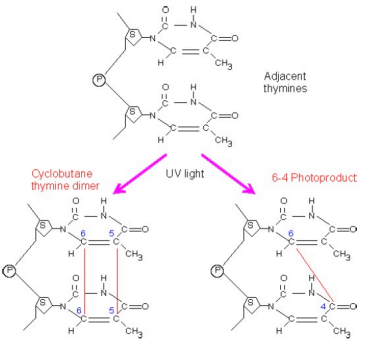
Both of these types of damage, as well as some other small base lesions, are repaired via base excision repair (BER). First, a damaged base is released from the deoxyribose by the action of a **glycosylase** enzyme, creating an abasic site (not necessary for spontaneous depurination). Then, an **AP (apurinic/apyrimidinic) endo/exonuclease** cleaves the phosphate backbone on either side of the abasic site. A **DNA polymerase** inserts the correct base(s) and **ligase** seals the final nick.

Nucleotide excision repair (NER)

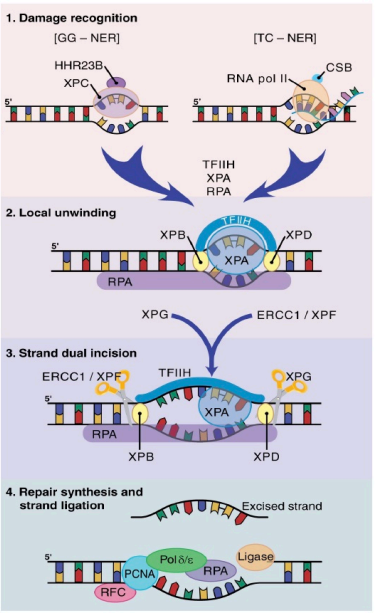


The NER pathway primarily recognizes so-called "bulky lesions," lesions that create significant distortions in the structure of the DNA double helix. One of the most important of these is damage resulting from UV light exposure, which causes dimerization of adjacent pyrimidines in one strand. This dimerization can involve one bond (forming a 6- 4 photoproduct, 6,4-PP) or two bonds to create a cyclobutane pyrimidine dimer, CPD.

NER is subdivided into two pathways based on the initial recognition of damage. In **global genome (GG)-NER**, a helix-distorting lesion is detected anywhere in the genome by damage surveillance proteins such as XPC. **In transcription-coupled (TC)-NER**, damage is recognized when it causes an elongating RNA polymerase II to stall.



As a result, this TC-NER pathway exclusively repairs damage to the transcribed template strand of active genes. Following damage recognition, the two sub-pathways merge. The damaged region is unwound by the XPB/XPD helicases, which are transcriptional helicases already present in the elongating RNA polymerase complexes; these helicases are recruited subsequent to damage recognition in GG-NER. XPF/ERCC1 and XPG are endonucleases that cleave the damaged strand to release ~30nt of DNA containing the damage. Replication machinery fills the gap.



"XP" stands for xeroderma pigmentosum, a disease characterized by severe photosensitivity and high rates of childhood malignancies, particularly malignant melanoma, squamous and

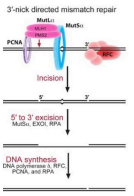
basal cell carcinomas, and leukemias.

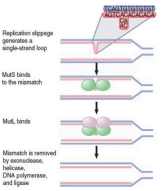
Most patients with XP have a defect in both GG- and TC-NER, and thus are unable to repair various bulky lesions. The failure to repair UV photolesions is particular evident in the high rates of skin damage and cancers. Interestingly, different mutations in the XPB/XPD helicases can cause XP, Cockayne syndrome, or trichothiodystrophy (TTD), depending on the particular mutation. Cockayne syndrome patients exhibit cognitive impairment, progressive neurological dysfunction, and microcephaly, and appear to have defects in TC-NER but not GG-NER. In TTD, repair is normal but transcription of specific genes is impacted; patients have sulfur-deficient brittle hair, and may also show cognitive impairment and reduced stature. Although the world-wide prevalence of XP is about 1 in a million, some populations have a much higher prevalence of about 1/30,000, possibly due to a genetic founder effect and genetic bottle-necking in isolated populations.

