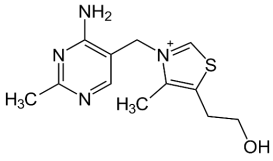
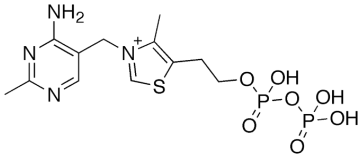
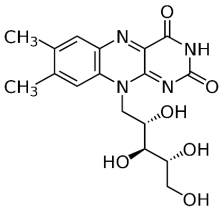
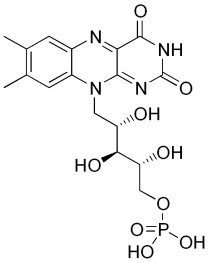
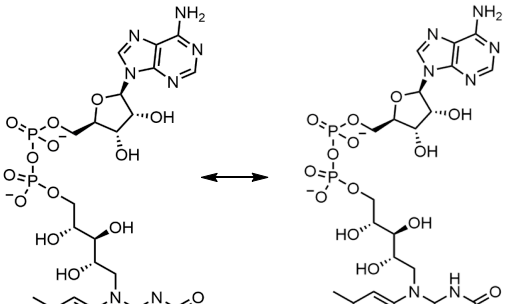
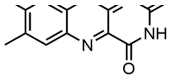
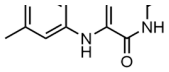
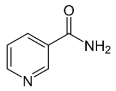
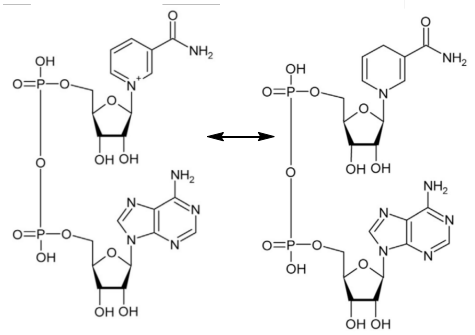
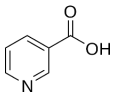
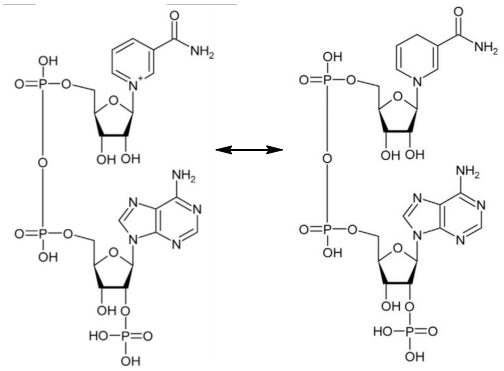
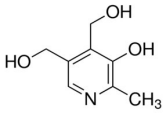
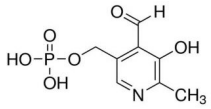
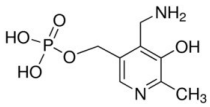
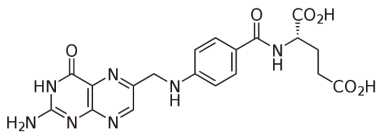
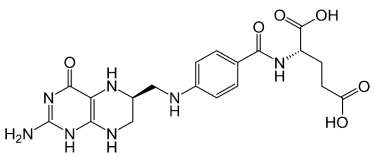
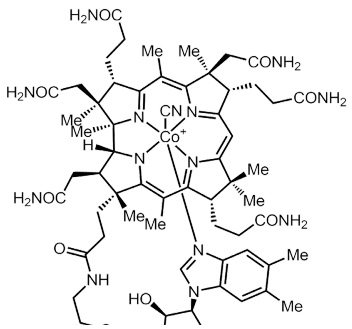
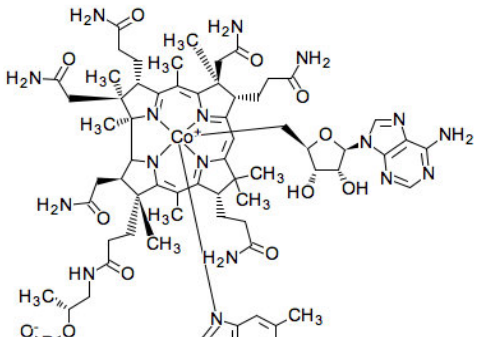
zinc, and molybdenum. ***Coenzymes*** are small organic molecules that are often derived from vitamins. ***Coenzymes*** can bind loosely with the enzyme and release from the active site. As such, they are also considered substrates for the reaction. Alternatively, they may be tight binding and cannot dissociate easily from the enzyme. In this case, after their initial participation in an enzyme-catalyzed reaction, the enzyme would no longer be able to use the cofactor in another round of catalysis until the initial state of the cofactor is reformed, which takes another chemical reaction and often an additional substrate.

Tight-binding coenzymes are referred to as ***prosthetic groups***. Enzymes not yet associated with a required cofactor are called ***apoenzymes***, whereas enzymes bound with their required cofactors are called ***holoenzymes.*** Sometimes, organic molecules and metals combine to form coenzymes, such as in the case of the heme cofactor (Figure 7.15). Coordination of heme cofactors with their enzyme counterparts often involves interactions with histidine residues, as shown in the succinate dehydrogenase enzyme shown in Figure 6.8.1.

Figure 6.8.1: The Heme Cofactor. The family of heme cofactors contains an iron metal coordinated with a porphyrin ring structure, as shown in the *left-hand panel* within the structure of Heme B. In the right-hand panel, Heme B is shown complexed with the succinate dehydrogenase enzyme from the Kreb Cycle. The structure of Heme B shown in the left-hand panel is from Yikrazuul, and the crystal structure of Succinate Dehydrogenase complexed with Heme B is from Richard Wheeler.

Many biological cofactors are vitamin B derivatives, as shown below in Table 6.8.1. Many vitamin deficiencies cause disease states due to the inactivity of ***apoenzymes*** that can not function without the correctly bound ***coenzyme.***

B Vitamins	Modified Enzyme Cofactors
Vitamin B1 (Thiamine)	Thiamine Diphosphate (TPP)
	
Vitamin B2 (Riboflavin)	Flavin Mononucleotide (FMN)
	
	Flavin Adenine Dinucleotide (FAD <-> $FADH_2$)
	

	
Vitamin B3 (Niacinamide) - amide form	**Nicotinamide adenine dinucleotide (NAD <-> NADH)**
	
Vitamin B3 (Niacin) - carboxylic acid form	**Nicotinamide dinucleotide phosphate (NADP <-> NADPH)**
	
Vitamin B6 (Pyridoxine)	**Pyridoxal5'-phosphate**
	
	Pyridoxamine 5'-phosphate
	
Vitamin B9 (Folic Acid)	**Tetrahydrofolate**
	
Vitamin B12 (Cyanocobalamin)	**Adenosylcobalamin**
	

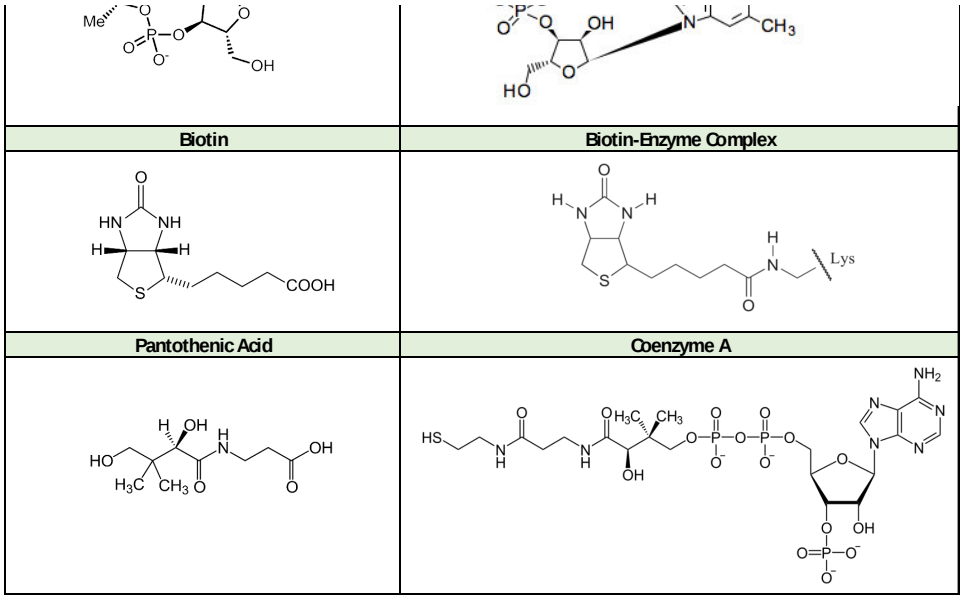


Table 6.8.1: Essential B-Vitamins and their Modified Enzyme Cofactors

Cofactors can help to mediate enzymatic reactions through the use of any of the different catalytic strategies listed above. They can serve as nucleophiles, mediate covalent catalysis, form electrostatic interactions with the substrate, and stabilize the transition state. They can also cause strain distortion or facilitate acid-base catalysis. Metal-aided catalysis can often use homolytic reaction mechanisms that involve radical intermediates. This can be important in reactions such as those occurring in the electron transport chain that requires the safe movement of single electrons.

We present plausible mechanisms for prototypical reactions using some of the cofactors shown in Table 6.8.1 above. Each shows the flow of electrons from a source to a sink. The source is often a pair of electrons on an anion, formed by the prior removal of a proton from the atom by a general base. A sink could be a carbonyl O, which receives a pair of electrons from one of the C=O bonds of the carbonyl. As a bond is made to the carbonyl, one of the double bonds must break with the electrons going (temporarily if the reaction is a nucleophilic substitution reaction) to the carbonyl O, an excellent sink since it is so electronegative. An even better sink is a positive N of an iminium ion; examples are shown below. Just the "business parts" of the cofactors are shown below.

To appreciate the mechanism used by cofactors and show a clear example of an electron source/sink, let's look at a reaction that doesn't require a cofactor, the spontaneous decarboxylation of a β-keto acid as shown in Figure 6.8.1.

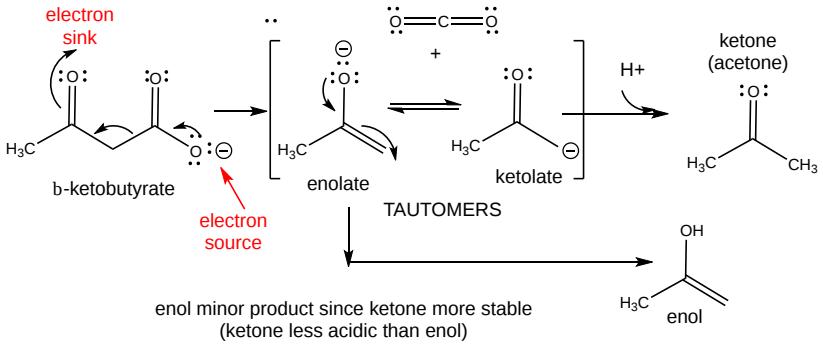


Figure 6.8.1:spontaneous decarboxylation of a β-keto acid

Even though no cofactor is required, nucleophilic catalysis by an amine through Schiff Base formation would speed up the reaction (as we will see below). Now let's look at how some of the cofactors listed in Table 6.8.1 above facilitate electron flow in reactions.

6.8.2: Thiamine pyrophosphate - decarboxylation of α-keto acids

Thiamine pyrophosphate (TPP) facilitates the decarboxylation of α-keto acids. TPP is a derivative of thiamine, vitamin B1, whose deficiency causes beriberi. TPP is covalently attached to the enzyme, such as in pyruvate dehydrogenase and alpha-ketoglutarate dehydrogenase, two enzymes that catalyze the decarboxylation of α-keto acids. The structure and "business" end of TPP and its catalytic activity are shown in Figure 6.8.2.

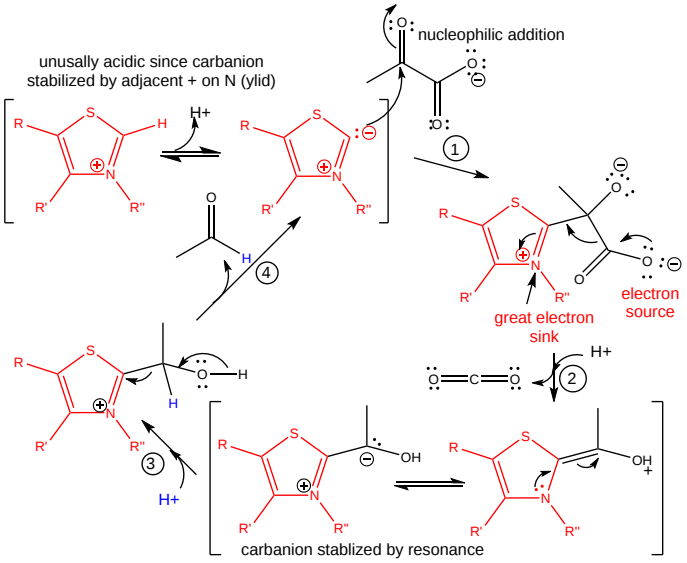


Figure 6.8.2: Role of TPP in the decarboxylation of pyruvate (step 1) and release of acetaldehyde (step 4)

The number of arrows leading to the product does not reflect the number of steps.

Figure 6.8.3 shows an interactive iCn3D model of the thiamin diphosphate-dependent enzyme pyruvate decarboxylase from the yeast Saccharomyces cerevisiae (1pvd).

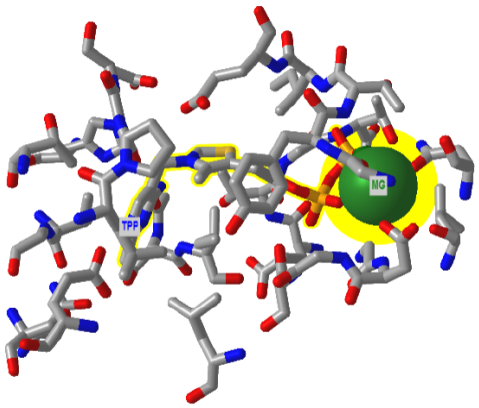


Figure 6.8.3: Thiamin diphosphate-dependent enzyme pyruvate decarboxylase from the yeast Saccharomyces cerevisiae (1pvd) (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...cCq9EJrquwggj8

6.8.3: Flavin Adenine Dinucleotide (FAD) - hydride transfer

FAD and its reduced form, $FADH_2$, are tightly or covalently attached to an enzyme, so FAD must be regenerated in each catalytic cycle. Figure 6.8.4 shows an example of how this cofactor facilitates the transfer of a $:H^-$ hydride ion to the "business end" of FAD. In contrast to a transfer of protons (H^+), an acid/base reaction, hydride transfer removes 2 electrons from the substrate (in this case, succinate) along with a proton in an oxidation reaction as FAD is reduced.

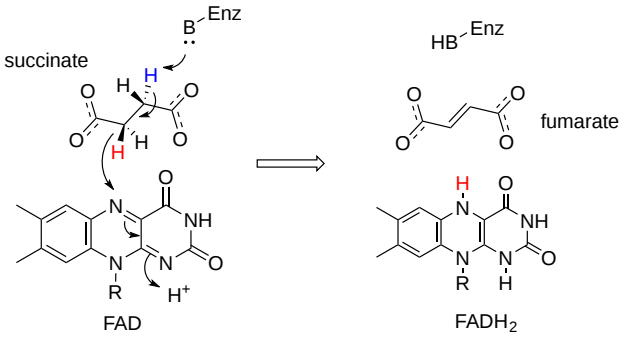


Figure 6.8.4: Oxidation of succinate by FAD

Figure 6.8.5 shows an interactive iCn3D model of the FAD-binding domain of cytochrome P450 BM3 from Priestia megaterium in complex with $NADP^+$ (4DQL)

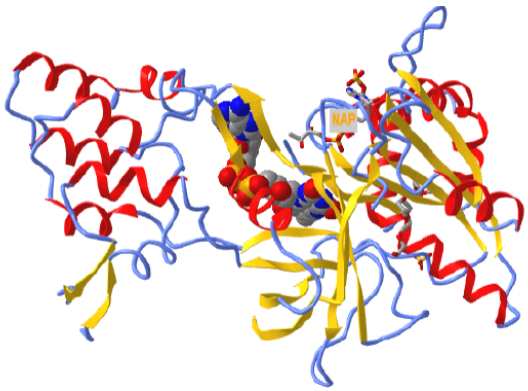


Figure 6.8.5: FAD binding domain of cytochrome P450 BM3 in complex with $NADP^+$ (4DQL). (Copyright; author via source).

Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/icn3d/share.html?hoT1WDCUv1wZFMyRA

FAD is shown in spacefill. $NADP^+$, which reoxidizes the reduced $FADH_2$ back to FAD, is shown in sticks and labeled NAP.

6.8.4: Nicotinamide Adenine Dinucleotide (FAD) reactions

NAD^+ and a phosphorylated form, $NADP^+$, are one of nature's most widely used oxidizing agents and are used as dissociable substrates/cofactors for many different types of enzyme-catalyzed oxidation reactions. The free enzyme is continually active since the NAD^+ or $NADP^+$ cofactor binds (as a substrate) and dissociates (as a product) after each catalytic cycle. The biological synthesis of NAD^+ requires the vitamin nicotinic acid, also called niacin (nicotinic acid), an absence of which causes pellagra.

Oxidation of an alcohol to an aldehyde: The oxidation of ethanol to acetaldehyde by NAD^+, catalyzed by the enzyme alcohol dehydrogenase, is shown in Figure 6.8.6.

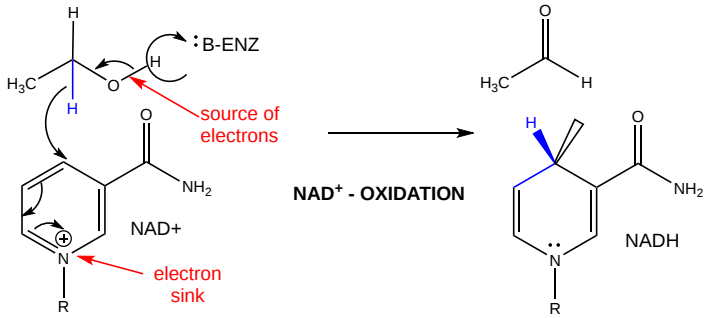


Figure 6.8.6: Oxidation of ethanol by NAD^+

The product acetaldehyde contributes to hangovers after ethanol consumption. Note that this reaction is a hydride transfer, which would not be expected to occur in the aqueous environment of a cell, given the extreme reactivity and basicity of a $:H^-$ hydride ion. This transfer happens in the enzyme's active site, which is anhydrous after binding substrates.

Oxidative decarboxylation of an alcohol: A two-step mechanism for this reaction is shown in Figure 6.8.7

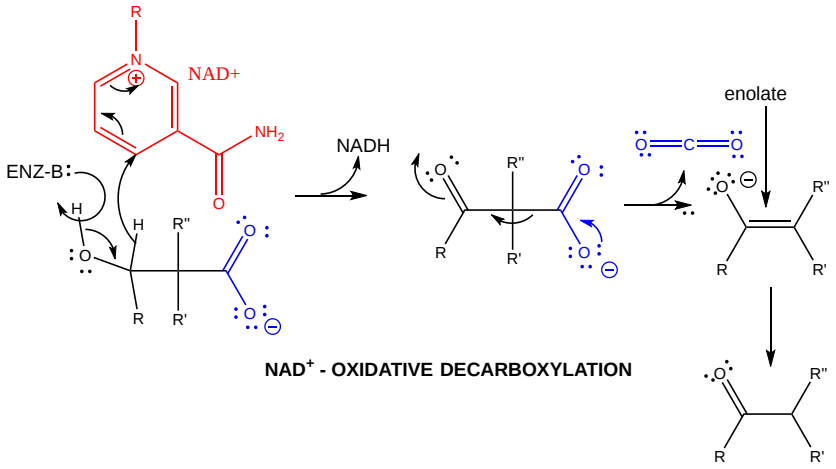


Figure 6.8.7: Oxidative decarboxylation of an alcohol

After the first step, an electron sink (the oxygen of the carbonyl) is present at the β-carbon, facilitating the decarboxylation step.

Oxidative deamination of an amine: A two-step reaction, a hydride transfer to form a Schiff base, followed by hydrolysis of the Schiff base, is shown in Figure 6.8.8.

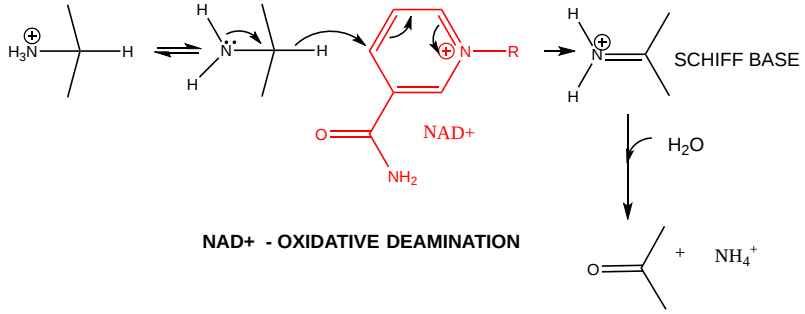


Figure 6.8.8: **Oxidative deamination of an amine**

We will discuss Schiff base chemistry in more detail below.

6.8.5: Pyridoxal Phosphate Enzymes

Pyridoxal phosphate (PLP) is a derivative of vitamin B6 or pyridoxal. Deficiencies cause convulsions, chronic anemia, and neuropathy. It assists in many reactions (catalyzed by PLP-dependent enzymes). The PLP is bound covalently to lysine residues in a Schiff base linkage (aldimine). This form reacts with many free amino acids (as substrates) to replace the Schiff base to Lys of the enzyme with a Schiff base to the amino acid substrate. First, we will review Schiff base (an imine) formation by the reaction of an aldehyde or ketone with an amine, as shown in Figure 6.8.9.

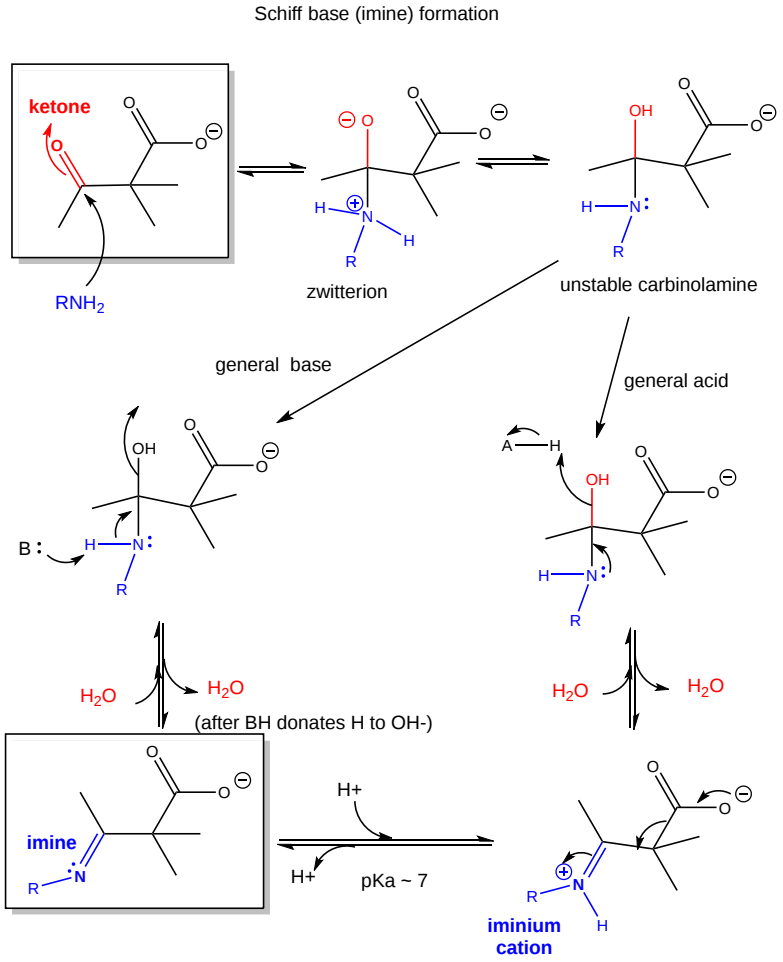


Figure 6.8.9

The reaction is essentially a nucleophilic attack of a carbonyl carbon of an aldehyde or ketone by an amine, followed by a dehydration step. Note that the net effect is to replace one electron sink, a carbonyl (C=O), with an **imine** (C=NH ↔ $C=NH_2^+$), with pKa around 7.0. Hence, 50% of the imine is protonated at neutral pH to form the **iminium cation**, a much better electron sink than the starting carbonyl!

The structure of pyridox**al** phosphate, which contains a reactive **al**dehyde, is converted to an imine by reaction with the ε-amino side chain of a lysine in the active site of a PLP-dependent enzyme, is shown in Figure 6.8.10.

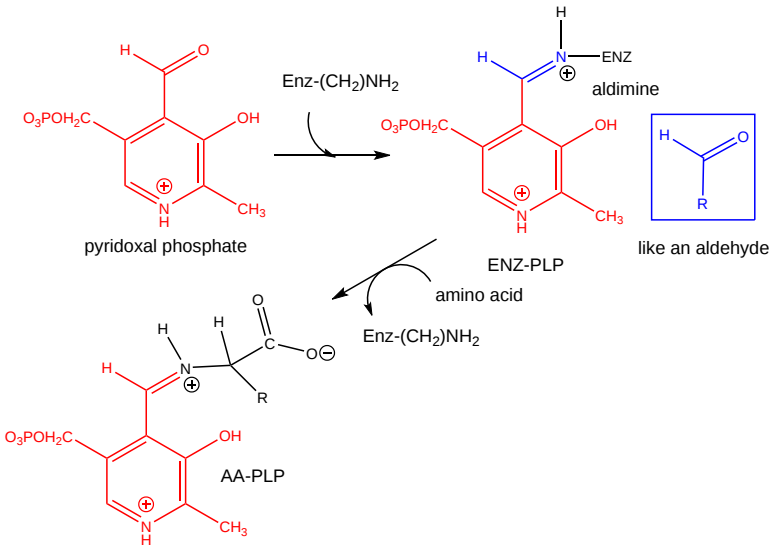


Figure 6.8.10

The figure also shows the replacement of the enzymes' lysine ε-NH_2-PLP bond to that of a free amino as an incoming substrate, a process which should proceed with a ΔG^0 of approximately 0. This occurs in PLP-dependent enzymes with free amino acids as substrates (we will discuss several examples below).

Figure 6.8.11 shows an interactive iCn3D model of the E. Coli Aspartate aminotransferase, W140H mutant, maleate complex (1ARI).

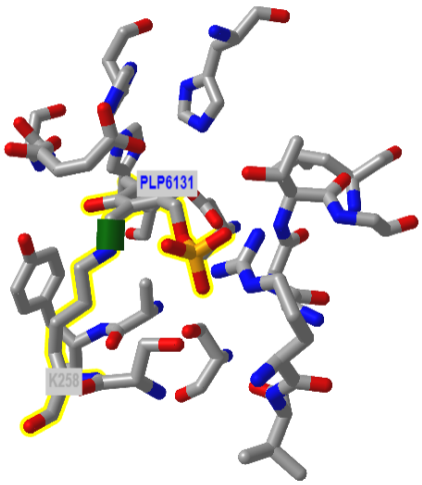


Figure 6.8.11: **Aspartate aminotransferase, W140H mutant, maleate complex** (1ARI). (Copyright; author via source).

Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...ZQixXhCwX5EEQ9

Note that the PLP is in Schiff base linkage with the ε-NH_2 group of a lysine in the enzyme's active site.

PLP is quite impressive!

From a chemistry perspective, PLP is an ideal molecule to facilitate electron flow in biochemical reactions. William Jencks noted this in his classic text, Catalysis in Chemistry, in which he wrote this elegant description of its properties:

"It has been said that God created an organism especially adapted to help the biologist find an answer to every question about the physiology of living systems; if this is so, it must be concluded that pyridoxal phosphate was created to provide satisfaction and enlightenment to those enzymologists and chemists who enjoy pushing electrons, for no other coenzyme is involved in such a wide variety of reactions, in both enzyme and model systems, which can be reasonably interpreted in terms of the

chemical properties of the coenzyme. Most of these reactions are made possible by a common structural feature. That is, electron withdrawal toward the cationic nitrogen atom of the imine and into the electron sink of the pyridoxal ring from the alpha carbon atom of the attached amino acid activates all three of the substituents of this carbon for reactions that require electron withdrawal from this atom."

We'll present three examples of the reaction of an amino acid with a PLP-dependent enzyme. In each case, a different bond to the α-carbon of the amino acid substrate is broken.

α-decarboxylation of an amino acid: Figure 6.8.12 shows a plausible reaction mechanism.

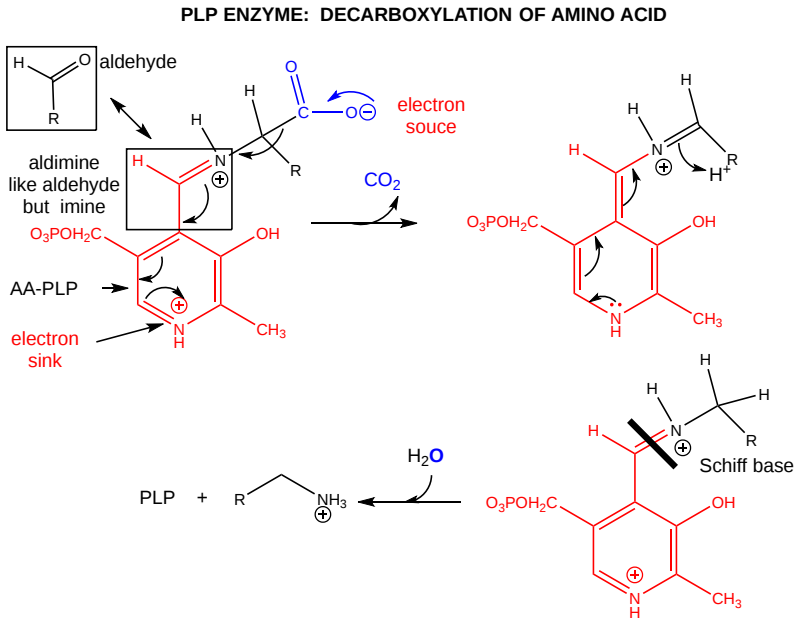


Figure 6.8.12: PLP-dependent decarboxylation of an amino acid.

β-elimination from serine: The enzyme serine dehydratase catalyzes the reaction shown in Figure 6.8.13

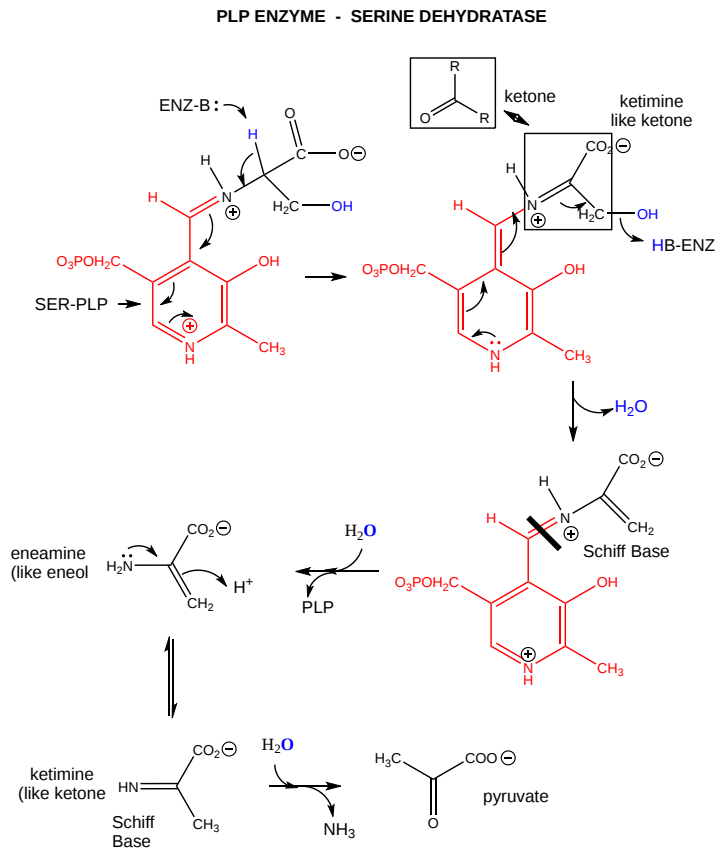


Figure 6.8.13: PLP-dependent β-elimination from serine

Racemization of amino acids: Amino acid racemases use PLP as a cofactor using a mechanism shown in Figure 6.8.14

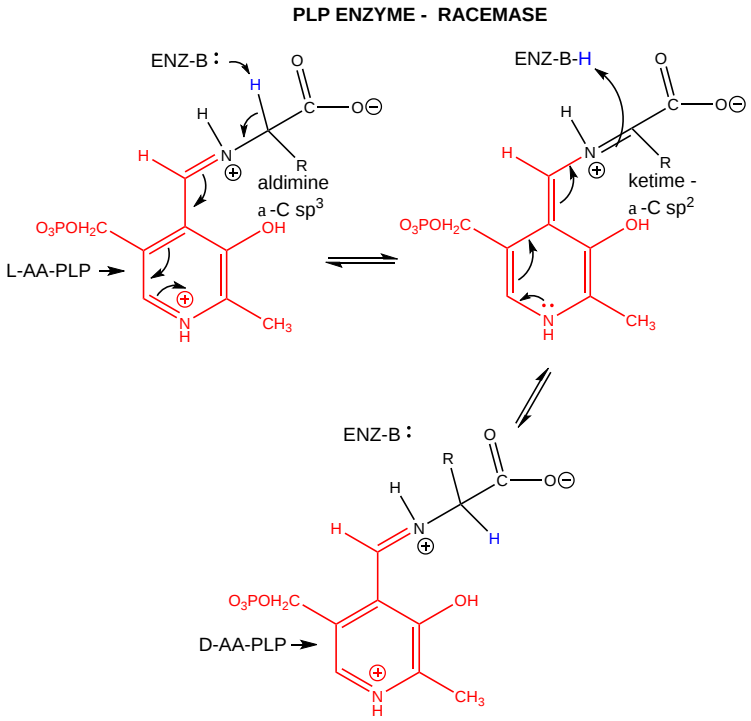


Figure 6.8.14: PLP-dependent racemization of amino acids

Why do racemases exist since the biological world consists of only L-amino acids? There are two possible reasons. Some D-amino acids are found, such as in bacterial cell walls. In addition, amino acids spontaneously racemize on their own, albeit slowly.

Racemases with oxygen atoms in the beta-carbon racemize at a much higher rate since they can stabilize the carbanion intermediate formed when the alpha proton is removed during racemization. The concentration of D-Asp and D-Asn can also be used to date biological material.

Transamination reactions: PLP enzymes also catalyze the transamination reaction, which is shown in Figure 6.8.15

Amino Acid 1 + α-keto acid 1 ↔ α-keto acid 2 + Amino Acid 2 For example: Figure 6.8.x

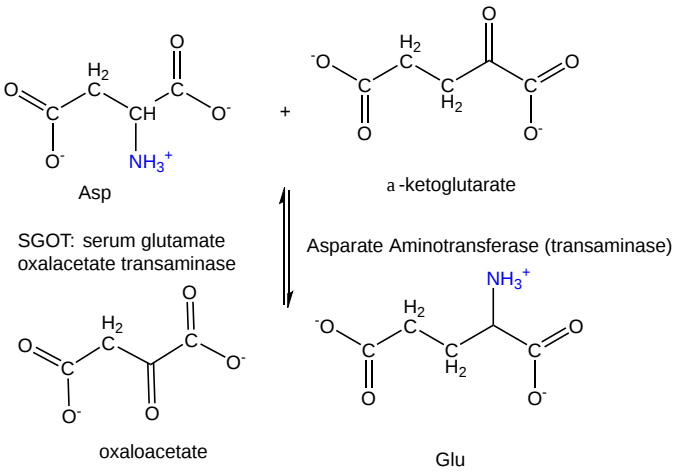


Figure 6.8.15: PLP-dependent transamination reaction

First, Asp, bound to PLP through a Schiff base link, loses the α-H and forms a ketimine through a tautomerization reaction, which ultimately hydrolyzes to form the released oxaloacetate and pyridoxamine. The pyridoxamine reacts with α-ketoglutarate in the reverse of the first three reactions to form Glu.

See Chapter 18.2 for a great description of the role of PLP in transamination reactions. Here is a link to more mechanistic details of the reactions of PLP.

We will explore other cofactors in future chapters.

6.8.6: Summary

This chapter explores how enzymes harness cofactors to facilitate the movement of electrons—a crucial process in the making and breaking of chemical bonds during catalysis. The discussion begins by establishing the fundamental concept of electron flow, where electrons are transferred from “sources” (electron-rich groups) to “sinks” (electron-deficient centers). This electron-pushing mechanism underpins many enzyme-catalyzed reactions.

A major focus is placed on the two primary classes of cofactors: metal ions and coenzymes. Metal cofactors, such as iron, magnesium, copper, and zinc, are essential for stabilizing charged intermediates and often participate directly in redox reactions. In contrast, coenzymes are small organic molecules—many of which are derived from vitamins—that can bind transiently to enzymes. Depending on their binding characteristics, these coenzymes may act as loosely bound substrates or as prosthetic groups that remain tightly attached throughout the catalytic cycle.

The chapter details several prototypical cofactors and their roles in catalysis:

- **Thiamine Pyrophosphate (TPP):** Derived from vitamin B1, TPP is critical for the decarboxylation of α-keto acids in enzymes such as pyruvate dehydrogenase. Its unique structure allows it to stabilize carbanion intermediates during bond cleavage.
- **Flavin Adenine Dinucleotide (FAD) and Nicotinamide Adenine Dinucleotide (NAD^+/$NADP^+$):** These coenzymes play central roles in oxidation-reduction reactions. FAD facilitates hydride transfer, as seen in the oxidation of succinate, while NAD^+/$NADP^+$ acts as a reversible electron carrier in numerous dehydrogenase reactions.
- **Pyridoxal Phosphate (PLP):** A derivative of vitamin B6, PLP is perhaps the most versatile coenzyme, participating in a range of reactions including decarboxylation, β-elimination, racemization, and transamination. Through the formation of Schiff base intermediates with amino acid substrates, PLP acts as an efficient electron sink, thereby lowering the activation energy for these transformations.

Throughout the chapter, the emphasis is on how cofactors enable precise control over electron movement, ensuring that reactions proceed efficiently under physiological conditions. Structural examples—illustrated with interactive models and diagrams—demonstrate how specific amino acid residues, such as histidines coordinating a heme group, facilitate these interactions.

Ultimately, this chapter underscores the critical role of cofactors in enzyme catalysis. By providing both the chemical functionality and the structural framework for electron transfer, cofactors bridge the gap between the inherent reactivity of small molecules and the exquisite specificity required for biological regulation. This understanding is fundamental for appreciating how enzymes achieve their remarkable catalytic efficiencies and for designing therapeutic strategies that target these processes.

CHAPTER OVERVIEW

7: Carbohydrates and Glycobiology

Return to Fundamentals of Biochemistry Search Fundamentals of Biochemistry

---

7.1: Monosaccharides and Disaccharides

Learning Goals (ChatGPT o1, 1/30/25)

- **Define Key Terminology:**
  - Differentiate between “sugar,” “carbohydrate,” and “glycan” and explain how these terms are used to describe molecules ranging from simple monosaccharides to complex glycan polymers.
- **Understand Monosaccharide Structures:**
  - Describe the chemical structure of monosaccharides as polyhydroxy-aldehydes or ketones, and explain the significance of stereochemistry in sugars.
  - Convert and interpret various structural representations (Fischer projections, Haworth projections, chair, and wedge/dash forms) for common monosaccharides like D-glucose, D-ribose, and D-fructose.
- **Cyclic Conformations and Anomer Formation:**
  - Explain the process by which monosaccharides cyclize to form furanose or pyranose rings, including the formation of hemiacetals.
  - Distinguish between α- and β-anomers in cyclic sugars and understand the factors that influence their stability in solution versus in polysaccharides.
- **Isomerism in Sugars:**
  - Identify and differentiate among configurational isomers (enantiomers, diastereomers, epimers, and anomers) and conformational isomers (chair and boat forms) of sugars.
- **Formation of Glycosidic Bonds:**
  - Describe the chemical mechanism of hemiacetal and acetal formation in monosaccharides leading to the formation of disaccharides and polysaccharides, including the significance of glycosidic linkages (e.g., $1 \rightarrow 4$, $1 \rightarrow 6$).
- **Reducing vs. Nonreducing Sugars:**
  - Explain the concept of reducing sugars in terms of the reversible opening of the cyclic form to expose an aldehyde group, and contrast these with nonreducing sugars such as sucrose.
- **Monosaccharide Derivatives:**
  - Recognize common chemical modifications of monosaccharides (e.g., oxidation, phosphorylation, amination, acetylation, lactonization) and discuss their biological significance.
- **Biological and Clinical Implications:**
  - Discuss how the structural diversity of glycans affects their function in cellular recognition, stability, and signaling.
  - Explain the molecular basis of alpha-gal syndrome, including how the disaccharide Gal(α1,3)Gal, found in tick saliva and red meat, triggers an immune response leading to allergic reactions.

Achieving these goals will equip students with the knowledge to navigate the complexity of glycan structures and their multifaceted roles in biological systems.

7.1.1: Introduction

Carbohydrate or glycan biochemistry is very complex and challenging owing to the stereochemical complexity of simple sugars, the large number of positions on the sugars used to form linkages between other sugars to create polymers, the large number of chemical modifications to base sugars, and the lack of a genetic template to instruct glycan polymer formation. No wonder our understanding of complex glycans has developed after that of the chemically simpler polymers like nucleic acids and proteins.

In addition, the terminology used to describe them varies as well. We use these general descriptions of them:

Sugar: usually refers to low molecular weight carbohydrates like glucose, lactose, and sucrose, but it can also refer broadly to any carbohydrate.

Carbohydrate: a general term that applies to simple sugars to complex sugar polymers like glycogen, starch, and cellulose. The name derives from the formula for simple sugars like glucose ($C_6H_{12}O_6$), which can be written as $C_6(H_2O)_6$ - a carbo (C) - hydrate (H_2O).

Glycan: a general term for molecules containing simple sugars and sugar derivatives linked in a polymer, either standalone molecules or attached to other molecules like proteins.

7.1.2: Monosaccharides Structures

The above definition of sugar needs some further nuance. From a chemical perspective, sugars can be defined as polyhydroxy-aldehydes or ketones. The simplest sugars contain at least three carbon atoms, and the most common are the aldo- and keto-trioses, tetroses, pentoses, and hexoses. The 3C sugars are glyceraldehyde and dihydroxyacetone, as shown in Figure 7.1.1.

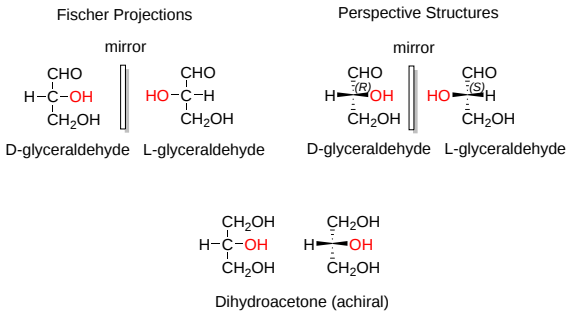


Figure 7.1.1: Three-carbon sugars

Glucose, an aldohexose, is a central sugar in metabolism. It and other 5C and 6C sugars can cyclize through intramolecular nucleophilic attack of one of the free hydroxyl groups on the carbonyl carbon of the aldehyde or ketone. Such intramolecular reactions occur if stable 5- or 6-member rings can form. The resulting rings are labeled furanose (5-member) or pyranose (6-member) based on their similarity to furan and pyran. On nucleophilic attack to form the ring, the carbonyl O becomes an OH that points either below (α anomer) or above (β anomer) the ring.

Figure 7.1.2 shows different representations of the linear and cyclic forms of the sugars D-glucose, D-ribose, and D-fructose

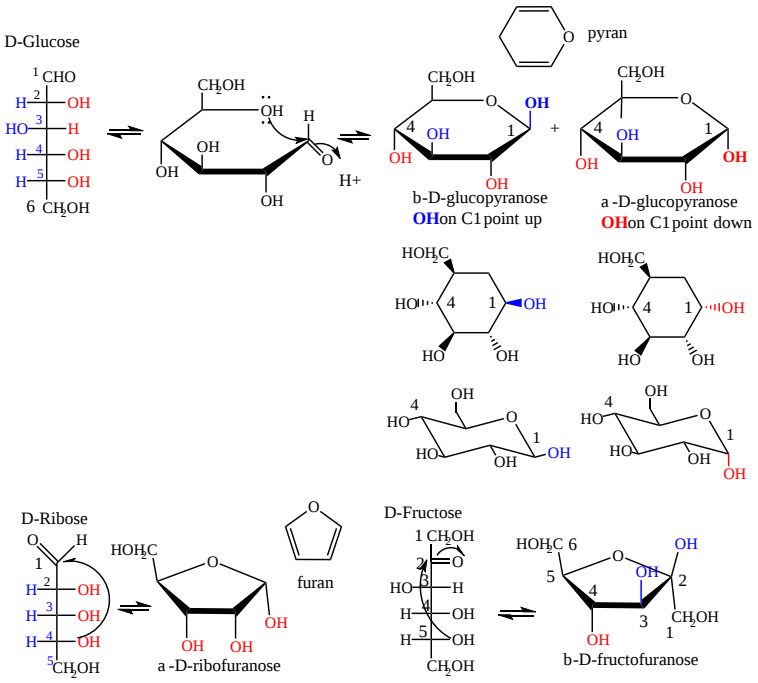


Figure 7.1.2: Linear and cyclic forms of D-glucose, D-ribose and D-fructose

Monosaccharides in solution exist as equilibrium mixtures of the straight and cyclic forms. In solution, glucose (Glc) is mainly in the pyranose form, fructose is 67% pyranose and 33% furanose, and ribose is 75% furanose and 25% pyranose. However, in polysaccharides, Glc is exclusively pyranose, and fructose and ribose are furanoses.

Sugars can be drawn in the straight chain form as Fischer projections or perspective structural formulas.

In the Fisher projection, the vertical bonds point down into the plane of the paper. That's easy to visualize for 3C sugars but more complicated for larger ones. For those, draw a wedge and dash line drawing of the molecule. When determining the orientation of the OHs on each C, orient the wedge and dash drawing in your mind so that the C atoms adjacent to the one of interest are pointing down. Sighting towards the carbonyl C, if the OH is pointing to the right in the Fisher project, it should be pointing to the right in the wedge and dash drawing, as shown below for D-threose and D-glucose. Figure 7.1.3 shows how to convert Fisher projections to wedge dash representations.

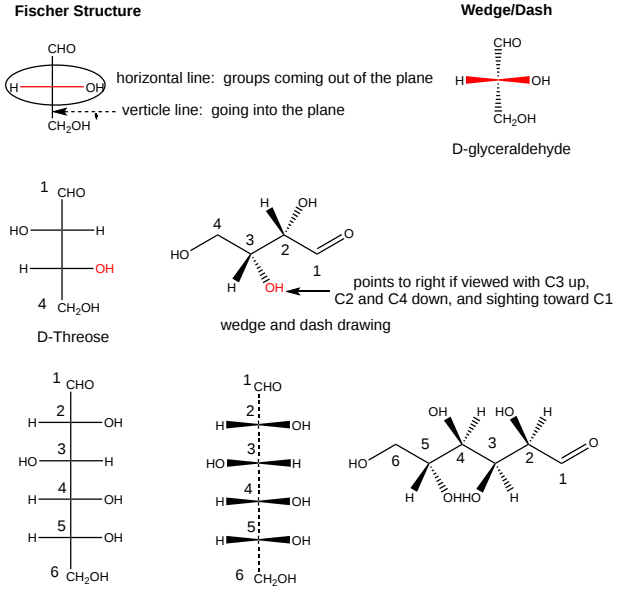


Figure 7.1.3: Converting Fisher projections to wedge dash representations.

Figure 7.1.4 shows an interactive iCn3D model of D-glucose in a linear form.

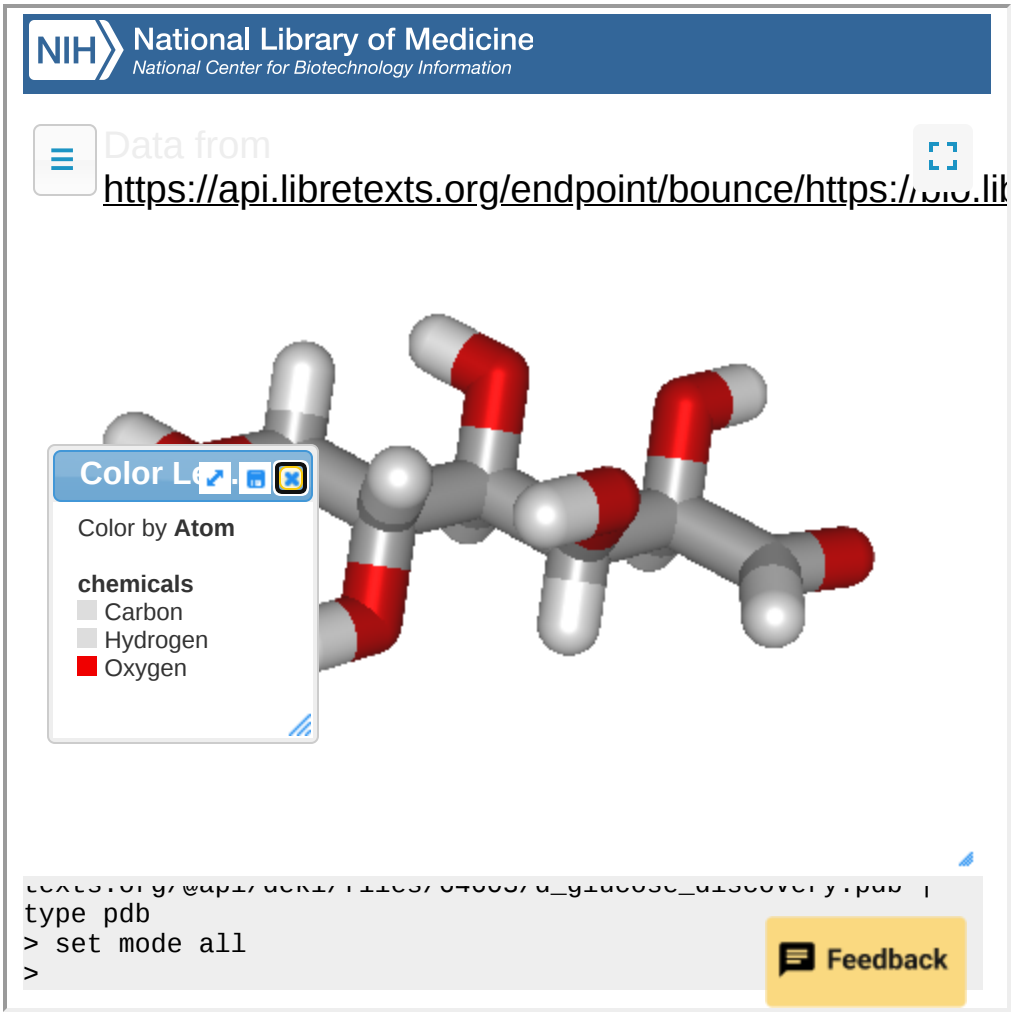


Figure 7.1.4: D-glucose (Copyright; author via source).

Orient the molecule as shown in Figure 7.1.5 below, with the carbonyl oxygen pointed to the far right, and compare it to the orientation shown in Figure 7.1.5 to reinforce your understanding of Fisher and wedge/dash projections.

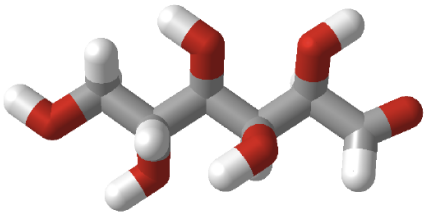


Figure 7.1.5:

Cyclic forms can be drawn either as the Haworth projections, which show the molecule as cyclic and planar with substituents above or below the ring) or the more plausible bent forms (showing glucose in the chair or boat conformations). β-D-glucopyranose is the only aldohexose that can be drawn with all its bulky substituents (OH and CH_2OH) in equatorial positions, which probably accounts for its widespread prevalence in nature. Figure 7.1.6 shows four different representations of glucose.

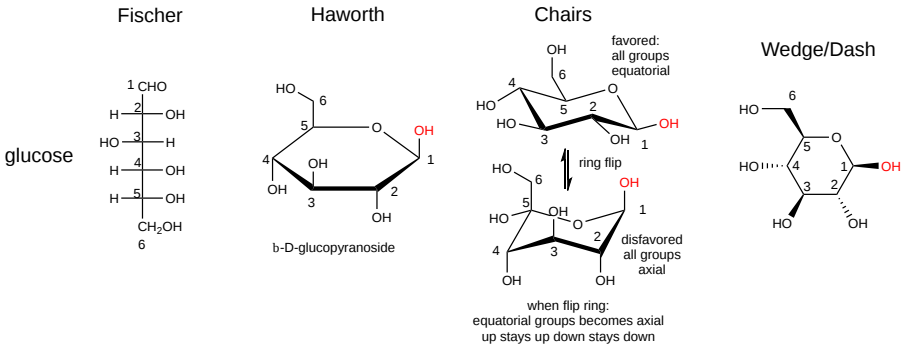


Figure 7.1.6: Fisher and cyclic Haworth, chair and wedge/dash representations of glucose

Haworth projections are more realistic than the Fisher projections, but you should be able to draw both structures. Generally, if a substituent points to the right in the Fisher structure, it points down in the Haworth. If it points left, it points up. Generally, the OH on the **α**-anomer points down (**α**nts down) while on the **β**-anomer, it points up (**β**utterflies up) as illustrated in Figure 7.1.7

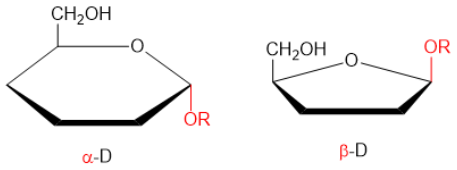


Figure 7.1.7: Alpha and beta Haworth representations of sugars

In the Haworth projections, the bulky R group of the next carbon after the carbon whose OH group was the nucleophile for ring formation is pointed up if the OH engaged in the attach was on the right-hand side in the straight chain Fisher diagram (as in α-D-glucopyranose above when the CH_2OH group is up). It is pointed down if the OH engaged in the attack was on the left-hand side in the straight chain Fisher diagram (as in α-D-galactofuranose above when the $(CHOH)CH_2OH$ group is down). The rest of the OH groups still follow the simple rule that if they point to the right in the Fisher straight chain form, they point down in the Haworth form.

The Fisher structures of most common monosaccharides (other than glyceraldehyde and dihydroxyacetone), which you will encounter most frequently, are shown in Figure 7.1.8.

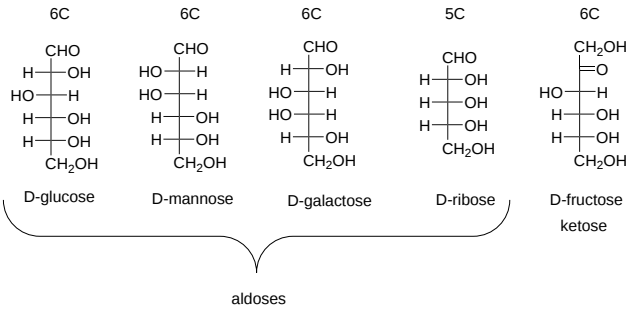


Figure 7.1.8: Most common monosaccharides discussed in this book

The mirror image of D-Glc is L-Glc. The D- and L- designations refer to the center of asymmetry most remote from the aldehyde or ketone. By convention, all chiral centers are related to D-glyceraldehyde, so sugar isomers related to D-glyceraldehyde at their last asymmetric center are D sugars.

Figure 7.1.9 shows multiple renderings of common hexoses.

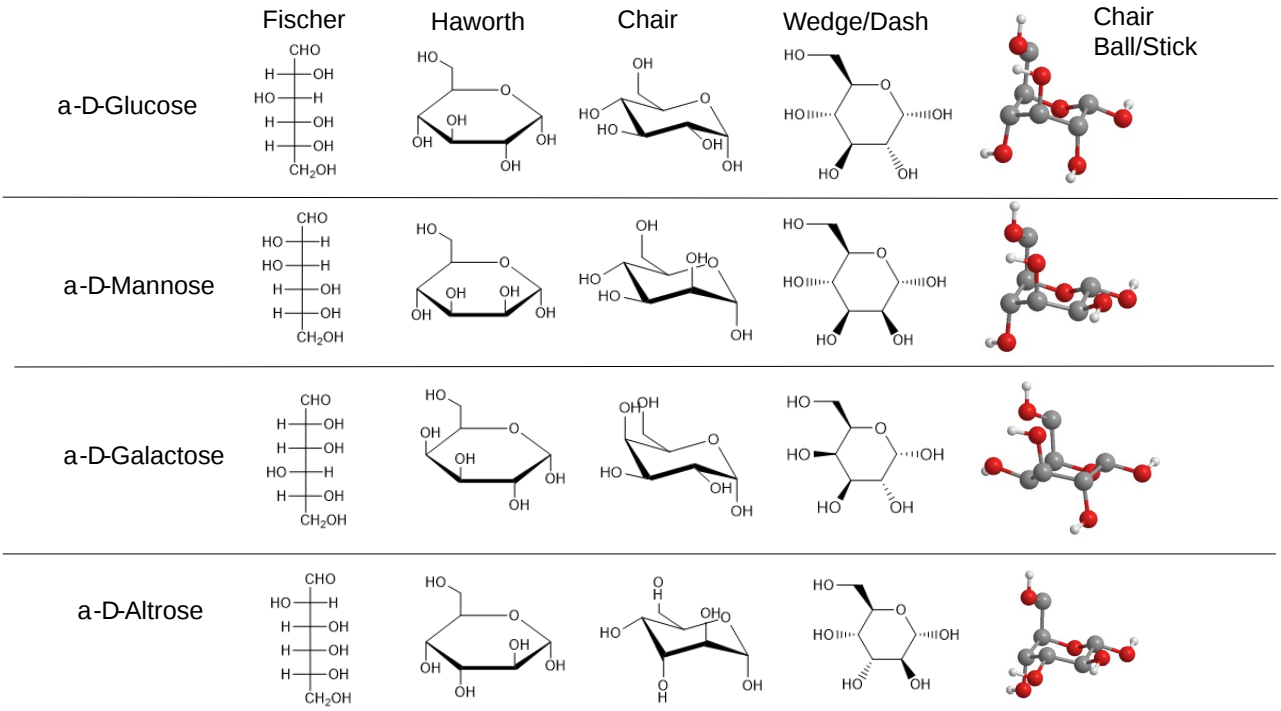


Figure 7.1.9: Multiple renderings of common hexoses

7.1.3: Isomers

Sugars can be configurational (interconverted only by breaking covalent bonds) or conformational isomers. Figure 7.1.10 reviews different configurational isomers.

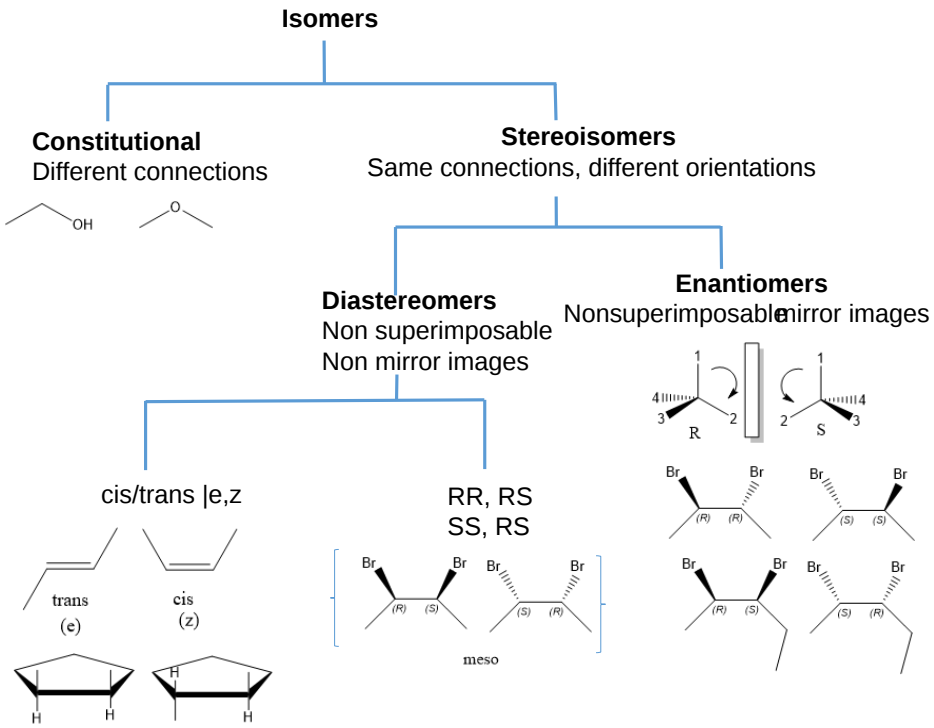


Figure 7.1.10: Types of isomers

The configurational isomers include enantiomers (stereoisomers that are mirror images of each other), diastereomers (stereoisomers that are not mirror images), **epimers** (diastereomers that differ at one stereocenter), and **anomers** (a particular form of stereoisomer, diastereomer, and epimer) shows enantiomers, diastereomers, epimers and anomers of 6 carbon sugars.

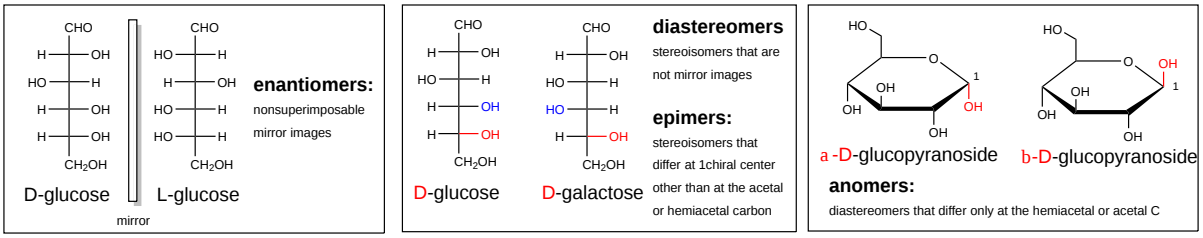


Figure 7.1.11: Enantiomers, diastereomers, epimers, and anomers of 6 carbon sugars.

Sugars can also exist as conformational isomers, which interchange without breaking covalent bonds. These include chair and boat conformations of the cyclic sugars as shown in Figure 7.1.12

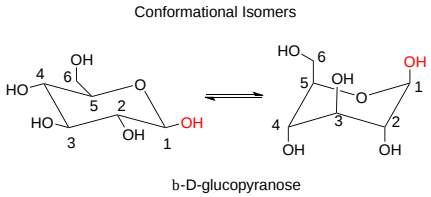


Figure 7.1.12: Conformational isomers of b-D-glucopyranose

7.1.4: Monosaccharide Derivatives

Many derivatives of monosaccharides are found in nature. These include

- oxidized forms in which the aldehyde and/or alcohol functional groups are oxidized to carboxylic acids
- phosphorylated forms in which phosphates are transferred from ATP to form phosphoester derivatives
- amine derivatives such as glucosamine or galactosamine
- acetylated amine derivatives such as N-Acetyl-GlcNAc (GlcNAc) or GalNAc
- lactone forms (intramolecular esters) in which an OH group attacks a carbonyl C that was previously oxidized to a carboxylic acid
- condensation products of sugar derivatives with lactate ($CH_3CHOHCO_2^-$) and pyruvate ($CH_3COCO_2^-$), both from the glycolytic pathway, to form muramic acid and neuraminic acids (also called sialic acids), respectively.

Figure 7.1.13some simple monosaccharide derivatives.

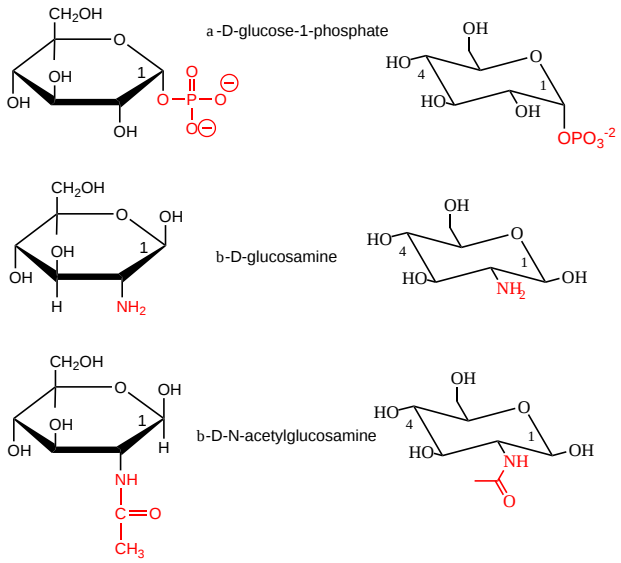


Figure 7.1.13: Monosaccharide derivatives

Figure 7.1.14 shows some additional oxidative derivatives of glucose shown in Fischer projections.

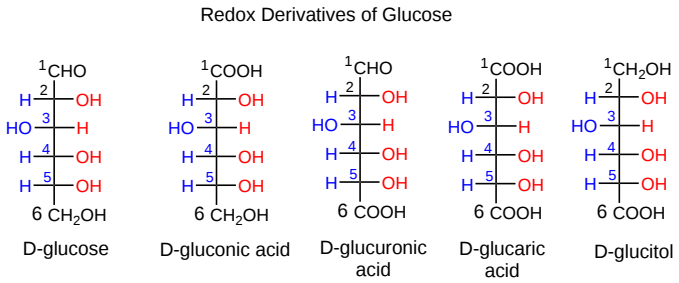


Figure 7.1.14: Redox derivatives of glucose

Other important derivatives of monosaccharides are **sialic acids**. **N-acetylmuramic acid**, found in bacterial cell walls, consists of GlcNAc in ether link at C3 with lactate, while **N-acetylneuraminic acid** results from an intramolecular cyclization of a condensation product of ManNAc and pyruvate. These sialic acids are shown in Figure 7.1.15.

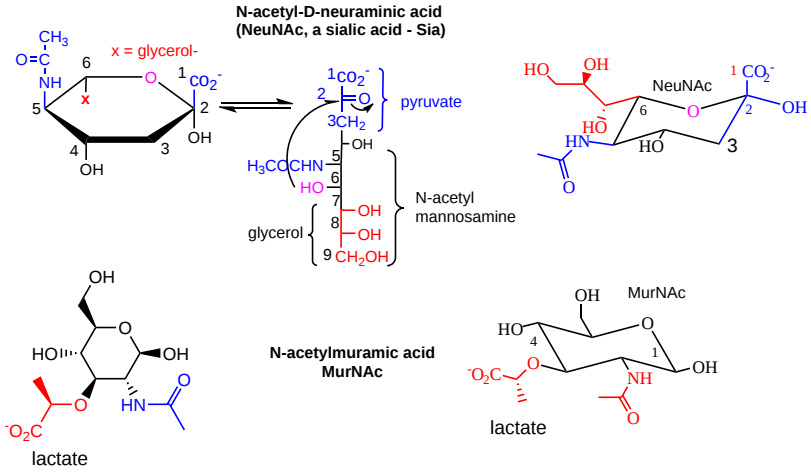


Figure 7.1.15: Sialic Acids

Sugars are very complicated as the linkages and substituents are so diverse. Figure 7.1.16 shows differences in sialic acids between humans and chimps.

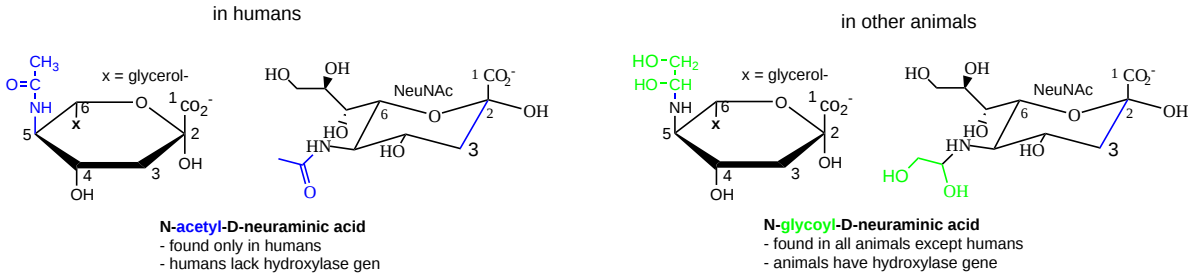


Figure 7.1.16: Sialic acids between humans and chimps.

What happens when non-vegan humans eat animal products (meat, milk) with N-glycoyl neuraminic acids (Neu5Gc)? Some get incorporated into human membrane glycans. Sialic acids on surface proteins can serve as "receptors" that allow the binding of self-cells as well as foreign cells or proteins that have evolved to bind them. A toxin, SubAB, secreted by E. Coli 0157, can bind Neu5Gc. Hence eating meat products can make us more susceptible to bacteria that recognize Neu5Gc.

7.1.5: Formation of Hemiacetals, Acetals, and Disaccharides

Monosaccharides that contain aldehydes can cyclize through an intramolecular nucleophilic attack of an OH at the carbonyl carbon in an addition reaction to form a **hemiacetal.** In the past, the group was called a **hemiketal** if the attack was on a ketone, but now they are also called hemiacetals. On the addition of acid (which protonates the anomeric OH, forming water as a potential leaving group), another alcohol can add, forming an **acetal** with water leaving. These reactions are shown in Figure 7.1.17.

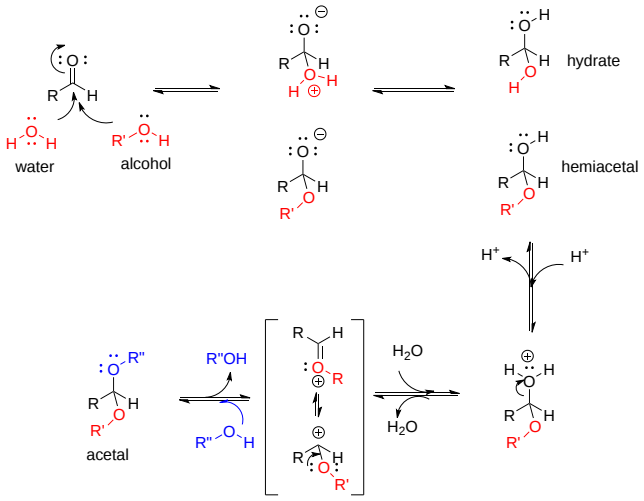


Figure 7.1.17: Hemiacetal and acetal formation

If the other alcohol is a second monosaccharide, a disaccharide results. The acetal link bonding to the two monosaccharides is called a **glycosidic link**. If the link between the two sugars involves an anomeric carbon, the newly formed OH group at the link can be designated either as α (if the O on the carbon involved in the glycosidic link is pointing down) or β (if the O is pointing up). For a 2-2 link between hexoses (i.e., between two non-anomeric carbons, the α/β designation is not used. Since sugars contain so many OH groups, which can act as the "second" alcohol in acetal formation, links between sugars can be quite diverse. These include α and β forms of 1-2, 1-3, 1-4, 1-5, 1-6, etc. links. Here are examples of disaccharides:

- maltose: Glc(α 1,4)Glc, which can be considered a disaccharide hydrolysis product of the polysaccharide glycogen or starch (discussed in the section)
- cellobiose: Glc(Glc(α 1,4)Glc 1,4)Glc, which can be considered a disaccharide hydrolysis product of cellulose.
- lactose: Gal(β 1,4)Glc, also known as milk sugar.
- sucrose: Glc(α 1,2)Fru. Since fructose is attached through the anomeric OH of this ketose, it is not in equilibrium with its straight-chain keto form; hence, sucrose is a nonreducing sugar. Note also that since the anomeric C-OH or each sugar is used, the α/β designation in the disaccharide is used. Hence, sucrose would be abbreviated as Glc(α1,2β)Fru

Figure 7.1.18 illustrates the differences between lactose and sucrose. Note that the β-D-fructofuranose ring is flipped (left to right as in turning one of your hands over) compared to Figure 7.1.16.

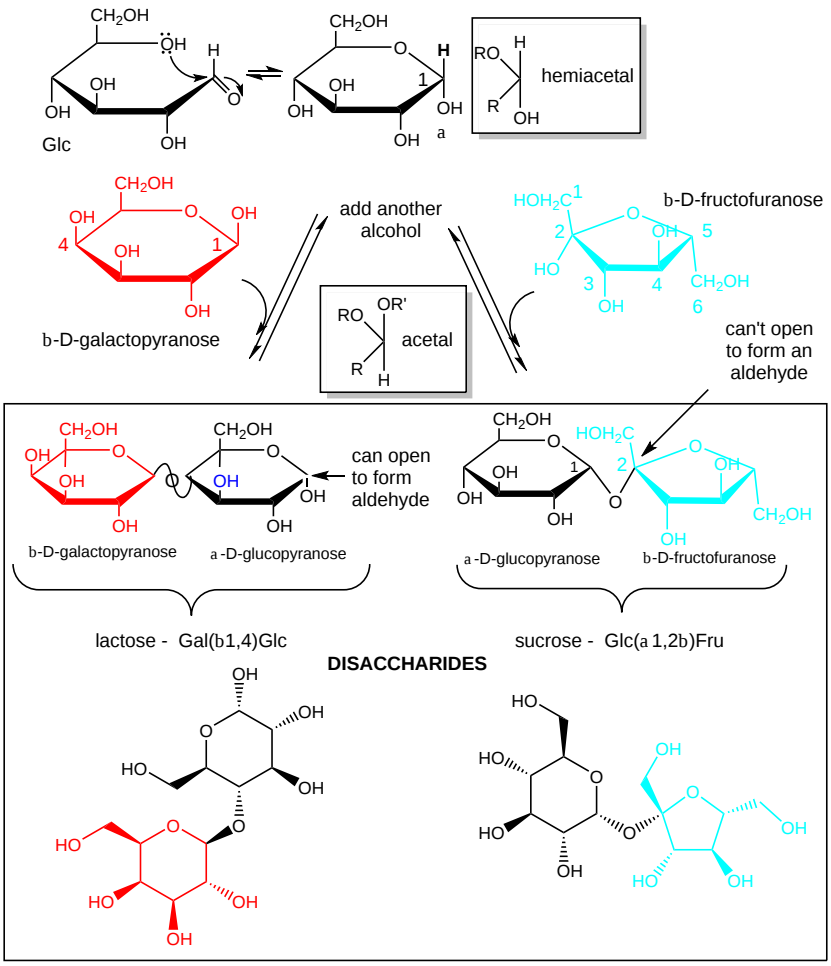


Figure 7.1.18: Structures of lactose and fructose

Acetal links between sugars in glycans can be hydrolyzed by water (catalyzed by H^+), as with the other key biological polymers, proteins, and nucleic acids.

The disaccharides described above that are linked through a 1,4 linkage are called reducing sugars since they can act as reducing agents in reactions in which they get oxidized. For example, in lactose, since galactose is attached to glucose through the OH on C4, the anomeric glucose carbon, C1, could revert to the noncyclic aldehyde form. This aldehyde is susceptible to oxidation by reagents (Benedict's Solution - with citrate, Fehling's reagent - with tartrate) as these reagents are subsequently reduced. In both reagents, reducing sugars reduce a basic blue solution of $CuSO_4$ (Cu^{2+}) to a brick-red precipitate of Cu_2O (Cu+). Sugars (monosaccharides, disaccharides and polysaccharides) that have a potentially open aldehyde at C1 or have an α-hydroxymethyl ketone group which can isomerize to an aldehyde under basic conditions (such as fructose) are called **reducing sugars**. These oxidizing agents are mild and react with aldehydes and not ketones.

If a monosaccharide, disaccharide, or even polysaccharide has at least one hemiacetal link (for instance, the second sugar in lactose), it is a reducing sugar, as the monomer with the cyclic hemiacetal can reversibly open to form an aldehyde. However, if the only links in sugars are full acetals (such as in sucrose when the link is between the two anomeric carbons), the sugar is not reducing.

Alpha-gal syndrome

Alpha-gal syndrome (AGS) is a relatively newly discovered disease caused by the bite of a tick. Tick saliva contains the disaccharide galactose-α-1,3-galactose (alpha-gal). After a tick bite, people develop an immune response to the disaccharide through IgE antibodies. Further bites could cause a mild rash up to an anaphylactic response.

What makes AGS worse is that red meat contains the disaccharide but is not found in fish, birds, or people. Hence, people who mount a strong IgE response to the disaccharide can also elicit the same response when they eat red meat or even drink cow's milk. Estimates that up to 450,000 people in the US may develop serious and even life-threatening symptoms after eating red meat.

The structure of Gal(α1,3)Gal is shown in Figure 7.1.19 below.

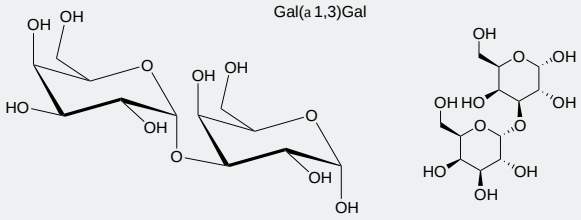


Figure 7.1.19: The structure of the disaccharide Gal(α1,3)Gal

Figure 7.1.20 shows an interactive iCn3D model of α-D-galactosyl-(1,3)-α-D-galactose (PubChem 9840966)

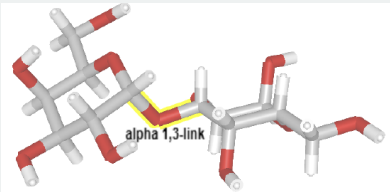


Figure 7.1.20: α-D-galactosyl-(1,3)-α-D-galactose (PubChem 9840966). (Copyright; author via source).
Click the image for a popup or use this external link: https://www.ncbi.nlm.nih.gov/Structu...5c4f77d7102ad6

Figure 7.1.21 shows an interactive iCn3D model of human anti-alpha-galactosyl antibody complex (7UEN)

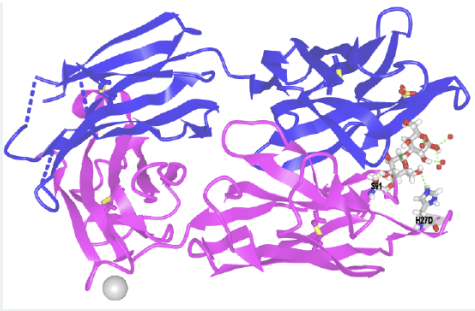


Figure 7.1.20: Human anti-alpha-galactosyl antibody complex (7UEN). (Copyright; author via source).
Click the image for a popup or use this external link: https://www.ncbi.nlm.nih.gov/Structu...57bc3eaca2a3f0

This antibody in this complex is an IgG, not an IgE. The Fab light chain fragment of the antibody is in magenta and the heavy chain Fab fragment in blue. Two side chains, H27 and S91 show interactions with the dissacharide.

Here is another iCn3D link that shows more interactions: https://www.ncbi.nlm.nih.gov/Structure/icn3d/share2.html?22f544b2836aadd2332a8ab3edd98be1. The gal residues are shown as yellow cubes in the SNFG coloring system. The abbreviation for Gal in alpha linkage in the PDB is **GLA.**

7.1.6: Summary

This chapter provides an in-depth exploration of the complexity of carbohydrate chemistry, emphasizing the structural, stereochemical, and functional diversity of sugars and glycans. It highlights why glycan biochemistry is particularly challenging due to the vast array of possible structures that arise from simple monosaccharide building blocks.

Key topics include:

1. **Terminology and Definitions:**
   The chapter begins by clarifying the terms "sugar," "carbohydrate," and "glycan." While "sugar" typically refers to small, low molecular weight carbohydrates like glucose, the term "carbohydrate" encompasses everything from simple sugars to complex polymers such as glycogen and cellulose. "Glycan" is used more generally to describe any polymer of sugars, whether free or attached to proteins and lipids.

2. **Monosaccharide Structures and Representations:**
   An overview of monosaccharide structures is provided, including the chemical basis of sugars as polyhydroxy-aldehydes or ketones. The chapter explains how simple sugars (e.g., trioses, tetroses, pentoses, hexoses) can cyclize to form stable 5-membered (furanose) or 6-membered (pyranose) rings. Various methods of structural representation are compared, such as Fischer projections, Haworth projections, and chair conformations, highlighting the stereochemical intricacies and the significance of α- and β-anomers.

3. **Isomerism and Structural Diversity:**
   The text reviews different types of isomers found in sugars, including configurational isomers (enantiomers, diastereomers, epimers, and anomers) and conformational isomers (chair and boat forms). This section underscores how subtle differences in stereochemistry can have profound effects on glycan function and recognition.

4. **Monosaccharide Derivatives and Modifications:**
   Beyond the basic sugar units, the chapter discusses various chemical modifications of monosaccharides, such as oxidation, phosphorylation, amination, acetylation, lactone formation, and complex condensation reactions that lead to the formation of important derivatives like sialic acids. These modifications are critical in determining the biological roles and properties of glycans.

5. **Glycosidic Bond Formation and Disaccharide Assembly:**
   The formation of glycosidic bonds via hemiacetal and acetal reactions is explained, along with the implications for disaccharide and polysaccharide synthesis. The chapter details how different linkages (e.g., α-1,4, β-1,4, α-1,2) lead to diverse structures, which in turn affect the reducing or nonreducing nature of the resulting carbohydrates.

6. **Biological Implications and Clinical Relevance:**
   Finally, the chapter explores the physiological significance of glycan diversity. An example is provided through alpha-gal

syndrome, an allergic response triggered by the disaccharide Gal(α1,3)Gal found in tick saliva and red meat. This section highlights how glycan structures can influence cell–cell interactions, pathogen recognition, and immune responses.

Overall, this chapter lays a foundation for understanding the intricate world of carbohydrate biochemistry by examining the molecular structures, stereochemical variations, and modifications that contribute to the functional complexity of glycans in biological systems.

7.2: Polysaccharides

Learning Goals (ChatGPT o1, 1/30/25)

- **Define and Differentiate Carbohydrates and Glycans:**
  - Distinguish among the terms “sugar,” “carbohydrate,” and “glycan,” and explain how these concepts encompass everything from simple monosaccharides to highly complex glycan polymers.
- **Understand Monosaccharide Structure and Representation:**
  - Describe the chemical structure of monosaccharides, including key features such as stereochemistry, the formation of cyclic (pyranose and furanose) structures, and the distinction between α- and β-anomers.
  - Convert between different structural representations (Fischer projections, Haworth projections, chair and wedge/dash forms) and understand how these representations inform our perception of 3D carbohydrate conformation.
- **Glycosidic Linkages and Polymer Structure:**
  - Explain the mechanism of glycosidic bond formation (hemiacetal and acetal formation) and describe how different linkage types (e.g., α 1,4, α 1,6, β 1,4) give rise to the structural diversity of polysaccharides.
  - Compare and contrast homopolymers such as starch, glycogen, dextran, cellulose, and chitin in terms of their linkage types, branching patterns, and overall 3D conformation.
- **Structure–Function Relationships in Storage vs. Structural Polysaccharides:**
  - Discuss how the type of glycosidic linkage (e.g., α in starch/glycogen versus β in cellulose/chitin) influences the polymer’s physical properties, biological roles (energy storage vs. structural support), and digestibility.
  - Analyze the role of branching (e.g., α 1,6 linkages in glycogen and amylopectin) in enhancing solubility and rapid mobilization of stored glucose.
- **Utilizing Symbolic Nomenclature for Glycans (SNFG):**
  - Learn and apply the SNFG system to accurately represent glycan structures using specific colors and shapes, facilitating clear communication of complex carbohydrate architectures.
- **3D Conformational Analysis and Modeling:**
  - Interpret interactive 3D models and computer-generated representations (e.g., iCn3D models) of polysaccharide fragments to understand their tertiary and quaternary arrangements, such as the helical structures of V-amylose or the fiber assembly in cellulose.
- **Exploration of Glycosaminoglycans (GAGs):**
  - Identify the structure and function of key GAGs (e.g., hyaluronic acid, keratan sulfate, chondroitin sulfate, dermatan sulfate, and heparin) and describe how variations in disaccharide repeat units and sulfation patterns contribute to their biological roles, such as in extracellular matrix composition and cell signaling.
- **Comparative Structural Analysis in Biological Contexts:**
  - Contrast the molecular structures of polysaccharides from different biological sources (e.g., cellulose in plants versus chitin in invertebrates) and discuss the impact of subtle structural differences (e.g., the presence of N-acetyl groups) on function.
- **Practical and Clinical Implications:**
  - Understand the physiological significance of polysaccharide structure in energy storage (glycogen and starch) and structural support (cellulose, chitin, and GAGs) and discuss how these structures influence their biochemical properties, such as solubility and reactivity.
  - Explore real-world applications, such as the use of iodine staining in starch detection, the role of glycogenin in glycogen synthesis, and the impact of dietary glycans on human health (e.g., Neu5Gc incorporation).

Achieving these goals will equip students with the fundamental and advanced concepts needed to analyze and interpret the structural and functional diversity of carbohydrates and glycans in biological systems.

Polysaccharides contain many monosaccharides in glycosidic links and may have many branches. They serve as either structural components or energy storage molecules. Polysaccharides consisting of single monosaccharides are homopolymers. Starch, glycogen, dextran, cellulose, and chitin are the most common. We'll discuss based on whether the acetal link is alpha or beta.

7.2.1: α 1,4 main chain links

Starch and **Glycogen**: These polysaccharides are polymers of glucose linked in α 1,4 links with α 1,6 branches. **Starch**, found in plants, is subdivided into **amylose**, which has no branches, and **amylopectin,** which does. Starch granules consist of about 20% amylose and 80% amylopectin. **Glycogen,** the main CHO storage in animals, is found in muscle and liver and consists of glucose residues in α 1,4 links with lots of α 1,6 branches (many more branches than in starch).

Here are various ways to render in 2D the chemical structure of a branched glycogen and starch fragment, as shown in Figure 7.2.1.

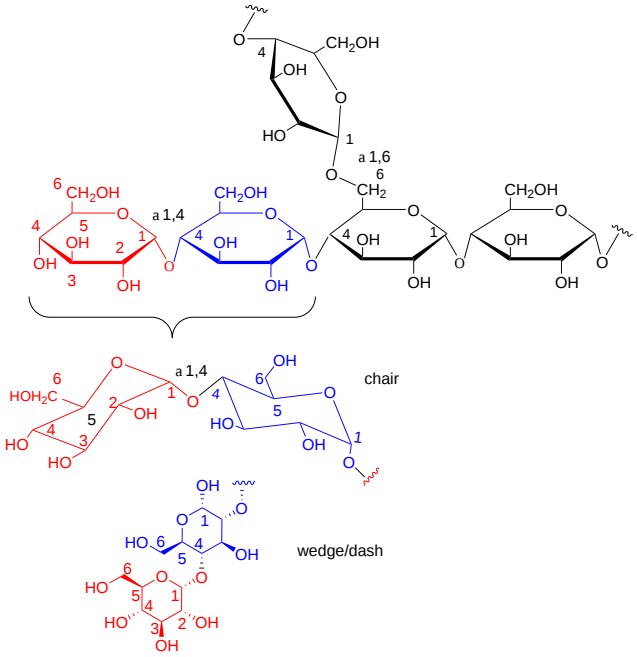


Figure 7.2.1: Structure of a branched glycogen and starch fragment

The top part of the figure shows the Haworth structure. The bottom part shows two glucose units in red and blue in the more structurally clear chair and wedge/dash representations.

Figure 7.2.2 shows an interactive iCn3D model of 10 glucose monosaccharides in an α-(1,4) linkage with five glucose units with α-(1,4) linkages attached to the main chain through α-(1,6) branch at glucose 6 of the main chain. The type of substructure would be found in starch (amylopectin) and glycogen.

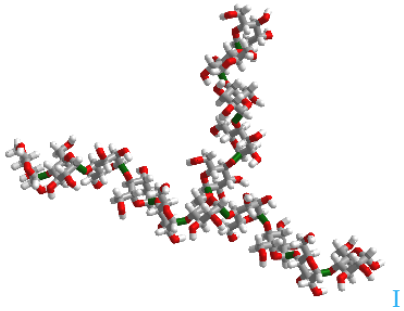


Figure 7.2.2: α-(1,4) linked glucose with an α-(1,6) branch (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...TXXy589xPNgmG9

Figure 7.2.3 the structure in the iCn3D model in a diagrammatic fashion in which glucose is represented as a blue circle with the acetal/glycosidic/glucosidic linkages between the monosaccharides written between the circles. The 14A label shows that the acetal linkage is an α-(1,4) link with a single α-(1,6) branch.

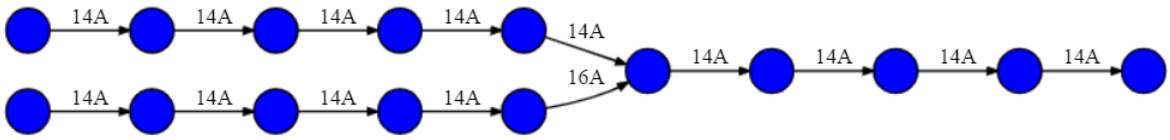


Figure 7.2.3: Symbolic Nomenclature For Glycans (SNFG) for an α-(1,4) linked glucose polymer with an α-(1,6) branch.

The linkages are written in a variety of conventions. These include 14A, 14α, 4A and 4α. The between many sugars is often a 1,x link where x is 2,3, 4, 5, or 6. In those cases, the number 1 can be omitted. The program-generated images in this text use numbers and A or B.

What makes carbohydrates so complex is their 3D structures. Like proteins and nucleic acids, they can adopt a myriad of conformations. As the monomeric units are so homogeneous, especially in homopolymers, it isn't easy to get crystal structures for them, so computer models are often used.

Studies have shown that the simple starch fraction amylose α 1,4 polymer of glucose, often envisioned as a straight chain, can adopt three main conformations. They are double-helical A- (found chiefly in cereals), double helical B-(found primarily on tubers) amyloses, and single-helical V-amylose (or simply A, B, and V structures). The A and B do **NOT** represent alpha or beta in this classification system. The A and B forms consist of double helices aligned in a parallel fashion with about six glucoses per turn. The helices appear to be left or right-handed, and this ambiguity might arise from a lack of crystal structures.

In contrast, a well-defined structure of the V helix is known. It folds into a left-handed helix with six glucoses per turn and a pitch of about 8Å. Unlike alpha helices of proteins and the double-stranded helix of DNA, the center of the helix is **NOT** packed tightly and can accommodate small molecules. One is iodine (triiodide, I_3^-), which exhibits a dark blue color when bound in amylose in starch. This is the basis for starch indicators that you may have used in titration reactions in biology and chemistry courses.

Alpha helices might self-associate during folding in proteins to form a 4-helix bundle. Likewise, the helices in V-amylose can associate into bundles. Figure 7.2.4 shows an interactive iCn3D model of the actual structure of a V-amylose, cycloamylose 26 (1C58). It consists of a linear cycloamylose strand of 26 glucose monomers, which has collapsed to form secondary structure with 6-residue helices packed together into a tertiary structure of 4-helix bundle. The blue sphere "cartoon" color coding of each glucose residue corresponds to the blue circles in the diagrammatic representation above.

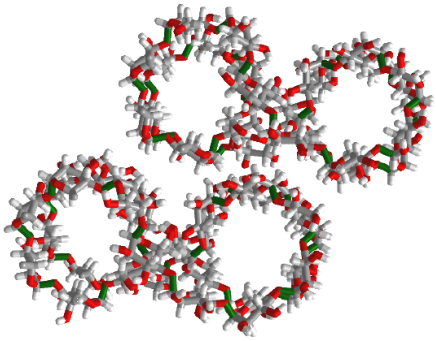


Figure 7.2.4: V-amylose, cycloamylose 26 (1C58). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...6tghM5HMTypm88

Rotate the model to explore it. Trace the chains by following the blue sphere symbolic representation of glucose as you trace the main chain. Rotate it to view down the helix axes to see the four holes that can each accommodate a I_3^-. In the menu button (☰), choose Style, Chemicals, Sphere to see a spacefilling model that shows the holes within each helix. Remember, NO unoccupied holes exist within either a protein alpha helix or a double-stranded DNA molecule.

The well-known macrocyclic compound cyclodextrins (for example, α-cyclodextrin) are structures equivalent to one turn of V-amylose. The V-amylose helix is partially stabilized by hydrogen bonds from donors and acceptors within the helix from the OH3 on the i^{th} glucose and the OH2 on the i^{th}+1 glucose as well as from the OH6 on i^{th} glucose and OH2 on the i^{th} + 6 glucose.

In vivo, glycogen is synthesized by attaching glucose monomers to a core protein called glycogenin. Figure 7.2.5 shows a model of a glycogen particle with glycogenin at its core.

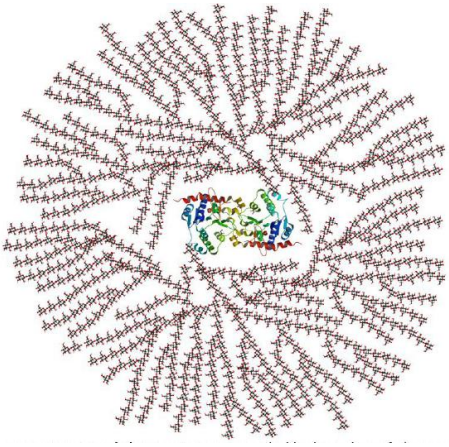


Figure 7.2.5: Glycogen particle with glycogenin at its core

The dimeric protein glycogenin is an enzyme that autoglucosylates itself stepwise. The first glucose is added at Tyr 195. At some point, the active site must get buried and the protein can no longer add more monomers.

Storing glucose residues as either glycogen or starch, one large molecule, makes chemical sense. A review of colligative properties would inform you that if glucose was stored as the monosaccharide, a great osmotic pressure difference would be found between the outside and inside of the cell. Glycogen, with its many branches, is a single molecule. When glucose is needed, it is cleaved one residue at a time from all the branches (at the nonreducing ends) of glycogen, producing a large amount of free glucose quickly.

Phi/Psi angles can also be used in Ramachandran plots to show the conformations around the acetal link for the starch/glycogen main chain in a fashion comparable to that for proteins (around the alpha carbon). The phi torsion angle describes rotation around the C1-O bond of the acetal link, and the psi angle describes rotation around the O-C4 bond of the same acetal link, with the glucopyranose ring considered as a rigid rotator (just as the 6 atoms in the planar peptide bond unit). The most extended form of a glucose polymer occurs when the glycosidic link is β 1,4 (as in cellulose), which forms linear chains. This would be analogous to the more extended parallel beta strand (phi/psi angles of -119^0, -113^0) and antiparallel beta strands (phi/psi angles of -139^0, $+135^0$) of proteins. The α 1,4 linked main chain of glycogen and starch causes the chain to turn and form a large helix. Iodine (or I_3^-) can fit into the helix, which turns a solution/suspension of starch blue, which turns starch purple. The less extended structure is analogous to the less extended protein alpha helix, which has phis/psi angles of -57^0,-47^0.

Figure 7.2.6 shows phi/psi angles for acetal/glycosidic linkage in maltose, a disaccharide of glucose, as shown below.

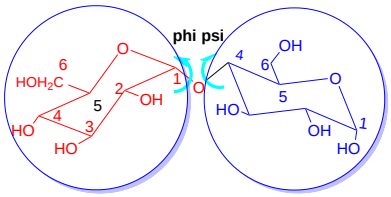


Figure 7.2.6: Phi/psi angles for acetal/glycosidic linkage in the glucoside disaccharide maltose

7.2.2: α 1,6 main chain links

Dextran is a branched polymer of glucose in α 1,6 links with α 1,2, α 1,3, or α 1,4 linked side chain. This polymer is used in some chromatography resins. Figure 7.2.7 shows chair structures (A) and wedge/dash structures (B) for dextran, showing the main chain α 1,6 link with one α 1,3 branch.

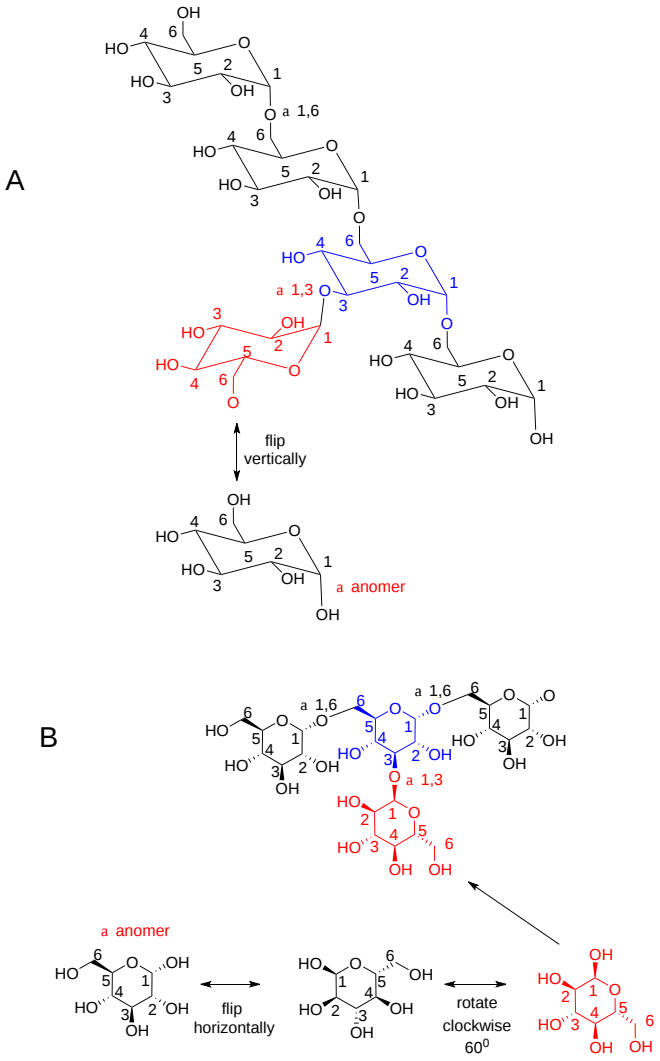


Figure 7.2.7: Chair structures (A) and wedge/dash structures (B) for dextran showing the main chain α 1,6 link with one α 1,3 branch

Depending on its molecular weight, it is soluble in water (forming viscous solutions) and organic solvents. It is also used as a food thickener and stabilizer. It is synthesized by lactic acid-forming bacteria using sucrose as an energy source. Most uses are commercial.

7.2.3: β 1,4 links

Cellulose, a structural homopolymer of glucose in plants, has β 1,4 main chain links without branching. Multiple chains are held together by intra and inter-chain H-bonds. It is the most abundant biological molecule in nature. Various renderings of 4 glucose residues in cellulose are shown in Figure 7.2.8. Haworth structures are not shown. Instead, more chemically informative chairs and wedge/dash structures are used. It's important to see the structures displayed in many ways since different representations of carbohydrate structures can be found in different sources.