Mismatch repair (MMR)

Replication errors occur at ~1 base change/cell division missed by DNA polymerase's proofreading capacity (without proofreading the error rate would be more like 100 bases/cell division). The mismatch repair system operates in conjunction with replication

to repair remaining errors. The human MutS (hMutS) complexes recognize the mismatches. hMutS-alpha (a dimer of MSH2 and MSH6) recognizes single mismatches, while human hMutS-beta (MSH2-MSH3) recognizes insertion/deletion loops resulting from replication slippage. The hMutL-alpha complex (MLH1 and PMS2) links hMutS to the replication machinery and can nick the DNA, although usually a preexisting nick between Okazaki fragments is used as an excision point. Replication machinery fills in the remaining gap after excision.

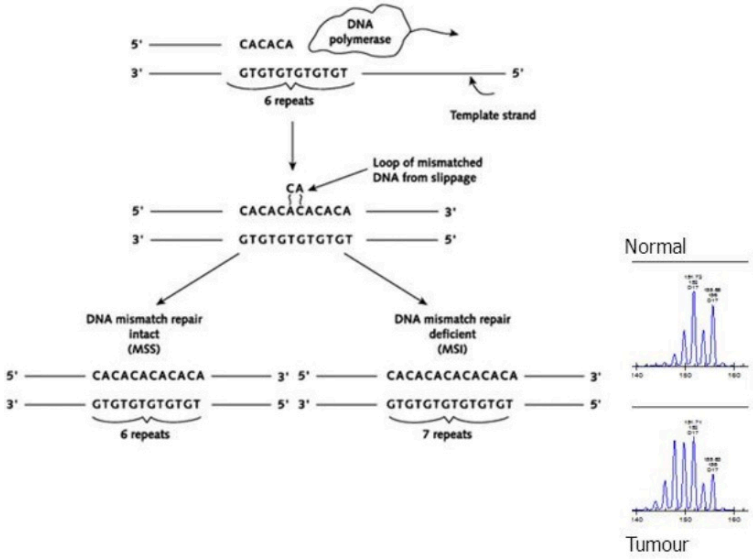




Inherited mutations in MMR machinery (most often in MSH2 and MLH1) result in **Lynch syndrome** (previously known as hereditary nonpolyposis colorectal cancer, HNPCC), a genetic syndrome with a dominant pattern of inheritance that results in a high risk of colon cancer, endometrial cancer, and several others. A hallmark of Lynch syndrome is microsatellite instability (MSI), caused by the failure to correct replication slippage at repeated sequences, resulting in changes in repeat numbers (increases or decreases) in tumor cells. Clinical testing of tumor tissues may include immunohistochemistry and PCR-based microsatellite instability analysis. However, a definitive diagnosis of Lynch syndrome requires identification of a mutation in one of the four human mismatch repair genes, **MLH1, PMS2, MSH2 and MSH6**, or potentially an inactivating methylation of one of these genes.

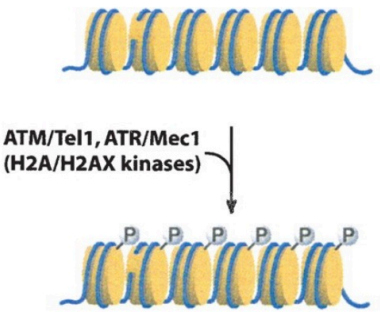
A newer use of identifying defects in MMR genes is in evaluating tumor mutational burden (TMB) (number of somatic mutations per megabase.) The efficacy of treatment with immune checkpoint inhibitors for multiple types of cancers may be estimated by evaluating the TMB, which will be elevated in tumors harboring defects in mismatch repair leading to microsatellite instability or defects in other DNA repair genes. Current research is directed at

evaluating the predictive value of high TMB on increased long-term survival of patients with different tumor types.