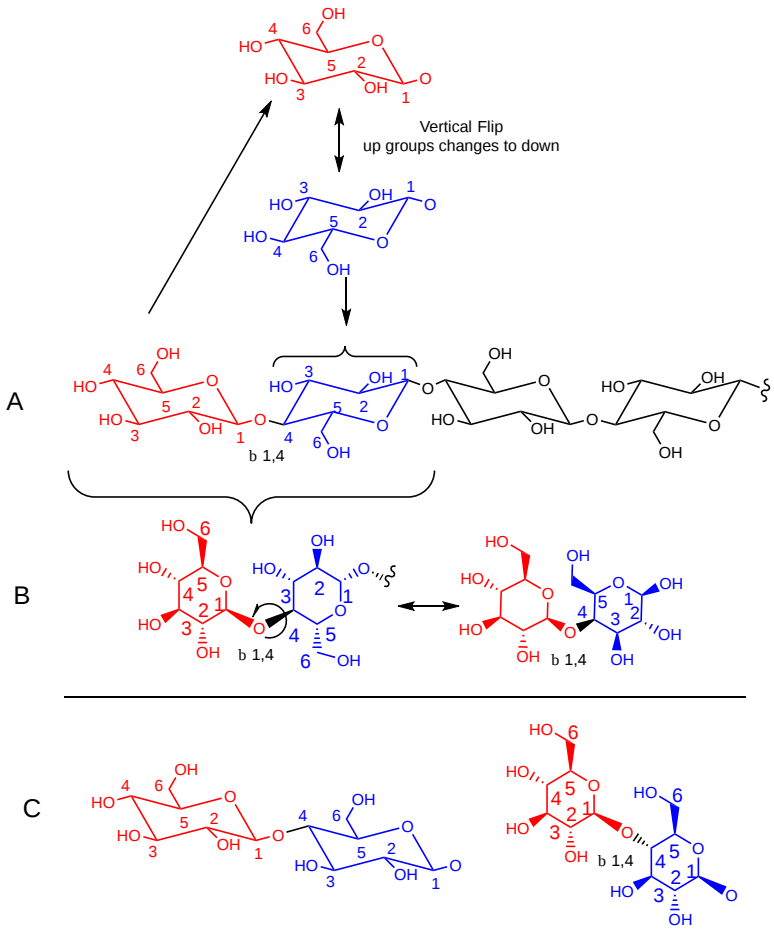


Figure 7.2.8: Rendering of 4 glucose residues in cellulose

In A, the most common chair representation, the 2nd and 4th residues from the right-hand end are flipped versions of residues 1 and 3. Residues 1 and 2 are colored red and blue for clarity. This unit is repeated to generate the full chain. The top part of A shows a simplified version of the flip of the red ring to produce the blue ring to help you see that they are indeed identical structures.

The same structure as in A is shown in the left part of B in wedge/dash from (looking down on the ring). The right-hand side of B shows a variant of B's left-hand side generated by simple 180^0 rotation around the bond indicated in the left of B.

The simple repeat is shown in C without the chain flips in A and B. The acetal/glycosidic/glucosidic bond seems to be shown in a straight line in the chair structures (a bit confusing and structurally deceptive) but more clearly in the adjacent wedge/dash structure.

All structures are correct, but the one shown in A is most often used.

One long chain can interact with other chains in a structure stabilized by intrachain and interchain hydrogen bonds. Different sources display different hydrogen bonds. Some common ones are shown below. These chains align in parallel and twist to form larger cellulose fibers. Figure 7.2.9 shows an interactive iCn3D model of cellulose chains.

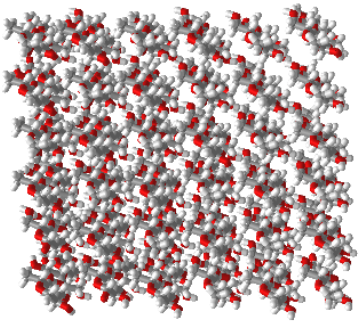
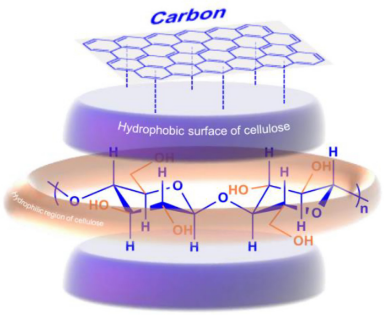


Figure 7.2.9: Cellulose chains (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...TUHKki2VN4jeTA

In addition, hydrophobic interactions occur between adjacent planes of cellulose strands. How can that be considering the strongly polar nature of glucose and a single cellulose strand? Figure 7.2.9*B* below, with two representations of cellulose, shows how axial hydrogens project vertically from the plane of the glucose monomer rings to form a weak but biologically significant hydrophobic surface.



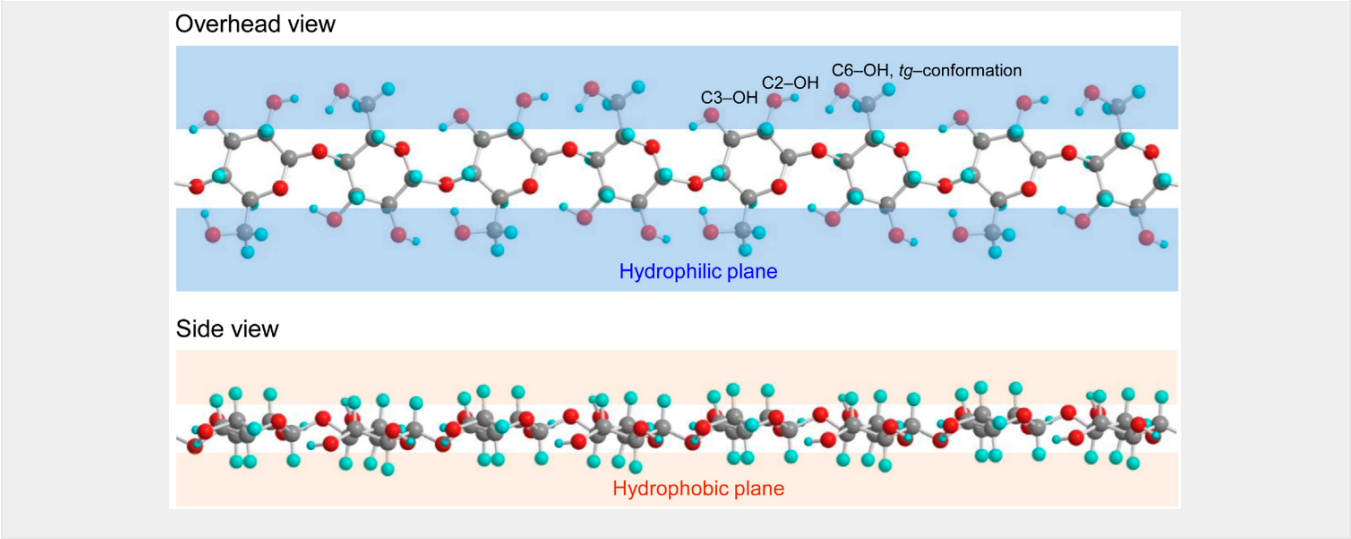


Figure 7.2.9: Two representations showing the nonpolar axial surface of cellulose strands that contribute to hydrophobic interactions stabilizing the cellulose assemblies. The top panel shows how cellulose could interact with a nonpolar planar substance, such as a layer of graphite. Axial H atoms are shown in green/cyan in the bottom panel.

Top Panel: Yang G, Luo X and Shuai L (2021) Bioinspired Cellulase-Mimetic Solid Acid Catalysts for Cellulose Hydrolysis. *Front. Bioeng. Biotechnol.* 9:770027. doi: 10.3389/fbioe.2021.770027. **Creative Commons Attribution License (CC BY).**

Bottom Panel: Uusi-Tarkka, E.-K.; Skrifvars, M.; Haapala, A. Fabricating Sustainable All-Cellulose Composites. *Appl. Sci.* **2021**, *11*, 10069. https://doi.org/10.3390/app112110069. Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Chitin

The glycan is the major component in the exoskeletons of anthropoids and mollusks. It is a β 1,4 linked polymer of N-acetylglucose (GlcNAc). Compare this to cellulose, which is a β 1,4 linked polymer of glucose. What a difference an N-acetyl substituent makes!

The basic chemical structure of chitin is shown in chair form in Figure 7.2.10 along with the symbolic nomenclature for glycans (SNFG).

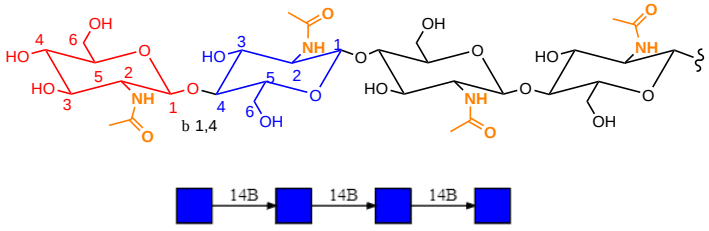


Figure 7.2.10: Chitin - Chain and symbolic nomenclature for glycans SNFG)

7.2.4: Symbolic nomenclature for glycans (SNFG) -

Before we go further into the complexities of glycan structure, let's explore the symbolic nomenclature for glycan structures. The Consortium for Functional Glycomics (2005) proposed a scheme based on specific colored geometric shapes for each, as shown for the example glycan shown in Figure 7.2.11 for a complex glycan.

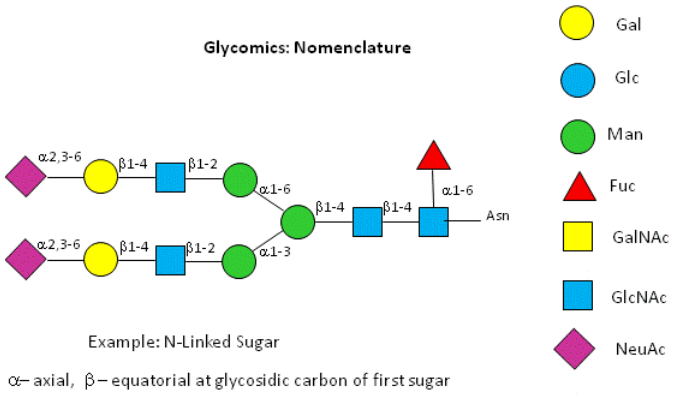


Figure 7.2.11: SNFG representation of a complex glycan

This nomenclature has recently been updated in Appendix 1B of Essentials of Glycobiology, 3rd Edition (Glycobiology 25(12): 1323-1324, 2015. doi: 10.1093/glycob/cwv091 (PMID 26543186) and is summarized in the Figure 7.2.12

SHAPE	White (Generic)	Blue	Green	Yellow	Orange	Pink	Purple	Light Blue	Brown	Red
Filled Circle	Hexose	Glc	Man	Gal	Gul	Alt	All	Tal	Ido	
Filled Square	HexNAc	GlcNAc	ManNAc	GalNAc	GulNAc	AltNAc	AllNAc	TalNAc	IdoNAc	
Crossed Square	Hexosamine	GlcN	ManN	GalN	GulN	AltN	AllN	TalN	IdoN	
Divided Diamond	Hexuronate	GlcA	ManA	GalA	GulA	AltA	AllA	TalA	IdoA	
Filled Triangle	Deoxyhexose	Qui	Rha			6dAlt		6dTal		Fuc
Divided Triangle	DeoxyhexNAc	QuiNAc	RhaNAc							FucNAc
Flat Rectangle	Di-deoxyhexose	Oli	Tyv		Abe	Par	Dig	Col		
Filled Star	Pentose		Ara	Lyx	Xyl	Rib				
Filled Diamond	Nonulosonate		Kdn				Neu5Ac	Neu5Gc	Neu	
Flat Hexagon	Unknown	Bac	LDManHep	Kdo	Dha	DDManHep	MurNAc	MurNGc	Mur	
Pentagon	Assigned	Api	Fru	Tag	Sor	Psi				

Figure 7.2.12: Symbolic Nomenclature for Glycans (SNFG) for the most common monosaccharides

7.2.5: Glycosaminoglycans - Heteropolysaccharides with disaccharide repeating units

Many polysaccharides consist of repeating disaccharide units. A major class of polysaccharides with disaccharide repeats include the **glycosaminoglycans** (**GAGs**), all of which contain one amino sugar in the repeat and in which one or both of the sugars contain negatively charged sulfate and/or carboxyl groups. The extent and position of sulfation vary widely between and within GAGs. GAGs are found in the vitreous humor of the eye and synovial fluid of joints, as well as in connective tissue like tendons, cartilage, etc., as well as skin. They are found in the extracellular matrix and are often covalently attached to proteins to form proteoglycans. From a bird's eye view, they are all elongated polyanions.

They and their structures are very complicated and exceedingly diverse. This makes it difficult for those who want unambiguous structures. From a biological perspective, they present in their local environment an incredibly diverse array of potential binding sites for ligands (both small and large). Because of these, they also have functions in cell signaling. In addition, some GAGs are free-standing, others are covalently attached to proteins (like glycogen is attached to glycogenin). These large molecules are called **proteoglycans**. We will discuss this later in Chapter 7.4 when we discuss the "carbohydrate code."

Here are the ring structures and descriptions of important GAGs. The common disaccharide repeat unit is shown twice for each structure, with the knowledge that sulfation patterns may differ for the disaccharide repeats in the actual chains. Note also that the first member of each disaccharide repeat shows the ring flipped vertically (top to bottom) as was shown in the structures for other beta-linked glycans (cellulose, chitin) above.

In a long chain, selecting which is the repeating disaccharide unit is a bit relative, as shown in Figure 7.2.13 for the repeating disaccharide sequence of N-acetylglucosamine (blue square) and N-acetylgalactosamine (yellow square).

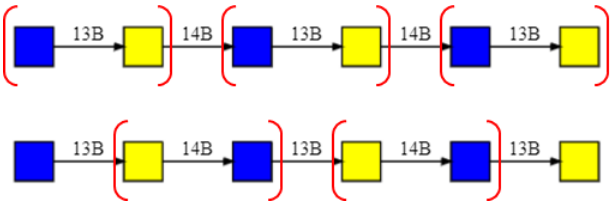


Figure 7.2.13: SNFG representation of a disaccharide sequence of N-acetylglucosamine (blue square) and N-acetylgalactosamine (yellow square)

At the top, the repeating units (blue-yellow) are connected through beta 1,4 links, while at the bottom, the connection of the repeating unit (yellow-blue) is beta 1,3. The best choice of annotating the repeating unit without knowing the full chain is elusive. What's most important, however, is to note the alternating acetal/glycosidic links throughout the whole sequence. In the figures below, different disaccharide repeats are highlighted.

Hyaluronic acid

This is a polymer of glucuronate (β 1,3) GlcNAc. It offers a backbone for the attachment of protein and other GAGs. It's the only GAG without sulfate. Figure 7.2.14 shows a tetrasaccharide fragment with two disaccharide repeats. The illustrated disaccharide repeat's internal acetal/glycosidic link is β 1,3, while the disaccharide's connection is β 1,4. For one last time, the vertical flip of the glucuronic acid is shown to allow a better understanding of its flipped presentation in the actual GAG.

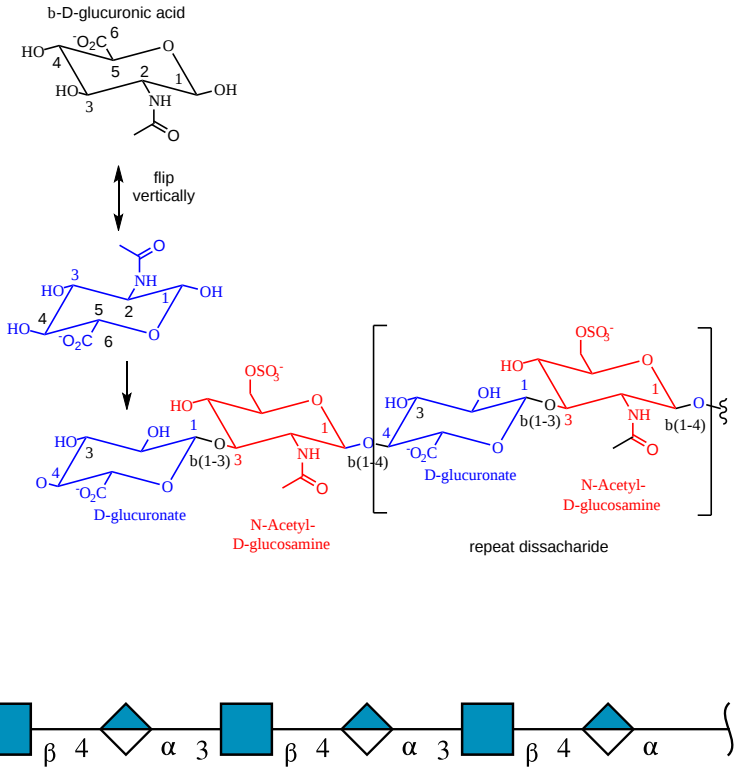


Figure 7.2.14: Hyaluronic acid - tetrasaccharide fragment with two disaccharide repeats

Hyaluronic acids are found in a variety of locations, including synovial fluid, the extracellular matrix, and skin, where they help control skin moisture. It is water soluble and displays twin antiparallel left-oriented helices. Covalent conjugates of the chemotherapeutic drugs doxorubicin and camptothecin linked to hyaluronic acid, whose overall structure is similar to "worm-like micelles", have been used successfully to treat skin cancers.

Figure 7.2.15 shows an interactive iCn3D model of hyaluronic acid (4HYA). The dotted lines represent hydrogen bonds.

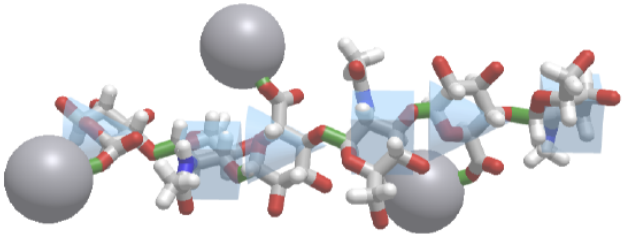


Figure 7.2.15: hyaluronic acid (4HYA). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...E7Y2GkhcxRR6n8

Three calcium ions are shown as well. The SNFG cartoon is also illustrated.

Keratan sulfate

This GAG contains repeats of N-acetyl-D-glucosamine-6-phosphate in β 1,3 link to D-galactose or D-galactose-6-sulfate. The link between Gal and the modified glucosamine is β 1,4. Keratin sulfate is most abundant in the cornea of the eye, but is also found in other connective tissues such as bone, cartilage, and tendon, as well as in the central and peripheral nervous system.

Figure 7.2.16 shows a tetrasaccharide containing two repeating disaccharides.

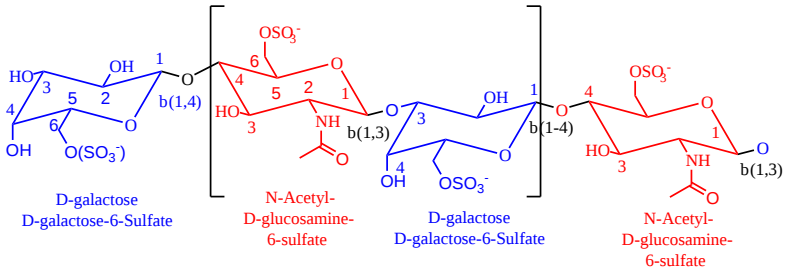
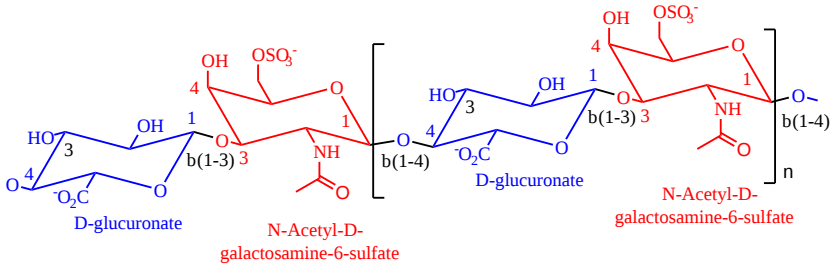


Figure 7.2.16: Keratan sulfate disaccharide repeats

Chondroitin sulfate

The repeat disaccharide unit is D-glucuronate β(1,4) GalNAc-4 or 6-sulfate. It's found in connective tissue matrix, the cell surface (in the form of proteoglycans), basement membranes, and intracellular granules. Figure 7.2.17 shows a tetrasaccharide showing two disaccharide repeats.



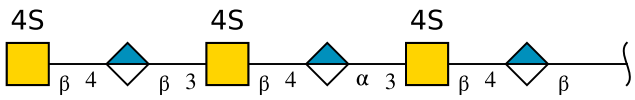


Figure 7.2.17: Chondroitin sulfate disaccharide repeats

Figure 7.2.18 shows an interactive iCn3D model of chondroitin-4-sulfate (1C4S). The dotted lines represent hydrogen bonds.

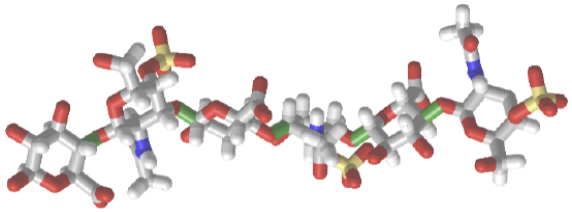


Figure 7.2.18: Chondroitin-4-sulfate (1C4S). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...vCXygXckaAVQV8

Dermatan sulfate

This glycosaminoglycan is similar to chondroitin sulfate. It is first made as a polymer of the disaccharide unit of D-gluconic and N-acetyl-D-galactosamine. The gluconic acid is epimerized to L-iduronic acid, followed by sulfation. Its structure is shown in Figure 7.2.19

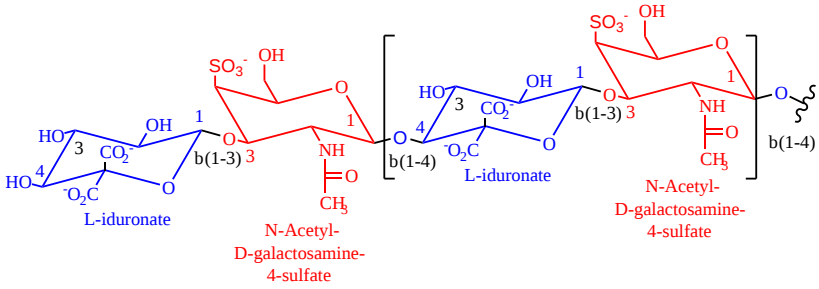


Figure 7.2.19: Dermatan sulfate disaccharide repeats

Heparin

This GAG contains a highly trisulfated disaccharide repeat, as shown in Figure 7.2.20. Note that the molecule can contain glucuronate or iduronate, and the degree of sulfation of the chains varies. Remember, no genetic code specifies these polymers' sequence or sulfation pattern.

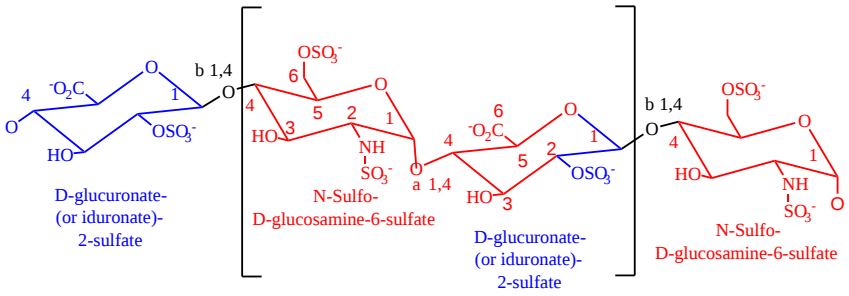


Figure 7.2.20: Heparin disaccharide repeats

Figure 7.2.21 shows an interactive iCn3D model of the solution structure of an 18-mer of heparin (3IRI). The dotted lines represent hydrogen bonds.

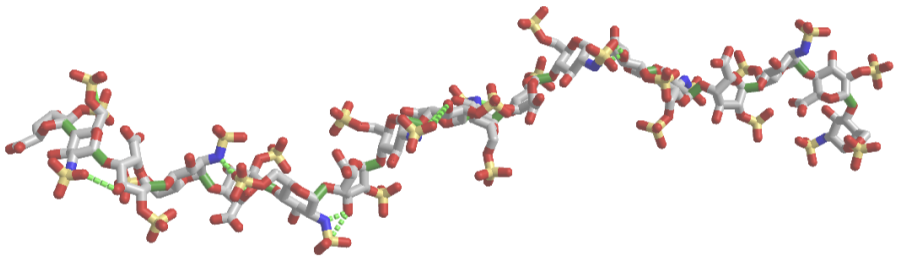


Figure 7.2.21: solution structure of an 18-mer of heparin (3IRI). (Copyright; author via source). Click the image for a popup or use this external link:

Most people are familiar with the anti-clotting properties of heparin administered as a drug. Heparin acts as a "catalyst" to accelerate the inhibition of the enzyme thrombin, which cleaves fibrinogen and activates platelets to form clots by the blood protein antithrombin III. Heparin works in two ways to facilitate thrombin inactivation. It has a specific binding site for antithrombin III, which causes a conformation in the protein, making it a more effective inhibitor. Thrombin, a positively-charged serine protease, can nonspecifically bind the heparin, a polyanion. When it does, it diffuses along the heparin chain, where it can find bound antithrombin III much quicker than if the inhibitor was free in the blood. Heparin effectively changes the search path of thrombin from a 3D to a 1D search.

Figure 7.2.22 shows an interactive iCn3D model of the amino acids in antithrombin III within seven angstroms of a bound heparin 5mer (1NQ9). Dotted lines represent hydrogen bonds and salt bridges between the two. Heparin is highlighted in yellow

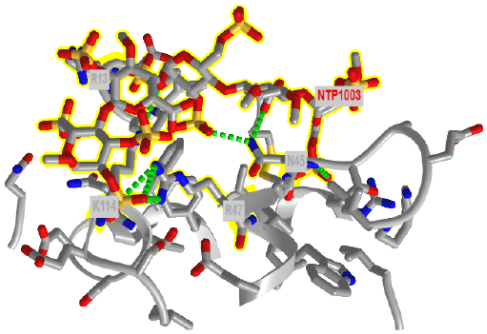


Figure 7.2.22: Heparin 5mer - antithrombin III complex (1NQ9. (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...fnvt6YfbJovat8

7.2.6: Agarose

Agarose is the main polysaccharide component derived from red algae. Agarose is a polymer of a disaccharide repeat of (1,3)-β-D-galactopyranose-(1,4)-3,6-anhydro-α-L-galactopyranose, is often used for a gelable solid phase for electrophoresis of nucleic acid and as a component of chromatography beads. As with starch, a mixture of amylose and amylopectin, agarose is often found with agaropectin, a sulfated galactan. A tetrasaccharide fragment with two disaccharide repeats is shown in Figure 7.2.23

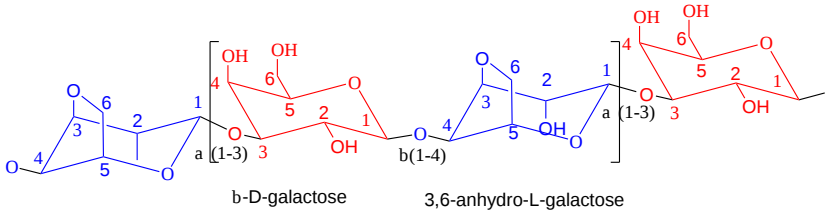


Figure 7.2.23: Agarose disaccharide repeats

Figure 7.2.24 shows an interactive iCn3D model of the agarose double helix (1AGA). The dotted lines represent hydrogen bonds within the polymer.

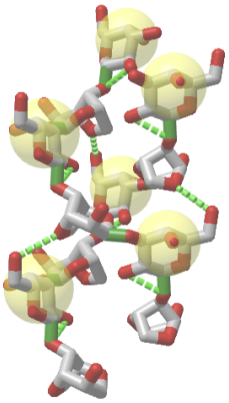


Figure 7.2.24: Agarose double helix (1AGA). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...5nnzm9XcPogMz7

In the iCn3d model, choose **Style, Glycan, Show Cartoon** to see the yellow sphere for galactose.

7.2.7: Summary

This chapter delves into the intricate world of carbohydrate biochemistry, emphasizing the remarkable complexity and diversity of polysaccharides. It highlights the challenges in studying glycans due to the stereochemical complexity of monosaccharides, the variability of glycosidic linkages, and the lack of a direct genetic template for glycan synthesis. Key themes include:

1. **Fundamental Building Blocks:**
   - The chapter begins by reviewing the structure of monosaccharides, including their linear (Fischer projections) and cyclic forms (Haworth, chair, and wedge/dash representations). Emphasis is placed on the formation of pyranose and furanose rings and the significance of α and β anomers, which influence both structure and function.
2. **Glycosidic Linkages and Polysaccharide Assembly:**
   - Glycosidic bonds are the chemical linkages that join monosaccharide units into polymers. The discussion covers the differences between α and β linkages and their consequences for the overall conformation of the polysaccharide.
   - Homopolymers such as starch, glycogen, dextran, cellulose, and chitin are introduced. Starch and glycogen, composed of glucose units linked by α 1,4 bonds with branching via α 1,6 linkages, serve as energy storage molecules. In contrast, cellulose and chitin, linked by β 1,4 bonds, form rigid, linear structures that provide mechanical support in plant cell walls and exoskeletons, respectively.
3. **Structural Representations and Models:**
   - Various methods for depicting carbohydrate structures are discussed, including two-dimensional diagrams (e.g., Haworth and Fischer projections) and symbolic nomenclature (SNFG) that uses colored shapes to represent different monosaccharides.
   - The chapter utilizes interactive 3D models to illustrate the spatial arrangement of branched glycogen and starch fragments, emphasizing how subtle differences in linkage and branching affect three-dimensional conformation and biological function.

4. **Functional Implications:**
    - The structural diversity of polysaccharides is directly linked to their biological roles. Energy storage polysaccharides like glycogen and starch must be compact and highly branched to allow rapid mobilization of glucose, while structural polysaccharides like cellulose and chitin form extensive networks stabilized by hydrogen bonding and hydrophobic interactions.
    - The chapter also touches on glycosaminoglycans (GAGs), which are heteropolysaccharides with repeating disaccharide units. These molecules, often attached to proteins as proteoglycans, play critical roles in cellular signaling and the formation of the extracellular matrix.
5. **Biological and Clinical Relevance:**
    - Understanding the complexity of carbohydrate structures is essential for insights into cellular function, tissue architecture, and even clinical conditions. For example, variations in sialic acid structures can influence cell recognition and immunity, while dietary components such as glycans impact overall health.

Overall, this chapter provides a comprehensive overview of the chemical, structural, and functional properties of polysaccharides. By integrating detailed structural representations with biological context, it lays the groundwork for further study in glycobiology and its applications in health and disease.

7.3: Glycoconjugates - Proteoglycans, Glycoproteins, Glycolipids and Cell Walls

Learning Goals (ChatGPT o1, 1/30/25)

- **Define Key Terminology and Concepts:**
  - Differentiate among the terms “sugar,” “carbohydrate,” and “glycan” and explain how each applies to both free and protein-conjugated carbohydrates.
  - Understand the structural diversity introduced by glycosylation, and why glycans are considered “information rich” compared to linear biopolymers like DNA and proteins.
- **N-linked Glycoprotein Structure and Diversity:**
  - Describe the core oligosaccharide structure (Man)3(GlcNAc)2(\text{Man})_3(\text{GlcNAc})_2 attached to Asn in the consensus sequence (X-Asn-X-Thr/Ser) and its significance in N-linked glycoprotein formation.
  - Differentiate between high-mannose, complex, and hybrid N-linked glycans by comparing their outer sugar compositions (e.g., presence or absence of mannose, lactosamine repeats, or branching patterns) and discuss their functional implications.
  - Utilize SNFG (Symbolic Nomenclature For Glycans) to interpret and represent glycan structures and variations.
- **O-linked Glycoproteins and Blood Group Antigens:**
  - Explain the process of O-linked glycosylation where carbohydrates (commonly Gal(β1,3)GalNAc) attach to Ser/Thr residues, and relate this to the structure and function of blood group antigens.
  - Recognize how variations in O-linked glycan structures contribute to cell–cell recognition and immune responses.
- **Proteoglycans and the Extracellular Matrix (ECM):**
  - Describe the structure of proteoglycans, including the linkage of glycosaminoglycans (GAGs) to core proteins, and differentiate between soluble and membrane-bound forms.
  - Explain the roles of proteoglycans in the ECM, including their contribution to tissue structure, cellular adhesion, and signal transduction.
- **Carbohydrate Structures in Cell Walls and Membranes:**
  - Compare and contrast the composition and structure of cell walls in different organisms:
    - For Gram-positive bacteria: Describe the peptidoglycan network, including the roles of NAG and NAM, pentapeptides, and teichoic acids.
    - For Gram-negative bacteria: Explain the arrangement of the peptidoglycan layer, outer membrane, and the role of lipopolysaccharides (LPS) in antigenicity and antibiotic resistance.
    - For plants: Summarize the components of primary and secondary cell walls (cellulose, hemicellulose, pectin, and lignin) and discuss how these polymers contribute to cell rigidity and growth.
  - Discuss the unique features of archaeal cell membranes and walls, including differences in lipid composition and the absence of peptidoglycan.
- **Biological and Clinical Implications of Glycosylation:**
  - Evaluate the functional roles of glycosylation in protein folding, stability, proteolytic protection, and in modulating protein half-life.
  - Analyze how specific glycan structures (e.g., Gal(α1,3)Gal) affect immune recognition and their implications in xenotransplantation and diseases such as alpha-gal syndrome.
  - Discuss how differences in glycosylation patterns contribute to cell signaling and tissue-specific functions, particularly within the ECM and basement membranes.

By achieving these goals, students will gain a comprehensive understanding of the multifaceted roles that carbohydrate modifications play in protein function and cellular organization, and they will be prepared to explore further applications of glycosylation in health, disease, and biotechnology.

Many proteins, especially those destined for secretion or insertion into membranes, are post-translationally modified by the attachment of carbohydrates. They are usually attached through either Asn or Ser side chains. Carbohydrate modifications on the protein appear to be involved in recognizing other binding molecules, preventing aggregation during protein folding, protecting from proteolysis, and increasing the proteins' half-life. In contrast to a protein sequence determined by a DNA template, sugars are attached to proteins by enzymes that recognize appropriate sites on proteins and attach the sugars. Since there are many sugars with many functional groups that can serve as potential attachment sites, the structures of the oligosaccharides attached to proteins are enormously varied, complex, and hence "information rich" compared to linear or folded polymers like DNA and proteins.

7.3.1: N-linked Glycoproteins

These contain carbohydrates attached through either a GlcNAc or GalNAc to an Asn in a X-Asn-X-Thr sequence of the protein. There are three types of N-linked glycoproteins, high mannose, complex, and hybrid. They all contain the same core oligosaccharide - (Man)3(GlcNAc)2 attached to Asn as shown in Figure 7.3.1.

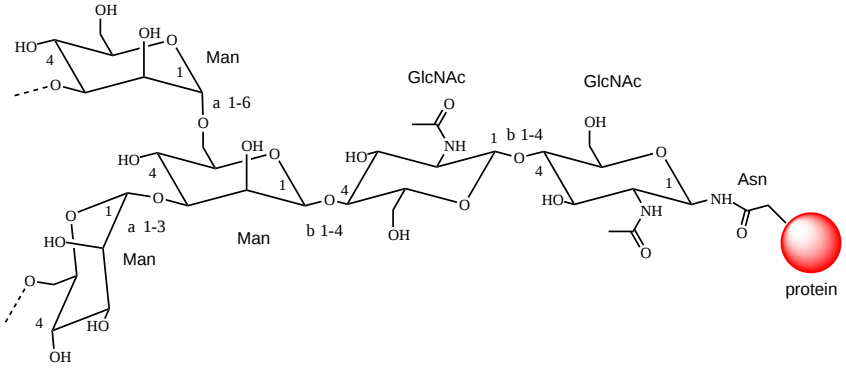
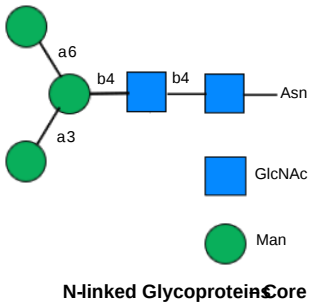
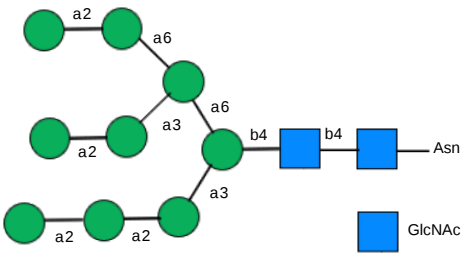


Figure 7.3.1: Core oligosaccharide - (Man)3(GlcNAc)2 attached to Asn in N-linked glycoproteins

Table 7.3.1 below shows the SNFG representation for the main core and variant glycans of **N-linked glycoproteins**. Note that the designation of α2 implies an α(1→2) linkage. Unless otherwise stated, the linkage is presumed to start from carbon 1.

Core	 N-linked Glycoproteins Core
High mannose	 N-linked Glycoproteins High Man

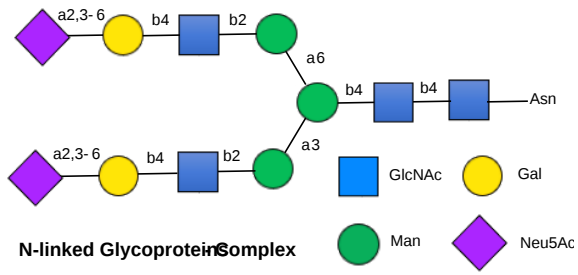
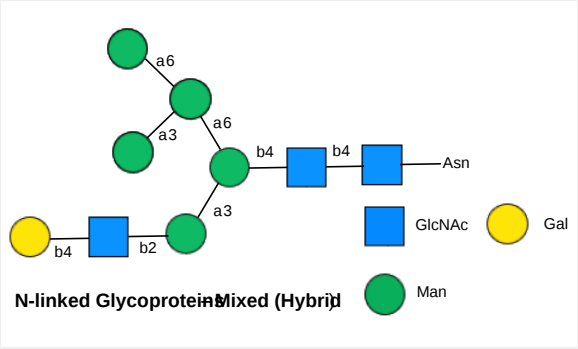
Complex	
Mixed hybrid	

Table 7.3.1: SNFG representation for the main core and variant glycans in **N-linked glycoproteins**

Complex N-linked glycans don't contain mannose outside the core glycan and have GlcNAc attached to the branching mannoses in the core structure. The complex glycan shown above has a Gal(β1,4)GlcNAc sequence, which could be named the disaccharide lactosamine. Often, lactosamine repeats in the sequence.

Hybrid glycans have both unsubstituted terminal mannoses (as in the high-mannose type) and substituted mannoses with an N-acetylglucosamine attached (as in the complex type. GlcNAc residues added to the core in the hybrid and complex N-glycoproteins are called antennae. Figure 7.3.2 shows an example of a biantennary N-linked glycan with two GlcNAc branches linked to the core. The core is outlined in red, and the two GlcNAcs are labeled 1 and 2.

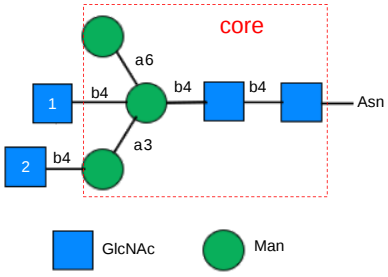


Figure 7.3.2: Biantennary N-linked glycan with two GlcNAc branches linked to the core

Complex glycans also have bi-, tri-, and tetraantennary forms and comprise most N-glycans. As shown in Table 7.3.1 above, complex N-linked glycans usually end with sialic acid residues. About 50% of the surface area of the COVID Sars-2 spike protein is covered with glycans, as shown in the model structure in Figure 7.3.3. The protein surface is gray, and the glycans (biantennary LacNAc N-glycans) in spacefill CPK with carbon in cyan.

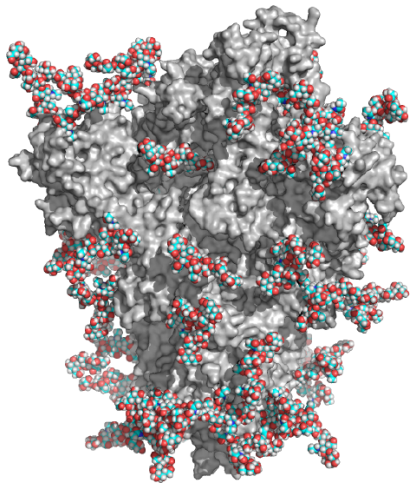


Figure 7.3.3: COVID Sars-2 spike protein is covered with glycans (colored spacefill). (PDB file coordinates 5.Swiss.3.M3F1.CYX.TER from Analysis of the SARS-CoV-2 spike protein glycan shield: implications for immune recognition. Oliver C. Grant, David Montgomery, Keigo Ito, Robert J. Woods. **doi:** https://doi.org/10.1101/2020.04.07.030445)

In the hybrid oligosaccharide shown above, one terminus contains Gal(β1,4)GlcNAc. However, in all other mammals except man, apes, and old-world monkeys, an additional Gal is often connected in an α1,3 link to the Gal to give a terminus of Gal(α1,3)Gal(β1,4)GlcNAc. These animals have an additional enzyme, an α1,3 Gal transferase. Bacteria also have this enzyme, and since we have been exposed to this link through bacterial infection, we mount an immune response against it. Why is this important? Pig hearts are similar to human hearts, so they might be good candidates for human transplantation (xenotransplants). However, the Gal-α1,3-Gal link is recognized as foreign, and we mount a significant immune response against it. Several biotech firms are trying to delete the pig α1,3 Gal transferase, which would prevent the addition of the terminal Gal and make them good donors for transplanted hearts.

Figure 7.3.4 shows an interactive iCn3D model of an N-linked glycoprotein, human beta-2-glycoprotein-I (Apolipoprotein-H) (1C1Z).

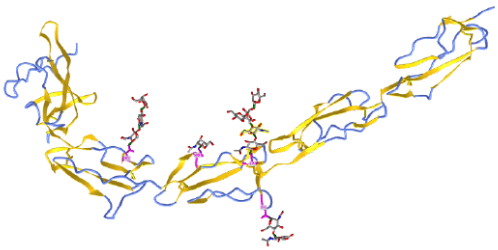
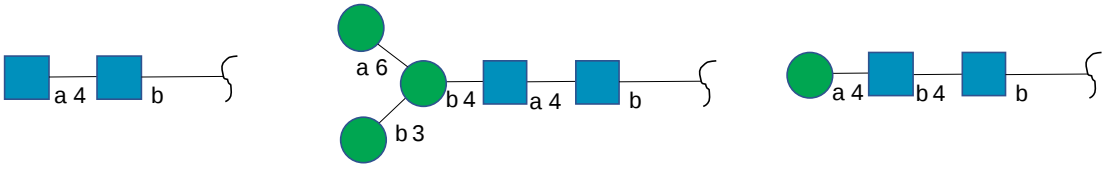


Figure 7.3.4: N-linked glycoprotein, human beta-2-glycoprotein-I (Apolipoprotein-H) (1C1Z). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...GkZSVFdciXiBT7

? Exercise 7.3.1

The glycan structures for the beta-2-glycoprotein-I are shown below. Identify the monosaccharides in each and specify to which asparagine they are linked.



Answer

Here is an interactive iCn3D image showing the protein and attached glycans using SNFG notation.

Figure 7.3.5 shows an interactive iCn3D model of the GP120 HIV protein that contains a high mannose, complex, and hybrid N-linked glycans. Most glycoproteins in the Protein Data Bank do not contain attached glycans. The glycans here were added with the program GlyProt at 3 of 17 possible Asn residues that would presumably have attached glycans. Use your mouse or keypad to hover over the monomers in the attached glycans. Abbreviations for the given residues in the model are: adm = alpha-D-Man, bdg= beta-D-Glc or Gal, adn = alpha-D-neuraminidase.

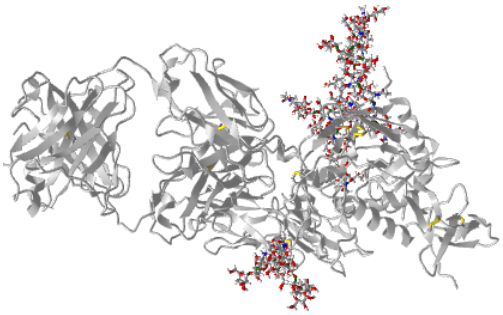


Figure 7.3.5: GP120 HIV protein with high mannose, complex, and hybrid N-linked glycans. (Copyright; author via source). No external link is available for this structure.

The coronavirus pandemic has been deadly (over 1.1 million deaths in the USA alone and 19-36 million around the world). However, the 1918 influenza pandemic was far worse per capita, with an estimated 650,000 deaths in the USA and 50 million around the world in a population in a population less than 1/3 of the present. An additional 3 million deaths in the USA were probably prevented by vaccination as of December 2022. Many deaths in the developing world would have been prevented if wealthy nations had allocated more resources to produce and distribute the vaccines. The evolution of the virus in nonvaccinated areas might come back to haunt wealthy countries if present vaccines become ineffective against the mutants. A worse pandemic might await us. An avian version of the influenza virus (H5N1), presently endemic in wild birds and now found in mink populations, has infected 240 people as of January 2023 and killed 56% of them. A quick note: the 1918 pandemic affected youth the most.

Influenza and the Avian Flu

Figure 7.3.6 shows the simple yet deadly influenza virus. It interacts with human cells through a surface protein, hemagglutinin (HA).

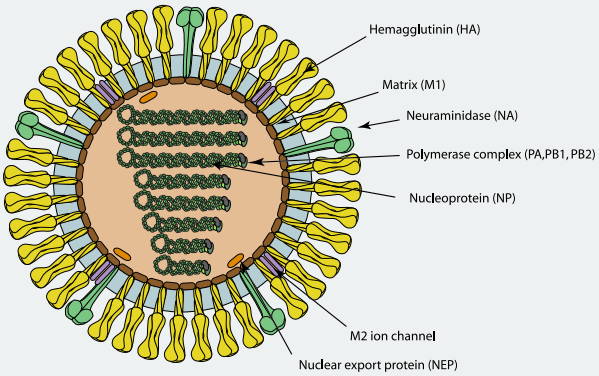
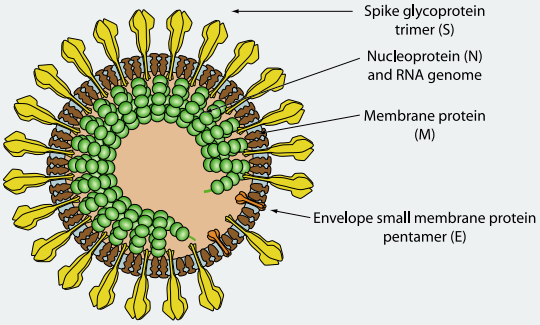


Figure 7.3.6: Alpha Influenza virus. https://viralzone.expasy.org/6. Creative Commons Attribution 4.0 International (CC BY 4.0) License.

Note the similarities and differences with the SARS-CoV-2 virus, shown below.



SARS-Covid-2 virus. https://viralzone.expasy.org/764. Creative Commons Attribution 4.0 International (CC BY 4.0) License.

The virus binds to host cells through the interaction of HA with cell surface carbohydrates. Once bound, the virus internalizes, ultimately releasing the viral RNA genome into the host cell.

The hemagglutinin protein is the most abundant protein on the viral surface. 15 avian and mammalian variants have been identified (based on antibody studies). Only three have adapted to humans in the last 100 years, giving pandemic strains H1 (1918), H2 (957), and H3 (1968). Three recent avian variants (H5, H7, and H9) have jumped directly to humans recently but have low human-to-human transmissibility.

The influenza hemagglutinin protein has the following characteristics:

- the mature form is a homotrimer (3 identical protein subunits), MW 220,000 with multiple sites for covalent attachment of sugars. Hemagglutinin is a glycoprotein.
- each monomer is synthesized as a single polypeptide chain precursor (HA0), which is cleaved into HA1 and HA2 subunits by the protease trypsin in epithelial cells of the lung.
- structure known for human (H3), swine (H9), avian (H5) subtypes.

Figure 7.3.7 below shows an interactive iCn3D model of an influenza hemagglutinin trimer (6HJQ). The white backbone traces on the membrane's extracellular (red) side are antibody molecules used to stabilize the structure for crystallization.

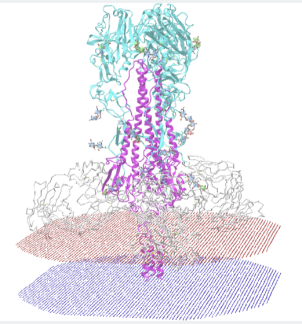


Figure 7.3.7: Influenza Hemagglutinin Trimer (6HJQ). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...EhePsF4BiEjwA8

The protein is a trimer of a heterodimer, $(HA1:HA2)_3$. Each heterodimer has two separate chains, HA1 and HA2. Together they form a globular head and stalk region, the latter crossing the membrane. HA1 interacts primarily with the sialic acid on host cells' surface proteins, while HA2 is more involved in the fusion of the virus with host cells and internalization.

Hemagglutinin binds to sialic acid (Sia) covalently attached to many cell membrane glycoproteins. The sialic acid is usually connected through an α(2,3) or α(2,6) link to galactose on N-linked glycoproteins. The subtypes found in avian (and equine) influenza isolates bind preferentially to Sia (α2,3) Gal, which predominates in the avian GI tract where viruses replicate. Human influenza isolates prefer Sia α(2,6)

Sia (α2,3) Gal which predominates in the avian GI tract where viruses replicate. Human influenza isolates prefer Sia α(2,6) Gal. The human virus of H1, H2, and H3 subtypes (causes of the 1918, 1957, and 1968 pandemics) recognize Sia α(2,6) Gal, the major form in the human respiratory tract. The swine influenza HA binds to Sia α(2,6) Gal and some Sia (α2,3) Gal, both found in swine. The structures of the Sia-Gal disaccharide are shown in Table 7.3.2 below.