Double strand break repair (DSBR)

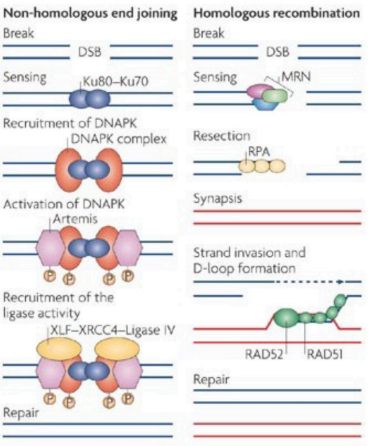
One of the most serious categories of DNA damage is DNA double-strand breaks (DSBR). A single break will arrest the cell cycle until repair is completed or the cell undergoes apoptosis. DSBR is divided into two subcategories: **non-homologous end joining (NHEJ) and homologous recombination (HR).** Both pathways involve immediate and extensive phosphorylation of the histone variant H2AX at the site of damage (and over megabases surrounding the damage site) by the checkpoint kinases ATM/ATR. This phosphorylation is involved in subsequent interactions with repair factors.



NHEJ is a rapid repair pathway that directly ligates broken DNA ends, sensed and bridged by the Ku8-Ku70 complex. NHEJ can lead to loss of sequence at the site of joining. The proteins involved in NHEJ are the same as those involved in somatic recombination in the immune system.

In contrast, **HR** (which has many subtypes) is a highly accurate repair system that uses homologous sequence (usually a sister chromatid) as a template for accurate repair of a damage site. Repair by HR counterintuitively begins with 5' resection of the broken ends to generate long 3' overhangs. These overhangs are then used to invade the homologous donor sequence, and DNA polymerase then extends from the 3' end using the homologous donor as a template.

Like NHEJ, HR is a repair system that is also used for another cellular process: **meiotic recombination during gametogenesis.** During meiosis, DSBs are deliberately induced, and then HR machinery produces meiotic recombinants. Errors can occur during this normal meiotic recombination process as well, which can lead to significant chromosomal aberrations, such



as interstitial deletions causing microdeletion syndromes, discussed further later in this chapter.

SLO 5. Describe how DNA repeat expansion relates to the presence and severity of specific disorders: Fragile X syndrome/ Fragile X-associated tremor/ataxia syndrome (FXTAS), Huntington disorder, myotonic dystrophy.

Trinucleotide Repeat Disorders

Some genetic disorders arise from a progressive expansion of a region of DNA containing multiple 3-nucleotide repeats (triplet repeats). The phenomenon of potential trinucleotide repeat expansion, from one generation to the next, is refered to as "genetic anticipation." In these disorders, healthy individuals have a variable number of repeats below a particular threshold; repeats beyond that threshold number present with a disorder, with the severity of the disorder typically correlated with the number of repeats (but there are exceptions). Repeat expansions are caused by strand slippage as discussed above in the context of mismatch repair. Individuals with repeats in the normal range are not at any increased risk of their offspring developing disease, but individuals in the premutation range have repeats approaching the threshold which are unstable and more likely to expand (or rarely, contract) in the next generation. Remarkably, the sex of the parent can affect the risk of repeat expansion.

Fragile X syndrome (FXS) is the second most common inherited form of intellectual and developmental disability (after Down syndrome). Fragile X syndrome is visible cytologically in affected patients as a constricted site in the X chromosome that appears "fragile" but is not (located at Xq27.3). The repeated sequence is a CGG triplet in a non-coding 5' untranslated (5'UTR) region of the FMR1 (Fragile X Messenger Ribonucleoprotein 1) gene. The repeats may interfere with gene function by providing a substrate for CpG methylation and silencing of FMR1 gene transcription or may involve other mechanisms such as hybridization of the CGG region in the

mRNA to the complementary DNA to form an RNA-DNA hybrid mediating epigenetic gene silencing. The FMRP protein encoded by this gene is normally involved in translation and trafficking of mRNAs in neurons and plays a role in learning and memory.