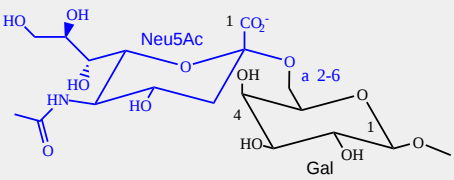
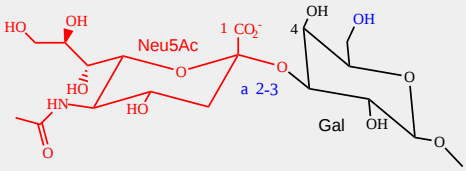
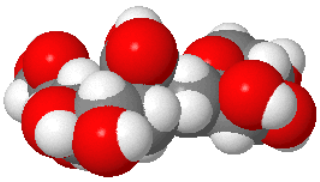
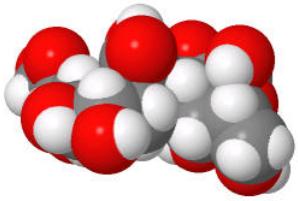
Sia α(2,6) Gal (Human)	Sia α(2,3) Gal (Avian and some Swine)
	
 (made with Sweet, with an OH, not AcNH on sialic acid on C5)	 (made with Sweet, with an OH, not AcNH on sialic acid on C5)

Table 7.3.2: Structures of Sia α(2,6) Gal (human) and Sia α(2,3) Gal (avian/swine)

The H5N1 avian flu H5N1) virus is deadly but presently lacks human-to-human transmissibility. Why? One reason is that it appears to bind deep in the lungs and is not released easily on coughing or sneezing. It appears that cell surface glycoproteins deeper in the respiratory tract have Sia (α2,3) Gal which accounts for this pathology.

Before it leaves the cell, the virus forms a bud on the intracellular side of the cell with the HA and NA in the cell membrane of the host cell. The virus in this state would not leave the cell since its HA molecules would interact with sialic acid residues in the host cell membrane, holding the virus in the membrane. Neuraminidase hydrolyzes sialic acid from cell surface glycoproteins, allowing the virus to complete the budding process and be released from the cell as new viruses. The drugs Oseltamivir (Tamiflu) and Zanamivir (Relenza) bind to and inhibit neuraminidase, whose activity is necessary for viral release from infected cells. Tamiflu appears to work against N1 of the present H5N1 avian influenza viruses. Governments across the world are hopefully stockpiling this drug in case of a pandemic caused by the avian virus jumping directly to humans and becoming transmissible from human to human.

7.3.2: O-linked Glycoproteins

The CHOs are usually attached from a Gal (β 1,3) GalNAc to a Ser or Thr of a protein, as shown in Figure 7.3.8.

Figure: O-linked Glycoproteins

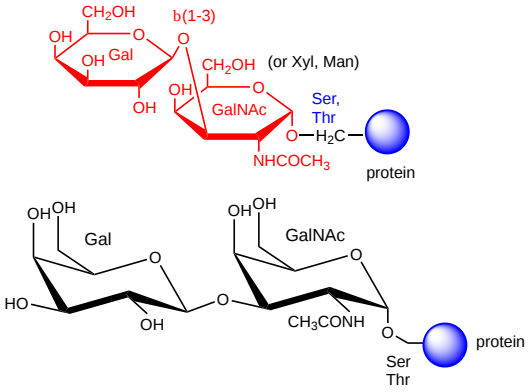


Figure 7.3.8: O-linked Glycoproteins

The blood group antigens (CHOs on cells attached to either proteins or lipids) are examples. The sugars shown as chairs (in contrast to structures found in many texts) in Figure 7.3.9 are the blood group antigens. They are attached to a core heterosaccharide (a red ellipse below), which is connected to either a membrane glycoprotein or glycolipid.

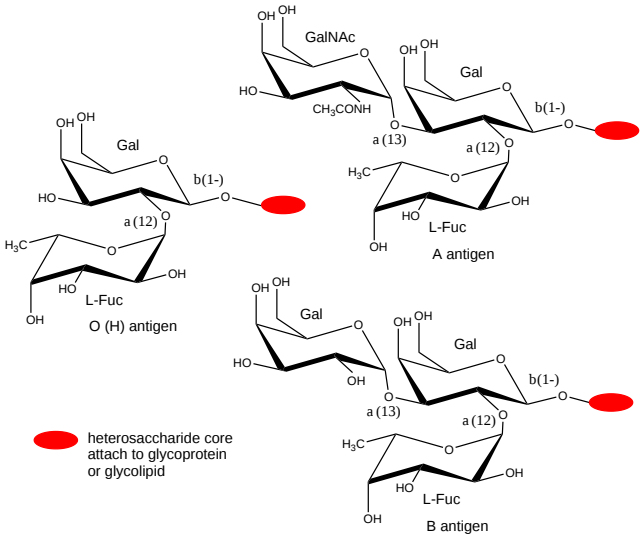


Figure 7.3.9 : Blood group glycans

Figure 7.3.10 shows the SNFG representation for the A antigen in the glycolipid form.

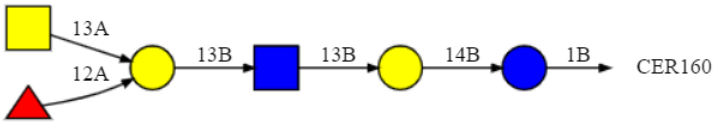


Figure 7.3.10: SNFG representation for the A antigen in the glycolipid form

The trimeric branched residues on the left-hand side represent the A antigen shown above. The red triangle is L-fucose. Yellow represents galactose or GalNac, while blue is glucose or GlcNAc.

7.3.3: Proteoglycans

Some proteins are so modified with CHOs that they contain more CHOs than amino acids. Proteins linked to glycosaminoglycans are together called proteoglycans (PGs). The consists of a core protein linked to one or more glycosaminoglycans. GAGs are linear sulfated glycans, which we described earlier. The structures of a few proteoglycans are known. The GAGs are O-linked to the protein, typically to a Ser of a Ser-Gly dipeptide, often repeated in the protein. Some of the proteoglycans also contained N-linked oligosaccharide groups. Figure 7.3.10 represents a proteoglycan structure.

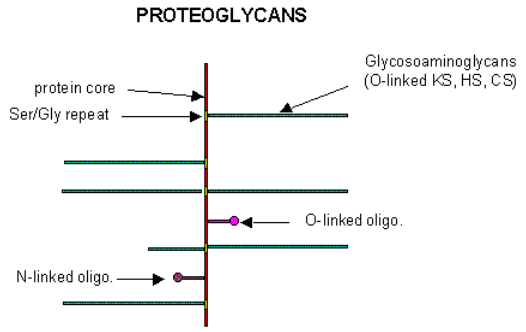


Figure 7.3.10: Representation of proteoglycan structure.

PGs can be soluble and are found in the extracellular matrix or as integral membrane proteins. There are about 43 genes for proteoglycans. Differential splicing of the RNA transcripts gives rise to soluble and transmembrane forms. Given the diversity of sugars and the varying extent of sulfation, the CHO part of PGs provides an incredible variety of binding structures at or near the cell surface. Figure 7.3.11 shows the variety of proteoglycans found in mammalian cells. PGs help form the extracellular matrix, which provides a rich binding environment between cells.

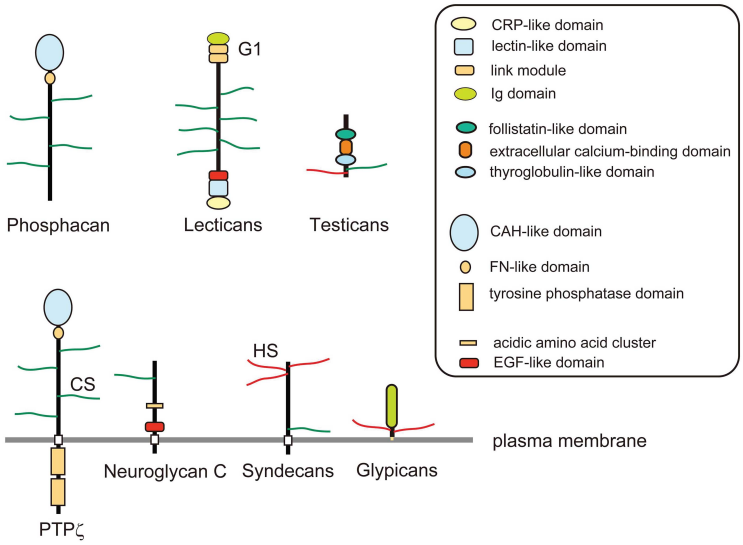


Figure 7.3.11: Proteoglycans found in mammalian cells

One PG, syndecan, binds through its intracellular domain to the internal cytoskeleton of the cell while interacting with another protein - fibronectin - in the extracellular matrix. Fibronectin also binds other molecules, regulating cellular growth and other interactions. PGs act like glue in connecting the extracellular and intracellular functions of the cell. There are four different core

syndecan proteins (SDCs 1–4), with SDC4 lacking the cytoplasmic and transmembranes and is a soluble form in the intracellular matrix. The glycan components of syndecans are mostly heparan sulfate, while SDC 1 and 3 also have two chondroitin sulfate chains.

Most proteins bind PGs through a PG binding motif of BBXB or BBBXXB where B is a basic amino acid. Some proteins bind to specific sequences in specific GAGs. For instance, antithrombin 3, an inhibitor of blood clotting, binds specifically to heparin. This enhances its interaction with clotting proteins such as thrombin and Factor Xa. Figure 7.3.12 shows an interactive iCn3D model of a five residue fragment of heparin interacting with the key amino acids side chains of Factor Xa (2gd4).

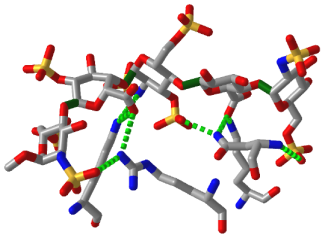


Figure 7.3.12: Heparin-5mer interactions with antithrombin III in a complex with Factor Xa. (2gd4) (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...5nZ99mUhULUi59

The extracellular matrix (ECM) might appear to be a nondescript mess for those more chemically oriented, since chemists are used to well-defined structures. Figure 7.3.13 shows a cartoon of the ECM and may clarify the components. Few structure files exist for them, given the inherent flexibility of the glycan components.

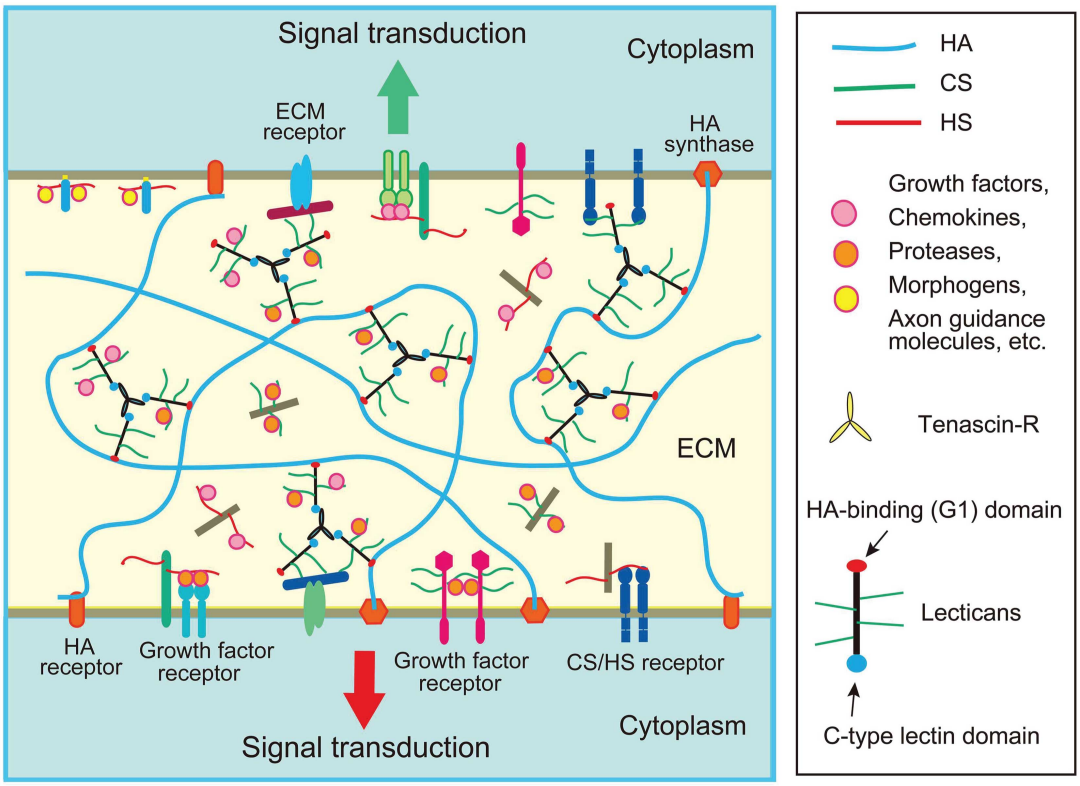


Figure 7.3.13: Cartoon representation of the extracellular matrix. Frontiers in Neuroscience 9(50):98 (2015). DOI: 10.3389/fnins.2015.00098. CC BY 4.0

7.3.4: Cell Walls and Glycolipids

In contrast to eukaryotic cells, bacteria, and plant cells have a cell wall in addition to a lipid bilayer membrane. These are essentially carbohydrate polymers that determine cell shape, affording protection from exterior pathogens, hypotonic conditions, and high internal osmotic pressures, preventing swelling and bursting of the cells. This is especially important in plants, which need strength and rigidity against the "turgor" pressure of the aqueous cytoplasm against the cell membrane. This prevents wilting in plants. The cell walls in plants and probably bacteria are involved in cell signaling across the cell membrane.

7.3.4.1: Bacteria Cell Walls

Two types of cell walls occur.

7.3.4.1.1: a. Gram-positive bacteria-

These bacteria can be stained with Gram stain. The wall consists of a GlcNAc (β 1,4) MurNAc repeat. (GlcNAc is often abbreviated as NAG, while MurNAc is abbreviated as NAM.) This is similar to the GlcNAc (β 1,4) GlcNAc homopolymer chitin, except that every other GlcNAc contains a lactate molecule covalently attached in an ether-linkage to the C3 hydroxyl to form the monomer N-Acetylmuramic acid. A pentapeptide (Ala-D-isoGlu-Lys-D-Ala-D-Ala) is attached through an amide link to the carboxyl group of the lactate in MurNAc. A pentaglycine bridge covalently connects the GlcNAc (β 1,4) MurNAc strands through the epsilon amino group of the pentapeptide Lys on one strand and the terminal D-Ala of a pentapeptide on another strand. A small part of the structure of a gram-positive bacterial cell wall is shown in Figure 7.3.14 It shows one repeating GlcNAc-MurNAc disaccharide unit in front (darker) and one in the back (lighter) connected through the peptides shown.

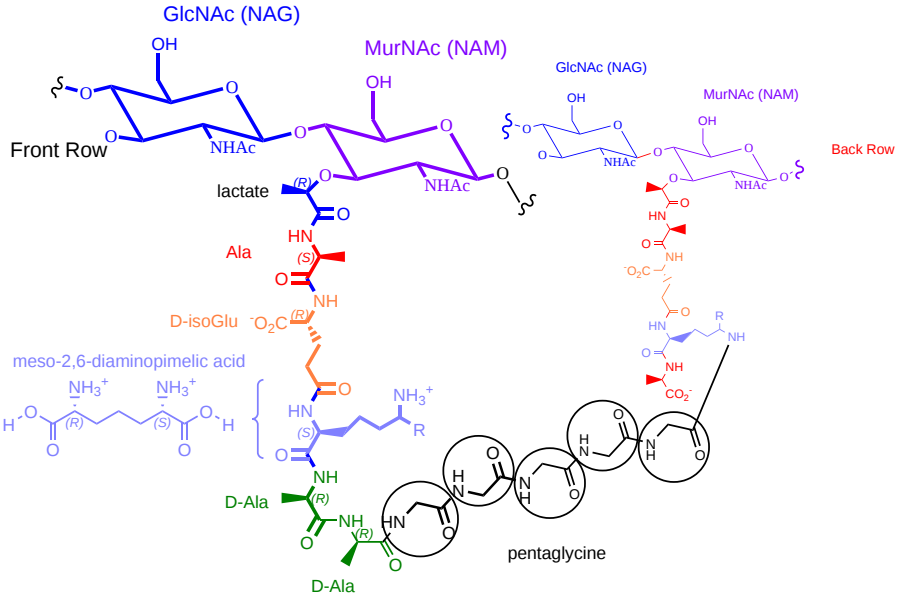


Figure 7.3.14: Part of the structure of a gram-positive bacterial cell wall

The SNFG representation of a larger section of the gram-positive cell well is shown in Figure 7.3.15

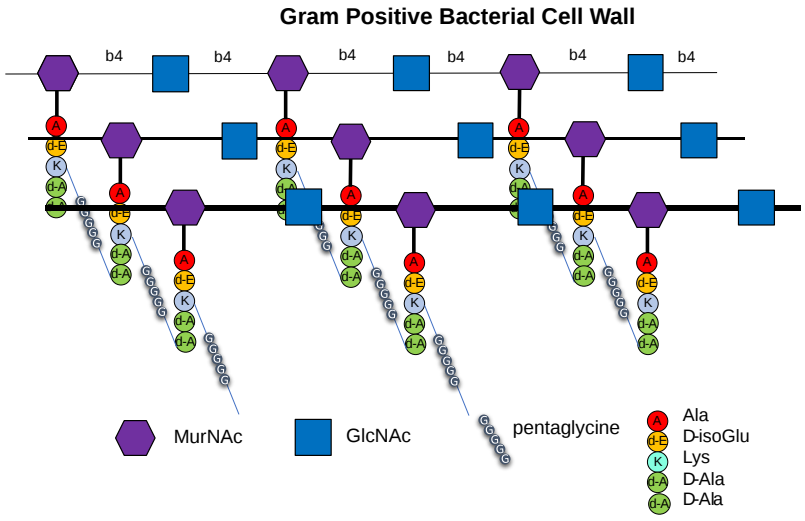


Figure 7.3.15: SNFG representation of a larger section of the gram-positive cell wall.

One final structure is found in Gram + peptidoglycan cell walls. Techioic acids are often attached to the carbon 6 of MurNAc. Teichoic acid is a polymer of glycerol or ribitol with alternative GlcNAc and D-Ala linked to the middle C of the glycerol. Multiple glycerols are linked through phosphodiester bonds. These teichoic acids often make up 50% of the dry weight of the cell wall and present a foreign (or antigenic) surface to infected hosts. These often serve as receptors for viruses that infect bacteria (called bacteriophages). Its structure is illustrated in Figure 7.3.16.

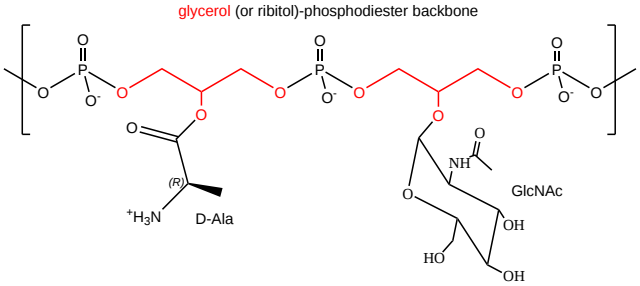


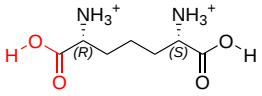
Figure 7.3.16 : Teichoic acid

Notice that all monomeric units of peptidoglycan and attached teichoic acid derivatives are covalently attached on form one large molecule comprising the entire cell wall! This structure, along with the Gram-negative cell wall structures, is the largest single macromolecule in nature.

7.3.4.1.2: *b. Gram-negative bacteria*

These bacteria can NOT be stained with Gram stain. The wall consists of the same structure as in Gram-positive bacteria. However, the GlcNAc (β 1,4) MurNAc strands are covalently connected through a direct amide bond between a derivative of Lys, meso-diaminopimelic acid (m-A2pm), on one peptide strand and to the last D-Ala of a pentapeptide on another strand. (i.e., there is no penta-Gly spacer). The connector peptide is Ala-D-isoGlu-m-A2pm-D-Ala-D-Ala

m-A2pm replaces Lys 3 of the peptide in most Gram-negative species and Gram-positive bacteria of the genus *Bacillus* and *mycobacteria*. The stereochemistries at each chiral center are different (R and S), but because the molecule has a plane of symmetry, it is an example of a meso-compound, a diastereoisomer of a molecule, which does not have a different enantiomeric version. The structure is shown in Figure 7.3.17.



meso diaminopimelic acid

Figure 7.3.17: Meso-diaminopimelic acid

A small part of the structure of a Gram-negative bacterial cell wall is shown in Figure 7.3.18.

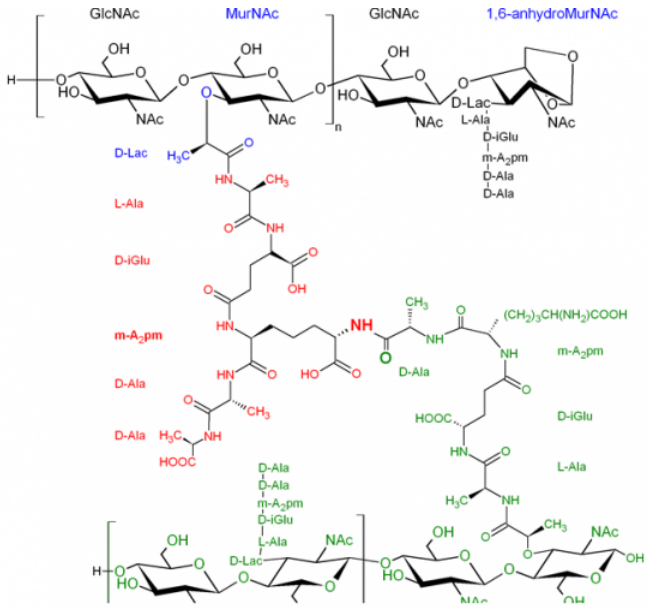


Figure 7.3.18: Part of the structure of a Gram-negative bacterial cell wall. http://www.glycopedia.eu/e-chapters/...ular-structure. https://gagdb.glycopedia.eu/license

Figure 7.3.19 shows an image of a computed model (not a crystal or NMR structure) of the Gram-negative peptidoglycan of E. Coli. The PDB coordinates were kindly provided by Jame Gumbart. The repeating $(GlcNAc\text{-}MurNAc)_n$ are shown in sticks, alanines in red spheres, m-A2-pm (meso-diaminopimelic acid) in gray spheres, and DiG in orange spheres. The glycine pentapeptide is not shown since it is not found in Gram-negative bacteria.

Figure 7.3.19: Part of a Gram-negative peptidoglycan of E. Coli. PDB coordinates kindly provided by Jame Gumbart.

Follow these instructions to get an interactive iCn3D model of the structure.

- open iCn3D
- Download this file to your computer's download folder. **IMPORTANT**: If the file opens as an image in a new browser window, right-click the image and save the file to download it!
- **File, Open File, iCn3D png (appendable),** and choose the downloaded file (it takes a while).

In addition, Gram-negative bacteria don't have teichoic acid polymers. Rather, they have a **second, outer lipid bilayer**. The cell wall peptidoglycan (PG) is sandwiched between the inner and outer bilayers. The space between the lipid bilayers is called the periplasm. The outer leaflet of the outer membrane is coated with a **lipopolysaccharide (LPS)** (a glycolipid) of varying composition. The LPS determines the antigenicity of the bacteria. The different LPS are called the O-antigens. Figure 7.3.20 shows the structure of the Gram-negative bacterial membrane organization. (In the figure, PS is LPS, PG is peptidoglycan). The LPS in the outer leaflet is amphiphilic, with nonpolar acyl chains forming a more classic bilayer, and the inner leaflet phospholipid and

polar/charged sugars form the LPS fringe. The extra membrane of Gram-negative bacteria makes them the major source of antibiotic resistance.

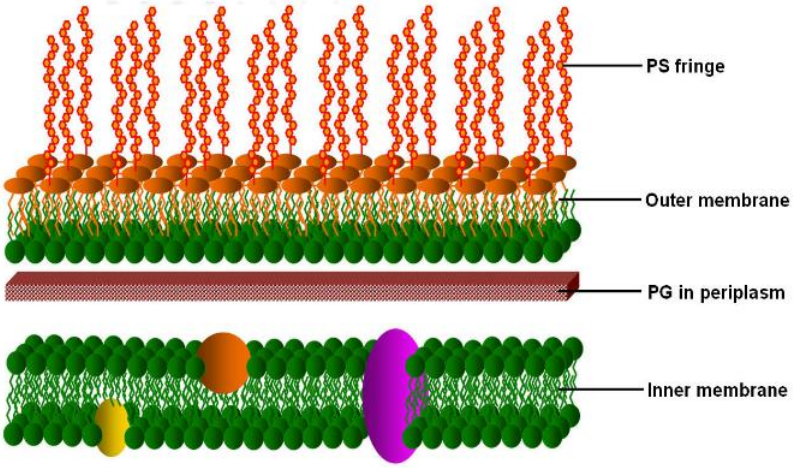


Figure 7.3.20: Overall structure of the Gram-negative bacterial membrane organization. https://cronodon.com/BioTech/Bacteria_envelope.html

A detailed view of the structure of the lipopolysaccharide (LPS) from *Salmonella tryphimurium* is shown in Figure 7.3.21 below.

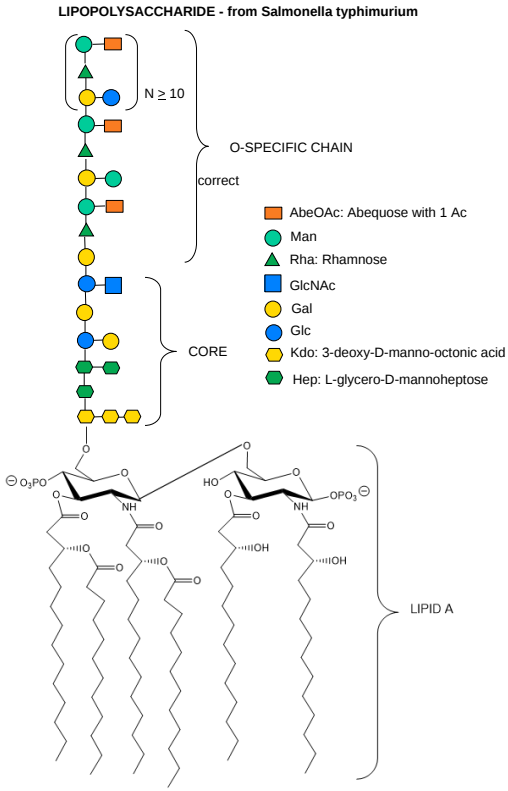


Figure 7.3.21: Lipopolysaccharide (LPS) from *Salmonella Typhimurium*

Recent Updates: 11/8/24

LPS is synthesized in the inner leaflet of the inner membrane and must translocate all the way to the outer leaflet of the outer membrane. Movement requires an ATP-binding cassette transporter $LptB_2FG$. Figure 7.3.22 shows the machinery used to move

LPS to the outer membrane. We'll discuss membrane protein and transport thoroughly in Chapter 11.

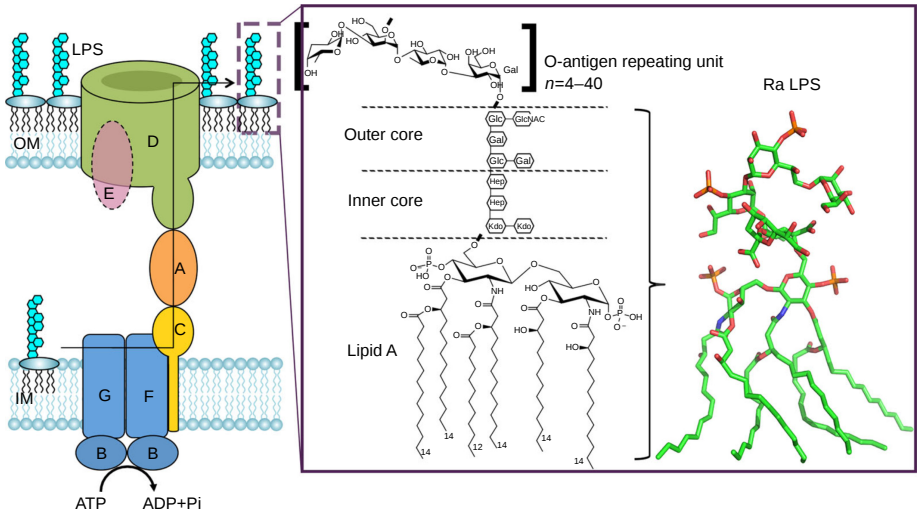
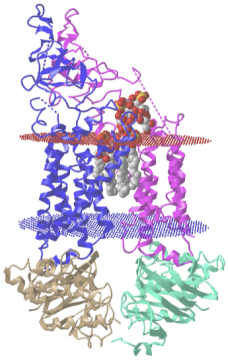


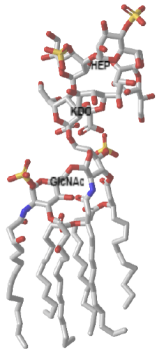
Figure 7.3.22: LPS transport from the IM to the OM by the trans-envelope complex LptABCDEFG. Dong, H., Zhang, Z., Tang, X. *et al.* Structural and functional insights into the lipopolysaccharide ABC transporter $LptB_2FG$. *Nat Commun* **8**, 222 (2017). https://doi-org.ezproxy.csbsju.edu/1...67-017-00273-5. Creative Commons Attribution 4.0 International License. http://creativecommons.org/licenses/by/4.0/.

Legend: LPS is extracted from the periplasmic side of the IM by the ABC transporter $LptB_2FG$, and is delivered to an IM protein LptC, which forms a complex with $LptB_2FG$. LptC comprises a single membrane-spanning domain and a large periplasmic domain, forming a periplasmic bridge with LptA and the N-terminal domain of LptD. LPS is then inserted into the OM by LptD/E complex. LPS contains O-antigen, core oligosaccharide, and lipid A components, of which the O-antigen has 4-40 O-antigen repeat units. Ra-LPS, rough LPS, Kdo, 3-deoxy-D-manno-oct-2-ulosonic acid, Hep, L-glycero-D-manno-heptose, Glc, D-glucose, Gal, D-galactose.

Figure 7.3.23 shows an interactive iCn3D model of the E. Coli lipopolysaccharide ABC transporter LptB2FG (6MHU). The left panel shows the membrane protein that binds and transports the LPS (spacefill) from the inner to the outer membrane. The right panel shows the bound LPS. Just part of the inner core is shown since other parts were likely to flexible to be observed. The outer O antigen is not shown since a mutation in the protein prevented it. 6 acyl chain are evident. The phosphorylated GlcNAc, KDO and HEP "layers" are labeled in the right panel.



Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...CoPtdH2RVY82P6. Click Style, Background, Transparent in the iCn3D window for a better view.



Click the image for a popup (instructions below)

Figure 7.3.23: E. Coli lipopolysaccharide ABC transporter LptB2FG (6MHU) (left panel) and the bound LPS (right panel). (Copyright; author via source).

To see an interactive iCn3D model of the E. Coli LPS from the right panel, follow these steps:

- open iCn3D
- Download this file to your computer's download folder. **IMPORTANT**: If the file opens as an image in a new browser window, right-click the image and save the file to download it!
- **File, Open File, iCn3D png (appendable),** and choose the downloaded file.

7.3.4.1.3: c. Archaeal Cell Membranes and Walls

We have already discussed that the lipids in Archaeal cell membranes contain L (instead of D) glycerol derivatives and that ether links (more stable in reactive environments) replace ester links with isoprenoid (sometimes branched) chains, replacing fatty acid chains. The cell wall is also quite different, and some don't have one. The type of cell wall depends on the environmental need for stability. They don't contain peptidoglycans. Figure 7.3.24 shows four different types.

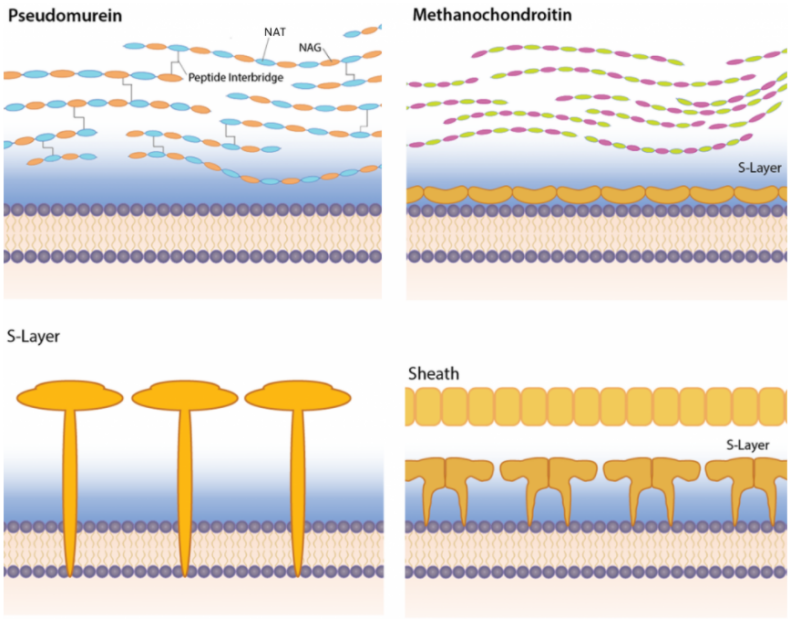


Figure 7.3.24:Archaeal cell walls https://open.oregonstate.education/g...apter/archaea/. https://creativecommons.org/licenses/by-nc/4.0/

Some differences include the presence of

- pseudomurein - This is the closest to the peptidoglycans presented above. Instead of repeating disaccharide units of (NAM-NAG)$_n$, they have a repeating disaccharide unit of N-acetylalosaminuronic acid (NAT)-NAG. The structure of NAT is shown in Figure 7.3.25

Figure 7.3.25: N-acetylalosaminuronic acid (NAT)compared to N-acetylmuramic acid (NAM)

- methanochondroitin - This is similar to the glycosaminoglycan chondrotin sulfate
- S-Layer
- Sheath/S-Layer

7.3.4.1.4: d. Plant Cell Wall

If you thought bacterial cell walls were complicated, wait until you see plant cell walls! There are about 35 different types of plant cells, and each may have a different cell wall depending on the local needs of a given cell. Cells synthesize thin cell walls that

extend and stay thin as the cell grows.

Figure 7.3.26 shows the primary cell wall of plants. The primary cell wall contains cellulose microfibrils (no surprise) and two other polymers, pectin and hemicellulose. The middle lamella, consisting of pectins, is somewhat analogous to the extracellular matrix discussed above.

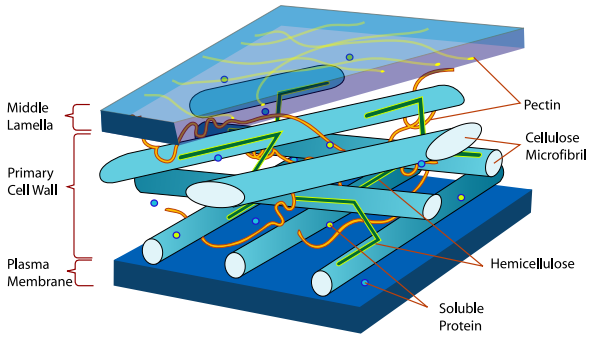


Figure 7.3.26: Primary cell wall of plants. https://commons.wikimedia.org/wiki/F...diagram-en.svg

After cell growth, the cell often synthesizes a secondary cell wall thicker than the first for extra rigidity. The enzymatic machinery for its synthesis is in the cytoplasm and the cell membrane. It is deposited between the cell membrane and the primary cell wall, as shown in the animated image in Figure 7.3.27.

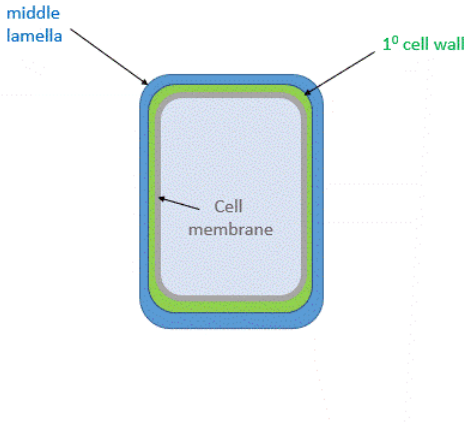


Figure 7.3.27: Primary and secondary cell wall of plants

Figure 7.3.28 shows a structural representation of both the primary and secondary cell walls.

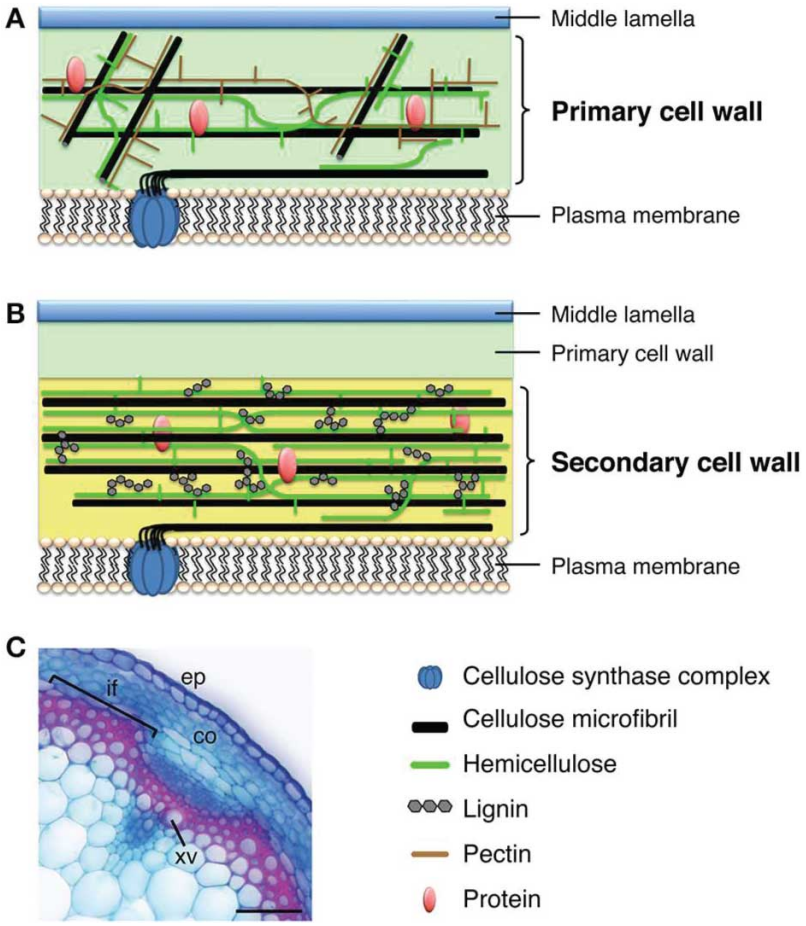


Figure 7.3.28: A structural representation of both the primary and secondary cell walls of plants. Nakano Yoshimi et al. Frontiers in Plant Science (6), 288 (2015) https://www.frontiersin.org/article/...015.00288/full. Creative Commons AttributionLicense (CC BY).

The middle lamella, which contains pectins, lignins and some proteins, helps "glue together" the primary cell walls of surrounding plants.

Primary Cell Wall:

The main component of the primary plant wall is the homopolymer cellulose (40% -60% mass) in which the glucose monomers are linked β(1→4)-linked into strands that collect into microfibrils through hydrogen bond interactions. Two other groups of polymers, **hemicellulose** and **pectin,** make up the plant cell wall.

Hemicellulose can make up to 20-40% by the mass These polymers have β(1,4) backbones of glucose, mannose, or xylose (called xyloglucans, xylans, mannans, galactomannans, glucomannans, and galactoglucomanannans along with some β(1,3 and 1,4)-glucans. The most abundant hemicellulose in higher plants are the xyloglucans with a cellulose backbone linked at O6 to α-D-xylose. **Pectin** consists of linked galacturonic acids forming homogalacturonans, rhamnogalacturonans, and rhamnogalacturonans II (RGII) [12] [13]. Homogalacturonans (α1→4) linked D-GalA making up more than 50% of the pectin

Figure 7.3.29 shows some variants of the cell wall components of a plant.

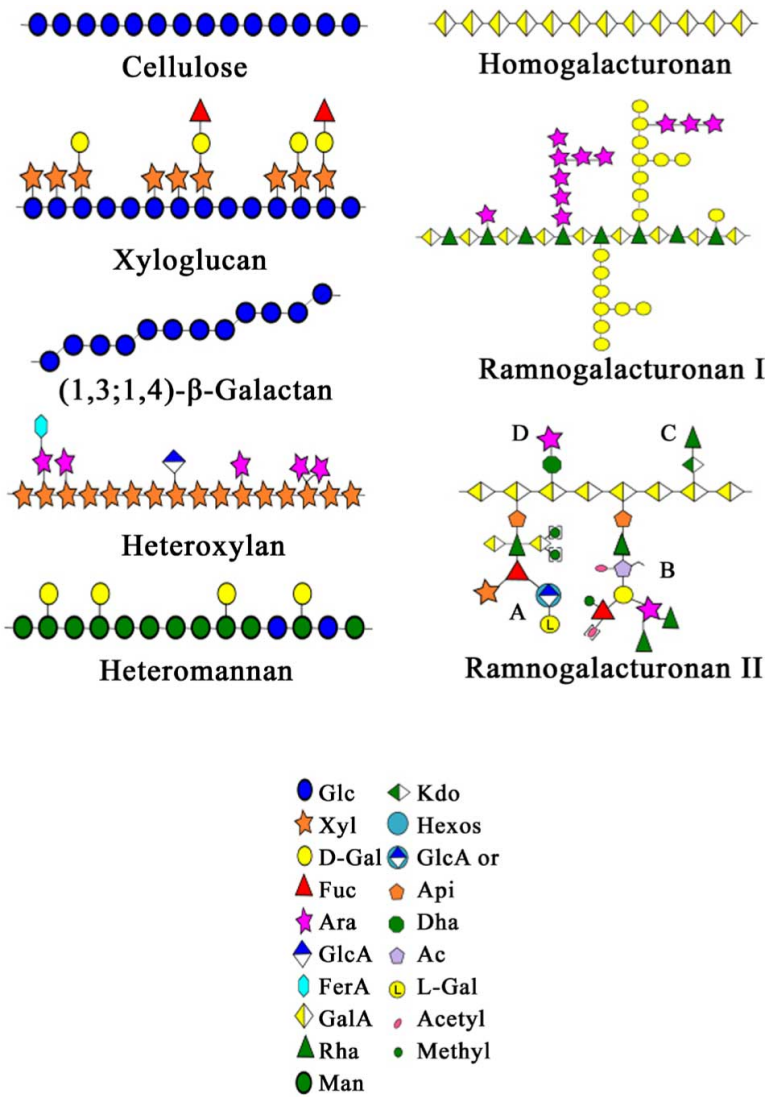


Figure 7.3.29: Variant of the cell wall components of a plant. Costa and Plazanet. Advances in Biological Chemistry 06(03):70-105. DOI: 10.4236/abc.2016.63008, License CC BY 4.0

Secondary Cell Wall

The secondary cell wall's structure depends on the cell's function and environment. It contains cellulose fibers, hemicellulose, and, in addition, a new polymer, **lignin**. It is abundant in xylem vessels and fiber cells of woody plants. It gives the plant extra stability and new functions, including the transport of fluids within the plant through channels.

Lignins, which can make up to 25% of the biomass weight, are made from phenylalanine derivatives but more directly from cinnamic acid. This derivative is made from phenylalanine, which is hydroxylated and converted through other steps to hydroxycinnamyl alcohols called monolignols, as shown in Figure 7.3.30. Three common monomeric (M) derivatives, p-coumaryl, coniferyl, and sinapyl alcohols, can polymerize into lignins, with the units in the polymer (P) named hydroxyphenyl, guaiacyl, and syringyl, respectively.

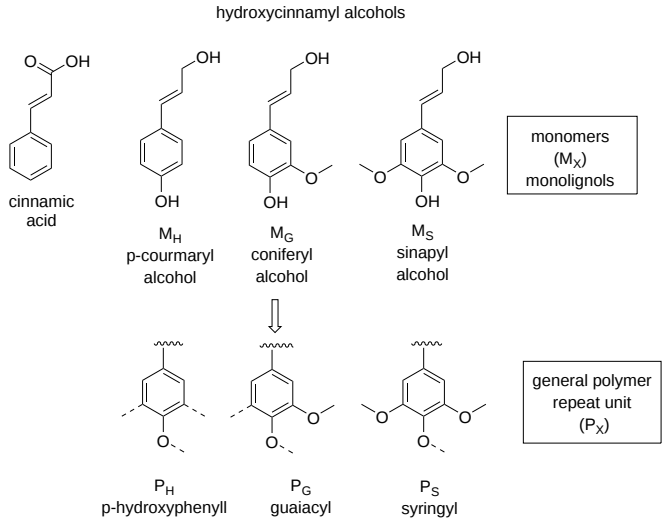


Figure 7.3.30: Monolignols and their polymers

Lignols are activated phenolic compounds that form phenoxide free radicals (catalyzed by peroxidase enzymes), which can attack other lignols to form covalent dimers. Reaction mechanisms for dimerizing the M_S sinapyl alcohol free radical are shown as an example in Figure 7.3.31.

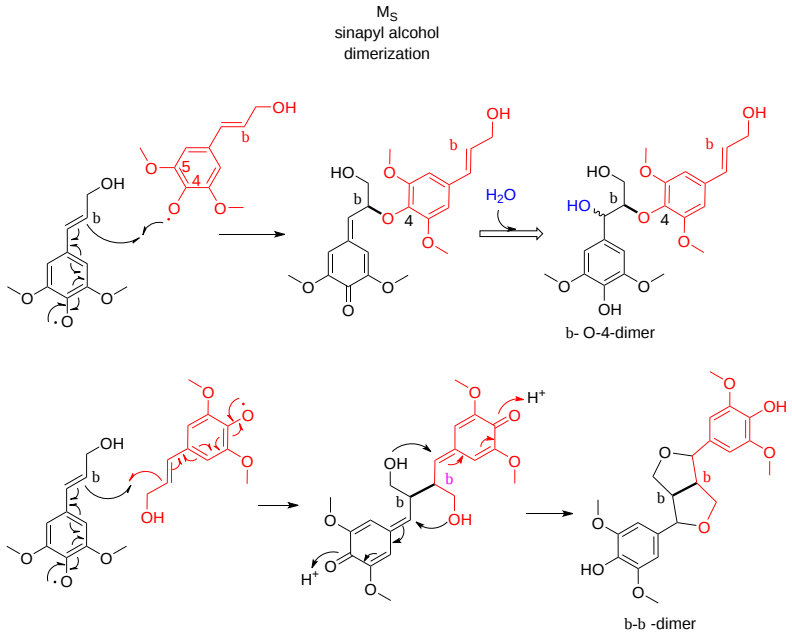


Figure 7.3.31: Dimers of lignols

Now imagine this polymerization continuing by forming additional phenolic free radicals and coupling at many sites to form a huge covalent lignin polymer. Figure 7.3.32 shows one example of a larger lignin.

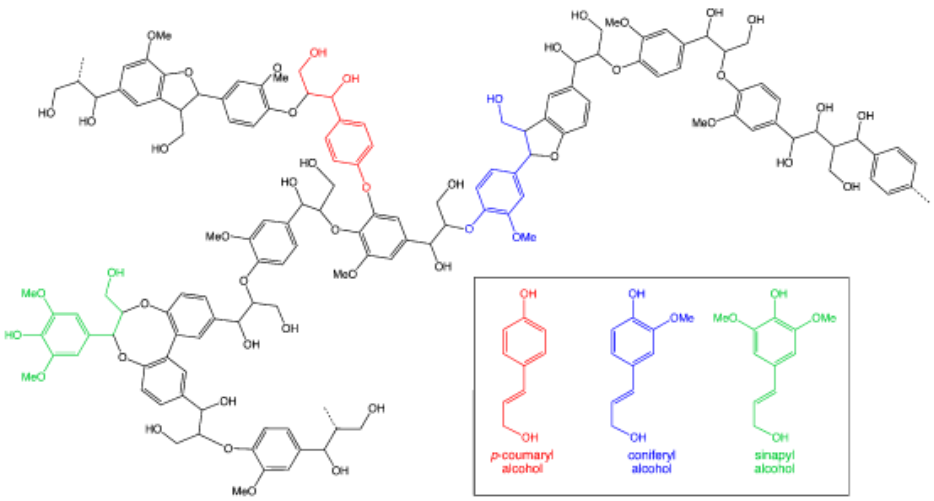


Figure 7.3.32: A larger lignin. https://commons.wikimedia.org/wiki/C...ile:Lignin.png . By Smokefoot - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=16022799

Finally, Figure 7.3.33 shows an image of a poplar tree cell wall, made using surface Raman scattering, showing lignin, cellulose, and lipids in secondary xylem cell walls.

Figure 7.3.33: Surface Raman scattering (SRS) images of lignin, cellulose, and lipids in xylem cell walls. **d-e are** SRS images of lignin, cellulose, and lipids in the secondary xylem cell walls of poplar, respectively. The 3D surface plots are shown in **g–i**. Xu, H., *et al.* A label-free, fast and high-specificity technique for plant cell wall imaging and composition analysis. *Plant Methods* **17,** 29 (2021). https://doi.org/10.1186/s13007-021-00730-9. http://creativecommons.org/licenses/by/4.0/.