Disease severity generally correlates with CGG trinucleotide repeat length and ranges from 55-200 repeats for the premutation alleles and >200 to several hundred to several thousand repeats for the full mutation, as compared with 5-44 CGG repeats in a normal allele. Because the disorder is X-linked, more males are affected than females, however females heterozygous for a full mutation are also at risk for intellectual disability. X-inactivation in the female will play a role in presentation of the disorder, depending on which X is inactivated in which tissues. In addition to mild to moderate intellectual disability, males typically present with characteristic facies of a long narrow face, large ears and macroorchidism (large testes). Males with premutations may develop an adult-onset related neurodegenerative disorder called Fragile X tremor-ataxia syndrome (FXTAS).

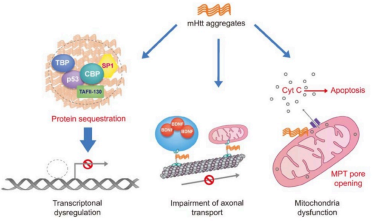
An allele in the premutation range may be stable for generations or increase in size when inherited from the mother (or rarely regress). However, sperm carry only premutation alleles, even if a male has a full mutation. Consequently, males with a premutation or full mutation pass the premutation to all of their female children (and of course not to their male children, who do not inherit the father's X chromosome.) Furthermore, in father-to-daughter transmission, the size of a premutation contracts in about one third of female offspring. The mechanism underlying these observations has not been established.

Myotonic dystrophy (aka dystrophia myotonica, DM) is one of the most common inherited forms of progressive muscle disease. Age of onset decreases and severity increases with larger repeat sizes. The image to the right shows three generations of individuals with DM: the infant has >1000 repeats, and the mother and grandmother have ~100.

There are two types of autosomal dominant DM (1 and 2), both caused by repeat expansions in the transcribed portion of the genes. DM Type 1 is caused by the expansion of the CTG triplet repeat in the 3'-UTR of the DMPK gene, encoding a protein kinase that is important in muscle, heart and brain, though its precise role is not clear. Type 2 is milder and is caused by a tetranucleotide repeat expansion in the first intron of the CNBP gene. In both types of DM, the pathology appears to be primarily due to the excessively long mRNAs produced, which form inclusions that interfere with cellular translational machinery and/or other functions.

Huntington Disorder (Huntington Disease) (HD) is an autosomal dominant progressive disorder causing cognitive, motor, and behavioral changes. The mean age of onset is 35-44 years (with earlier onsets with increased repeat lengths) with a survival time

of 15-18 years from onset. The earliest manifestations are usually subtle motor defects including involuntary movements (chorea) and irritability. Cognitive decline ultimately leads to dementia.



The repeated sequence is a CAG codon for glutamine in the first exon coding region of the huntingtin (HTT) gene. The resulting Huntingtin protein contains a polyglutamine (polyQ) tract. Huntingtin is a poorly understood, ubiquitously expressed protein (highest expression in brain and testes) with multiple cellular functions. It is not clear which loss of function might contribute to HD. However, one key feature of the polyQ tracts is that turnover of mutant huntingtin (mHTT) leads to accumulation of undegraded polyQ fragments, which, because of their polar nature, tend to form aggregates (with mHTT and other proteins), leading to large inclusions in neurons. One of the proteins that becomes trapped in these inclusions is histone acetyltransferase CBP (which itself has an 18-Q tract), which is a coactivator for many genes. This trapping of CBP has been implicated in the neural pathology of HD, however there are other poly Q proteins, such as transcription factors, that could be affected and other mechanisms operating in the pathogenicity of Huntington disorder.