7.3.5: The Extracellular Matrix (ECM) and Basement Membranes

We won't formally discuss cell membranes until Chapter 11, but since anyone reading this book has previously seen biological membranes (including the Gram-negative and positive bilayers discussed above), let's explore a term that most chemistry students, but perhaps not biology students, will find very confusing. That topic is the basement membrane. The basement membrane is encountered so often that we will explore its overall structure here, even though it is not a lipid bilayer. It fits well here since it is a complex structure consisting of proteins and proteoglycans. It's very amorphous, making its structure difficult for those hoping for crystal structures or complex bilayers. It is somewhat similar to the cell wall in functionality. We will offer a cursory explanation. Please visit Introduction to Extracellular Matrix and Cell Adhesion in BioLibre texts for a great overall introduction. Some of the images (when noted) below come from that Cell Biology book chapter.

The extracellular matrix (ECM) is a general term for the large protein and polysaccharide network formed on secretion by some cells in a multicellular organism. They act as connective material to hold cells in a defined space. Cell density can vary greatly

between different tissues of an animal, from tightly-packed muscle cells with many direct cell-to-cell contacts to liver tissue, in which some of the cells are only loosely organized, suspended in a web of extracellular matrix, shown in Figure 7.3.34

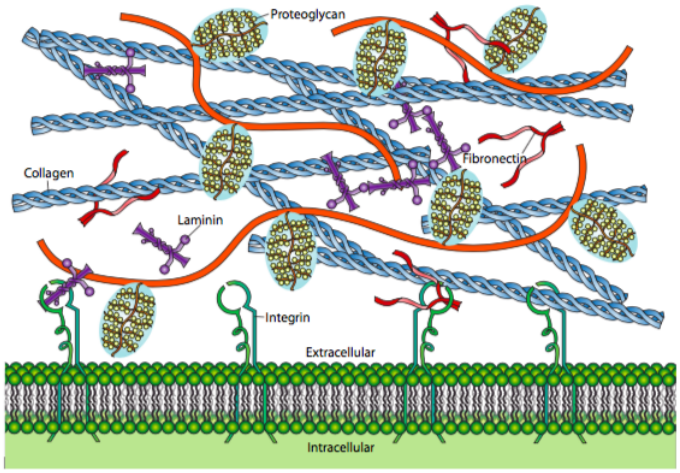


Figure 7.3.34: Extracellular matrix (ECM). Typical components include collagen, proteoglycans (with hydration shells depicted around sugars), bronectin, and laminin. The cellular receptors for several of these ECM components are integrins, although the exact integrin αβ *pair may differ.*

The ECM is a generic term encompassing mixtures of polysaccharides and proteins, including collagens, bronectins, laminins, and proteoglycans, all secreted by the cell. The proportions of these components can vary greatly depending on tissue type. Two quite different examples of ECM are the basement membrane underlying the epidermis of the skin, a thin, almost two-dimensional layer that helps to organize the skin cells into a nearly impenetrable barrier to most simple biological insults, and the massive three-dimensional matrix surrounding each chondrocyte in cartilaginous tissue. The ability of the cartilage in your knee to withstand the repeated shock of your footsteps is due to the ECM proteins in which the cells are embedded, not to the cells that are rather few and sparsely distributed. Although both types of ECM share some components in common, they are distinguishable not just in function or appearance but in the proportions and identity of the constituent molecules

Figure 7.3.35 shows a general structure of the basement membrane. Think of it as an amorphous polymer mixture (somewhat similar to a polyacrylamide gel).

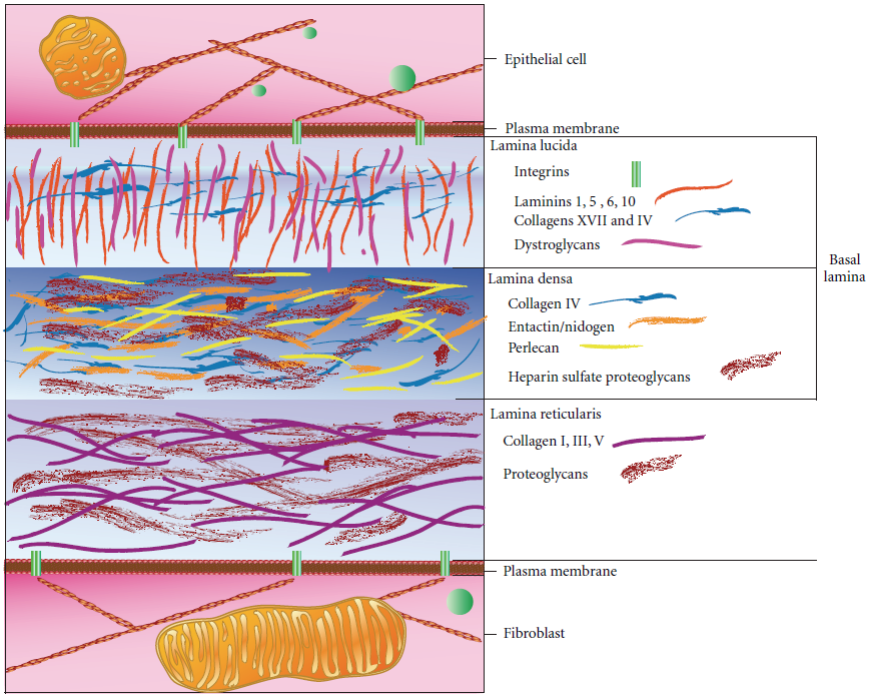


Figure 7.3.35: General structure of the basement membrane. Mentor and DuBois. Journal of Cell Biology · February 2012. DOI: 10.1155/2012/723419 · Source: PubMedCreative Commons License. https://creativecommons.org/licenses/by/3.0/

7.3.6: Summary

This chapter explores the multifaceted world of protein glycosylation, highlighting the complex nature and functional significance of carbohydrates attached to proteins. It emphasizes how post-translational modifications by glycans greatly expand the informational content and functional diversity of the proteome, far beyond what is encoded in the genome.

Key Themes:

1. **Glycosylation and Its Biological Importance:**
   The chapter begins by explaining that many proteins destined for secretion or membrane insertion undergo glycosylation. This modification, typically attached to asparagine (N-linked) or serine/threonine (O-linked) side chains, plays critical roles in:
   - Mediating recognition and binding interactions.
   - Preventing aggregation during protein folding.
   - Protecting proteins from proteolysis.
   - Increasing the half-life of proteins in circulation.
2. **N-linked Glycoproteins:**
   A central focus is placed on N-linked glycosylation, where a core oligosaccharide (Man)3(GlcNAc)2(\text{Man})_3(\text{GlcNAc})_2 is covalently attached to asparagine in a specific consensus sequence. Variants of N-linked glycans include:
   - **High Mannose:** Retaining terminal mannose residues.
   - **Complex:** Featuring GlcNAc branches and terminal sialic acid residues.
   - **Hybrid:** Displaying characteristics of both high mannose and complex types. The chapter uses SNFG (Symbolic Nomenclature For Glycans) to depict these structures and discusses their impact on protein function and immune recognition—as seen, for example, on the heavily glycosylated surface of the SARS-CoV-2 spike protein.
3. **O-linked Glycoproteins and Blood Group Antigens:**
   O-linked glycosylation, which attaches sugars to serine or threonine residues (often starting with a Gal(β1,3)GalNAc moiety), is examined next. This section explains how such modifications give rise to blood group antigens and affect cell–cell recognition and immune responses.

4. **Proteoglycans and the Extracellular Matrix:**
   The chapter outlines how extensive glycosylation can result in proteoglycans—proteins heavily modified with glycosaminoglycan (GAG) chains. Proteoglycans are major components of the extracellular matrix (ECM) and basement membranes, where they contribute to structural integrity, cell adhesion, and signal transduction.
5. **Carbohydrate Structures in Cell Walls and Membranes:**
   The discussion is broadened to cover the roles of carbohydrates in cell wall structures across different organisms:
   - **Bacterial Cell Walls:** Differences between Gram-positive and Gram-negative bacteria are highlighted, including the structure of peptidoglycan, the role of teichoic acids, and the unique organization of the outer membrane in Gram-negative species.
   - **Plant Cell Walls:** The synthesis and structure of primary and secondary cell walls, with components such as cellulose, hemicellulose, pectin, and lignin, are described, underscoring their importance in plant rigidity and growth.
   - **Archaeal Cell Walls:** Distinct features of archaeal membranes and cell walls are noted, emphasizing the unique lipid composition and structural adaptations in extreme environments.
6. **Clinical Relevance and Glycan Diversity:**
   The chapter concludes by discussing how variations in glycan structures, such as the presence of Gal(α1,3)Gal, have profound implications for immune recognition and xenotransplantation. It also touches on the role of glycosylation in pathogen-host interactions and cell signaling.

Overall, this chapter provides a comprehensive overview of glycosylation as a key post-translational modification, illustrating how complex carbohydrate structures enrich protein function and influence a wide range of biological processes—from cellular communication to structural support.

7.4: The Sugar Code and Lectin Decoding

Learning Goals (ChatGPT o1, 1/30/25)

- **Foundational Concepts in Glycosylation and Glycan Diversity:**
  - Define key terms (sugar, carbohydrate, glycan) and explain why glycan structures are extraordinarily diverse compared to proteins and nucleic acids.
  - Discuss the absence of a genetic template for glycan synthesis and how this contributes to the structural heterogeneity of glycans.
- **N-linked Glycoprotein Structure and Variation:**
  - Describe the core oligosaccharide (Man)3(GlcNAc)2(\text{Man})_3(\text{GlcNAc})_2 attached to asparagine in a consensus sequence and explain the differences among high-mannose, complex, and hybrid N-linked glycans.
  - Interpret glycan structures using the Symbolic Nomenclature for Glycans (SNFG) and explain how variations in branching and terminal modifications (e.g., sialic acid, additional galactose) affect function and immunogenicity.
- **O-linked Glycoproteins and Their Roles:**
  - Explain how O-linked glycans are attached to serine or threonine residues and discuss their biological roles, such as in blood group antigens and cell–cell recognition.
- **Proteoglycans and the Extracellular Matrix (ECM):**
  - Define proteoglycans and describe their structure (core protein with covalently attached glycosaminoglycans) and their role in forming the ECM.
  - Analyze how the diversity of glycan structures on proteoglycans contributes to the binding, signaling, and structural integrity of the ECM.
- **Carbohydrate Structures in Cell Walls and Membranes:**
  - Compare and contrast the carbohydrate-based cell walls of bacteria (Gram-positive and Gram-negative) and plants, including the roles of peptidoglycan, teichoic acids, and lignin.
  - Describe the unique features of archaeal cell walls and plant cell walls, and explain how glycan structure underlies functions such as rigidity, protection, and signaling.
- **Glycan-Binding Proteins (GBPs) and the Glycan Code:**
  - Explain the concept of the glycan code and describe how glycan-binding proteins (lectins) serve as “readers” of this code.
  - Differentiate among various classes of GBPs, including viral capsid proteins, bacterial adhesins, and animal lectins, and provide examples such as C-type lectins, galectins, and siglecs.
- **Mechanisms of Glycan Recognition:**
  - Analyze how GBPs interact with specific glycan epitopes through both continuous (linear) and discontinuous (conformational) binding sites.
  - Interpret structural data (including interactive 3D models) to identify key domains (e.g., C-type lectin domains, immunoglobulin-like domains) and motifs (e.g., EPN or WND) involved in glycan binding.
- **Biological and Clinical Relevance:**
  - Discuss the role of glycan modifications in mediating immune responses, cell adhesion, and pathogen recognition (e.g., the role of sialylated glycans in influenza virus binding or the recognition of Gal(α1,3)Gal in xenotransplant rejection).
  - Explore how alterations in glycosylation patterns influence processes such as inflammation, cancer progression, and cell signaling in the ECM and basement membranes.

By achieving these goals, students will develop a detailed understanding of the structural and functional diversity of glycans, the mechanisms by which glycan-binding proteins decode these structures, and the critical roles these processes play in cellular communication, immune responses, and disease.

7.4.1: Introduction

By now, you should be convinced that the structures of glycans are extraordinarily complex and, in many ways, much more complicated than proteins and nucleic acids. Their structure diversity is staggering, given the number of different sugar monomers, stereocenters, linkages, lengths, conformers, dynamic flexibility, and chemical modifications. Yet evolution has allowed this astronomical diversity, which must serve more than just simple functions such as protecting proteins from degradation. Much of the diversity derives from lacking an equivalent genetic code for glycan synthesis.

Since all events in biology start with a binding interaction, let's ponder the binding interaction of glycans with partner "ligands" such as proteins, lipids, nucleic acids, etc. A binding site on a glycan could be a single monosaccharide to a much larger and more complex interface. Figure 7.4.1 shows an interactive iCn3D model of one of the few glycoproteins with pdb coordinates, the unliganded simian immunodeficiency virus (SIV) gp120 core glycoprotein (3fus).

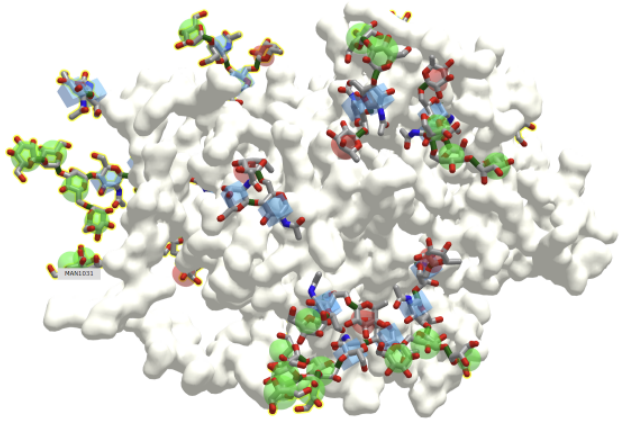


Figure 7.4.1: Simian immunodeficiency virus (SIV) gp120 core glycoprotein (3fus). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/icn3d/share.html?DzZYhLp8JL1vPYYw6

The protein surface is shown in ivory, and the glycans are shown in color sticks with the correct symbolic color-coded spheres or cubes around them.

Now let's convert in our imagination an image file showing one face of the protein to a black and white QR code as shown in Figure 7.4.2.

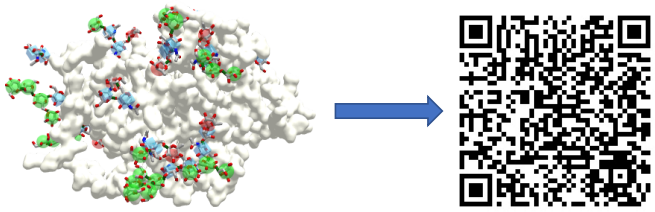


Figure 7.4.2: An imaginary QR code for the surface of a glycoprotein.

Computers can recognize information encoded in the QR codes and decode it into another form of information, such as a menu at a restaurant. Likewise, organisms have evolved "readers" to decode the **glycan code** written by enzymes (glycan synthases, hydrolases, and modifying enzymes). The glycan code is written onto the 3D surfaces of polysaccharides, glycoproteins, glycolipids, and proteoglycans. It should be no surprise that the biological readers of the glycan code are mostly proteins, which locate and bind to the correct "QR" code displayed on the glycan surface.

Luckily, the QR code metaphor for the glycan code is a bit exaggerated since the readers of the glycan code, glycan-binding proteins, seem to recognize just small sections of a glycan. They can be compared to protein antibodies that bind to foreign molecules, such as proteins. The binding site on a foreign protein recognized by an antibody is called an **epitope**. Epitopes can be continuous (linear) stretches of the foreign protein sequence or discontinuous (conformational), made of some continuous stretches of amino acid and some further away in the sequence but close in the 3D folded protein. The average continuous epitope is often 5-6 amino acids long. Yet that might be an underestimation since an analysis of all contact residues (within a conservative 4 A° distance) for target proteins and their bound antibodies found in the Protein Data Bank is around 18-19 amino each (Stave and

Lindpaintner). Glycan-binding proteins presumably also bind a mixture of continuous and discontinuous glycan sequences. Linear one would be much easier to determine and study.

Now, let's explore the family of these glycan-binding proteins (GBP), the readers of the glycan code.

7.4.2: Glycan-Binding Proteins (GBPs)

There appear to be nine types of glycan-binding proteins (GBPs). These include nonenveloped capsid virus GBPs,enveloped-virus GBPs (ex. influenza and coronaviruses), eukaryotic microbial GBPs (ex yeast), and bacterial toxin GBPs (ex botulinum toxin). Bacterial **adhesins** (parts of organelles like flagella), **lectins** (soluble proteins) and lectin domain-containing proteins are also examples of **glycan-binding proteins (GBP).** We will discuss in more detail three other types: **C-type lectins, galectins, and siglecs**.

In the broadest sense, if a **lectin** is a protein that binds a specific carbohydrate motif (i.e., a glycan code) without modifying the motif, then any glycan-binding protein could be called a lectin. This excludes enzymes that synthesize, degrade, or modify glycans and antibodies that recognize foreign or self-glycan sequences. Table 7.4.1 below shows some lectins and their target glycan ligand from plants, animals, viruses, and bacteria.

Lectin Family/Lectin	Abbreviation	Ligand(s)
Plants		
Concanavalin A	ConA	Man α1- OCH3
Griffonia simplicifolia lectin 4	GS4	Lewis b (Leb) tetrasaccharide
Wheat germ agglutinin	WGA	Ner5Ac(α 2,3)Gal(β 1,4)GlcGlcNAc(β1,4)GlcNAc
Ricin		Gal(β 1,4)Glc
Animals		
Galectin-1		Gal(β 1,4)Glc
Mannose-binding protein	MBP-A	High Mannose Octasaccahride
Viral		
Influenza Virus hemagglutinin	HA	Neu5Ac(α 2,6)Gal(β1,4)Glc
Polyoma virus protein 1	VP1	Neu5Ac(α 2,3)Gal(β1,4)Glc
Bacterial		
Enterotoxin	LT	Gal
Cholera toxin	CT	GM1 pentasaccharide

Table 7.4.1: Lectin families and their ligands.

In animals, lectins facilitate cell-cell interactions by forming multiple but weak interactions (also called **multivalent interactions**) between the protein and many sugars on the ligand to which it binds.

Now, let's consider the other three classes of glycan-binding proteins (or lectins), **C-type lectins, galectins,** and **siglecs** in more detail. Focus on the very different structures of the carbohydrate-binding domains.

7.4.3: C-Type Lectins

C-type lectins comprise the largest number of glycan-binding proteins. These proteins have a glycan or carbohydrate recognition domain that depends on the Ca^{2+} ion. They bind self-glycans and those on pathogens, which can target viruses to specific cells. Many are on the surface of immune cells. They have an N-terminal glycan-binding domain called a C-lectin (CLECT) or carbohydrate recognition domain (CRD). However, some proteins with the domain do not appear to bind either Ca^{2+} or glycans. They serve as adhesion molecules and are also involved in cell signaling. Some residues in the lectin binding domain appear

critical for binding lectins. These include an EPN motif, which interacts with Man, GlcNAc, Fuc, and Glc) and a WND motif involved in binding to Gal and GalNAc.

Let's look now at one example of a C-type lectin, the **selectins**.

P-Selectins

These are involved in the interaction of immune cells in the blood with endothelial cells that line the blood vessel wall. Think of the challenges an immune cell faces as it moves from the blood into a tissue where an infection might occur! Blood flow in vessels at a rate inversely proportional to the total cross-sectional area of the blood vessel. That rate is about 5-20 cm/sec in arteries, 1.5-7 cm/sec in veins, and about 1 mm/sec (1000 μm)/sec) in capillaries. Assuming a lymphocyte's average diameter of 10 μm, the cell would move about 100 cell lengths per second. An equivalent speed for a human with an arm span of 6 feet (approximately fitting into a circle of diameter = 6 feet as drawn by Leonardo da Vinci) would be around 600 feet/second. The cell must go from its typical circulating speeds to a stop before moving through the blood vessel wall into tissues. Nature has solved this by providing a way to slow down the moving cell until its final capture. The cells roll along the endothelial cells, making transient low-affinity interactions, which slow it down enough until high-affinity ones effectively stop it (unless it dissociates first).

Also, you wouldn't want immune cells to stop and move into tissue without an infection signaled by mediator molecules. Another problem solved! P-selections are stored in the intracellular granules of platelets (the source of the name P-selectin) and endothelial cells, so moving immune cells are not spuriously captured without some signal. In the presence of the right chemical signal, endothelial and platelets get active, and P-selectin is transported rapidly to the cell surface, where "capture" occurs before the cell can move into the underlying tissue. P-selectin mediates the first transient interactions and subsequent rolling of immune cells on activated platelets and endothelial cells.

Figure 7.4.3 is a video animating the rolling and "capture" of a lymphocyte by endothelial cells. (See the video for the reference.) Note that cancer cells can also move through the endothelial cells of blood vessels in forming metastases.

Figure 7.4.3: Video animation of a lymphyocyte rolling and being captured by endothelial cells lining blood vessel walls

P-selectins, hence, are receptors for molecules on immune cells. They bind Ca^{2+} ions, which helps create an active conformation. Their binding ligands are glycan codes and nearby sections of protein connected to the glycan. The glycan ligand on the surface of a circulating immune cell is the **sialyl-Lewis X (SiaLew^X)** glycan or a derivative of it. One of the immune membrane proteins with SiaLew^X is the **P-selectin glycoprotein ligand 1** (**PSGL-1**, the gene name), also called **SELPLG**. It mediates rapid rolling of leukocytes over vascular surfaces during the initial steps in inflammation through interaction with SELPLG"

P-selectin is a mediator of cell adhesion (to other cells). As such, it could also be classified as an adhesion protein. The three main types of selectins:

1. L-selectins: found on leukocytes ("white" blood cells that are circulating immune cells).

2. P-selectins: found on activated platelets (which can aggregate to form a type of blood clot) and activated endothelial cells. Activation occurs during the inflammatory response, which can lead to the quick movement of pre-formed selectins stored within the cytoplasm to the membrane. In addition, their expression can be induced.
3. E-selectins: found on activated endothelial cells only after the cells have been induced to form them by certain immune hormones called cytokines released by immune cells during an inflammatory response.

Figure 7.4.4 shows the domain structure of human P-selectin.

Figure 7.4.4: Domain structure of human P-selectin (from https://smart.embl.de/)

It contains an N-terminal C-Lectin (CLECT) domain, which is also called the carbohydrate-recognition domain (CRDs) or the C-type lectin domain (CTLD). In addition, it has an epidermal growth factor domain (EGF), 9 complement control protein (CCP) domains and the blue transmembrane domain.

Figure 7.4.5 shows the structure of the $SLew^x$ glycan along with its symbol nomenclature for glycans (SNFG) representation.

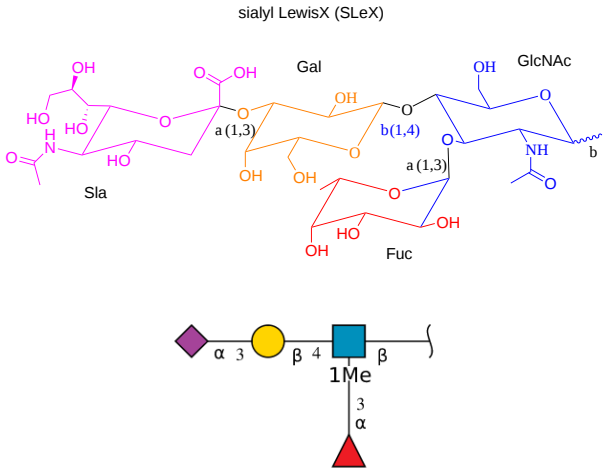


Figure 7.4.5: Structure of the $SLew^x$ glycan

Figure 7.4.6 shows an interactive iCn3D model of the crystal structure of P-selectin lectin/EGF domains complexed with SLeX (1g1r) to which P-selectin binds with weak affinity. Fucose interacts with the Ca^{2+} ion. The glycan interacts with the CLECT domain.

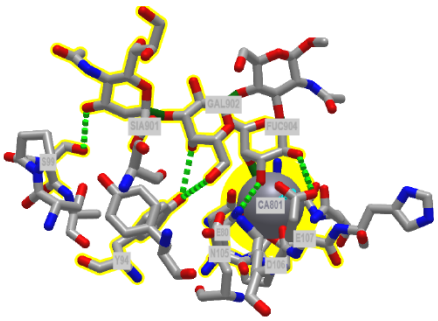


Figure 7.4.6: P-selectin lectin/EGF domains complexed with SLeX (1g1r). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i..s1GYbAfXfnuSz9

$SLew^x$ is not present in isolation but rather attached to a membrane protein on an immune cell, which serves as a ligand for the P-selectin on activated endothelial cells or platelets. (The $SLew^x$ can also be part of a glycolipid.) Now let's contrast the interactions of the P-selectin LE domain with "naked" $SLew^x$ with those present between P-selectin LE with a higher affinity natural binding ligand, human P-selectin glycoprotein ligand 1 (PSGL-1), an immune cell integral membrane protein. PSGL-1 is expressed on neutrophils, monocytes, and most lymphocytes. The P-selectin:PSGL-1 complex has a much lower K_D (higher

affinity) than binding of the unmodified SLe^X (1g1s). PSGL-1 is a disulfide-linked homodimer. When sulfated on a specific Tyr (48), the protein displays high affinity for P-selectin. In contrast, when sulfated on a different Tyr (51), it displays a high affinity for L-selectin instead.

The $SLex^X$ type glycan O-linked to the peptide is a bit more complicated than the simple $SLex^X$ ligand as it is connected to a protein through an O-linked bond at a threonine. The SNFG is shown in Figure 7.4.7.

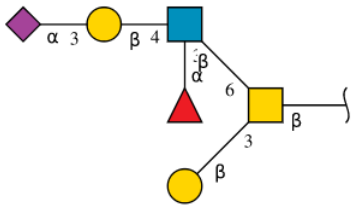


Figure 7.4.7: SNFG if $SLex^X$ type glycan O-linked to a peptide

The crystal structure of a trisulfated, $SLew^x$-modified peptide from the N terminal region of PSGL-1 (1G1S) bound to P-selectin lectin and EGF domains (P-LE) has been solved. Figure 7.4.8 shows an interactive iCn3D model that shows some of the interactions between the PSGL-1 peptide (green backbone) and P-LE (magenta backbone).

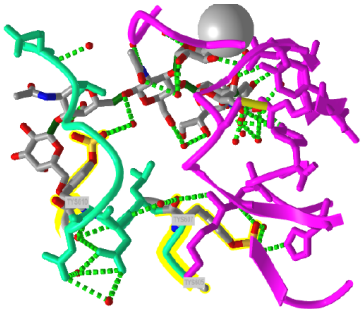


Figure 7.4.8: P-selectin Lectin/EGF domains and bound PSGL-1 peptide (1G1S). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i..E2mVULZc8WqVR9

In the crystal structure, the peptide from the P-selectin ligand (which again is a membrane protein) contains three sulfated tyrosine residues (605, 607, and 610), which correspond to amino acids 5, 7, and 10 in the peptide). No electron density for the side chain of Tyr 605 was seen. Tyr 607 binds through multiple interactions to the P-selectin LE domain and is most likely responsible for the high-affinity interaction of P-selectin with the P-selectin glycoprotein ligand (again represented by the green chain). In contrast, Tyr 610 interacts through an intermediary water molecule with the glycan $SLex^X$ of the peptide.

Figure 7.4.9 shows the electrostatic surface potential map of one of the dimers of the P-select LE domains. Blue represents positive potential and red negative. The backbone of the P-selectin ligand peptide is green with all of the negatively charged side chains (Tyr, Asp, and Glu) shown in stick with CPK colors. Note that these amino acids are all bound in blue (positive) regions of P-selection. The glycan portion attached to the peptide (stick, CPK color) is positioned mostly over negative potential, allowing hydrogen bonding between the hydrogen bond donors of sugar OHs with the protein.

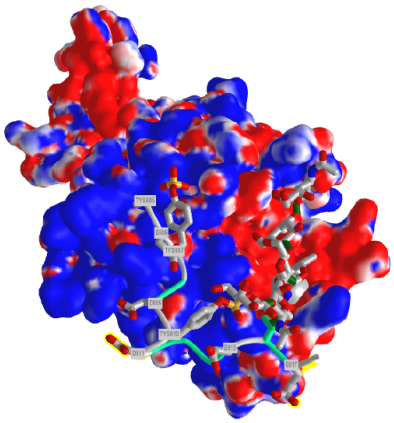


Figure 7.4.9: Electrostatic surface potential map of one monomer of the P-selectin LE domains with bound P-selectin ligand peptide.

You could surmise that the blue region of positive potential could also bind other strongly negatively charged ligands (such as heparin and other glycosaminoglycans), which could inhibit the function of this protein as it would prevent binding of the PSGL-1.

Figure 7.4.10 shows an interactive iCn3D model which shows the surface electrostatic potential of the P-selectin Lectin/EGF domains and bound PSGL-1 peptide

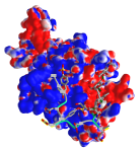


Figure 7.4.10: Electrostatic potential of the P-selectin Lectin/EGF domains and bound PSGL-1 peptide (1G1S). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i..X2CSQQzgbFn2VA

The blue represents positive potential, and the red represents negative. The backbone of the P-selectin ligand peptide is green, with all of the negatively charged side chains (Tys, Asp and Glu) shown in stick with CPK colors.

There are also nonpolar interactions that are not shown in the figure and model above. The aromatic ring of Tyr 607 (7) interacts with the nonpolar parts of a Ser ($-CH_2$) and Lys $(-CH_2)_4$ side chains, and the ring of Tys 610 (10) interacts with two leucine side chains.

The selectins are also part of a class of molecules called **adhesion** molecules. As described for the selectins, adhesion molecules contain

- an extracellular CHO binding domain (the lectin domain), which mediates binding to adjacent cells or the extracellular matrix;
- a transmembrane domain;
- and a cytoplasmic domain, which often interacts with the cytoskeleton within the cell.

This initial binding mediated by selectin-CHO interactions activates the expression of another adhesion molecule on the leukocyte, **integrin**, a heterodimer with an a and b chain. These cause strong leukocyte-endothelial cell interactions, leading to the movement of the leukocytes through the vessel wall. Other classes of adhesion molecules (in addition to selectins and integrins) are **cadherins** (calcium-dependent adhesion molecules) and the immunoglobulin-like superfamily (ICAM1, ICAM2, VCAM). VCAM (Vascular Adhesion Molecule) binds to integrin expressed on activated lymphocytes, leading to the lymphocyte's passage from the vessel's lumen into the tissues. Integrins appear to bind proteins in the extracellular matrix through RGD (Arg-Gly-Asp) and also through LDV (Leu-Asp-Val) motifs on the proteins, including fibronectin (RGD), thrombospondin (RGD & LDV), fibrinogen (RGD & LDV), van Willebrand Factor (RGD), vitronectin (RGD). They also bind other matrix proteins with an "alpha domain," including collagen and laminin. Integrin/Adhesion molecule interactions involve protein/protein interactions.

A fertilized egg (in the blastocyst stage, which is ready for implantation in the uterine cell wall) expresses L-selectin, which allows a low affinity (rolling-type) interaction of the fertilized egg with the uterine **epithelial** cells. These cells expressed the CHO ligands on their surface, which bind to the L-selectin on the blastocyst. The CHO ligands are only

transiently expressed on the surface of the epithelial cells of the uterus, presumably only when the uterus is primed for implantation. After the initial interaction of the blastocyst and epithelial cells, further expression of integrins on the blastocyst surface might result. Problems in any of these molecular steps could result in infertility. Figure 7.4.11 shows endothelial cell/leukocyte interactions mediated through selectins, integrins, and ICAMs.

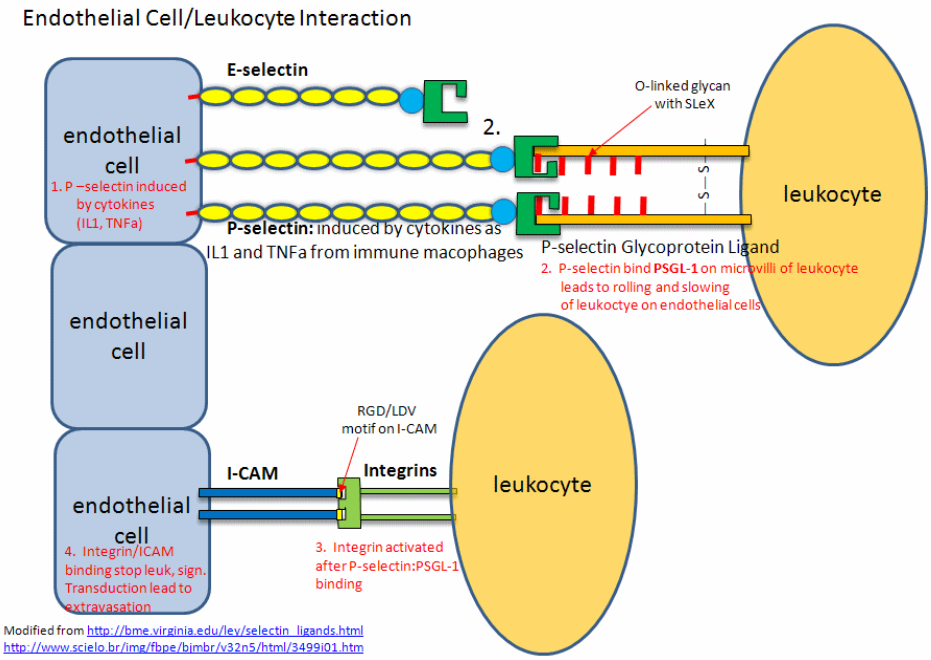


Figure 7.4.11: Endothelial cell/leukocyte Interactions mediated through selectins, integrins, and ICAMs

Post-translational modifications of protein modification (like glycosylation) can confer new binding and biological functions to a protein. Site-directed mutagenesis can replace surface amino acids with cysteine or methionine with nonnatural amino acid analogs that contain azide or alkyne groups. These modified groups could then direct the location of chemical modifying reagents (such as sugars) to these sites. A protein completely unrelated to PSGL-1 has been selectively modified to contain covalently attached glycans and sulfated tyrosine side chains. The unrelated protein bound to P-selectin.

7.4.4: Mannose Receptor.

What do you do with a protein that no longer has the correct structure(s) to perform its designed function(s)? Proteins, as with any molecule, undergo chemical changes during their biological lifetime. They must be recognized as aberrant and then removed from "service," ultimately being degraded into component amino acids for reuse. There are no repair enzymes for proteins as there are for DNA. One modification that changes glycoproteins and signals the need for their removal is the removal of terminal sialic acid residue, forming asialoglycoproteins, whose glycans end in galactose, as you can envision from Figure 7.4.12, which shows a typical structure of a N-linked glycoprotein.

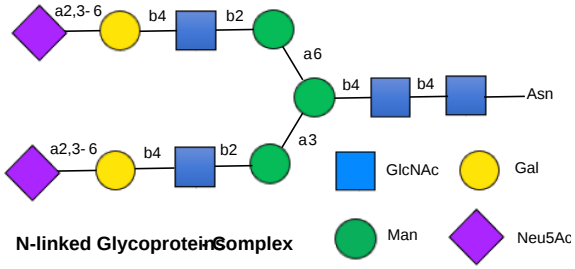


Figure 7.4.12: Typical structure of a N-linked glycoprotein

The **asialoglycoprotein receptor**, a member of the C-Type lectin family, is a transmembrane protein that binds terminal galactose and N-acetylgalactosamine sugars on the end of circulating asialoglycoproteins, leading to their endocytosis into the cell. It is expressed on the surface of hepatocytes (liver cells). Receptors of this type are also called scavenger receptors, as they remove proteins from circulation.

The **mannose receptor** (also called CD206), also expressed in liver endothelial cells, is another C-type lectin involved in the binding and removal of glycoproteins from the circulation. It binds both sulfated and non-sulfated glycans. It is also a receptor that allows the binding and phagocytosis of bacterial and fungal pathogens by a type of immune cell, namely macrophages or dendritic cells. Unfortunately, tumor cells can use the same process for uptake into macrophages, promoting tumor cell growth. The protein binds and scavenges sulfated glycoprotein hormones, mannose-bearing glycoproteins released during inflammation, lysosomal enzymes released from cells on injury, and collagen fragments.

Figure 7.4.13 shows the domain structure of the human mannose receptor.

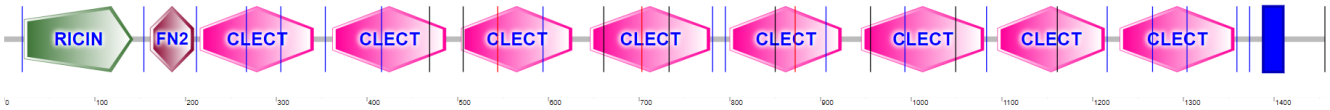


Figure 7.4.13: Domain structure of the human mannose receptor.

Given the large number of CLECT domains, you might surmise that this protein could bind a number of different target glycans from both self and pathogens. What is different about the domain structure compared to P-selectin is the presence of an N-terminal Ricin and a Fibronectin type 2 (FN2) domain. The FN2 domain has two cystines from the 4 conserved cysteines involved in disulfide bonds. What's so interesting about the mannose receptor is that it binds glycans both in the CLECT domains and in the FN2 domain.

Glycan binding at the CLECT domain: The CLECT domain binds targets containing mannose, fucose, and N-acetylglucosamine with a preference for Man(α1,2)Man or fucose. Figure 7.4.14 shows an interactive iCn3D model. of the **CLECT** 4 domain of the mannose receptor complexed with Man(α1,2)Man (7jue). Interactions of fucose lin ligands such as Lewis-a-trisaccharide strengthen the binding.

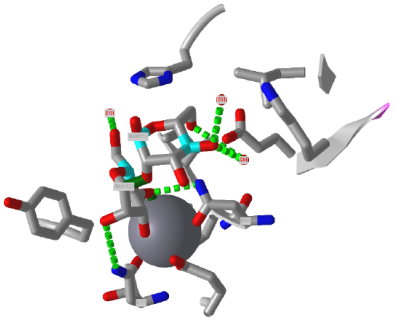


Figure 7.4.14: **CLECT** 4 domain of the mannose receptor complexed with Man(α1,2)Man (7jue). (Copyright; author via source).
Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i..jhn5eMLMrG6mX7

The receptor can bind a variety of glycans. Both mannose and N-acetylglucosamine interact with bound Ca^{2+} through equatorial OHs on carbon 3 and 4 of the ring, while fucose uses OHs on carbon 2 and 3, or 3 and 4.

The interactions with fungal pathogens are medically important. Fungi, like yeast, have an outer structure composed of a membrane bilayer and a mixture of glycans, which deploys an incredibly complex "glycan code" to host infected by them, as illustrated in Figure 7.4.15.

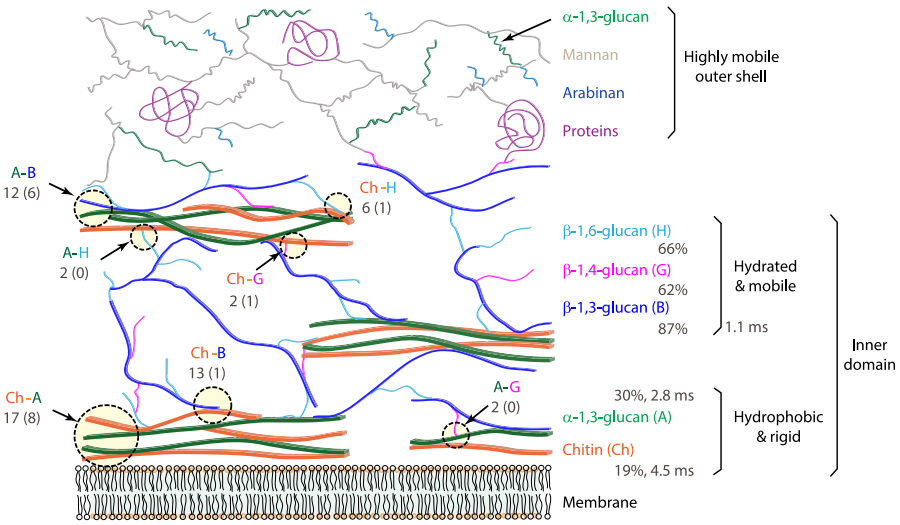


Figure 7.4.15: Structure of fungal cell wall. Kang, X., Kirui, A., Muszyński, A. *et al.* Molecular architecture of fungal cell walls revealed by solid-state NMR. *Nat Commun* **9,** 2747 (2018). https://doi.org/10.1038/s41467-018-05199-0. Creative Commons Attribution 4.0 International License: http://creativecommons.org/licenses/by/4.0/

Mannans, polymers of just mannose, differ widely in structure. Their main backbone can be DMan(α-1,6)DMan or DMan(β-1,4)DMan with many branches.

Glcyan binding at the FN2 (Cysteine-Rich) Domain (1FWU): The mannose receptor can also bind non-mannose sulfated glycans, such as 3-SO4-LEWIS(X), for which the SNFG representation is shown in Figure 7.4.16.

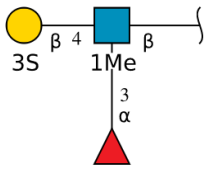


Figure 7.4.16: 3-SO4-LEWIS(X) non-mannose sulfated glycans for mannose receptor

The mannose receptor binds this glycan, which does not even contain mannose, through the FN2 domain (which contains four disulfide bonds) and not the CLECT calcium-dependent carbohydrate-binding domain. Hence, the protein can bind both sulfated and nonsulfated glycans.

Figure 7.4.17 shows an interactive iCn3D model of the complex of the FN2 domain of the mannose receptor with the non-mannose containing 3-SO4-LEWIS(X) glycan (1fwu).

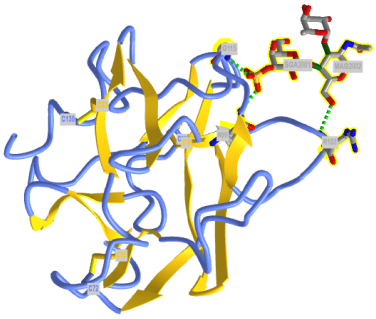


Figure 7.4.17: FN2 domain of the mannose receptor with the non-mannose containing 3-SO4-LEWIS(X) glycan (1fwu). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i..kXn2Rz8qMzebo9

Look at the number of CLECT domains in the domain structure diagram for the mannose receptor above. Along with interactions of sulfated glycans at the FN2 domain, these would enable the binding of widely diverse glycan structures. Reported ligands for the

mannose receptor include those with high mannose content released during inflammation (lysosomal hydrolases, collagen peptides, and tissue plasminogen activator), and sulfated ones (including the pituitary hormones lutropin and thyrotropin.

7.4.5: Galectins

This family of glycan-binding proteins contains a common carbohydrate recognition domain (CRD) of about 130 amino acids, which bind Galβ1,3GlcNAc or Galβ1,4GlcNAc disaccharides (hence the name galectins) as well as other glycan motifs. They are expressed in almost all cells and multicellular organisms. There are 15 different types grouped in how the CRD is functionally expressed (as dimers, tandem repeats, or chimeras), as illustrated in Figure 7.4.18. The figure also shows their role in cancer biology.

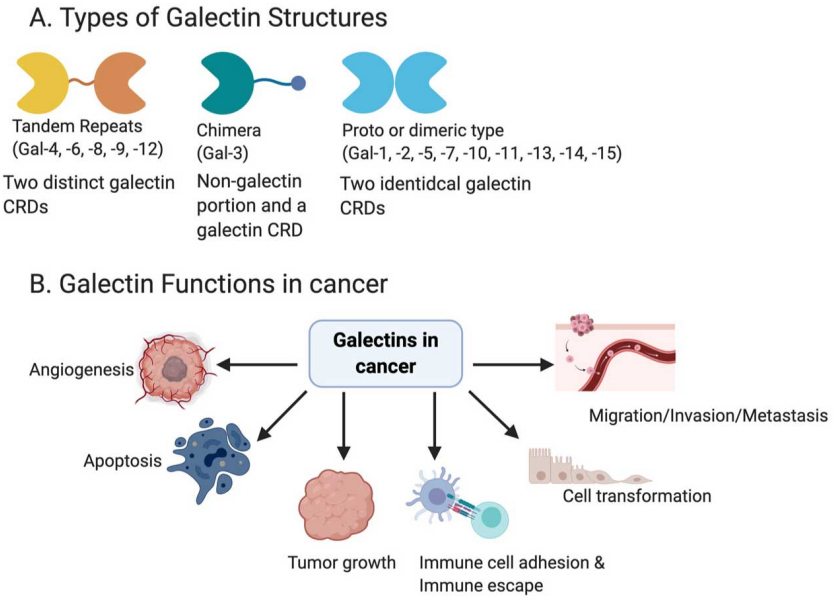


Figure 7.4.18: Galnectins structure and function Shimada, C.; Xu, R.; Al-Alem, L.; Stasenko, M.; Spriggs, D.R.; Rueda, B.R. Galectins and Ovarian Cancer. *Cancers* **2020**, *12*, 1421. https://doi.org/10.3390/cancers12061421. **Creative Commons Attribution License**

The carbohydrate-binding domain of the galactins has a jellyroll-like protein architecture with two anti-parallel β-sheets forming a β-sandwich.

Galectin I

This protein is secreted and is found in the extracellular matrix, as well as in the cytoplasm. It induces apoptosis in T-cells. It binds beta-galactosides as well as other glycans. The main ligand of galectin-1 has a Galβ1-4GlcNAc (or LacNAc) structure. Figure 7.4.19 shows an interactive iCn3D model of Human Galectin-1 in Complex with Type 1 N-acetyllactosamine (Gal(β1,3)GlcNAc), which binds less tightly than Galβ1,4GlcNAc (Type 2)(4XBL)

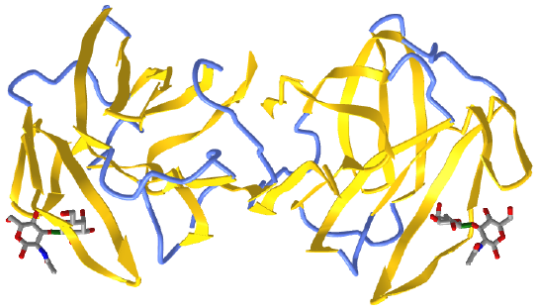


Figure 7.4.19: Human Galectin-1 in Complex with Type 1 N-acetyllactosamine (Gal(β1,3)GlcNAc) (4XBL). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i..MTyV7QvPKqqibA

A comparison of the crystal structures shows different phi/psi angles for bound Type I (135°) versus the more tightly bound Type 2 (-108°), which shows the nuance in binding conformations in the interactions of glycans with glycan-binding proteins.

7.4.6: Siglecs

The proteins are sialic acid-binding immunoglobulin (Ig)-like lectins found on immune cells like basophils, macrophages, mast cells, and eosinophils. One type (Siglec-4)is found in myelinated structures in the central and peripheral nervous systems. They all have an N-terminal extracellular immunoglobulin domain (abbreviated as IG or V-Set) and a differing number of IG-like domains, also called C2-set Ig domains. The glycan binding epitope recognized by Siglecs are sialylated oligosaccharides on a protein section containing a conserved arginine. Figure 7.4.20compares the domain structures of the human Siglec family.

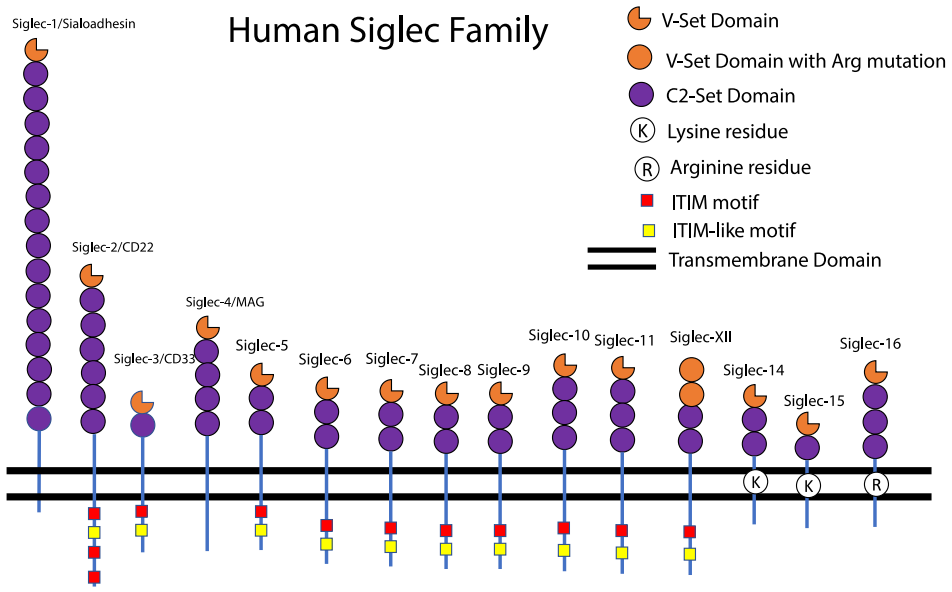


Figure 7.4.20: Domain structures of the human Siglec family Siddiqui, S.S.; Matar, R.; Merheb, M.; Hodeify, R.; Vazhappilly, C.G.; Marton, J.; Shamsuddin, S.A.; Al Zouabi, H. Siglecs in Brain Function and Neurological Disorders. *Cells* **2019**, *8*, 1125. https://doi.org/10.3390/cells8101125. Open access article distributed under the Creative Commons Attribution License

Here is one example of a Siglec.

Siglec-8

This protein is expressed on the surface of immune cells like basophil, mast cells and eosinophils. When activated by infection and prolonged inflammation, they release the contents of intracellular granules, which have potent physiological effects that can lead to allergic and asthmatic responses. On infection and other inflammatory states, immune cytokines are released that, in a signaling process, lead to the release of sialoglycans that act as ligands, binding to the Siglec-8 on the surface of the immune cells. One type of sialoglycan released is mucins, which are very large glycoproteins with many 6′S sLe^x glycans attached. These "multivalent" glycan epitopes can bind to Siglec-8 lead to signaling in the cells and ultimate inhibition of cell function (including by death or apoptosis). The mucins in mucus (cross-linked mucins), which cover epithelial cells or airways, also act as a first line of defense as they can bind viruses through mulitple-contact (multivalent) binding sites, effectively trapping the viruses. The glycan structure recognized by Siglec-8 is sialic acid and sulfate (NeuAcα2-3[6S]Galβ1-4G[Fucα1-3]GlcNAc-). Given their role in inhibiting and inducing apoptosis in immune cells, the family of siglecs are likely involved as checkpoints, which are important in cancer and inflammatory conditions.