Disorder	Normal range	Premutation range	Affected range	Repeat	Impact of repeat expansion
Fragile X	6-40	61-200	>200	**CGG** in 5'-UTR of *FMR1* gene	CpG methylation and heterochromatin formation trigger silencing of *FMR1* gene
Myotonic dystrophy	5-37	38-49	50-150 (mild) 100-1000 (classical) 1000-5000+ (congenital)	**CTG** in 3' UTR of *DMPK* gene; **CCTG** (type 2 DM) in first intron of *CNBP* gene	Long mRNAs bind to RNA binding proteins and form aggregates
Huntington Disorder	10-26	27-41	36-121	**CAG** in first exon of coding region of *HTT* gene	Encodes a polyglutamine tract in Huntingtin protein (HTT); CAG expansion beyond the normal length creates non-degraded polyQ fragments and "sticky" aggregates

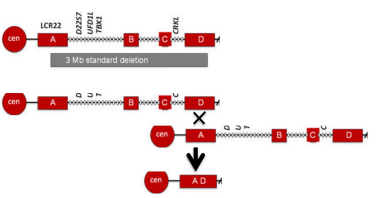
SLO 6. Describe how repeated DNA sequences and homologous recombination contribute to the appearance of interstitial deletion syndromes.

Microdeletion syndromes

The introduction of double strand breaks (DSB) are part of the normal process of meiotic recombination during gamete development. The homologous recombination repair system is used in producing the meiotic recombinants. Homologous chromosomes may align out of register in areas where there are blocks of repeated sequences. If a crossover occurs when the wrong blocks pair, one of the resulting recombinant chromosomes will have a deletion between the repeats, and the other chromosome will carry a duplication. As this process is occurring in gametogenesis, the result will be a germline mutation affecting the next generation.

Small deletions (microdeletions) may be **terminal** (end of chromosome is deleted) or **interstitial** (internal region of chromosome is deleted). Frequently the deleted area is in between segmental duplications, which are multiple copies of tandemly repeated sequences at the breakpoints for the deleted region. This process is considered a nonallelic homologous recombination because it involves recombination between different genetic areas that may be similar but nonhomologous, as illustrated in the generic figure below.

To relate these concepts to the most common microdeletion syndrome, the segmental duplications involved in 22q11.2DS (DiGeorge syndrome) are illustrated below, showing only the final deletion product after the nonhomologous recombination. Regions A and D have high sequence identity but are nonhomologous.



Microdeletion syndromes are designated by the chromosomal region affected or by their eponymous name. The "q" followed by a number indicates the chromosomal region affected on the long arm, and the "p" refers to the short arm.

Some of the disorders that have been studied more extensively are listed below, but epidemiological studies have included several other microdeletion syndromes not described here. Because these disorders occur at a greater frequency than one might expect, the estimates reported by the National Organization for Rare Disorders (NORD) for the incidences in live births, however inaccurate because of misdiagnoses and underreporting, are included below to highlight the more common microdeletion disorders and those represented in USMLE First Aid book or in lecture.

- **22q11.2: DiGeorge syndrome (Velocardiofacial syndrome)**
    - o Common interstitial microdeletion syndrome presenting with many different congenital anomalies
    - o Interstitial microdeletion syndrome with incidence of ~1/3000-1/6000 live births

- **7q11.23: Williams-Beuren syndrome (Williams syndrome)**
    - o Interstitial microdeletion syndrome with incidence of ~1/7500 live births
    - o Associated with supravalvular aortic stenosis and renal artery stenosis
    - o 7q micro deletion involves 27 genes, including elastin gene, may be included in diagnostic tests for haploinsufficiency
    - o Sometimes refered to as “Happy syndrome” because of the highly social personality; some have hypothesized a connection between syndrome and folklore tales of people with elfin qualities and magical powers
    - o Armellino Center of Excellence established at University of Pennsylvania in June 2022, launched by Dr. Jocelyn Krebs (WWAMI-AK), former president of the Board of Trustees for the Williams Syndrome Association and Williams syndrome researcher.

- **15q11.2-q13: Prader Willi (PWS) and Angelman (AS) syndromes**

- o Interstitial microdeletion syndrome with incidence of ~1/10,000-30,000 for PWS and ~1/12,000-20,000 for AS live births
- o Low copy tandemly repeated sequences are present at the common breakpoints (BP1, BP2, BP3) flanking the deletion regions (rare deletions use additional deletion breakpoints)
- o Presentation of disorder from interstitial deletions is affected by parent of origin imprinting discussed in other sessions of the WWAMI foundations curriculum

- **5p deletion: Cri du chat syndrome**
  - o Incidence of ~ 1/15,000 – 50,000 live births
  - o Terminal or interstitial deletions with many different break points
  - o No known correlation between deletion size and severity of disorder

- **4p deletion Wolf-Hirschhorn syndrome** (terminal microdeletion)
  - o Incidence of ~ 1/50,000 live births
  - o Deletion of most terminal portion of 4p (about ½ of short arm)

11. Cell Cycle

PAM LANGER

Cell Cycle

Session Learning Objectives

SLO 1: Summarize the cell cycle and the events that occur in Go, G1, S, M, G2, and M phases.

SLO 2: Describe multiple regulators of the cell cycle, including cyclin, cyclin dependent kinase (CDK), and CDK inhibitors.

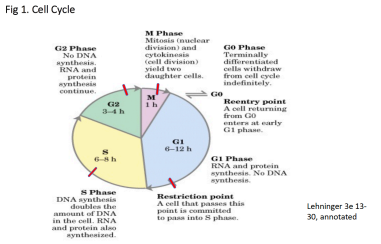
SLO 3: Describe the roles of the retinoblastoma protein (Rb) and the transcription factor p53 in cell cycle regulation and the cancers associated with defects in these genes.

SLO 1: Summarize the cell cycle and the events that occur in Go, G1, S, M, G2, and M phases.

Cell Cycle Overview

Regulation of the cell cycle is crucial for normal development and proper maintenance of tissues. Loss of this regulation is one of the fundamental defects leading to the onset of cancer. Either overstimulation of cell cycle progression or avoidance of cell death via apoptosis can lead to uninhibited cell growth and propagation. Cell signaling pathways regulate molecular choices that tip the fates of cells. For an interactive review, visit Howard Hughes Medical Institute (HHMI) BioInteractive module "Cell Cycle Regulators and Cancer" at: https://www.hhmi.org/biointeractive/eukaryotic-cell-cycle-and-cancer.

Fig. 1. Cell Cycle (Lehninger ed. 3, Fig. 13-30)



- Length of cell cycle phases varies among different cells and conditions
- Interphase = G1 + S + G2

phases
- Cell may enter into a G0 phase if it is terminally differentiated or in the absence of growth signals
- M phase is divided into prophase, metaphase, anaphase, telophase, cytokinesis (not shown in figure)
- Four main regulatory checkpoints are indicated as bars crossing the circle

SLO 2: Describe multiple regulators of the cell cycle, including cyclin, cyclin dependent kinase (CDK), and CDK inhibitors.

Molecular processes in regulation of the cell cycle

A network of regulatory proteins mediates an ordered series of switches that promote or inhibit cell cycle progression. It is an evolutionarily conserved system that differs in the details among organisms. Regulation relies heavily on protein phosphorylation and dephosphorylation as well as timely synthesis of proteins and degradation via the ubiquitin/proteosome system (UPS). Cell proliferation is stimulated by various signal transduction pathways that promote the progression of the cell cycle or inhibit processes that would lead to apoptosis. There are many layers of details understood with respect to the cell cycle, however we are focusing here on the larger picture of regulatory events and defects related to promoting cancer.

Cell cycle checkpoints

The most critical checkpoint in the cell cycle is the G1, G1/S or "Restriction point." Progression through the G1/S checkpoint requires stimulation, otherwise the cell will remain in G0 (G zero), a state that does not progress to the S phase. In order to pass the restriction or R point, the cellular environment must be favorable for DNA synthesis to begin. That is, growth factors or other external or internal signals should be present to stimulate progression of the cycle, and the cell should be the appropriate size and contain proteins and other components required for DNA synthesis. If

cellular DNA is damaged, the cell cycle will pause until it is repaired or undergo apoptosis if the damage is severe. Once the G1/S checkpoint is passed, the cell is committed to synthesizing DNA and preparing for cell division. However, the cell is not committed to divide at this point because later checkpoints can also stop cell cycle progression.