Figure 7.4.21shows the domain structure of Siglec-8

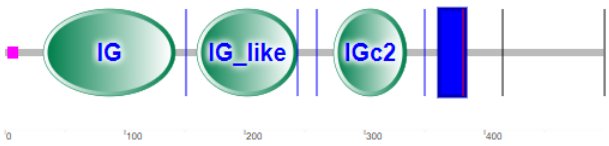


Figure 7.4.20: Domain structure of Siglec-8

Note that there is no CLECT domain, but rather immunoglobulin- (IG) or IG-like domains, which seems logical given their role in binding glycan "epitopes". The IG domain is also called the immunoglobulin V-set domain (V-Set). The blue rectangle represents the transmembrane domain (single helix). The cytoplasm contains a tyrosine-inhibitory motif (ITIM) involved in transducing the signal on binding 6′S sLe^x glycans to the IG domains.

As discussed above, humans lack a hydrolase gene necessary for the hydroxylation of Neu5Ac to Neu5Gc, which is found in chimps that possess the enzyme. Chimp's immune systems seem to confer protection from acquiring simian versions of AIDS, cirrhosis, and other diseases which humans acquire when they are infected with the human versions of the HIV, hepatitis B or C, or other viruses. These diseases and others associated with overactive T cells (rheumatoid arthritis, asthma, type-I diabetes) are uncommon in chimps. It turns out that there is a link between the type of sialic acid and the expression of siglecs that influences the difference for our disease propensity. Varki et al. have shown that chimps and gorillas show much higher levels of expression of siglecs on T cells, which are critical regulatory and effector cells in the immune system. When siglecs on T cells are activated, T-cell responses are down-regulated. Although the HIV virus ultimately kills T helper cells, the virus initially activates them on infection, leading to their proliferation and production of a larger number of cells for the virus to infect.

Figure 7.4.21 shows an interactive iCn3D model of human Siglec-8 lectin domain in complex with 6'sulfo sialyl Lewisx (2N7B)

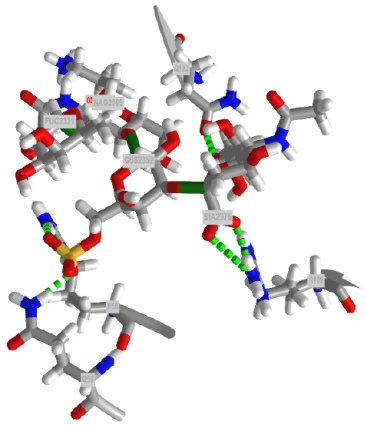


Figure 7.4.19: Human Siglec-8 lectin domain in complex with 6'sulfo sialyl Lewisx (2N7B). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i..Gaa93cu86jMPA9

7.4.7: Summary

This chapter delves into the complex world of protein glycosylation and glycan interactions, highlighting the extraordinary structural diversity and biological significance of carbohydrate modifications. Although the genetic code provides instructions for protein synthesis, glycan structures are not directly encoded, resulting in a vast array of possible modifications that contribute to the functional complexity of the proteome.

Key topics include:

- **Protein Glycosylation and Its Roles:**
  The chapter introduces how carbohydrates are post-translationally attached to proteins—most commonly via N-linked glycosylation (through Asn in a consensus sequence) and O-linked glycosylation (typically via Ser or Thr). These modifications enhance protein folding, stability, and protection from proteolysis while also mediating critical binding interactions.
- **Diversity in Glycan Structures:**
  N-linked glycans share a common core structure $(Man)_3(GlcNAc)_2$ but diverge into high-mannose, complex, or hybrid types based on further modifications and branching patterns. Similarly, O-linked glycans, exemplified by blood group antigens, show a wide range of structures that affect cell–cell recognition and immune responses.
- **Glycan-Binding Proteins (GBPs):**
  The chapter explains how specific proteins—often termed lectins—serve as “readers” of the glycan code. GBPs, including C-type lectins (like selectins), galectins, and siglecs, bind to defined glycan epitopes and mediate essential functions such as immune cell adhesion, signaling, and pathogen recognition. Structural insights, including interactive 3D models and symbolic nomenclature (SNFG), help illustrate these interactions.
- **Glycans in Cellular Architecture:**
  Beyond glycoproteins, the role of glycans in forming complex extracellular matrices, bacterial cell walls, and plant cell walls is

examined. The unique composition and organization of these carbohydrate-rich structures underpin critical functions—from maintaining cell shape and rigidity to modulating cellular interactions and immune responses.

- **Biological and Clinical Implications:**
  The chapter also highlights the clinical importance of glycan structures. For example, variations in terminal glycan epitopes can trigger immune responses (e.g., in xenotransplantation or alpha-gal syndrome) and influence the infectivity of viruses such as influenza and SARS-CoV-2.

Overall, this chapter underscores that glycosylation is a fundamental, yet highly complex, modification that plays vital roles in protein function, cell signaling, and organismal physiology. It prepares the reader to appreciate how the dynamic "glycan code"—decoded by specialized glycan-binding proteins—contributes to both normal biological processes and disease states.

7.5: Working with Carbohydrates

Learning Goals (ChatGPT o1, 1/30/25)

- **Describe the Core Structure and Diversity of N-Glycans:**
  - Explain the structure of the core pentasaccharide (Man)3(GlcNAc)2(\text{Man})_3(\text{GlcNAc})_2 found in N-linked glycoproteins.
  - Distinguish among the three main types of N-glycans—high mannose, hybrid, and complex—and describe how additional modifications (e.g., sialylation, core fucosylation, bisecting GlcNAc) contribute to their structural diversity.
- **Explain the Biosynthesis and Functional Roles of N-Glycans:**
  - Outline the process of N-glycosylation from the attachment of the high-mannose precursor in the ER to the modification in the Golgi apparatus.
  - Discuss how N-glycans affect protein folding, stability, half-life, and protection from proteolysis.
  - Analyze the biological roles of N-glycans in cellular recognition, immune response, and signal transduction.
- **Evaluate Methods for the Synthesis and Isolation of N-Glycans:**
  - Compare chemical synthesis, enzymatic synthesis, and isolation from natural sources as strategies to obtain pure, homogeneous N-glycans.
  - Assess the advantages and limitations of each method in the context of functional studies.
- **Analyze Glycan–Lectin Interaction Techniques:**
  - Summarize the principles behind glycan array technology and how it enables high-throughput analysis of glycan–lectin interactions.
  - Describe the use of STD-NMR (Saturation Transfer Difference NMR) in mapping glycan epitopes and determining the atomic-level details of glycan–lectin binding.
  - Explain how these analytical techniques contribute to understanding structure–activity relationships.
- **Relate Glycan Structure to Biological Function and Disease:**
  - Evaluate how differences in glycan branching, chain length, and terminal modifications influence lectin recognition and downstream cellular responses.
  - Discuss examples such as the role of sialic acids in immune regulation via Siglecs and the implications of glycan structures in cancer metastasis.
- **Explore Applications in Drug Development and Therapeutics:**
  - Investigate how homogeneous glycoproteins with defined N-glycan structures can be used to study glycan functions and improve the pharmacokinetics of therapeutic proteins.
  - Explain the concept of glycan editing and its potential for designing next-generation vaccines and targeted drug delivery systems.
- **Integrate the “Glycan Code” Concept:**
  - Understand how the vast structural diversity of glycans acts as an “information-rich” code on the cell surface.
  - Describe how glycan-binding proteins (lectins) “read” this code to mediate cellular recognition and signal transduction.

By achieving these goals, students will gain a comprehensive understanding of the synthesis, structural diversity, functional implications, and modern analytical methods used in the study of N-linked glycosylation and glycan–lectin interactions.

synthesize of N-glycans, but the same principles would be used for study of O-glycans, proteoglycans, etc. Text in boxes have been added to offer addition simplifying information when necessary for undergraduate students. References can be found in the original paper linked above.

7.5.1: Introduction

Glycosylation is the most common post-translational modification of proteins. Over 60% of proteins are linked to glycans. Asparagine-linked oligosaccharides (*N*-glycans) have a core pentasaccharide composed of mannose and glucosamine and are classified into three types: high-mannose, hybrid, and complex as shown in **Figure 7.5.1** below. In the biosynthesis of *N*-glycan-modified proteins, the high-mannose type *N*-glycan consisting of 14 residues ($Glc_3Man_9GlcNAc_2$) is first attached to proteins in the endoplasmic reticulum (ER). The initial high-mannose *N*-glycans play an important role in protein folding in the ER. Glycoproteins then migrate to the Golgi apparatus and are subsequently converted into complex-type *N*-glycans. Complex *N*-glycans have diverse structures due to differences in their associated synthesizing enzymes, resulting in different functions for each structure. For example, polylactosamine, consisting of a repeating structure of galactose and glucosamine, is involved in cancer metastasis and immune response. Sialic acids, in contrast, control immunity via recognition by Siglecs expressed in immune cells. Core fucose, which is a fucose linked to the glucosamine 6 position at the reducing end, and bisecting glucosamine, which is a glucosamine linked to the branched mannose 4 position, also play various roles and are closely related to many diseases. Hybrid *N*-glycans have both high-mannose and complex-type structures. Thus, *N*-glycans have diverse structures and are involved in a variety of biological phenomena. However, the molecular bases of their modes of action are yet to be fully elucidated.

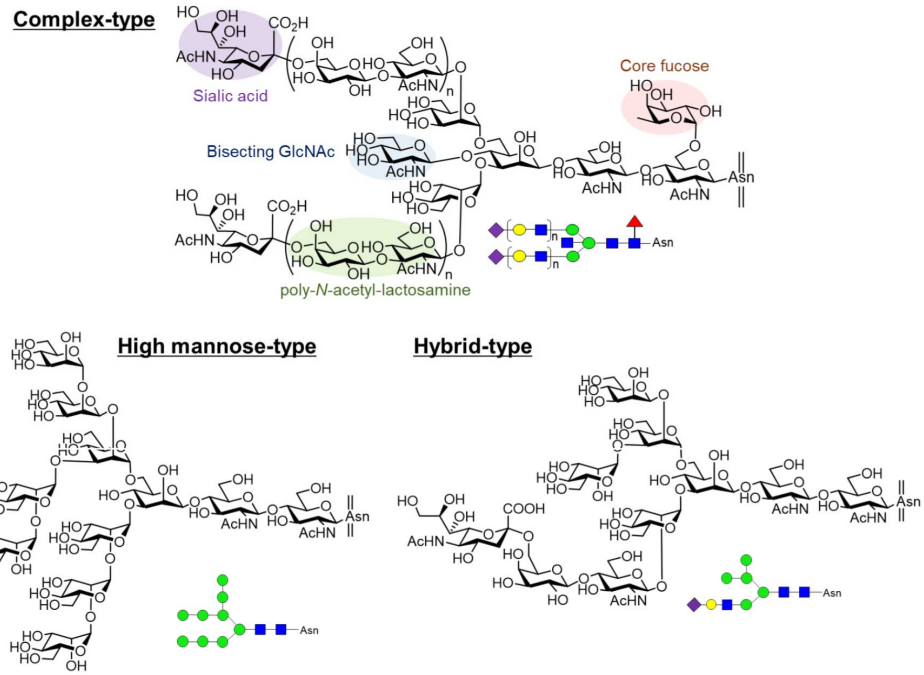


Figure 7.5.1: Structures of *N*-glycans. Complex-type *N*-glycans have diverse structures with/without sialic acid, poly-*N*-acetyl-lactosamine, bisecting GlcNAc, core fucose and so on. High-mannose-type *N*-glycan is composed of 14 residues ($Glc_3Man_9GlcNAc_2$) containing 3 glucoses, 9 mannoses, and 2 GlcNAc. Hybrid-type *N*-glycans have both high-mannose and complex-type structures.

Chemical synthesis, enzymatic synthesis, and isolation of diverse, pure *N*-glycans have been vigorously investigated for analyzing *N*-glycan functions at the molecular level. Chemical synthesis is an extremely potent approach that allows the de novo construction of glycan structures. Any desired glycan structures can be constructed, including partial and artificial structures. Danishefsky et al. successfully synthesized various *N*-glycans with multiantennary structures. Unverzagt et al. achieved convergent synthesis of complex-type *N*-glycans with bisecting glucosamine and/or core fucose. We have also reported the synthesis of *N*-glycans. In addition, Ito et al, Boons et al., Wang et al., Wong et al., and Schmidt et al. have achieved *N*-glycan synthesis. Since the chemical synthesis of *N*-glycans with complex structures is a challenging process that requires multiple steps, the isolation of *N*-glycans from natural resources has also been explored. Kajihara et al. established an efficient method for the isolation of *N*-glycans from egg yolk, which has become a standard method for *N*-glycan preparation. In recent years, the preparation of *N*-glycans using enzymatic reactions has also been extensively investigated. Ito et al., Boons et al., Wang et al., and Wong et al. successfully constructed a

wide range of *N*-glycan libraries via enzymatic synthesis using isolated or chemically synthesized glycans as substrates. Thus, over the years, the technical basis for a sufficient supply of various *N*-glycans has been established.

Owing to the increased availability of pure *N*-glycans, their functional elucidation has advanced considerably in recent years as shown on **Figure 7.5.2** below. *N*-Glycan functions have mainly been analyzed using molecular biological techniques, including knockout of biosynthetic enzymes. However, it is difficult to determine the precise structure–activity relationship using these methods. Although interaction analysis of lectins using relatively small glycan fragments, such as disaccharides and trisaccharides, has been used to study function, it is not possible to estimate the conformational effect or multivalent interactions of the complex structure of *N*-glycans. Recent interaction analysis of lectins using various *N*-glycans has elucidated the significance of such complex structures. The increased availability of *N*-glycans also allows one to prepare glycoproteins with homogeneous glycoforms, enabling the elucidation of *N*-glycan function on the distinct protein. Furthermore, *N*-glycans have the potential to be used in the development of novel drugs. This review provides an overview of the recent chemical biology study of *N*-glycans.

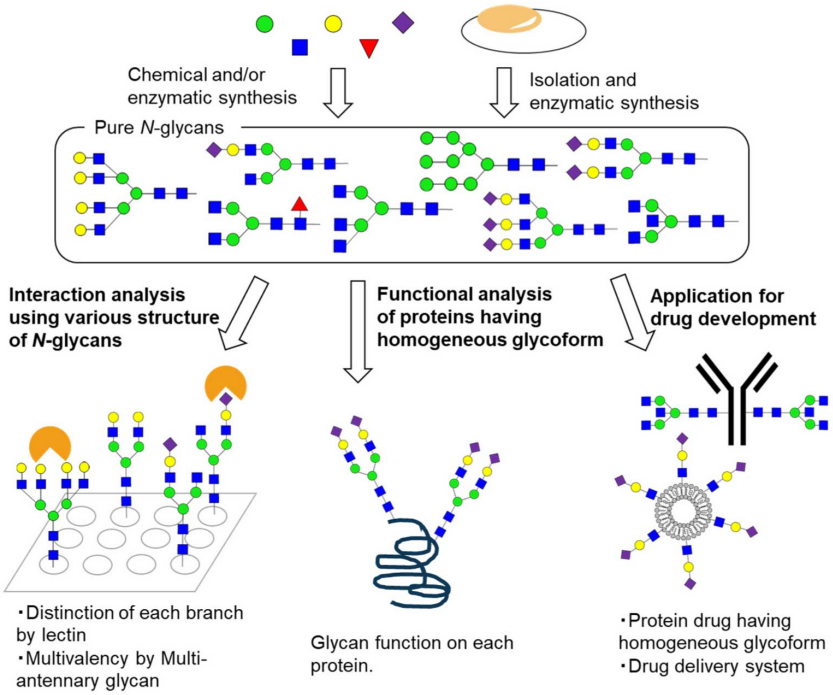


Figure 7.5.2: The chemical biology study of homogeneous *N*-glycans. Chemical synthesis, enzymatic synthesis, and isolation of diverse, pure *N*-glycans enable their functional analysis at the molecular level. Interaction analysis using various *N*-glycans revealed the significance of complex *N*-glycan structures—for example, distinction of each branch and multivalent interaction in lectin recognition. Functional analysis using proteins with homogeneous glycoforms clarifies glycan function on each protein. Application of *N*-glycans for drug development is also investigated

7.5.2: Elucidation of the Molecular Basis of *N*-Glycan Recognition by Lectins

Glycans control various biological phenomena through their recognition by lectins. Thus, interaction analysis between *N*-glycans and lectins is essential to elucidate *N*-glycan functions .

7.5.2.1: Methods for the Glycan–Lectin Interaction Analysis

The glycan–lectin interaction analysis methods include analyses using glycan arrays, nuclear magnetic resonance (NMR), isothermal titration calorimetry (ITC), surface plasmon resonance (SPR), fluorescent polarization (FP), and X-ray crystallography. ITC gives thermodynamic parameters, whereas SPR provides kinetic parameters. FP realizes a simple and easy assay system. X-ray crystallography provides precise structural information. In this review, we focus on studies using glycan arrays and NMR, which are effective methods for elucidating the interaction between glycans and lectins.

Glycan arrays are used to detect the binding of lectins to immobilized glycans. An advantage of a glycan array is that a large number (dozens and hundreds) of samples can be examined in a high-throughput manner using a small amount of glycans. Interaction analysis using various structures of glycans provides insights into precise structure–activity relationships. Glycan immobilization methods are divided into two categories: noncovalent and covalent. Noncovalent immobilization utilizes hydrophobic interactions, charge interactions, and biotin–streptavidin interactions among others. Covalent immobilization methods

typically use coupling of an amino group introduced at the reducing end of a glycan to the plate surface activated with *N*-hydroxysuccinimide. Many other methods, including thiol-maleimide coupling and alkyne-azide click reactions, have been reported. As for the detection, fluorescence is usually used to realize high-throughput analysis.

NMR can be used to analyze interactions at the atomic level. Saturation transfer difference (STD) NMR is a particularly powerful method for the analysis of glycan–lectin interactions. In this method, saturation transfer from the protein to the ligand is observed as STD signals after the saturation of protein by radio frequency as shown in **Figure 7.5.3** below. The closer the protons are to the protein, the stronger the STD signals observed. STD-NMR was originally developed as a method for screening ligands from mixture systems but is now widely used for the analysis of protein–ligand binding modes. This method works when the affinity is not high (K_D is 10^{-3} to 10^{-8} M), because STD signals are measured when the protein and ligand are in an equilibrium state of binding and dissociation. Since glycan–lectin interactions are usually weak (K_D in mM-μM), STD-NMR is highly effective. NMR is also a powerful tool for the conformational analysis of glycans. Importantly, this method does not require labeled proteins. In addition, a small amount of receptor is necessary (typically micromolar range). However, an excess of the ligands is used (typically the millimolar range), thus, low solubility of the ligand causes a problem. While conformation is an important factor for glycan recognition, the flexibility of glycans makes conformational analysis difficult. In addition to analysis based on coupling constants and the nuclear Overhauser effect (NOE), an analysis using pseudocontact shift (PCS) by paramagnetic metals has recently been developed, and its efficiency has been demonstrated.

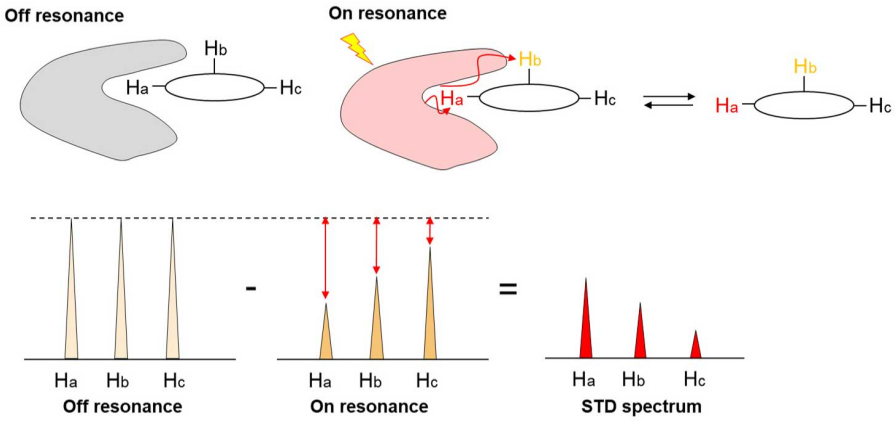


Figure 7.5.3: Mechanism of Saturation transfer difference nuclear magnetic resonance (STD-NMR). “Off resonance” experiment gives a reference spectrum. Under “on resonance” conditions, the saturation is transferred from protein to ligand by spin diffusion through intermolecular nuclear Overhauser effects (NOEs). The closer the protons are to the protein, the stronger the STD signals that are observed.

Examples of the analysis of glycan–lectin interactions using glycan arrays and NMR are introduced below.

7.5.2.2: Analysis of Sugar–Lectin Interactions Using Glycan Arrays

Glycan arrays are excellent tools for the comprehensive analysis of glycan–lectin interactions. Previous interaction analysis using small fragments, such as disaccharides and trisaccharides, revealed the minimum structure (epitope) required for recognition of individual lectins. Meanwhile, recent advances in the preparation of the whole structure of various *N*-glycans have allowed the full realization of structure–activity relationships, elucidating the significance of complexity of *N*-glycan structures as shown in **Figure 7.5.4** below. For example, these advances have provided insights into the differences in the lectin recognition of each branch, the improvement of affinity due to the inclusion of multiple recognition units (multivalent effect), the influence of chain length on affinity, and remote (heterovalent) recognition.

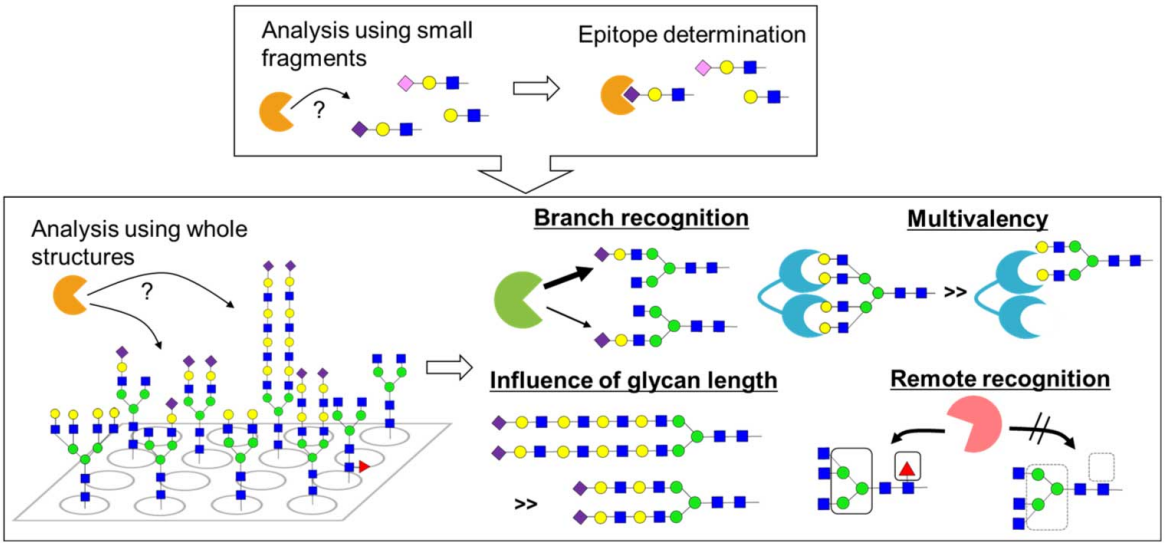


Figure 7.5.4: Analysis of glycan–lectin interaction using glycan arrays. Interaction analysis using small fragments, such as disaccharides and trisaccharides, revealed the epitope required for lectin recognition, whereas interaction analysis using whole structures of *N*-glycans revealed the significance of complexity of *N*-glycan structures; these analyses provided the insights into the differences in the lectin recognition of each branch, the improvement of affinity due to the inclusion of multiple recognition units (multivalent effect), the influence of chain length on affinity, and remote (heterovalent) recognition

Wang et al. demonstrated the differences of each *N*-glycan branch in lectin recognition by comprehensive interaction analysis of various *N*-glycans with several lectins using a glycan array. Plant-derived *Sambucus nigra* lectin (SNA), which recognizes sialic acid, recognized the sialic acid on the α1,3-branched chain more strongly than the sialic acid on the α1,6-branched chain. Meanwhile, plant-derived *Maackia amurensis* lectin (MAL-I) and virus-derived lectin hemagglutinin (HA) strongly bind to sialic acid on the α1,6-branched chain. MAL-I also interacts with terminal galactose; in this case, MAL-I strongly recognizes galactose on the α1,3-branched chain, suggesting that MAL-I has two distinct glycan recognition domains. In addition, *Erythrina cristagalli* lectin (ECL), which recognizes the lactosamine structure, has a higher affinity to lactosamine on the α1,3-branched chain than on the α1,6-branched chain. *Phaseolus vulgaris* erythroagglutinin (PHA-E) prefers terminal galactose on the α1,6-branched chain and terminal glucosamine on the α1,3-branched chain, whereas wheat germ agglutinin (WGA) strongly interacts with glucosamine on the α1,3-branched chain.

Branch selective binding of C-type lectins and monoclonal antibodies was also revealed by using glycan array including *N*-glycan positional isomers prepared by chemo-enzymatic method. DC-SIGN, C-type lectin recognizing glycan on bacteria and viruses, showed strong binding to hybrid- and complex-type glycans and *N*-glycans presenting Lex epitopes. DC-SIGN showed preferential binding to the biantennary glycans with terminal galactose or *N*-acetylgalactosamine on the α1,6-branched chain, whereas DC-SIGNR showed the opposite binding behavior. L-SECtin showed the preference to GlcNAc1,2-Man residues on the 3-arm of the complex and hybrid *N*-glycans.

N-glycans have symmetric structures on the nonreducing end side, and these studies indicate that glycan structures on each branched chain have distinct functions.

The structural redundancy of *N*-glycans plays an important role in enhancing their affinity to lectins because of their multivalency. Interaction analysis of Siglec-1, -2, -9, and -10 with sialic acid-containing *N*-glycans using a glycan array showed a higher affinity for four-branching *N*-glycans than for two-branching *N*-glycans. Multivalent effects were also confirmed in ECL, which recognizes the lactosamine structure, and *Ricinus communis* agglutinin (RCA120), which recognizes terminal galactose. Similarly, interactions with galectin were also enhanced as the number of recognition units increased.

The structure at the remote positions of the lectin recognition unit can affect its interaction. *Lens culinaris* agglutinin (LCA), which binds to core fucose, only recognizes core fucosylated biantennary and triantennary *N*-glycans with particular branching patterns, but did not recognize triantennary *N*-glycans with other branching patterns or tetraantennary *N*-glycans. These results indicate that the branching structure away from the core fucose affected recognition by LCA, although its recognition site is core fucose. On the other hand, the affinity between HA from H3N3 and sialic acids at the nonreducing end was increased by chain elongation; the insertion of a polylactosamine repeating structure enhanced the affinity. These studies revealed that both the epitope and the whole glycan structure are important for the recognition of *N*-glycans.

Glycan arrays, comprising *N*-glycans along with glycolipids and *O*-glycans, have been used to investigate the host–pathogen interactions in diagnostic and therapeutic applications. For example, the inhibition of human anti-N9 antibodies to influenza neuraminidases was analyzed by glycan array. The binding study of the H3N2 influenza viruses using glycan microarrays demonstrated the changes in virus hemagglutinin that affect the receptor binding properties of the viruses.

Glycan arrays can also be used to explore artificial glycoligands as new drug candidates that target lectins. High-affinity ligands for Siglecs or several C-type lectins, which are involved in immune regulation, are expected to be lead compounds for drug development. However, glycan–lectin interactions are usually weak, which is a major issue in the utilization of glycans as bioactive molecules. Thus, the synthesis of glycans and derivatization of artificial molecules, followed by high-throughput screening using glycan arrays, is expected to be a powerful approach to address this limitation.

7.5.2.3: Analysis Using NMR

NMR analysis can provide insights into glycan–lectin interactions at the atomic level. STD-NMR can be used for high-resolution epitope mapping. Similar to the results obtained from glycan arrays, NMR analysis also reveals that not only epitopes but the whole structure of *N*-glycans plays an important role in glycan–lectin interactions. The conformational analysis of *N*-glycans using NMR is a powerful approach that provides a rational explanation for the molecular basis of the recognition of complexity of *N*-glycan structures.

STD-NMR allows for a detailed analysis of glycan–lectin interactions. Many researchers have analyzed the interactions between sialic acid containing *N*-glycans and Siglecs . Silipo et al. analyzed the interaction between Siglec-2 and sialyl *N*-glycans using STD-NMR and molecular dynamics (MD) simulations. The Siglec-2 epitope was clearly shown by STD-NMR, and the conformation of sialyl *N*-glycans was predicted by NMR analysis and MD simulations. When biantennary sialyl *N*-glycans were recognized, Siglec-2 only interacted with the sialyl disaccharide at the nonreducing end, and the other part was expected to protrude from the protein surface. These results suggest that multiantennary *N*-glycans with multiple sialic acids can interact with several Siglec-2 and induce the formation of Siglec-2 oligomers on B cells.

STD-NMR analysis using the whole structure of *N*-glycans has demonstrated that lectins not only recognize small units, such as disaccharides and trisaccharides, but also interact with *N*-glycans in a more complex manner. In *N*-glycan recognition by *Pisum sativum* agglutinin (PSA), a mannose-recognition lectin, core fucose was shown to alter its binding mode. When biantennary *N*-glycans without core fucose were used for the interaction analysis with PSA, the mannose on each branch gave comparable STD signals. While for the core fucose containing *N*-glycans, the STD signals of mannose on the α1,6-branched chain were weakened, and instead, an interaction with the methyl group of the core fucose was observed. STD-NMR using a fluorine derivative (2D STD-TOCSYreF) indicated that the mannose on the α1,3-branched chain was more strongly recognized by PSA than the mannose on the α1,6-branched chain. In addition, dectin-1, which recognizes fungal β-glucan, was found to recognize core fucose on immunoglobulin (IgG) . STD-NMR analysis indicated that dectin-1 interacted not only with core fucose but also with an Fmoc group attached to the amino group of asparagine introduced at the reducing end. These results suggest that dectin-1 recognizes amino acids with aromatic side chains, such as phenylalanine and tyrosine, together with core fucose. On the other hand, STD-NMR is also effective for the analysis of substrate recognition by glycosyltransferases. The STD-NMR analysis of FUT8, a fucosyltransferase that builds core fucose structure, revealed the precise interaction between FUT8 and *N*-glycan . FUT8 recognizes not only glucosamine at the reducing end (reaction point) but also the whole glycan structure. In particular, FUT8 strongly interacted with the α1,3-branched chain at the nonreducing end.

Advanced STD-NMR methods have been developed. Saturation transfer double difference (STDD)-NMR is useful for the direct observation of ligands binding on the surfaces of living cells. Clean-STD can avoid accidental saturation to give improved detection of ligand–protein interactions at low concentration of protein. Second dimension STD-NMR, i.e., STD-TOCSY, STD-HSQC, STD-NOESY, can overcome the problems of proton overlapping typical of glycan NMR analysis.

Conformation analysis of glycans using NMR provides important insights into complex glycan–lectin interactions. The *N*-glycan conformation can be predicted by combining PCS-based NMR analysis and MD simulations as shown in **Figure 7.5.5** below. Kato et al. analyzed the conformation of high-mannose glycans by PCS-based NMR analysis using ^{13}C-labeled compounds and Tm^{3+} as a paramagnetic metal ion tag. They elucidated the conformational change caused by mannose trimming during the *N*-glycan biosynthetic process. Unverzagt and Barbero et al. distinguished each branch of tetraantennary *N*-glycan based on the PCS method and analyzed the differences in the recognition of each branch by lectins . *Datura stramonium* seed lectin (DSL), which recognizes the lactosamine structure, interacts more strongly with the lactosamine on the α1,6-branched chain than with that on the α1,3-branched chain. On the other hand, no differences in the strengths of STD signals of each branch were observed with *Ricinus communis* agglutinin (RCA120), which recognizes terminal galactose, indicating that RCA120 recognizes all branches without

distinction. In a similar analysis between sialic acid containing biantennary *N*-glycans and HA, STD signals from both sialic acids were observed, suggesting the contribution of two sialic acids in a multivalent effect. Furthermore, interesting results have been reported showing that the *N*-glycan conformation directly affects lectin recognition. *N*-Glycans have three back-fold conformations and two extended conformations, in which the α1,6-branched chain is folded toward the reducing end or extended, respectively. The addition of core fucose or bisecting glucosamine significantly changes their conformational equilibria and reduces the number of major conformations from five to four and five to two, respectively. Crystal structure analysis and transferred NOE (TrNOE) analysis revealed that *Calystegia sepium*-derived calsepa and *Phaseolus vulgaris*-derived phytohemagglutinin (PHA-E), which recognize bisecting glucosamine containing *N*-glycans, recognize *N*-glycans in the back-fold conformation induced by bisecting glucosamine addition

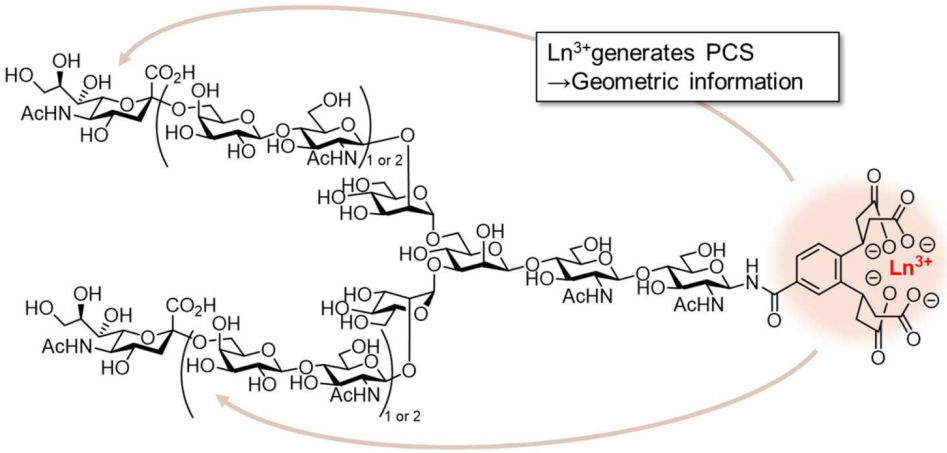


Figure 7.5.6*a*: Conformation analysis of *N*-glycan using pseudocontact shift (PCS). Chelation with paramagnetic metals can induce PCS to give geometric information of *N*-glycan.

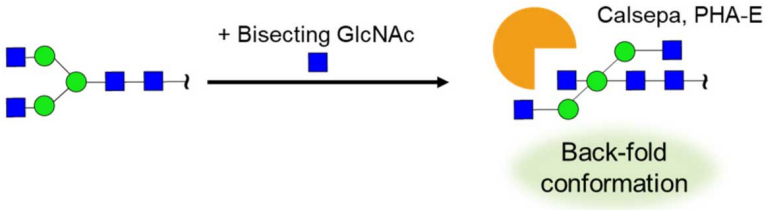


Figure 7.5.6*b*: Recognition of bisecting GlcNAc containing *N*-glycan by Calsepa and *Phaseolus vulgaris* erythroagglutinin (PHA-E). Attachment of bisecting GlcNAc enhances back-fold conformation, which is recognized by Calsepa and PHA-E.

7.5.3: Functional Analysis of *N*-glycans on Glycoproteins

Analysis of *N*-glycan functions on glycoproteins needs to be considered with proteins. In recent years, improvements in the techniques for the synthesis of peptides and proteins, as well as glycans, have enabled the preparation of glycoproteins with homogeneous glycans. *N*-Glycans on glycoproteins can be modified by Endo-β-*N*-acetylglucosaminidases (ENGases). Synthesized glycoproteins with homogeneous glycan structures have helped elucidate precise glycan functions.

A series of synthetic studies of glycoproteins and glycoprotein mimics by Ito and Kajihara et al. revealed the precise function of *N*-glycans in a quality-control mechanism for glycoproteins in the endoplasmic reticulum (ER). ER has a quality control system that promotes the correct folding of ribosome-produced proteins. In the case of *N*-glycosylated proteins, high-mannose *N*-glycans work as tags for protein folding. A common dolichol-linked oligosaccharide precursor containing terminal glucose trisaccharide is first synthesized in the ER and is transferred to proteins by the oligosaccharyltransferase (OST). The folding process then starts. The first glycosidase (GCSI) cleaves the terminal glucose and the second glycosidase (CGSII) further cleaves glucose residues to afford monoglucosylated or nonglucosylated glycoproteins. The folded nonlucosylated glycoproteins are then transferred to the glycan modification process. The UDP-glucose:glycoprotein glucosyltransferase (UGGT) complex distinguishes misfolded glycoproteins and transfers glucose to the nonreducing end of the high-mannose glycan. This monoglucosylation serves as a marker for misfolded glycoproteins and the chaperone proteins calnexin/calreticulin (CNT/CRT) promotes folding. CGSII then cleaves glucose residue to transfer the glycoproteins for the glycan modification process. This cyle is illustrated in **Figure 7.5.7** below.

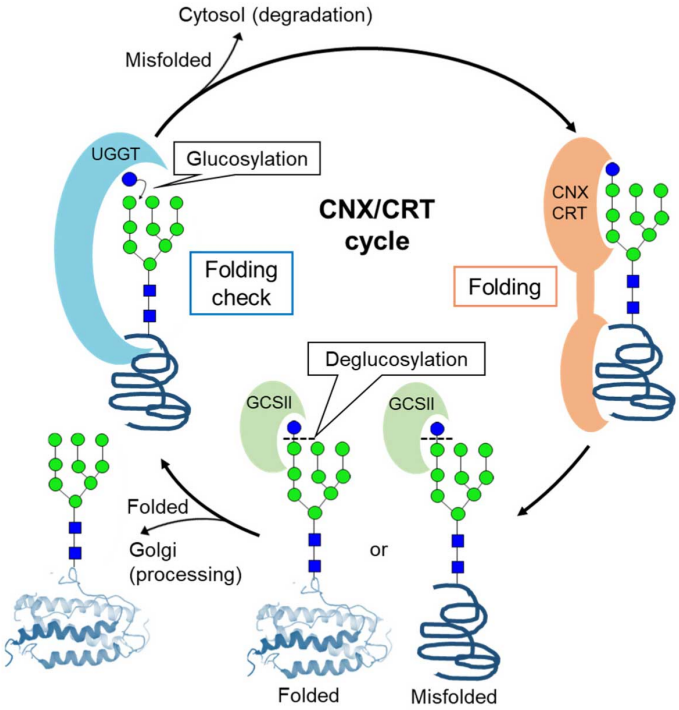


Figure 7.5.7: Protein quality control utilizing high-mannose-type *N*-glycan as a tag. UDP-glucose:glycoprotein glucosyltransferase (UGGT) complex distinguishes misfolded glycoproteins to transfer glucose to the nonreducing end of the high-mannose glycan. This monoglycosylation serves as a marker for misfolded glycoproteins and the chaperone proteins calnexin/calreticulin (CNT/CRT) promotes folding.

The defects in this process cause congenital disorders of glycosylation (CDGs), which are severe genetic diseases. CDG is classified into Type I and Type II. In Type I, the enzymes are mutated in synthesis and transfer a common dolichol-linked oligosaccharide precursor and enzyme substrates. Type II defects the modification process of *N*-glycans in the ER and Golgi. Lack of GCS1 causes CDG-IIb. Unfolded proteins lead to ER stress and cause CDGs .

Ito et al. introduced methotrexate (MTX) at the reducing end of high-mannose *N*-glycans and prepared a complex with dihydrofolate reductase (DHFR), which recognizes MTX. Such glycoprotein mimics were used to analyze the interaction with UGGT. They also investigated various aglycone structures as substrates of UGGT. In addition, chemically synthesized glycoproteins were used for the analysis of substrate recognition by UGGT. UGGT showed higher enzymatic activity against high-mannose *N*-glycans on misfolded interleukin-8 (IL-8) than against those on the folded one. Furthermore, they synthesized several glycoproteins and isotope-labeled glycopeptides and revealed that UGGT recognizes hydrophobic patches on misfolded proteins. As shown above, they elucidated the molecular basis of the quality-control mechanism based on high-mannose *N*-glycans using glycoprotein mimics and chemically synthesized glycoproteins.

Maintaining the appropriate folding is also critical for in the degradation process. Mutations in human *N*-glycanase 1 (NGLY1) cause the congenital disorder of deglycosylation (CDDG). Suzuki revealed that *N*-GlcNAc proteins are accumulated by the action of Endo-β-*N*-acetylglucosaminidase (ENGase) in Ngly1-defective cells . During ER-associated degradation, *N*-GlcNAc proteins form aggregates that seem to be toxic. Suzuki also revealed that lethality of Ngly1-KO mice is partially rescued by the additional deletion of the Engase gene, suggesting that ENGase inhibitors are targets for CDDG.

In recent years, the influence of *N*-glycan modifications on the bioactivity of proteins has been gradually elucidated using synthetic glycoproteins. Hematopoietic hormone erythropoietin (EPO), which is used to treat renal anemia, has three *N*-glycan-modification sites. EPO with various glycoforms is used as a drug. Several groups have reported the synthesis of EPO with homogeneous glycoforms, and the effect of *N*-glycans on their biological activities has been investigated. In addition, various neoglycoprotein analogues of EPO have been reported. Kajihara et al. synthesized five types of EPO, which is introduced sialic acid containing *N*-glycans into three *N*-glycosylation sites with different patterns, and showed the relationship between glycosylation sites and hematopoietic activities. Increasing the number of sialic acids containing *N*-glycans on EPO improved the stability in blood, leading to an improvement in hematopoietic activity. Moreover, the metabolic stability of EPO was highly correlated with hydrophobicity, suggesting that glycan modifications enhance the in vivo stability by covering hydrophobic sites on the protein

surface. Kajihara et al. also synthesized two types of interferon-β (IFN-β) with sialic acid-containing and noncontaining (asialo) *N*-glycans, and their activities were evaluated. IFN-β modified with sialic acid-containing *N*-glycans exhibited higher activity than that modified with asialo *N*-glycans, suggesting that sialic acid extended the in vivo half-life of IFN-β. Thus, *N*-glycans are closely related to the stability of glycoproteins in vivo. Indeed, Tanaka et al. demonstrated the effect of *N*-glycans on protein metabolic stability by positron emission tomography (PET) imaging using glycodendrimers as pseudoglycoproteins. On the other hand, *N*-glycosylation can also affect binding affinity to a receptor. Okamoto et al. synthesized two types of chemokine CCL1 with and without *N*-glycan, in which *N*-glycosylation reduced the activity of CCL1, suggesting that CCL1 biological activity can be regulated by *N*-glycan modification. Thus, it should be noted that the role of *N*-glycan modifications can be different between proteins. We reported that dectin-1 specifically recognized core fucosylated IgG and did not interact with other core fucosylated proteins, suggesting that core fucose on IgG has specific physiological functions. The role of *N*-glycans on distinct proteins is an important topic for future work.

7.5.4: Use of *N*-glycans for Drug Development

The increased supply of *N*-glycans has led to an increase in the use of *N*-glycans for drug development. Because *N*-glycans are endogenous molecules, they are unlikely to be toxic or immunogenic and, thus, are expected to have high safety profiles.

7.5.4.1: Next-Generation Protein/Peptide Drugs Modified with Homogeneous N-Glycans

Controlling the glycan structure is an important issue in the preparation of glycoprotein and glycopeptide drugs. Biopharmaceuticals, including antibodies, are common pharmaceuticals. Although many proteins utilized in biopharmaceuticals are glycoproteins, their actual glycan structures are often neglected or ignored. However, the significance of the role of glycans on the function of glycoproteins has recently been illuminated, and the importance of the glycan structure has been highlighted. The preparation of glycoproteins with homogeneous glycans is also important from the viewpoint of quality control.

IgG antibodies have *N*-glycans at Asn297 in the Fc region of the heavy chain, and their structures affect activity, dynamics, and safety (**Figure 7.5.8**). The importance of core fucose on these *N*-glycans is well known. The removal of the core fucose from IgG antibodies dramatically enhances antibody-dependent cellular cytotoxicity (ADCC) activity . Mogamulizumab, the antibody without core fucose, is actually in current use. Bisecting glucosamine and terminal galactose have been reported to affect ADCC and complement-dependent cytotoxicity (CDC) activities. Therefore, modifications of *N*-glycans on IgG antibodies have been extensively investigated. ENGase provides a powerful tool. *N*-Glycans on antibodies can be trimmed, and other *N*-glycans can be introduced by ENGase. Antibody–drug conjugates (ADCs) have also been prepared using this method in which *N*-glycans were changed into a structure with a tag for subsequent reactions, and small molecular drugs were introduced via bio-orthogonal reactions. This approach allows for the introduction of drugs into the Fc region without affecting antigen recognition. Furthermore, the *N*-glycan structure can be made homogeneous. Wong et al. introduced *N*-glycans with 3-position fluorinated sialic acids into antibodies. Because this fluorinated *N*-glycan was not degraded by sialidase and modification with sialic acid containing *N*-glycan can enhance the metabolic stability of proteins, this antibody is expected to show a significant improvement in pharmacokinetics.

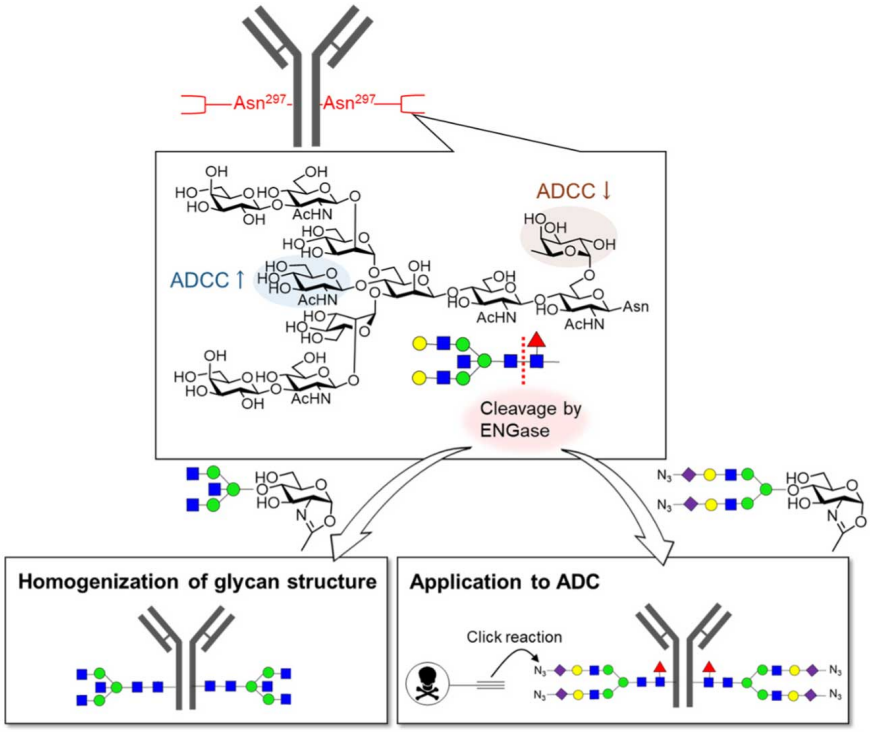


Figure 7.5.8: Glycan editing of immunoglobulin (IgG) antibodies. *N*-Glycans at Asn297 of IgG affect their activity. Core fucose reduces the antibody-dependent cellular cytotoxicity (ADCC) activity, whereas bisecting GlcNAc enhances the ADCC activity. *N*-Glycan editing using Endo-β-*N*-acetylglucosaminidase (ENGase) can give IgG as a homogeneous glycoform or can be applied for the preparation of Antibody–drug conjugates (ADCs).

N-Glycans play important roles not only in antibodies, but also in many other glycoproteins. As described above, the structure–activity relationship study of *N*-glycans on EPO demonstrates the importance of the *N*-glycan structure on the bioavailability and bioactivity of proteins. Hossain and Wade et al. reported that the physical properties of insulin can be improved by adding *N*-glycan to insulin, which originally has no glycans. Introduction of sialic acid containing *N*-glycans to insulin successfully inhibited problematic fibril formation. In addition, *N*-glycan-modified insulin bound to its receptor with almost the same affinity as the natural form, and further improvements in its metabolic stability were observed. Currently, PEGylation has been generally used to enhance the bioavailability of proteins; however, PEG is not without adverse effects. Considering that *N*-glycans are endogenous glycans and are expected to be extremely safe, "*N*-glycan modification" has the potential to become a common strategy for improving the protein/peptide bioactivity.

The structure of *N*-glycans is also important for vaccine development. Viruses use host biosynthetic systems to synthesize proteins. Consequently, viral proteins are subjected to glycan modification. Therefore, glycoproteins and glycopeptides are candidate antigens for vaccine development, and their glycan structures influence their functions. HIV vaccine candidates containing *N*-glycans have been designed and synthesized . Wang et al. reported that the glycan structure on the antigen was critical for the neutralization activity of antibodies, clearly demonstrating the importance of the glycan structure in vaccine design. In addition, Wang showed the importance of glycan structures in the development of influenza HA-based vaccines. For the development of vaccines against COVID-19, the spike protein is a promising antigen candidate. This protein is heavily glycosylated, but *N*-glycan modifications of spike proteins have been reported to reduce their antigenicity. However, *N*-glycan-modified antigens may induce antibodies against endogenous *N*-glycans, which should be carefully examined. Overall, glycans are likely to be important for developing highly efficient and safe vaccines.