At the S checkpoint, DNA damage or DNA replication errors will also cause the cycle to pause here until the DNA is repaired. The protein kinase ataxia telangiectasia mutated (ATM) protein is involved in halting the cell cycle in S phase, and defects in this protein are associated with increased risk of certain cancers. The process may also involve the breast cancer antigen 1 (BRCA1) protein, a tumor suppressor protein that is defective in some breast cancers.

At the G2 checkpoint, the DNA must be finished replicating, and if there is DNA damage, the cycle will once again pause until the damage is repaired or undergo apoptosis if the damage is too great.

The M checkpoint is at the metaphase to anaphase transition within the M (Mitosis) phase of the cell cycle. It is also called the "Spindle checkpoint" because sister chromatids must be attached to microtubule spindle fibers from opposite poles of the cell in preparation for chromatid separation. Anaphase-promoting complex/cyclosome (APC/C) is a ubiquitin protein ligase that promotes the degradation of proteins holding chromatids together at the centromere in order to allow chromatid separation during cell division. To avoid confusion, it is important to note that the anaphase-promoting complex (APC/C) is a cell cycle regulatory complex, and a similar acronym is attributed to a different protein, namely the adenomatous polyposis coli (APC) protein that is encoded by a tumor suppressor gene that is often found mutated in colorectal cancer.

Cyclin-CDK regulation

A family of proteins called cyclins, and a family of protein kinases called cyclin-dependent protein kinases (CDKs), regulate progression through the cell cycle. Each CDK must be associated

with a cyclin in order to be active. CDK activity is further regulated by phosphorylation. Signal transduction pathways, stimulated by external growth factors or other stimuli, or intracellular events, will lead to expression of different cyclin genes. Periodic synthesis of different cyclins regulates activity of their CDK partner, and degradation of cyclins via the ubiquitin proteasome system is an essential part of cell cycle progression. The activities of cyclin/CDK complexes are also regulated by specific CDK inhibitors (CKIs).

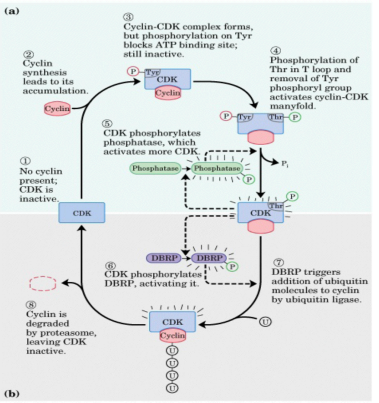


Fig. 2. Cyclin-CDK regulation summary

Figure is provided to illustrate cyclic nature of the changes in cyclin/CDK activity and is not meant to be memorized.

Signaling pathways and the cell cycle

Many different signaling pathways affect cell cycle regulation. For example, activation of a receptor tyrosine kinase (RTK) can stimulate a PI3 kinase or Ras/MAPK "pro-survival" pathway. Apoptosis of a cell is inhibited when it is receiving pro-survival growth signals. A defect in the normal apoptotic response could result in unregulated cell growth and division. A signaling pathway map in frequent use is provided here to illustrate this concept and is not meant to be memorized! However, you should note that

stimulation of RTK receptors leads to multiple pathways and that this map includes pro-survival as well as pro-apototic pathways (Fig. 2).

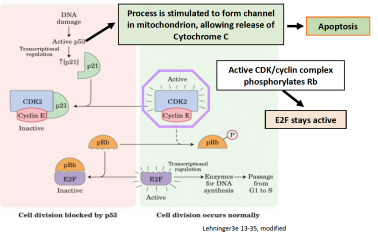
Defects in signaling pathways can lead to unregulated cell growth and cancer. For example, defects in Ras genes can cause overstimulation of a pro-survival signal and are one of the most common oncogenes in cancers. Alternatively, a defect in a GTPase activating protein (GAP), promoting hydrolysis of the GTP that activates the Ras protein, can cause constitutive activation of the Ras/MAPK pathway if the GTP stays bound to Ras and is not hydrolyzed to GDP. One such GAP protein is the NF1 (neurofibromin 1) protein which negatively regulates the Ras/MAPK pathway normally but is most commonly associated with development of neurofibromatosis type 1 (NF1 or von Recklinghausen syndrome) when both alleles for the NF1 gene are defective in a cell. Mutations in the NF1 gene are also associated with other disorders, including some breast cancers. (Mutations in NF1, a tumor suppressor gene, are recessive, but the inheritance pattern is considered to be autosomal dominant because of the high likelihood of a second acquired NF1 mutation in cells that start out having a single NF1 mutation.)

SLO 3: Describe the roles of the retinoblastoma protein (Rb) and the transcription factor p53 in cell cycle regulation and the cancers associated with defects in these genes. Retinoblastoma (Rb) and p53 proteins

Rb (pRb protein) and p53 (P53, TP53 proteins) play critical roles in cell cycle regulation. Rb is a type of brake on cell cycle progression. When Rb is phosphorylated, the brake is released and the cell proceeds to synthesize components required for DNA synthesis. P53 plays a dual role in cell cycle regulation in that it stimulates a process that will cause the cell cycle to pause until DNA damage is repaired or stimulates an apoptotic response if the damage to DNA is extensive. P53 and Rb are both tumor suppressor proteins, and some regulatory mechanisms mediated by P53 affect Rb activity.

In brief, the increase in p53 activity after DNA damage causes a decrease in cyclin/CDK activity which results in maintenance of the Rb brake on the cell cycle. While the cell cycle is paused, the DNA damage may be repaired, thus reducing the chance of propagating errors in DNA to the next cell generation. A summary of these pathways is shown below (Fig.3.)

Fig. 3 Cyclin, CDK, pRb, E2F, p53, p21 in Cell Cycle Regulation



Hereditary Retinoblastoma

Hereditary retinoblastoma is an inherited form of eye cancer with an incidence of about 200 -300 cases per year in the US. Most cases arise in individuals who have an inherited or de novo mutation at birth in the RB1 gene encoding the Rb protein (pRb) and then acquire a second pathogenic RB1 mutation somatically. Although RB1 is a tumor suppressor gene and is recessive, the disorder follows an autosomal dominant pattern of inheritance because the risk of developing the second RB1 defect in any cell is so high. Consequently, hereditary retinoblastoma is one of the "cancer predisposition" syndromes. A child with hereditary retinoblastoma may present with a retinoblastoma in one or both eyes and has an increased risk of developing other cancers later in life. Non-heritable retinoblastoma cases, where both pathogenic RB1 alleles have developed spontaneously during a person's life, are more likely to present as unilateral retinoblastoma.

Li-Fraumeni syndrome

Li-Fraumeni syndrome is also a rare, cancer predisposition syndrome, that results from inheritance of a single pathogenic allele of the tumor suppressor p53 gene and acquisition of a second mutation in the other, normal p53 allele in a lifetime. Individuals with Li-Fraumeni syndrome are at high risk for developing one or more independent tumors in different tissues and may elect prophylactic surgery to reduce the risk of breast or ovarian cancer

for example. Patients are also advised to avoid diagnostic X-rays and any other exposure to a potential mutagen that could increase risk of DNA damage.

While individuals with Li-Fraumeni syndrome are rare, p53 mutations are some of the most common mutations found in cancers. Because p53 plays such an important role in cell cycle regulation, it is not surprising that a defect in p53 could contribute to unregulated cell growth and division, possibly even in a heterozygous state where only one p53 allele is defective.

It should be noted that as with all defective alleles associated with disorders, the clinical presentation may vary with the specific mutation that an individual has. In the case of p53, research has been aimed at correlating specific p53 mutations with the type, severity, and prognosis of cancers that develop (genotype/phenotype correlations). Furthermore, not all p53 mutations reduce or inhibit the function of the P53 protein ("loss-of-function" mutations). Some p53 "gain-of-function" mutations result in a P53 protein with increased activity that can also disrupt normal cell cycle regulatory balance.

CF resources

Additional information about CF can be found at Dynamed (under "Top Resources" on the UW Health Sciences Library website), and at the Cystic Fibrosis Foundation.