7.5.4.2: Drug Delivery Systems (DDSs) Using N-Glycans

N-Glycans interact with various biomolecules, including many lectins, and thus show distinct dynamics in vivo. Therefore, DDSs using *N*-glycans have been investigated. Because glycan–lectin interactions are weak, multivalent materials, including polymers, dendrimers, and liposomes, are usually utilized to enhance their interactions.

We synthesized dendrimers of sialic acid containing *N*-glycans and evaluated their dynamics in vivo using PET imaging. We revealed that the structure of *N*-glycans affected the uptake of dendrimers into specific organs. In addition, Tanaka et al. developed

an *N*-glycan-based DDS using albumin as a multivalent scaffold (**Figure 7.5.9**). The albumins modified with *N*-glycans were used as carriers of metal catalysts to realize chemical reactions at the desired organ in vivo. It should be noted that they achieved metal-catalyzed reactions in vivo by utilizing the hydrophobic pocket of albumin.

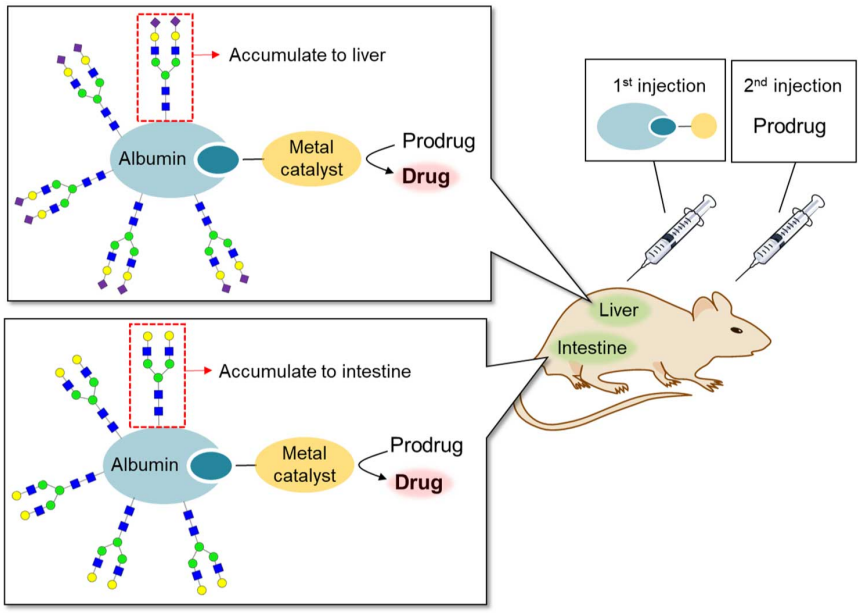


Figure 7.5.9: In vivo reaction using artificial glycosylated albumin metalloenzymes. Specific *N*-glycan conjugated albumin is specifically uptaken into the specific organs. Thus, the albumins conjugated with *N*-glycans were used as carriers of metal catalysts to realize chemical reactions for the activation of prodrug at the desired organ.

Siglecs, which recognize sialic acid, are expressed on immune cells and are involved in immune regulation. Immune cells can be targeted by utilizing sialyl glycan–Siglec interactions. Paulson et al. developed high-affinity Siglec ligands by the derivatization of sialic acid. They synthesized *N*-glycans containing these artificial structures, which exhibited a high affinity for Siglec-2. They achieved B cell targeting using liposomes displaying this *N*-glycan. Utilization of different sialyl glycans enables the targeting of various immune cells. In addition to Siglecs, DDSs targeting galectins, which recognize galactose, have also been investigated.

7.5.5: Future Perspectives

Glycans exist as polysaccharides in nature and are involved in multivalent interactions for pattern recognition. Conformational control via the formation of polysaccharides also plays an important role in glycan functions. In addition, many glycans function only when they are linked to proteins or lipids. Such emergent glycan functions can only be revealed by analysis using the whole glycan structure or glycoconjugates. As described herein, the increased availability of various *N*-glycans has led to the elucidation of the significance of complex *N*-glycan structures. The influence of *N*-glycan modification on some protein functions was also discussed. On the other hand, the molecular basis of glycan functions on membrane proteins remains to be elucidated, although glycans are attached to almost all membrane proteins and have diverse functions.

Recent advances in the engineering of cell-surface glycans are expected to provide a powerful approach to tackle this challenging issue. Bertozzi et al. developed metabolic labeling of cell surface glycan by incorporating unnatural sugar analogs using "**click chemistry,**" having the reaction tag followed by the **bioorthogonal reactions**. We'll discuss these techniques further below.

In addition to glycan function analysis, the therapeutic application of metabolic glycan labeling is being vigorously investigated. Glycan engineering by chemical and chemoenzymatic methods has also been investigated. In addition, de novo glycans on cell surfaces have also been reported, such as the direct introduction of defined glycan structures into plasma membranes by lipid insertion, liposomal fusion, and tag technology. Such glycan editing technique enables glycan functions to be explored on membrane proteins on living cell surfaces.

A major feature of glycans is their heterogeneity. Glycans attached to the same site on the same protein can have diverse structures. In addition, many proteins have multiple glycosylation sites to which various glycans can be added. Although studies using pure *N*-glycans have revealed the functions of individual *N*-glycans, little is known about their function in combination with each other. Kurbangalieva and Tanaka et al. prepared albumins labeled with several *N*-glycans and observed their dynamics in vivo. Interestingly, their dynamics were altered depending on the *N*-glycosylation pattern. These results suggest that the simultaneous

interaction of multiple *N*-glycans may result in the expression of functions different from those of individual *N*-glycans. Little is known about whether the interactions of glycans with multiple lectins work collaboratively or competitively. A bottom-up approach to the construction of controlled glycoforms is expected to be a powerful strategy to address this difficult issue.

Glycans are considered to be the third most important life chain and have attracted increasing attention in recent years. However, unlike nucleic acids and proteins, their functional analysis and regulation have been delayed due to the lack of simple preparation methods. Recent advances in the preparation of *N*-glycans are expected to accelerate functional studies.

7.5.6: Click Chemistry and Bioorthogonal Reactions

Click chemistry is a powerful way to connect two molecules covalently - hence the name click chemistry. Sharpless, Meldal, and Bertozzi were awarded the Nobel Prize in Chemistry in 2022 for its development and application. It has been used to synthesize active site inhibitors for enzymes as well as to label glycan and other biomolecules in vitro and in vivo.

In click chemistry, two molecules are "stitched" together to form a new molecule, like two activated amino acids condense to form a dipeptide. Click chemistry was developed to emulate the simple solutions found in nature to produce polymers. Click chemistry reactions should not be sensitive to water and oxygen, easy to purify products, and proceed with a favorable ΔG (< - 20 kcal/mol, -84 kJ/mol). Azides and acetylenes were the first used, but adding Cu^{1+} as a catalyst made the reaction very fast. The reaction can take place easily in blood and even urine.

Drugs that inhibit enzymes typically bind to the enzyme's active site where catalysis occurs. The binding of an inhibitor precludes binding of the normal reactants (substrates) for the enzyme, inhibiting its activity. Using click chemistry, two small reactive molecules selected to bind independently in the active site can covalently react with each other to form a new drug with very high specificity and very high binding affinity (low K_D). This has been used to synthesize noncovalent inhibitors of the enzyme acetylcholinesterase (Barry Sharpless Lab, Scripps Lab). The reactive groups chosen in the example below are azide and acetylene derivatives, which, when held in close approximation in the binding site of the enzyme, undergo a cycloaddition reaction to form a triazole.

Figure 7.5.10 shows the reaction of an azide and acetylene in solution (without a "binding template"), which leads to equal amounts of the syn and anti products.

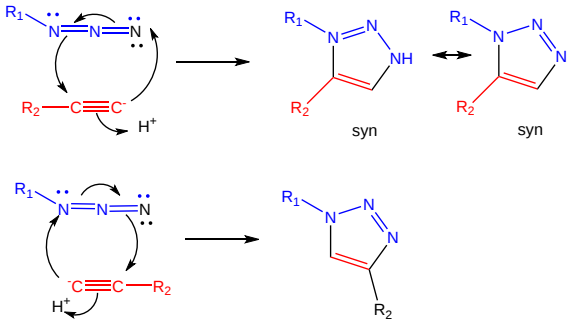


Figure 7.5.10: Click chemistry reaction between azide and acetylide without a directing template

The actual mechanism (not the simplified version shown) requires catalysis by copper ions (Cu^{1+}), which form a complex with the acetylide (deprotonated acetylene). This decreases the pKa of the acetylene functional group, making it a better nucleophile. A dicopper intermediate is suggested in which the azide interacts with the second copper. Subsequent rearrangement led to the triazole products.

The reaction of an azide and acetylide in an extended active site binding site leads to the production of only the syn product as shown for a click inhibitor of the enzyme acetylcholinesterase in Figure 7.5.11. The product would be a potent inhibitor of the enzyme.

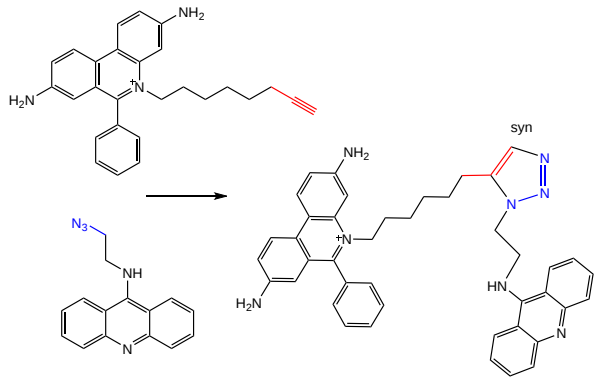
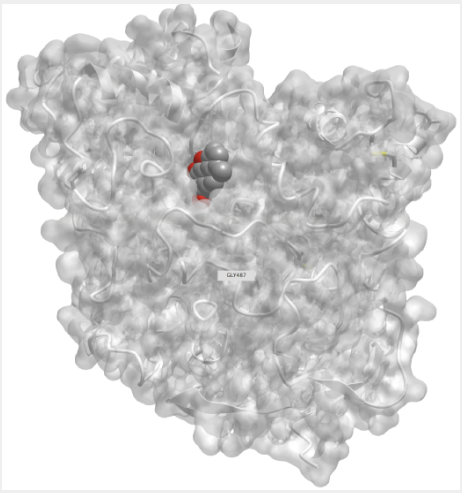
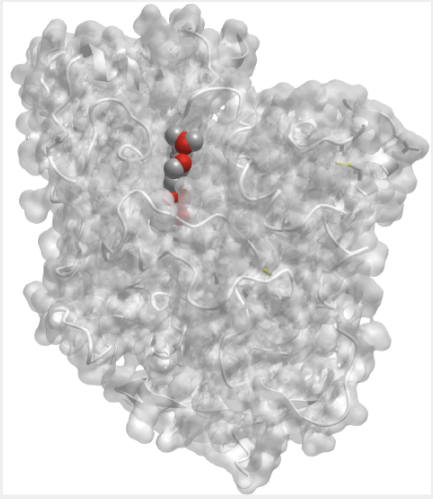


Figure 7.5.11: Click chemistry reaction between azide and acetylide with a directing template

This target-guided synthesis creates a bivalent inhibitor (one that binds at two different regions of an extended binding site). It would have a lower K_D than either of the separate inhibitors.

The enzyme has a catalytic site at the end of a very deep (20 Å) and narrow pocket. It also has a peripheral site near the surface of the extended binding pocket. Hence, it is a great potential target for click-chemistry inhibitors.

Figure 7.5.12 shows interactive iCn3D models of human acetylcholinesterase in complex with peripheral site inhibitor dihydrotanshinone I (4M0E) and the protein in complex with peripheral and active site-spanning inhibitor territrem B (4M0F) to illustrate the structural features of the enzyme that make it ideal for click chemistry inhibitors.

Human acetylcholinesterase in complex with the peripheral site inhibitor dihydrotanshinone I (4M0E)	Human acetylcholinesterase in complex with the peripheral and active site-spanning inhibitor territrem B (4M0F)
	
(Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...y8btPa8WMYNj4A	(Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...veBZM1DLSxzKX6

Click chemistry can be used in vivo in **bioorthogonal reactions**. These are specific reactions that take place in potentially reactive biological environments but without interfering with normal biological functions and activities. The reaction is different than when it occurs in a test tube in that a "bioorthogonal reporter such as a fluorophore" (that can be used, for instance, to track a target molecule in vivo when linked to it) covalently links to a target biomolecule (such as a surface glycan) without influencing its activity. Instead of using Cu ions as catalysts as in *in vitro* click chemistry, normal catalytic mechanisms of the cell are used.

Some more classical condensation reactions can be considered early examples of bioorthogonal reactions. Azides are reactive and are rare biologically, but biomolecules (lipids, proteins, nucleic acids, etc) can be easily labeled with azides or alkynes, facilitating their reaction with click chemistry. Copper ions, however, can generate ROS, which limits its potential for in vivo click reactions. Here are some classical and more modern bioorthogonal condensing reactions.

- carbonyls with hydrazines and alkoxylamines to form oximes and hydrazones.
- carbonyls and amines to form Schiff bases
- triarylphosphines and organic azides (Staudinger ligation)
- copper(I)-catalyzed azide–alkyne cycloaddition (CuAAC) described above
- strained cyclooctynes and azides in strain-promoted azide–alkyne cycloadditions (SPAAC) which occur fast enough and don't require a copper ion catalyst
- 1,2,4,5-tetrazine and an alkene or alkyne dienophile ([4 + 2]/retro [4 + 2]) cycloaddition to form a dihydropyridazine or pyridazine conjugate in an inverse electron-demand Diels–Alder reaction (IEDDA)

Figure 7.5.13illustrates newer generation click reactions for connecting two molecules

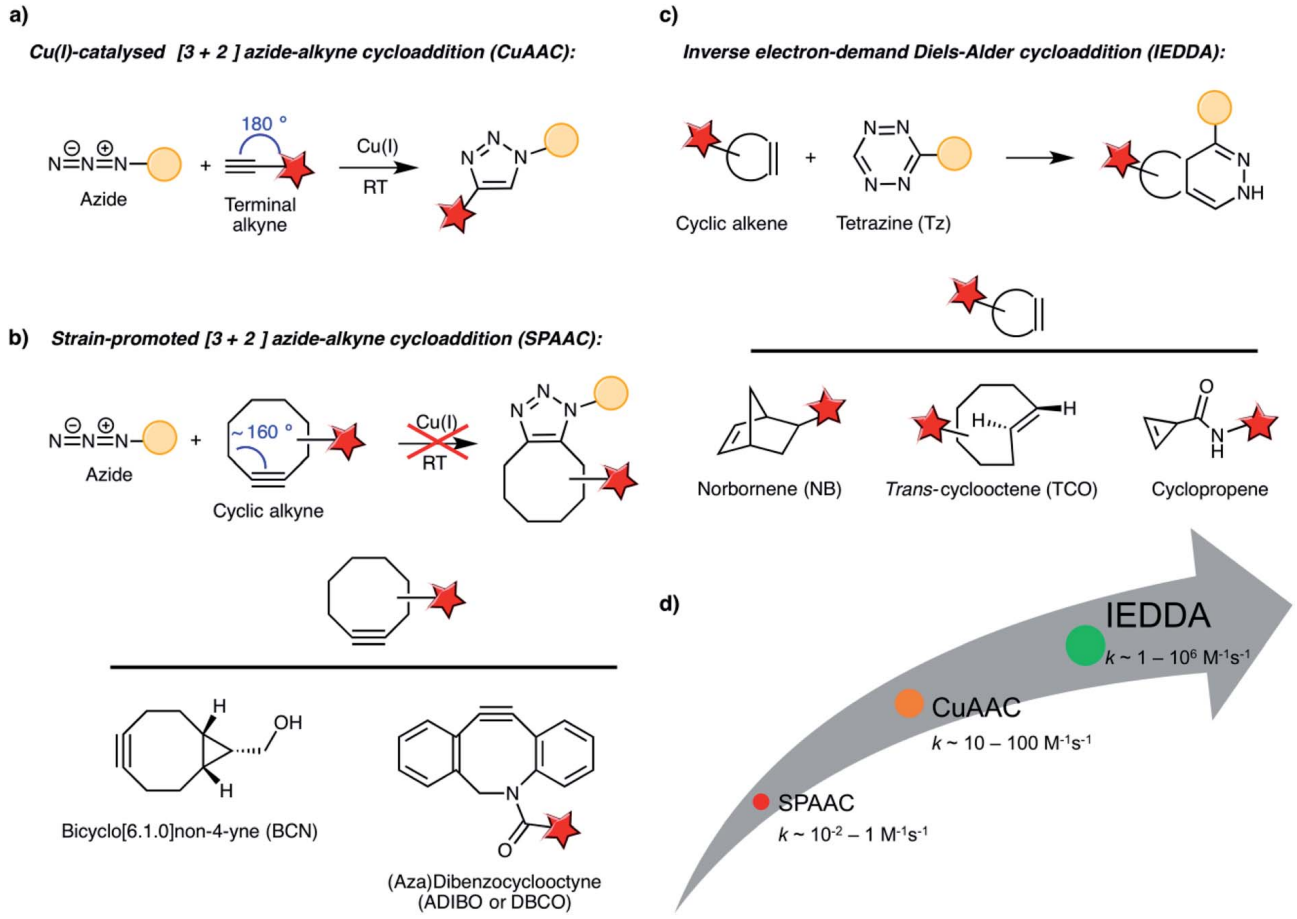


Figure 7.5.13: Newer generation click/bioorthogonal reactions Idiago-L'opez et al., Nanoscale Adv., 2021, 3, 1261. DOI: 10.1039/d0na00873g. Creative Commons Attribution-NonCommercial 3.0 Unported

Panels (a–c) show the main click chemistry reaction used in biochemical labeling reactions. Panel (d) compares reaction kinetics of CuAAC, SPAAC and IEDDA.Figure

Figure 7.5.14shows some applications of bioorthogonal reactions

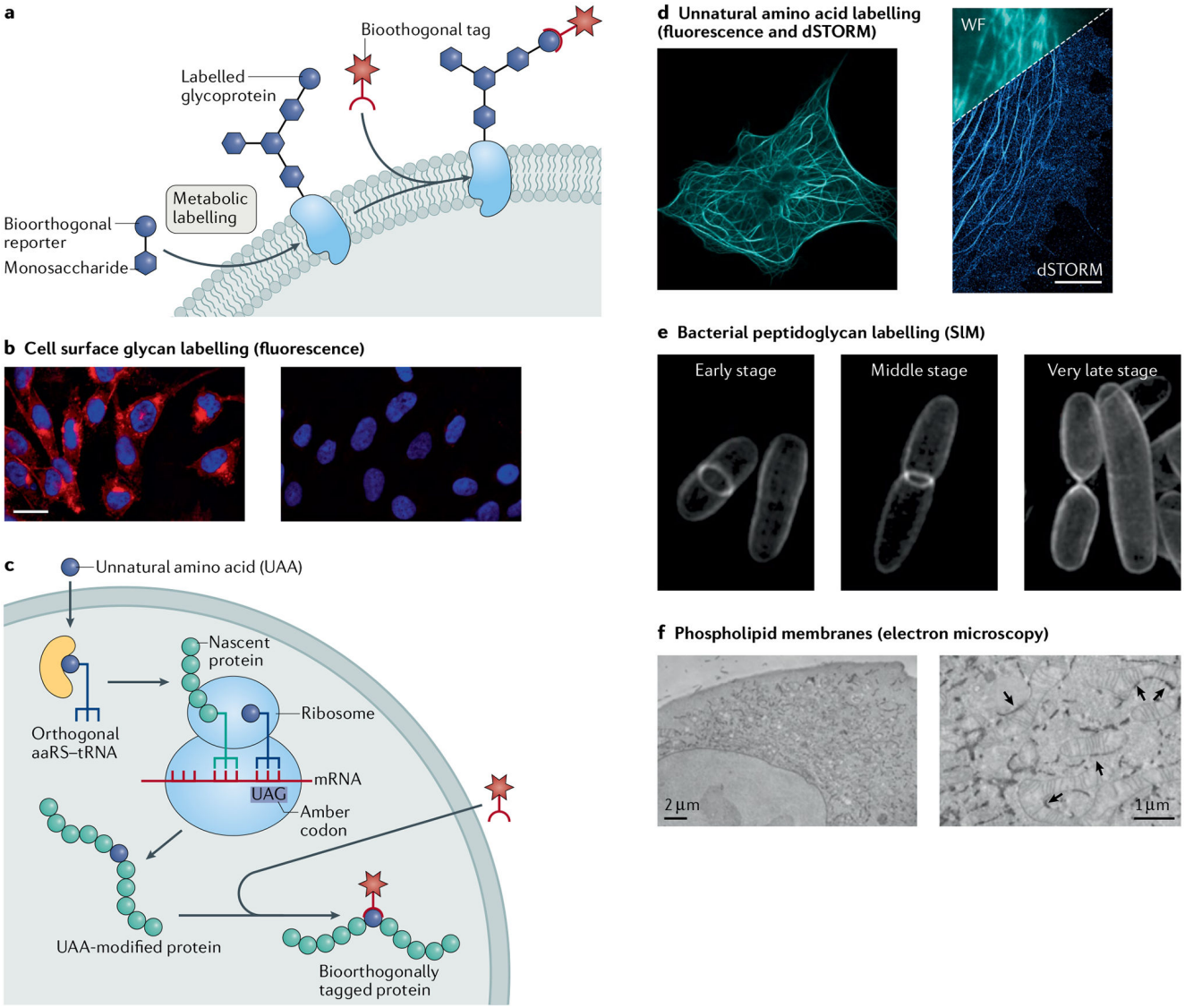


Figure 7.5.14: Applications for labeling different molecule types in cells. Scinto et al. Nat Rev Methods Primers. 2021 ; 1: . doi:10.1038/s43586-021-00028-z. Creative Commons (https://creativecommons.org/licenses/by/4.0/).

Panel A shows a model for metabolic engineering for cell labeling and imaging.

Panel B shows fluorescence microscopy of CHO cells incubated in the presence (left) or absence (right) of peracetylated N-azidoacetylmannosamine (Ac4ManNAz) and labeled with a fluorophore by the Staudinger ligation.

Panel C shows a model for genetic code expansion as a strategy for cell labeling and imaging.

Panel D shows fluorescence and direct stochastic optical reconstruction microscopy (dSTORM) super-resolution images of COS-7 cells where microtubule- microtubule-associated protein was encoded with an unnatural trans-cyclooctene (TCO) amino acid and tetrazine ligation was used to attach a microscopy dye.

Panel E shows structured illumination microscopy (SIM) images of Escherichia coli, where N-azidoacetyl-muramic acid (NAM) was metabolically incorporated into the bacterial peptidoglycan and fluorophore-labeled by copper(I)-catalyzed azide–alkyne cycloaddition (CuAAC).

Panel F shows electron microscopy images of HeLa cells, where azido-choline was metabolically incorporated, and cyclooctyne/azide click chemistry was used to conjugate electron microscopy imaging agents. The arrows indicate sites of endoplasmic reticulum–mitochondria contacts. aaRS, aminoacyl-
aminoacyl-tRNA synthetase.

Figure 7.5.15 shows the application of click chemistry to label surface glycoproteins called integrins, which we will explore in greater detail in Chapter 12.10.

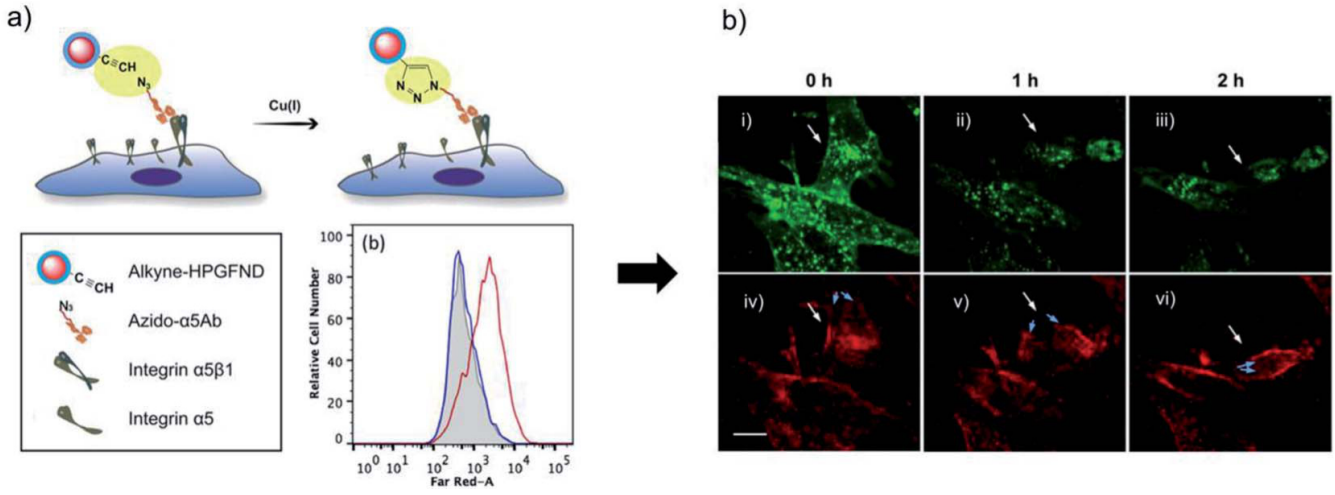


Figure 7.5.15: Bioorthogonal labeling of integrin a5 membrane proteins using azide-modified antibodies and alkyne-HPGFNDs: (a) flow cytometry analysis of fluorescence signals from HFW cells preincubated with Azido-5aAb (red) and control cells (blue). (b) HFW cells labeled with Alexa
Fluor 488-conjugated wheat germ agglutinin (i–iii) and 100 nm alkyne-HPGFNDs (iv–vi). White arrows indicate the cell migration route and blue
arrows show the migration of integrin a5 on cells filopodia. Scale bars: 20 mm. Idiago-L'opez et al., ibid

7.5.7: Summary

This chapter provides a comprehensive overview of N-glycans, emphasizing their structural diversity, biosynthetic pathways, methods for chemical and enzymatic synthesis, and roles in biological processes and therapeutic applications. It is designed to equip junior and senior biochemistry majors with a deep understanding of how glycosylation—particularly N-linked glycosylation—affects protein function and cellular communication.

---

1. **Introduction to N-Glycosylation**

- **Prevalence and Types:**
  Over 60% of proteins are glycosylated, with N-glycans being the most common modification. They are attached as a preformed high-mannose structure ($Glc_3Man_9GlcNAc_2$) in the endoplasmic reticulum (ER) and then processed in the Golgi apparatus into high-mannose, hybrid, or complex forms.
- **Structural Diversity and Function:**
  Complex N-glycans feature additional modifications—such as sialic acid, polylactosamine, core fucose, and bisecting glucosamine—that regulate protein folding, immune recognition, and cell signaling. Despite extensive studies, the precise molecular basis of their activity remains under investigation.
- **Synthesis Strategies:**
  The chapter reviews chemical synthesis (including de novo assembly and convergent strategies), enzymatic synthesis, and isolation from natural sources. These approaches have expanded the toolkit for preparing pure, structurally defined N-glycans, enabling detailed functional studies.

---

2. **Molecular Recognition: Glycan–Lectin Interactions**

- **Analytical Techniques:**
  Two main approaches are highlighted:
  - **Glycan Arrays:**
    These high-throughput platforms allow simultaneous analysis of many glycan structures to determine binding specificities

and structure–activity relationships with lectins. They reveal how branching patterns and multivalent interactions enhance affinity.

- **NMR Spectroscopy:**
  Techniques like Saturation Transfer Difference (STD-NMR) offer atomic-level insights into how glycans interact with lectins, map binding epitopes, and assess conformational changes. Advanced methods (e.g., PCS, transferred NOE) further detail how glycan flexibility influences recognition.

- **Key Findings:**
  Detailed studies have shown that lectins often recognize not only minimal epitopes (like disaccharides) but also the entire glycan structure. The relative binding strengths of different glycan branches and the effects of remote modifications (e.g., core fucose, chain elongation) have been elucidated.

---

3. **Functional Implications of N-Glycans on Glycoproteins**

- **Protein Quality Control:**
  In the ER, high-mannose N-glycans serve as tags for proper protein folding. Enzymatic trimming and reglucosylation (mediated by UGGT and glycosidases) are central to the ER quality control system. Defects in these processes can lead to congenital disorders of glycosylation (CDGs).

- **Influence on Protein Bioactivity:**
  Studies using synthetic glycoproteins (e.g., erythropoietin, interferon-β, and chemokines) have demonstrated that glycosylation can impact protein stability, receptor binding, and in vivo half-life. For instance, increasing sialylation generally improves metabolic stability and bioactivity, while glycan modifications can either enhance or reduce receptor interactions.

---

4. **Applications in Drug Development and Therapeutics**

- **Antibody Engineering:**
  Modifications of IgG N-glycans (e.g., removal of core fucose, addition of bisecting GlcNAc) significantly alter immune effector functions such as antibody-dependent cellular cytotoxicity (ADCC). Techniques like glycan editing using ENGase are used to generate homogeneous glycoforms and facilitate the creation of antibody–drug conjugates (ADCs).

- **Drug Delivery Systems (DDS):**
  N-glycans can be exploited for targeted delivery. Multivalent platforms (dendrimers, glycosylated albumin) have been developed to harness glycan–lectin interactions for organ-specific drug uptake and controlled in vivo reactions.

- **Vaccine Design:**
  The glycosylation of viral proteins influences antigenicity and neutralizing antibody responses. The chapter discusses how synthetic glycan modifications on vaccine candidates (e.g., for influenza, HIV, and COVID-19) are critical for optimizing immunogenicity while minimizing potential autoimmunity.

---

5. **Click Chemistry and Bioorthogonal Reactions in Glycobiology**

- **Fundamentals of Click Chemistry:**
  Click chemistry involves rapid, high-yield, and selective covalent reactions (often using azide–alkyne cycloadditions) that mimic natural condensation reactions. These reactions have been pivotal for synthesizing enzyme inhibitors and labeling biomolecules.

- **Bioorthogonal Strategies:**
  In living systems, bioorthogonal reactions (e.g., SPAAC, IEDDA) allow for the labeling and tracking of glycans without interfering with native biological processes. These techniques have broad applications in imaging, diagnostics, and therapeutic targeting.

- **Case Studies:**
  Examples include the development of click-based inhibitors for acetylcholinesterase and the labeling of integrin proteins on cell membranes. These strategies enhance our ability to probe glycan functions in complex biological environments.

Conclusion and Future Perspectives

The chapter concludes by emphasizing the rapid advancements in N-glycan synthesis, analytical techniques, and functional studies. It suggests that further progress in cell-surface glycan engineering and the construction of controlled glycoforms will deepen our understanding of glycan-mediated biological processes. With glycans now recognized as a “third major biopolymer,” their role in disease and therapeutic development continues to expand, promising innovative strategies in drug design and targeted therapies.

This summary encapsulates the multifaceted nature of N-glycans—from their biosynthesis and structural complexity to their emerging roles in medicine and technology—providing a detailed yet accessible overview for advanced undergraduate students in biochemistry.

7.6: Chapter 7 Problems - Answer Key

7.6.1: Chapter 7: Carbohydrates and Glycobiology - Answer Key for Problems

Note: Tables, graphs, diagrams, and figures can be copied from the research papers, the papers must have Creative Commons licenses. Science Advances, Nature Communications, JBC and PLOS journals usually have to right kind of CC licenses.

1. jdkfjdkf

2. djkfjdkjf

3. kjfkdjfkjdk

7.7: Chapter 7 Problems

Chapter 7: Carbohydrates and Glycobiology - Problems

Note: Tables, graphs, diagrams, and figures can be copied from the research papers, the papers must have Creative Commons licenses. Science Advances, Nature Communications, JBC and PLOS journals usually have to right kind of CC licenses.

1. jdkfjdkf

2. djkfjdkjf

3. kjfkdjfkjdk

CHAPTER OVERVIEW

8: Nucleotides and Nucleic Acids

Return to Fundamentals of Biochemistry Search Fundamentals of Biochemistry

---

8.1: Nucleic Acids - Structure and Function

Learning Goals (ChatGPT o1, 1/30/25)

1. **Differentiate Between DNA and RNA Structures:**
   - Explain the basic composition of nucleic acids (sugar–phosphate backbone and nitrogenous bases).
   - Describe how the sugar (deoxyribose in DNA vs. ribose in RNA) and the specific bases (T in DNA vs. U in RNA) influence overall structure and function.
2. **Describe Canonical and Noncanonical Base Pairing:**
   - Illustrate the principles of Watson–Crick base pairing and how hydrogen bonding between complementary bases stabilizes the double helix.
   - Define and compare alternative (noncanonical) base pairing modes, including reverse Watson–Crick, wobble, Hoogsteen, and reverse Hoogsteen base pairs, and explain how these variations affect nucleic acid function.
3. **Understand DNA Conformations and Structural Dynamics:**
   - Compare the three main double-helical forms of DNA (A-DNA, B-DNA, and Z-DNA), focusing on differences in handedness, base pair spacing, twist angles, and groove dimensions.
   - Analyze how environmental factors (e.g., hydration, salt concentration) and molecular interactions induce transitions between these forms.
4. **Explore Higher-Order Nucleic Acid Structures:**
   - Describe the formation and biological significance of G-quadruplexes, including how guanine tetrads and metal ions (e.g., K^+) stabilize these structures.
   - Explain the mechanisms of triple helix formation, both intermolecular and intramolecular (H-DNA), and discuss their potential roles in gene regulation.
   - Recognize the characteristics of four-way junctions and their role in nucleic acid folding and interactions with small molecules.
5. **Analyze the Role of Base Stacking in Nucleic Acid Stability:**
   - Explain how base stacking and hydrophobic interactions contribute to the stability of the DNA double helix, in addition to hydrogen bonding.
   - Evaluate the effects of nucleotide sequence (e.g., GC vs. AT content) on the melting temperature (Tm) and overall stability of nucleic acids.
6. **Interpret Structural Models and Parameters:**
   - Utilize interactive molecular visualization tools (e.g., iCn3D) to explore nucleic acid structures.
   - Identify key structural parameters such as propeller and twist angles and understand how they influence nucleic acid conformation and dynamics.
7. **Relate Nucleic Acid Structure to Biological Function:**
   - Connect the structural features of DNA (e.g., major and minor grooves) with the mechanisms of protein recognition and binding.
   - Discuss how conformational flexibility (e.g., local “breathing” and dynamic unfolding) enables critical cellular processes like replication, transcription, and repair.
8. **Evaluate Experimental Approaches in Nucleic Acid Research:**
   - Summarize the historical and modern techniques used to elucidate nucleic acid structure, including X-ray crystallography, NMR, and cryo-electron microscopy.
   - Analyze how advances in structural biology have deepened our understanding of nucleic acid dynamics and their roles in cellular function.

These goals will help you integrate knowledge of nucleic acid chemistry with its biological implications, preparing you for advanced study and research in biochemistry.

8.1.1: Introduction to Nucleic Acids

Alongside proteins, lipids, and complex carbohydrates (polysaccharides), *nucleic acids* are one of the four major types of macromolecules that are essential for all known forms of life. The nucleic acids consist of two major macromolecules, *Deoxyribonucleic acid (DNA)* and *ribonucleic acid (RNA)* that carry the genetic instructions for the development, functioning, growth, and reproduction of all known organisms and viruses. Both consist of polymers of a sugar-phosphate-sugar backbone with organic heterocyclic bases attached to the sugars. The sugar in DNA is deoxyribose while in RNA it is ribose. DNA contains four bases, cytosine and thymine (pyrimidine bases) and guanine and adenine (purine bases). DNA in vivo consists of two antiparallel strands intertwined to form the iconic DNA double-stranded helix. RNA is single-stranded but may adopt many secondary and tertiary conformations not unlike that of a protein. Figure 8.1.1 shows a low-resolution comparison of the structure of DNA and RNA.

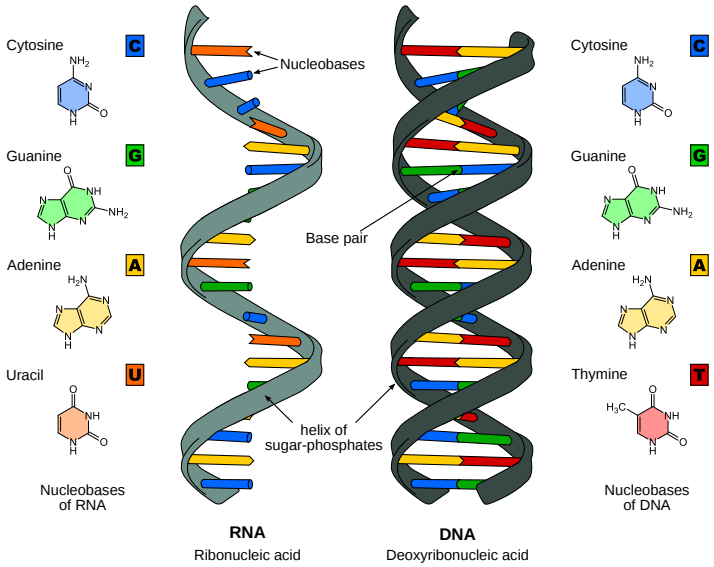


Figure 8.1.1: Low-resolution comparison of the structure of DNA and RNA. https://commons.wikimedia.org/wiki/F...DNA_RNA-EN.svg. Creative Commons Attribution-Share Alike 3.0 Unported license.

The biological function of DNA is quite simple: to carry and protect the genetic code. Its structure serves that purpose well. In the next section, we will study the functions of RNA, which are much more numerous and complicated. The structure of RNA has evolved to serve those added functions.

The core structure of a nucleic acid monomer is the *nucleoside*, which consists of a sugar residue + a nitrogenous base that is attached to the sugar residue at the 1′ position as shown in Figure 8.1.2. The sugar utilized for RNA monomers is ribose, whereas DNA monomers utilize deoxyribose that has lost the hydroxyl functional group at the 2′ position of ribose. DNA contains four nitrogenous bases, including the *Purines*, Adenine (A), and Guanine (G), and the *Pyrimidines*, Cytosine (C), and Thymine (T). RNA uses the same nitrogenous bases as DNA, except for Thymine. Thymine is replaced with Uracil (U) in the RNA structure.

When one or more phosphate groups are attached to a nucleoside at the 5′ position of the sugar residue, it is called a *nucleotide*. *Nucleotides* come in three flavors depending on how many phosphates are included: the incorporation of one phosphate forms a *nucleoside monophosphate*, the incorporation of two phosphates forms a *nucleoside diphosphate*, and the incorporation of three phosphates forms a nucleoside triphosphate as shown in Figure 8.1.2.

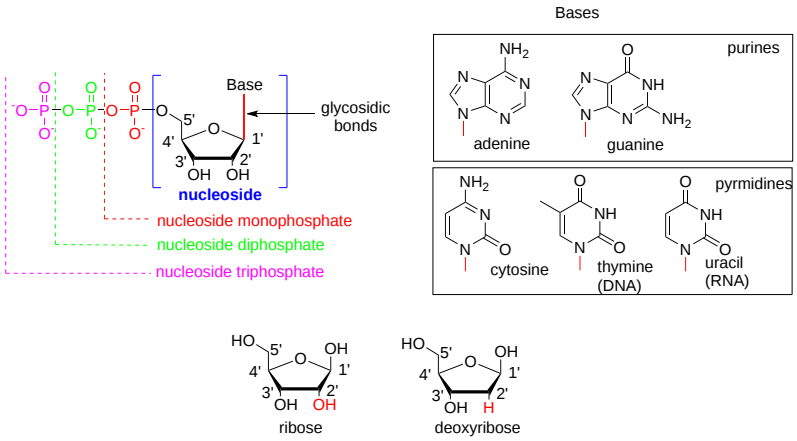


Figure 8.1.2: The Monomer Building Blocks of Nucleic Acids. The site of the nitrogenous base attached to the sugar residue (glycosidic bond) is shown in red.

8.1.2: DNA and RNA Hydrogen-bonded structures

Figure 8.1.3 below shows a "flattened" structure of double-stranded B-DNA that best shows the backbone and hydrogen-bonded base pairs between two antiparallel strands of the DNA. Unlike the protein α-helix, where the R-groups of the amino acids are positioned to the outside of the helix, in the DNA double-stranded helix, the nitrogenous bases are positioned inward and face each other. The backbone of the DNA is made up of repeating sugar-phosphate-sugar-phosphate residues. Bases fit in the double-helical model if pyrimidine on one strand is always paired with purine on the other. From Chargaff's rules, the two strands will pair A with T and G with C. This pairs a keto base with an amino base and a purine with a pyrimidine. Two H-bonds can form between A and T, and three can form between G and C. This third H-bond in the G:C base pair is between the additional exocyclic amino group on G and the C2 keto group on C. The pyrimidine C2 keto group is not involved in hydrogen bonding in the A:T base pair.

Furthermore, the orientation of the sugar molecule within the strand determines the directionality of the strands. The phosphate group that makes up part of the nucleotide monomer is always attached to the 5′ position of the deoxyribose sugar residue. The free end that can accept a new incoming nucleotide is the 3′ hydroxyl position of the deoxyribose sugar. Thus, DNA is directional and is always synthesized in the 5′ to 3′ direction. Interestingly, the two strands of the DNA double helix lie in opposite directions or have a head-to-tail orientation.

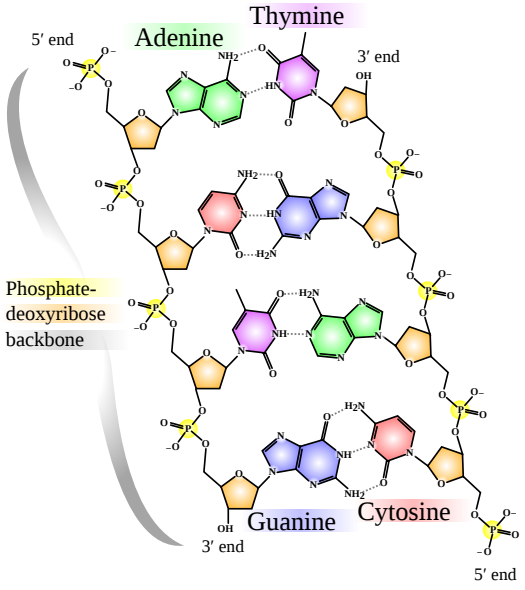


Figure 8.1.3: "Flattened" Structure of DNA Madeleine Price Ball. https://en.Wikipedia.org/wiki/File:D..._structure.svg. Wikimedia Commons

By analogy to proteins, DNA and RNA can be loosely thought to have primary and secondary structures. For a single strand, the primary sequence is just the base sequence read from the 5' to 3' end of the strand, with the bases thought of as "side chains" as illustrated in Figure 8.1.4 for an RNA strand which contains U instead of T.

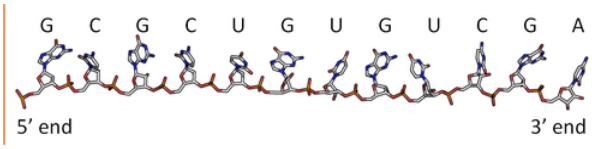


Figure 8.1.4: "Primary" sequence of a single RNA strand. https://en.Wikipedia.org/wiki/Nucleic_acid_sequence. Creative Commons Attribution-ShareAlike License

Since it is found partnered with another molecule (strand) of DNA, the double-stranded DNA, which consists of two molecules held together by hydrogen bonds, might be considered to have secondary structure (analogous to alpha and beta structure in proteins). Of course, the hydrogen bonds are not between backbone atoms but between side chain bases in double-stranded DNA.

Figure 8.1.5 shows an interactive iCn3D model of the iconic structure of a short oligomer of double-stranded DNA (1BNA).

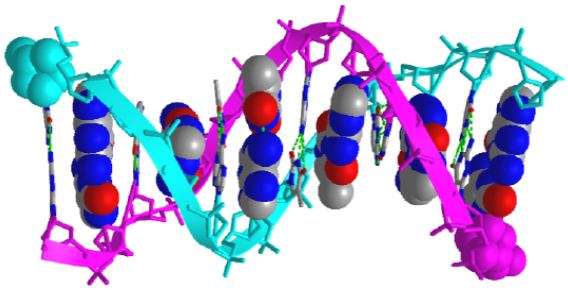


Figure 8.1.5: Iconic structure of a short oligomer of double-stranded DNA (1BNA). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...b5HUbmuQrCobg8

The backbones of the antiparallel strands are magenta (chain A) and cyan (chain B). Each chain's 5' sugar-phosphate end is shown in spacefill and colored magenta (chain A) and cyan (chain B). The hydrogen-bonded interstrand base pairs are shown alternatively in spacefill and sticks to illustrate how the bases stack over each other.

Figure 8.1.6 shows types of "secondary (flat representations) and their 3D or tertiary representations found in nucleic acids.

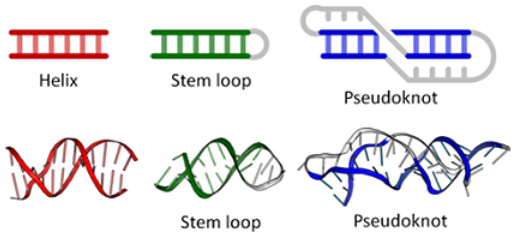


Figure 8.1.6: https://en.Wikipedia.org/wiki/Nucleic_acid_sequence. Creative Commons Attribution-ShareAlike License

Figure 8.1.7 shows an interactive iCn3D model of the tertiary structure of the T4 hairpin loop on a Z-DNA stem (1D16).

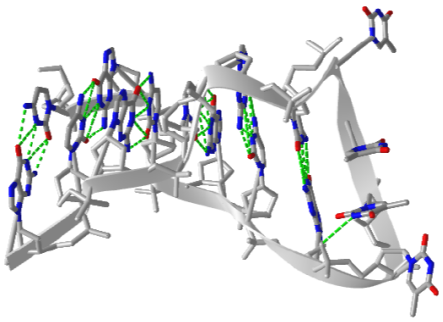


Figure 8.1.7: T4 hairpin loop on a Z-DNA stem (1D16). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...8C7qBqgh8ZTJH9

The hairpin shown is from a synthetic DNA oligomer C-G-C-G-C-G-T-T-T-T-C-G-C-G-C-G, which adopts an alternative Z-DNA conformation (which we will explore below) with a loop at one end. The thymine bases 7, 8, and 9 are generally perpendicular to one another and stack together, along with the ribose of T7.

Figure 8.1.8 shows an interactive iCn3D model of pseudoknot in RNA (437D).

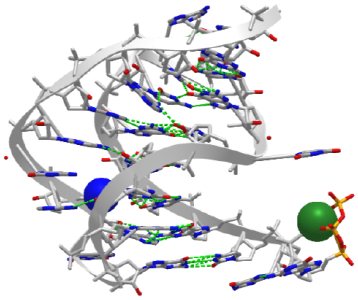


Figure 8.1.8: Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot (437D). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...ZtdeJqQXvjCKfA

The pseudoknot has two stems that form a "helix" and two loops. The knot consists of a hairpin in the nucleic acid structure with the loop between the helices paired to another part of the nucleic acid. Pseudoknots can be found in mRNA and ribosomal RNA and affect the translation of the RNA (decoding to instruct the synthesis of a protein sequence). RNA viruses have pseudoknots which likewise affect protein synthesis and RNA replication. Pseudoknots also occur in DNA.

8.1.3: Synthesis and structure of DNA

The nucleotide that is required as the monomer for the synthesis of both DNA and RNA is *nucleoside triphosphate*. During the incorporation of the nucleotide into the polymeric structure, two phosphate groups (P_i-P_i , called *pyrophosphate*) from each triphosphate are cleaved from the incoming nucleotide and further hydrolyzed during the reaction, leaving a *nucleoside monophosphate* that is incorporated into the growing RNA or DNA chain as shown in **Figure 8.1.9** below. The nucleophilic attack of the 3′-OH of the growing DNA polymer mediates the attack on the incoming nucleoside triphosphate. Thus, DNA synthesis is directional, only occurring at the 3′-end of the molecule.

The further hydrolysis of the pyrophosphate (P_i-P_i) releases a large amount of energy, ensuring that the overall reaction has a negative ΔG. Hydrolysis of P_i-P_i $\leftrightarrow$ $2P_i$ has a ΔG = -7 kcal/mol (-29 kJ/mol) and is essential to provide the overall negative ΔG (-6.5 kcal/mol, -27 kJ/mol) of the DNA synthesis reaction. Hydrolysis of the pyrophosphate also ensures that the reverse reaction, pyrophosphorolysis, will not occur, removing the newly incorporated nucleotide from the growing DNA chain.

This reaction is mediated in DNA by a family of enzymes known as DNA polymerases. Similarly, RNA polymerases are required for RNA synthesis. A more detailed description of polymerase reaction mechanisms will be covered in Chapter 24, which deals with DNA replication and repair, and DNA transcription in Chapter 25.

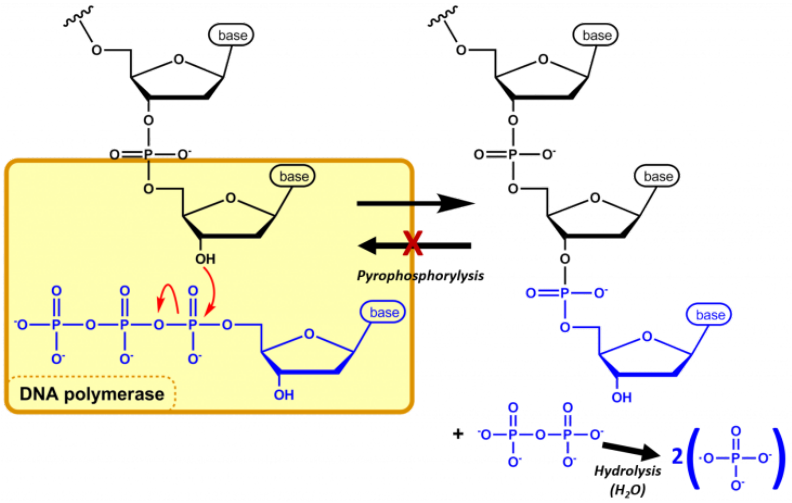


Figure 8.1.9: Nucleic Acid Synthesis: In nucleic acid synthesis, the 3' OH of a growing chain of nucleotides attacks the α-phosphate on the next NTP to be incorporated (blue), resulting in a phosphodiester linkage and the release of pyrophosphate (PPi). The DNA polymerase further mediates the hydrolysis of the pyrophosphate, preventing the reverse reaction from occurring and releasing enough energy to drive the reaction forward. The synthesis of DNA is shown in this diagram. *Image modified from Michal Sobkowski*

DNA was first isolated by Friedrich Miescher in 1869. The double-helix model of DNA structure was first published in the journal *Nature* by James Watson and Francis Crick in 1953 based upon the crucial X-ray diffraction image of DNA from Rosalind Franklin in 1952, followed by her more clarified DNA image with Raymond Gosling, Maurice Wilkins, Alexander Stokes, and Herbert Wilson, and base-pairing chemical and biochemical information by Erwin Chargaff. The prior model was triple-stranded DNA.

The realization that the structure of DNA is that of a double-helix elucidated the mechanism of base pairing by which genetic information is stored and copied in living organisms and is widely considered one of the most important scientific discoveries of the 20th century. Crick, Wilkins, and Watson each received one-third of the 1962 Nobel Prize in Physiology or Medicine for their contributions to the discovery. (Franklin, whose breakthrough X-ray diffraction data was used to formulate the DNA structure, died in 1958 and thus was ineligible to be nominated for a Nobel Prize.)

Watson and Crick proposed two strands of DNA – each in a right-hand helix – wound around the same axis. The two strands are held together by H-bonding between the complementary base pairs (A pairs with T and G pairs with C) as shown in **Figure 8.1.10** below. Note that when looking from the top view, down on a DNA base pair, the position where the base pairs attach to the DNA backbone is not equidistant, but that attachment favors one side over the other. This creates unequal gaps or spaces in the DNA known as the *major groove* for the larger gap and the *minor groove* for the smaller gap (Figure 4.5). Based on the DNA sequence within the region, the hydrogen-bond potential created by the nitrogen and oxygen atoms present in the nitrogenous base pairs causes unique recognition features within the major and minor grooves, allowing for specific protein recognition sites to be created.

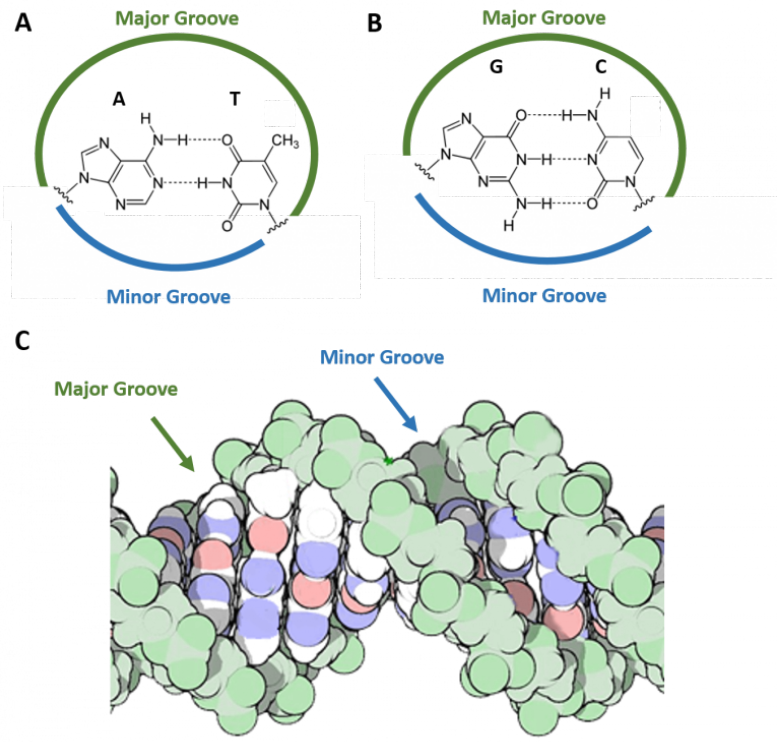


Figure 8.1.10: The Major and Minor Grooves of DNA. Top view of an (A) A-T base pair and a (B) G-C base pair showing the formation of the major and minor groove sides of the DNA. (C) Side view of the DNA double helix with the major and minor grooves indicated. The DNA backbone is green, potential nitrogen hydrogen-bonding locations are indicated in blue, and oxygen hydrogen-bonding locations are red. Figure C modified from *dullhunk*

Figure 8.1.1 shows a schematic representation of available hydrogen bond donors and acceptors in the major and minor grove for TA and CG base pairs.

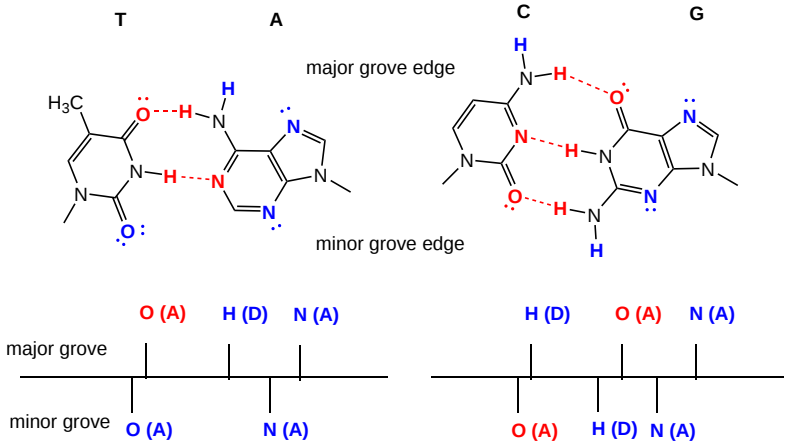


Figure 8.1.11: Available hydrogen bond donors and acceptors in the major and minor groove for TA and CG base pairs

Figure 8.1.12 shows an interactive iCn3D model of DNA showing the major and minor grooves.

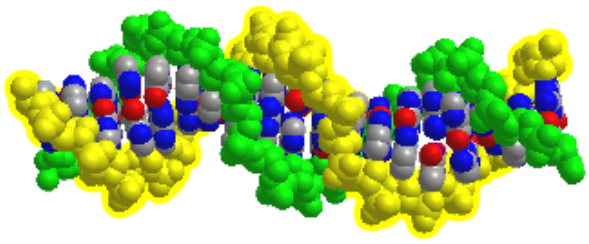


Figure 8.1.12: Major and minor groves of ds-DNA (1D66). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/icn3d/share.html?3xicqSv9ERCBPHvd6

The two sugar-phosphate backbones are shown in green and yellow. Some of the red (oxygen) and blue (nitrogen) atoms in the major grove (and to a much lesser extent in the minor groove) are not involved in inter-strand G-C and A-T base pairing and so would be available to hydrogen bond donors with specific binding proteins that would display complementary shape and hydrogen bonds acceptors and donors. (The white spheres are Cd ions.)

Figure 8.1.13 shows an interactive iCn3D model of the N-terminal fragment of the yeast transcriptional activator GAL4 bound to DNA (1D66).

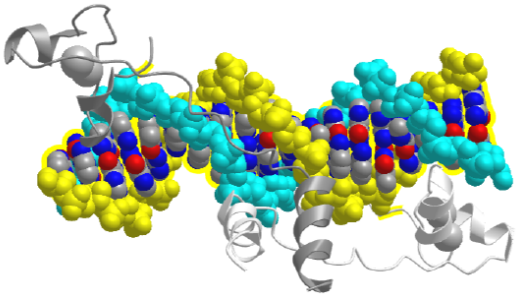


Figure 8.1.13: N-terminal fragment of the yeast transcriptional activator GAL4 bound to DNA (1D66). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...5kLYSSfG7rsmS9

The N-terminal fragment binds to conserved CCG triplets found at both ends of the DNA in the major grove. The protein shown is a dimer held together by a short coiled-coil interaction domain, so the site has 2-fold symmetry. A small Zn^{2+}-containing secondary structure motif in each member of the dimer interacts with the major grove. An extended chain connects the DNA binding and interaction domains of each protein.

In addition to the major and minor grooves providing variation within the double helix structure, the axis alignment of the helix, along with other influencing factors such as the degree of solvation can give rise to three forms of the double helix, the *A-form (A-DNA)*, the *B-form (B-DNA)*, and the *Z-form (Z-DNA)* as shown in Figure 8.1.14

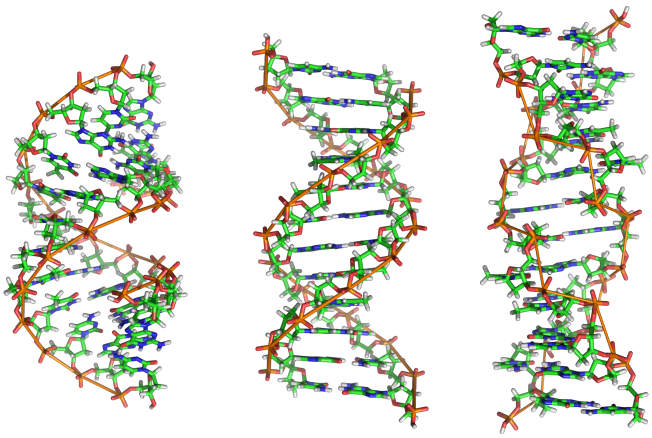


Figure 8.1.14: Structures of A-DNA (left), B-DNA (middle), and Z-DNA (right). https://en.Wikipedia.org/wiki/File:A..._and_Z-DNA.png. Creative Commons Attribution-Share Alike 3.0 Unported

Both the A- and B-forms of the double helix are right-handed spirals, with the B-form being the predominant form found *in vivo*. The A-form helix arises when conditions of dehydration below 75% of normal occur and has mainly been observed *in vitro* during X-ray crystallography experiments when the DNA helix has become desiccated. However, the A-form of the double helix can occur *in vivo* when RNA adopts a double-stranded conformation or when RNA-DNA complexes form. The 2′-OH group of the ribose sugar backbone in the RNA molecule prevents the RNA-DNA hybrid from adopting the B-conformation due to steric hindrance.

The third double helix type formed is a left-handed helical structure known as the Z-form or Z-DNA. Within this structural motif, the phosphates within the backbone appear to zigzag, providing the name Z-DNA. *In vitro*, the Z-form of DNA is adopted in short sequences that alternate pyrimidine and purines when high salinity is present. However, the Z-form has been identified *in vivo* within short regions of the DNA, showing that DNA is quite flexible and can adopt a variety of conformations. A comparison of features between A-, B-, and Z-form DNA is shown in Table 4.1.

Table 4.1 Comparisons of B-DNA, A-DNA and Z-DNA

	B-DNA	A-DNA	Z-DNA
helix sense	Right Handed	Right Handed	Left Handed
base pairs per turn	10	11	12
vertical rise per bp	3.4 Å	2.56 Å	19 Å
rotation per bp	+36°	+33°	-30°
helical diameter	19 Å	19 Å	19 Å

The double-stranded helix of DNA is not always stable. This is because the stair step links between the strands are noncovalent, reversible interactions. Depending on the DNA sequence, denaturation (melting) can be local or widespread, enabling various crucial cellular processes, including DNA replication, transcription, and repair.

Both sequence specificity and interaction (whether covalent or not) with a small compound or a protein can induce tilt, roll, and twist effects that rotate the base pairs in the x, y, or z axis, respectively, as seen in **Figure 8.1.15**, and can therefore change the helix's overall organization. Furthermore, slide or flip effects can also modify the geometrical orientation of the helix. Hence, the flip effects and (to a lesser extent) the other movements described above modulate the double-strand stability within the helix or at its ends. Indeed, under physiological conditions, local DNA 'breathing' has been evidenced at both ends of the DNA helix, and B-DNA to Z-DNA structural transitions have been observed in internal DNA regions. These locally open DNA structures are good substrates for specific proteins, which can also induce the opening of a 'closed' helix. The DNA replication and repair processes will be discussed in more detail in Chapter 24.

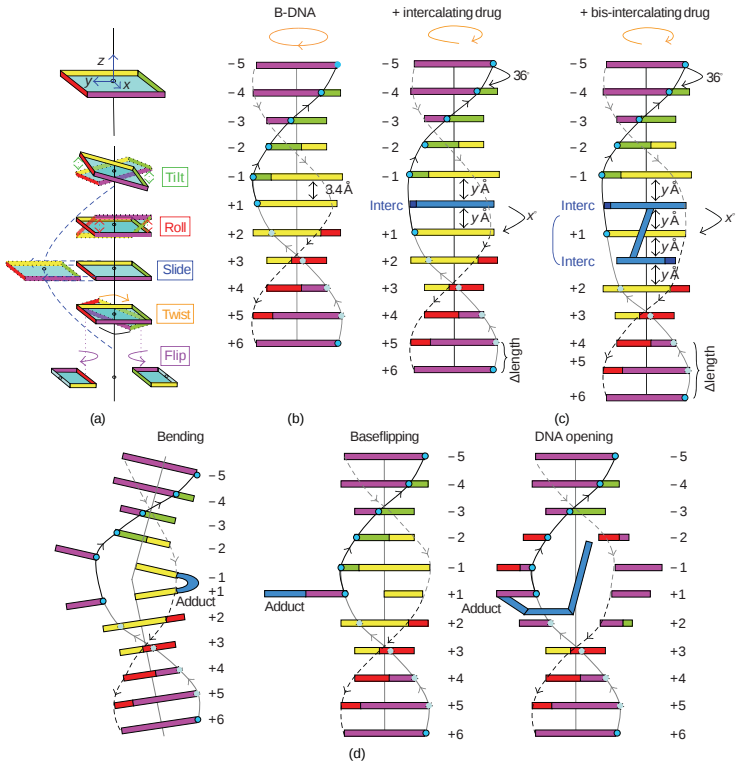
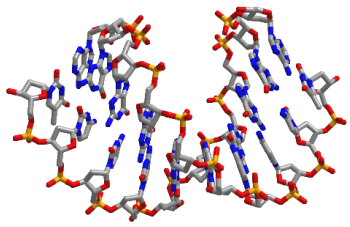
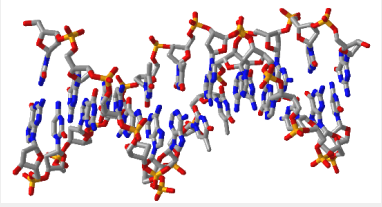
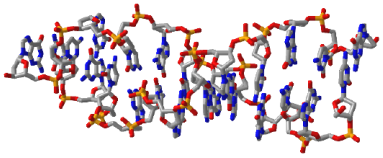


Figure 8.1.15: Localized Structural Modification of the DNA Double Helix. (a) Base pair orientations with the x, y, and z axes result in different kinds of rotation (tilt, roll, or twist) or slipping of the bases (slide, flip) regarding the central helix axis. (b) Matove B-DNA with nearly 11 base pairs within one helical turn. (c) Mono- or bis-intercalation of a small molecule (shown in blue) between adjacent base pairs resulting in an unwinding of the DNA helix (orange arrow on the top) and a lengthening of the DNA helix (ΔLength) depending on the X and y Å values that are specific for a defined DNA intercalating compound. (d) Representation of the DNA bending, base flipping, or double strand opening induced by some DNA destabilizing alkylating agents (adducts shown in blue). Adapted from Calladine and Drew's schematic box representation. Lenglet and David-Cordonnier (2010) Journal of Nucleic Acids, http://dx.doi.org/10.4061/2010/290935. Creative Commons Attribution License,

Figure 8.1.16 shows interactive iCn3D models of A-DNA (top), B-DNA (center), and Z-DNA (bottom). (Copyright; author via source). Click the image for a popup or use the external links in column 1.

A-DNA (440D)	
B-DNA (1BNA)	
Z-DNA (4OCB)	

NCBI iCn3D Figure 8.1.16: A, B, and Z-DNA. Click the image for a popup or use the links in column 1

We studied the structure of proteins in-depth, discussing resonance in the peptide backbone, allowed backbone angles φ, ψ, and ω, side chain rotamers, Ramachandran plots, and different structural motifs. We also explored them dynamically using molecular dynamic simulations. We also discussed the thermodynamics of protein stability and how stability could be altered by changing environmental factors such as solution composition and temperature.

In contrast, our understanding of the structural parameters and the dynamics of nucleic acids is less advanced. This may seem paradoxical, especially given the apparent simplicity of the iconic structure of DNA presented in textbooks. Yet, we should look at the types of secondary structures of nucleic acid presented and then the complicated tertiary and quaternary structures of RNA.

The nucleic acid backbone has a 5-membered sugar ring, which adds rigidity to the backbone, linked to another sugar ring by $CH_2O(PO_3)O$- connectors, adding some additional conformational freedom. We'll explore the effects of the pentose ring geometry in RNA and DNA in chapter section 8.3. To illustrate a yet unexplored complexity of nucleic acid structure, consider just the orientation of rings in double-stranded DNA and in regions of RNA where double-stranded structures form. The variants in the orientation of the hydrogen-bonded base pairs and the corresponding parameters that define them are shown in Figure 8.1.17.

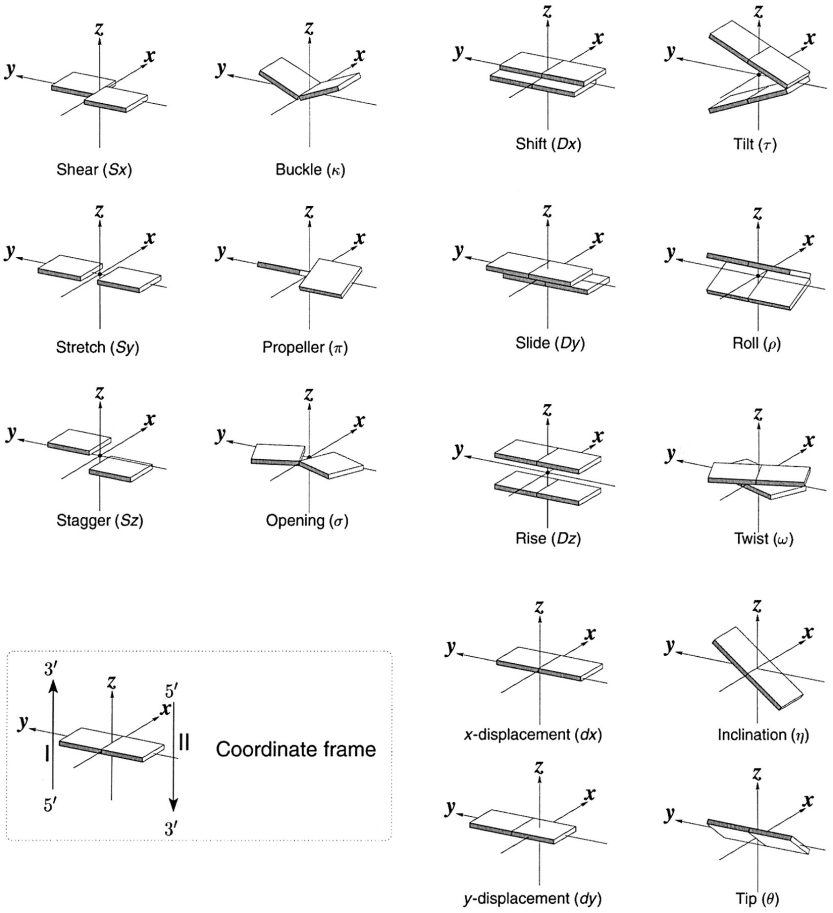


Figure 8.1.17: Base pair orientation and corresponding parameters in nucleic acids. http://x3dna.org/highlights/schemati...air-parameters (with permission). 2008 3DNA *Nature Protocols* paper (NP08), the initial 3DNA *Nucleic Acids Research* paper .

Consider just two of these: the **propellor** and **twist** angles. If you examine the iCn3D models of nucleic acids presented above, you will see that the base pairs are not perfectly flat but are twisted. Larger propeller angles are associated with increased rigidity. The propellor angles for A, B, and Z DNA are +18º, + 16 +/-7 º, and about 0º, respectively. The twist angles A, B, and Z DNA are +33º, +36 º, and -30º, respectively. The lower the twist angle, the higher the number of base pairs per turn. This, of course, affects the pitch of the helix (the length of one complete turn). These terms should be minimized to computationally determine the lowest energy state for a given double-stranded nucleic acid.

8.1.4: Alternative Base Pairing in DNA and RNA

A first glance at a DNA or RNA structure reveals a myriad of possible hydrogen bond donors and acceptors in the nucleic acid bases. Hence, it should come as no surprise that a variety of alternative or **noncanonical** (not in the canon or dogma) intermolecular hydrogen bonds can form between and among bases, leading to alternatives to the classical Watson-Crick base pairing. There are 28 possible base pairs with two hydrogen bonds between them. As structure determines function and activity, these alternative structures influence DNA/RNA function. We will consider four types of noncanonical base pairing: reverse Watson Crick, wobble, Hoogsteen, and reverse Hoogsteen base pairs.

These noncanonical base pairs can occur in DNA when bases become mismatched in double-stranded regions. In RNA, which we will explore more fully in Chapter 8.2, double-stranded molecules formed by separate RNA molecules aren't common. Instead, the molecule folds on itself in 3D space to form a complex tertiary structure containing regions of helical secondary structure. RNAs also form quaternary structures when bound to other nucleic acids and proteins. Larger RNAs have loops with complex secondary and tertiary structures, which often require noncanonical base pairing, stabilizing the alternative structures. The noncanonical structures are also important for RNA-protein interactions in the RNA region which binds proteins. As with protein-ligand interactions, protein binding to RNA might also induce conformation changes, specifically noncanonical base pairs, in the RNA. For example, the HIV Rev peptide binds to a target site in the envelop gene of HIV (which has an RNA genome) and leads to the formation of an RNA loop with hydrogen bonding between two purines.

Figure 8.1.18 shows an interactive iCn3D model of the REV Response element RNA complexed with REV peptide (1ETF).

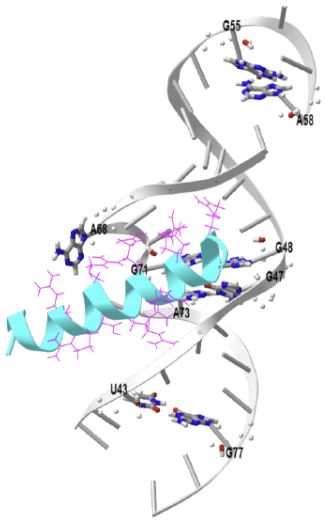


Figure 8.1.18: REV Response element RNA complexed with REV peptide (1ETF). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...T8CJ3pCe986Vx9

The peptide is shown in cyan, and its arginine side chains are shown as cyan lines. An extraordinary number of arginines form ion-ion interactions with the negatively charged phosphates in the major grove of this double-stranded A-RNA. The noncanonical base pairs are shown in CPK-colored sticks. A wobble base, U43-G77 (see below), can be seen, as well as three homopurine base pairs, G47-A73, G55-A58, and G48-G71. The solitary A68 base is shown projecting away from the RNA.

Figure 8.1.19 shows the Watson Crick and the first set of alternative non-canonical base pairs.

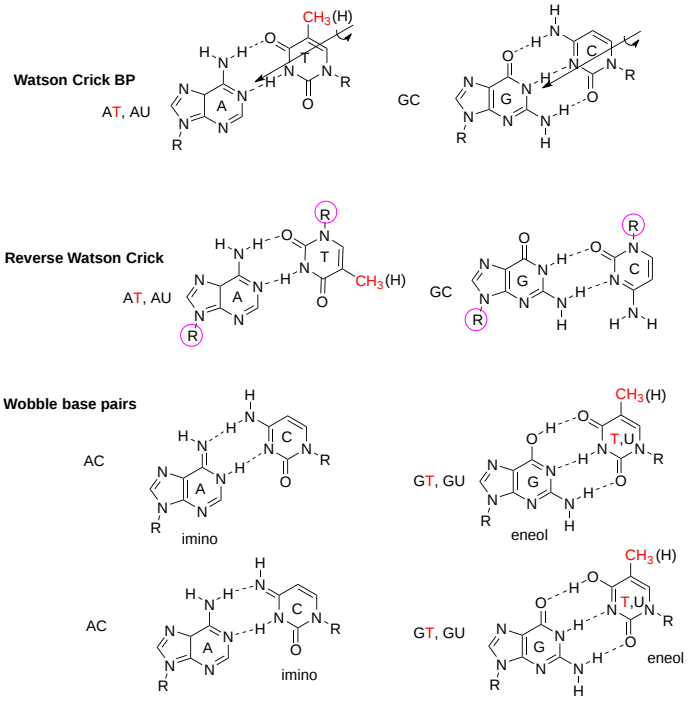


Figure 8.1.19: Some noncanonical base nucleic acid base pairs

Let's look at them in more detail.

Reverse Watson Crick: The reverse Watson-Crick AT (AU) and GC pairs can sometimes be found at the end of DNA strands and RNA. In forming the reverse base pairs, the pyrimidine can rotate 180^o along the axis shown and then rotate in the plane to align the hydrogen bond donors and acceptors, as shown in the top part of the figure. The glycosidic bond between the N in the base and the sugar (the circled R group) is now in an "antiparallel" arrangement in the reverse base pair.

Wobble Base Pairs

The bases in nucleic acids can undergo tautomerization to produce forms that can base pair noncanonically. They are termed **wobble** base pairs and include G-T(U) base pairs from keto-enol tautomerism and A-C base pairs from amino–imino tautomerism, as illustrated in Figure 18 above.

Figure 8.1.20 shows an interactive iCn3D model of the GT Wobble Base-Pairing in Z-DNA form of d(CGCGTG) (1VTT). Two such GT pairs are found in the structure.

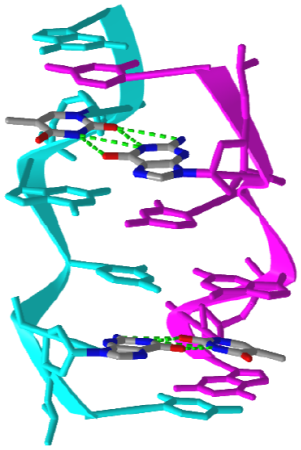


Figure 8.1.20: GT Wobble Base-Pairing in Z-DNA form of d(CGCGTG) (1VTT). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...LtwfzyeqDCaPEA

The water around the wobble base pairs can form hydrogen bonds and stabilize the pair if a hydrogen bond is missing.

Figure 8.1.21 shows an interactive iCn3D model of dsRNA with G-U wobble base pairs (6L0Y).

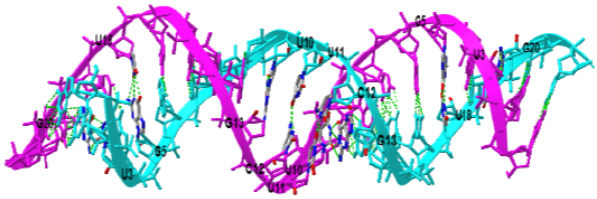


Figure 8.1.21: dsRNA with G-U wobble base pairs (6L0Y). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...reyeD6JQM1djq6

The structure contains many GU wobble base pairs and two CU base pairs between two pyrimidine bases.

Inosine, a variant of the base adenine, can be found in RNA. It is formed by the deamination of adenosine by the enzyme adenosine deaminase. A nucleotide having inosine is named hypoxanthine. Hypoxanthine can form the wobble base pairs I-U, I-A, and I-C when incorporated into RNA, as illustrated in Figure 8.1.22

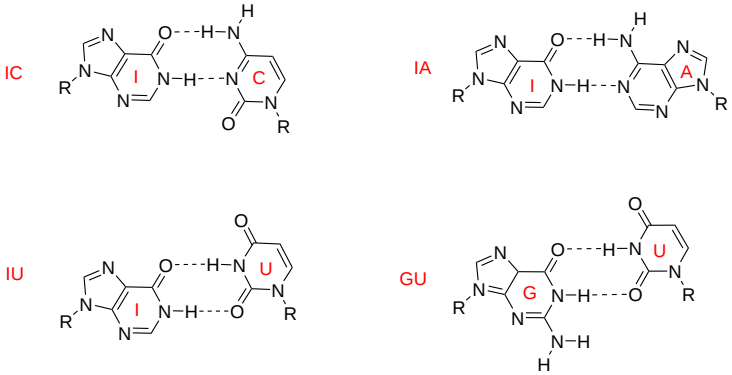


Figure 8.1.22 Wobble bases pairs using hypoxanthine with the base inosine

Wobble base pair interactions are especially important in translation when a protein sequence is made from a messenger RNA template (which will be discussed in Unit III). For that decoding process, two RNA molecules, messenger RNA (mRNA) and a transfer RNA (t-RNA) covalently attached to a specific amino acid like glutamic acid, must bind to each other through a 3-base pair interaction. The three bases on the mRNA are called the codon, and the three complementary bases on the tRNA are called the anticodon. The triplet base pairs are antiparallel to each other. The interaction between mRNA and tRNA is illustrated in Figure 8.1.23

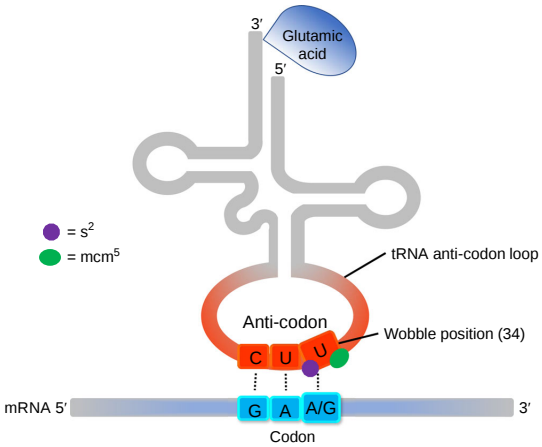


Figure 8.1.23: The wobble uridine (U34) of tRNA molecules that recognize both AA and AG-ending codons for Lys, Gln, and Glu, is modified by the addition of both a thiol (s2) and a methoxy-carbonyl-methyl (mcm5). This double modification enhances the translational efficiency of AA-ending codons. Goffena, J et al. *Nat Commun* **9,** 889 (2018). https://doi.org/10.1038/s41467-018-03221-z. Creative Commons Attribution 4.0 International License. http://creativecommons.org/licenses/by/4.0/.

The third 3' base on the mRNA is less restricted and can form noncanonical, specifically, wobble base pairs, with the 5' base in the anti-codon triplet of tRNA. The term wobble arises from the subtle conformational changes used to optimize the pairing of the triplets. Wobble bases occur much more in tRNA than other nucleic acids.

Hoogsteen base pairing

Flexibility in DNA allows rotation around the C1'-N glycosidic bond connecting the deoxyribose and base in DNA, allowing different orientations of AT and GC base pairs with each other. The normal "anti" orientation allows "Watson-Crick" (WC) base pairing between AT and GC base pairs, while the altered rotation allows "Hoogsteen" base pairs. Figure 8.1.24 shows the different orientations for an AT base pair.

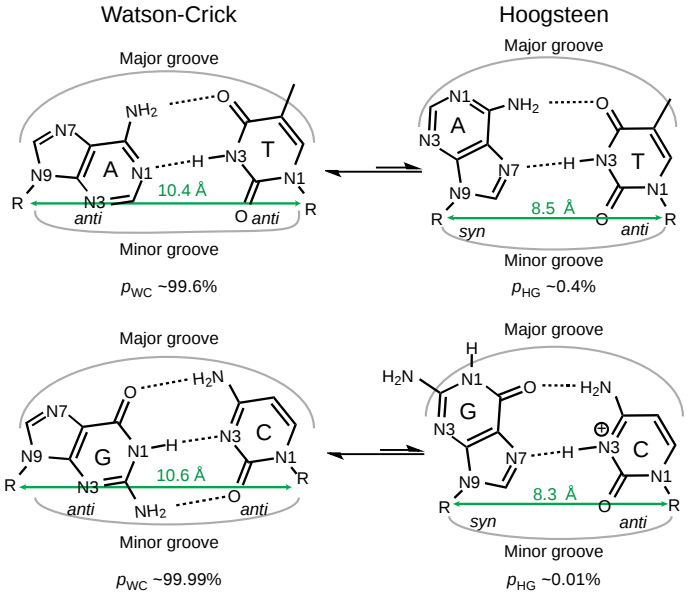


Figure 8.1.24: Xu, Y., McSally, J., Andricioaei, I. *et al.* Modulation of Figure 8.1.xaHoogsteen dynamics on DNA recognition. *Nat Commun* **9,** 1473 (2018). https://doi.org/10.1038/s41467-018-03516-1Creative Commons Attribution 4.0 International License. http://creativecommons.org/licenses/by/4.0/.

Hoogsteen base pairing is usually seen when DNA is distorted through interactions with bound proteins and drugs that intercalate between base pairs. Figure 8.1.25 shows an interactive iCn3D model of a Hoogsteen base pair embedded in undistorted B-DNA - MATAlpha2 homeodomain bound to DNA (1K61).

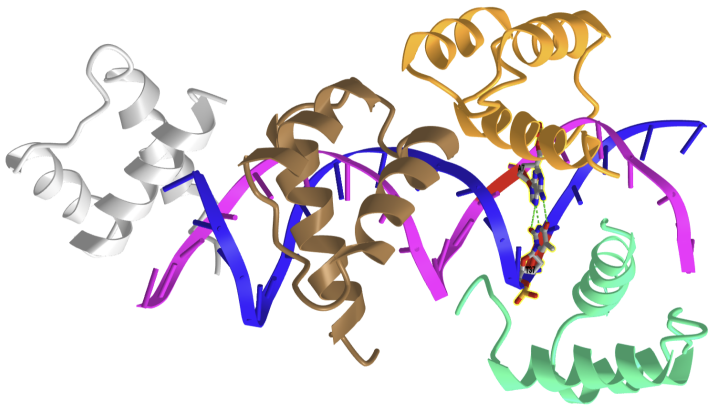


Figure 8.1.25: A Hoogsteen base pair embedded in undistorted B-DNA - MATAlpha2 Homeodomain bound to DNA (1K61). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...SLLRv1m8HQXKcA

The same DNA without bound protein has no Hoogsteen base pairs. To form Hoogsteen base pairs, a rotation around the glycosidic-base bond must occur. Hoogsteen base pairs between G and C can also occur on rotation, but in addition, the N3 of

cytosine is protonated, as shown in Figure 14 above.

Evidence suggests that Hoogsteen base pairing may be important in DNA replication, binding, damage, or repair. They can induce kinking of the DNA near the major grove.

There are also reverse Hoogsteen base pairing examples, as shown in Figure 8.1.26

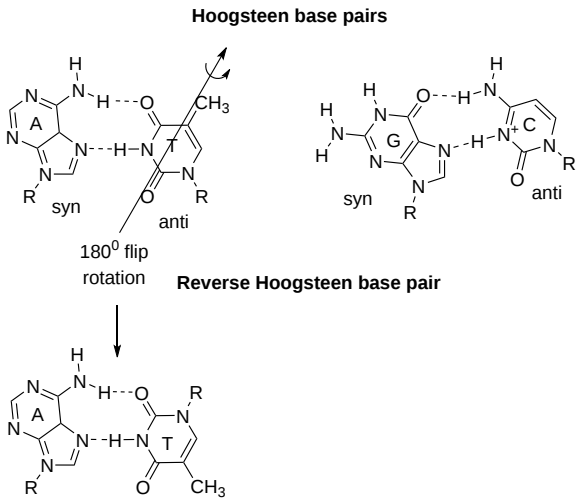


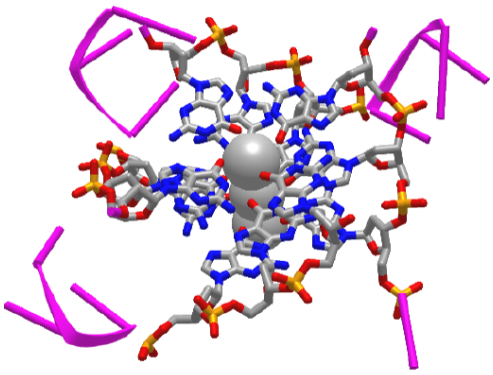
Figure 8.1.26: The reverse Hoogsteen AT base pair

8.1.5: Additional Alternative Structures: Quadruplexes, Triple Helices, and 4-Way Junctions

Quadruplexes

These can be formed in DNA and RNA from G-rich sequences involving tetrads of guanine bases that are hydrogen bonded. They are a bit hard to describe in words, so let's first examine one particular structure.

In human cells, telomeres (the ends of chromosomes) contain 300-8000 repeats of a simple **TTAGGG** sequence. The repetitive **TTAGGG** sequences in telomeric DNA can form quadruplexes. Figure 8.1.27 shows an interactive iCn3D model of parallel quadruplexes from human telomeric DNA (1KF1). The structure contains a single DNA strand **(**5'-AGGGTTAGGGTTAGGGTTAGGG-3') which contains four TTAGGG repeats.



NCBI iCn3D Figure 8.1.27: parallel quadruplexes from human telomeric DNA (1KF1). (Copyright; author via source).
Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...y5joFHDgWJQsQ6

Rotate the model to see three parallel layers of quadruplexes. In each layer, **4** **noncontiguous** guanine bases interact with a K^+ ion. Hover over the guanine bases in one layer, and you will find that one layer consists of guanines 4, 10, 16, and 22, which derive from the last G in each of the repeats in the sequence of the oligomer used (5'-AGGGTTAGGGTTAGGGTTAGGG-3'). These quadruplexes certainly serve in recognition and as binding sites for telomerase proteins. The guanine-rich telomere sequences, which can form quadruplexes, may also function to stabilize chromosome ends

A Quadruplex can be formed in 1 strand of nucleic acid (as in the above model) or from 2 or 4 separate strands. They also must have at least 2 stacked triads. As in the example above, single-stranded sections can form intramolecular G-quadruplex from a $G_mX_nG_mX_oG_mX_pG_m$ sequence, where m is the number of Gs in each short segment (3 in the structure above). A G might be in a loop if a segment is longer than others.

Triple Helices

These structures can occur in DNA (and also RNA) that contain homopurine and homopyrimidine sequences that have a mirror repeat symmetry. Hence, they can occur naturally. A mirror repeat contains a center of symmetry on a single strand. Here is an example: 5'-GCATGGTACG-3'.

They can also occur when a third single-strand DNA (a triplex-forming oligonucleotide or TFO) binds to a double-stranded DNA. The TFOs bind through Hoogsteen base pairing in the major grove of the ds-DNA. They can bind tightly and specifically in a parallel or antiparallel fashion. Specific and locally higher concentrations of divalent cations or positively charged polyamines like spermine act to stabilize the extra negative charge density from the binding of a third polyanionic DNA strand.

An example of a triple helix system that has been studied in vitro is shown in Figure 8.1.28

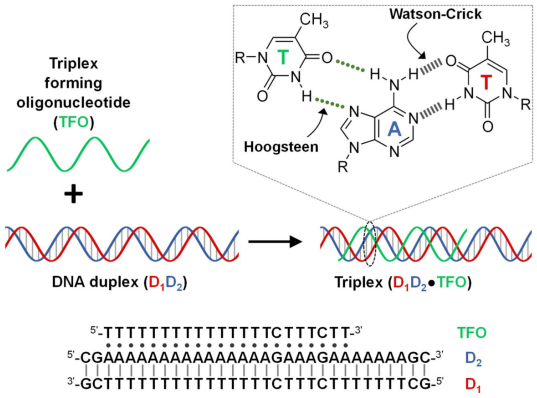


Figure 8.1.28 Intermolecular triplex formation and their oligonucleotide sequences (where "•" and "-" indicate Hoogsteen and Watson–Crick base pairings, respectively). Inset: chemical structure of a parallel T•AT triplet. Guerrini, L. and Alvarez-Puebla, R.A. Nanomaterials 2021, 11, 326. https://doi.org/10.3390/nano11020326. Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

The double-stranded canonical helix (D_1D_2) consists of 31 base pairs in which strand D_1 is pyrimidine-rich and D_2 is purine-rich strand (D_2). A 22-nucleotide Triple helix forming oligonucleotide (TFO) that is rich in pyrimidines binds the 19 AT and 2 C-GC base triplets. The TFO binds along the major grove of the D2 strand, which is purine-rich.

If the binding of the third strand in the major groove occurs at the site where RNA polymerase binds to a gene, then the third strand can inhibit gene transcription. Binding can also lead to a mutation or recombination at the site.

Figure 8.1.29 shows the base pairing of purine and pyrimidines of the third strand to the canonical AT dn GC base pairs of the original double-stranded DNA.

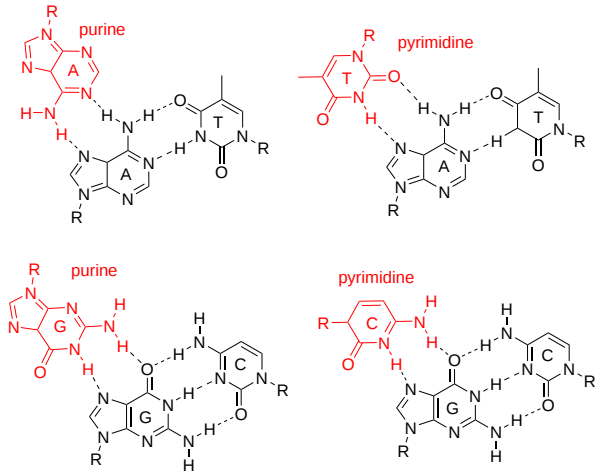


Figure 8.1.29: Base pairing in triple helix motifs. (after Jain et al. Biochimie. 2008. doi: 10.1016/j.biochi.2008.02.011

Figure 8.1.30 shows an interactive iCn3D model of a solution conformation of a parallel DNA triple helix (1BWG).

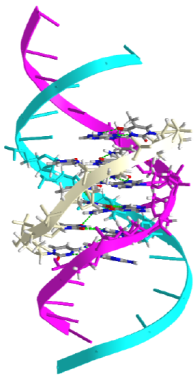


Figure 8.1.30: Solution conformation of a parallel DNA triple helix (1BWG). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...5JU813eNjND8E7

Triple helix formation can also occur within a single strand of DNA. The resulting structure is called **H-DNA**. An example is shown below. Note that the central blue, black, and red sequences are all mirror image repeats (around a central nucleotide). During processes that unravel DNA (replication, transcription, repair), the self-association of individual mirror repeats can form a locally stable triple helix, as shown in Figure 8.1.31.

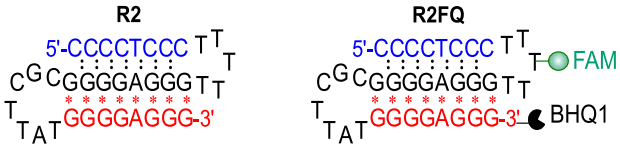


Figure 8.1.31: Schematic illustrations of (A) the H-DNA or intramolecular triplex structure used in this study; del Mundo et al. (2019) Nucleic acids research. 47. e73. 10.1093/nar/gkz237. Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/)

The * between in the G*G and A*A denote Hoogsteen hydrogen bonding (purine motifs) in this intramolecular triple helix. Reverse Hoogsteen hydrogen bonds can also occur.

Triple helices can form when single-stranded DNA formed during replication, transcription, or DNA repair with half of the required mirror symmetry folds back into the adjacent major grove and base pairs using Hoogsteen/reverse Hoogsteen bonding, which can be stabilized by Mg^{2+}.

Recent Updates: **Four-Way Junctions**

As we will see in the next section on RNA, nonhelical sections of DNA can bind small target molecules through noncovalent interactions. (RNA examples that we will see in the next chapter section include aptamers, ribozymes, and riboswitches.) One example is "lettuce" single-stranded DNA that can bind small fluorophores modeled after the intrinsic fluorophore of the green fluorescent protein. The fluorophore's fluorescence is dramatically enhanced when bound to the lettuce DNA. Figure 8.1.32 below shows the structure of extrinsic DNA fluorophores based on GFP that bind to the single-stranded "lettuce" DNA.

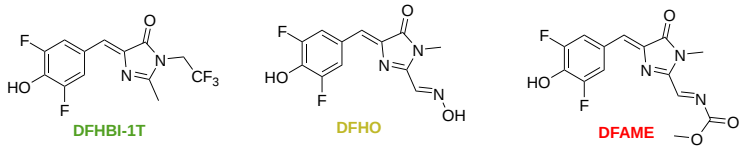


Figure 8.1.32 Structure of extrinsic DNA fluorophores based on GFP that bind to DNA. The font color of the names indicates the color of the emitted fluorescence.

Figure 8.1.33 shows an interactive iCn3D model of a solution conformation of a ssDNA:DFAME fluorophore complex (8FI0). The blue dotted lines show π-π stacking interactions, and the green dotted line a hydrogen bond.

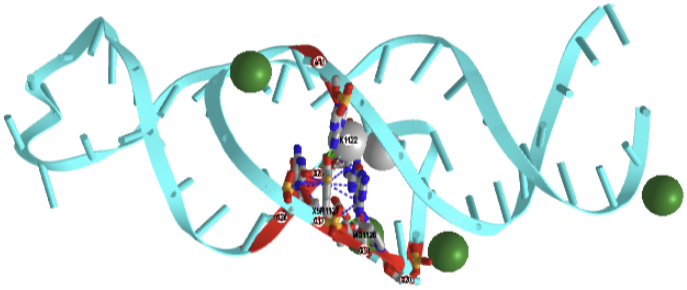


Figure 8.1.33: Solution conformation of a ss-53 mer DNA:DFAME fluorophore complex (8FI0). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...97eLqwNWmaTNC9

The DFAME ligand is shown as sticks.

Figure 8.1.34 shows a closeup of DFAME (colored spacefill) bound to the lettuce DNA.

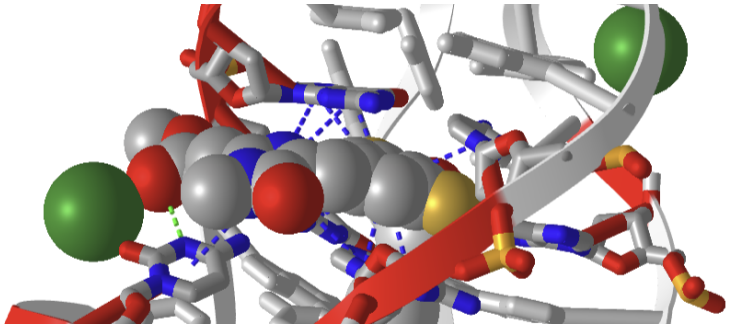


Figure 8.1.34: Pi stacking interactions (blue dotted lines) between the extrinsic fluorophore DFAME (spacefill) and ssDNA.

The DNA fold is characterized as a four-way junction (also seen in RNA, but they are more L or H-shaped). On either end are B-DNA duplexes, and the ssDNA between them forms stem-loops with odd base pairings in the stems. The overall structure is like a cloverleaf. Two coaxial stacks of nucleotides form a G-quadruplex where the fluorophore binds. Pi base stacking between diagonally packed bases and the binding of Mg^{2+} and K^{+} stabilize the structure.

8.1.6: Stability of nucleic acids

After looking at the myriad of structures showing the nearly parallel hydrogen-bonded base pairs and from ideas from most textbooks and classes you have taken, you probably think that double-stranded DNA is held together and stabilized by hydrogen bonds between the bases. It is well known that the greater the percentage of GC compared to AT, the greater the stability of the dsDNA. This translates into a higher "melting temperature (T_M)," the temperature at which the dsDNA is converted to ssDNA.

There is a linear relationship between GC content and T_M. The figures above show that GC base pairs have three inter-base hydrogen bonds compared to 2 in AT base pairs. These observations support the notion that inter-base hydrogen bonds are the source of dsDNA stability.

You would be, in general, correct in this belief. Still, you'd be missing the more important contributor to ds-DNA stability, base (π) stacking, and the noncovalent interactions associated with the stacking. The main contributors to stability are hydrophobic interactions in the anhydrous hydrogen-bonded base pairs in the helix. Given that the hydrogen bond donors and acceptors that contribute to base pairing exist in the absence of competing water, the donors and acce ptors are free to fully engage in bonding. The hydrogen bond interaction energy is more favorable in the stack. The stacking energy is similar for a AT-AT stack and a GC-GC stack (about -9.8 kcal/mol, 41 kJ/mol). Hence, the AT and GC base pairs contribute equally to stability. The excess stability of dsDNA enriched in GC base pairs can still be explained by the extra stabilization for an additional hydrogen bond per GC base pair

A myriad of interactions stabilizes proteins, but the folded state is marginally more stable than the ensemble of the unfolded state. Marginal stability is important as protein conformation often must be perturbed by binding and ensuing functions. The same must be true of double-stranded DNA, which must "unfold' or separate on replication, transcription, and repair. It is well known that dsDNA structure is sensitive to hydration (see the section on A, B, and Z DNA). As we saw with proteins, small molecules like urea can also denature DNA into single strands.

DNA must be stable enough to carry genetic information but dynamic enough to allow events that require partial unfolding. Other water-soluble molecules like ethylene glycol ethers (polyethylene glycol-400) and diglyme (dimethyl ether of diethylene glycol), which are more hydrophobic than water, appear to reduce base stacking interactions while maintaining them and, at the same time, allow a longitudinal extension or breathing of the helix. This dynamic extension may be required for transitions of B-DNA to Z-DNA, for example. The extensions also allow transient "holes" to appear between base pairs, which might assist in binding intercalating agents like some transition metal complexes. The extension caused by these ethers and natural extensions would decrease base stacking but also strengthen the hydrogen bonding between bases.

Longitudinal helical extensions might be important when homologous genes recombine. In that process, the homologous DNA strand is exchanged with a paired homolog. This processing is associated with strand extension and disruption of base pair at every third base. Recombination also must allow chain extension as it maintains base-pairing fidelity.

DNA structures get more complicated as they pack into the nucleus of a cell and form chromosomes, as shown in Figure 8.1.35. We will study the packing of DNA in other sections.

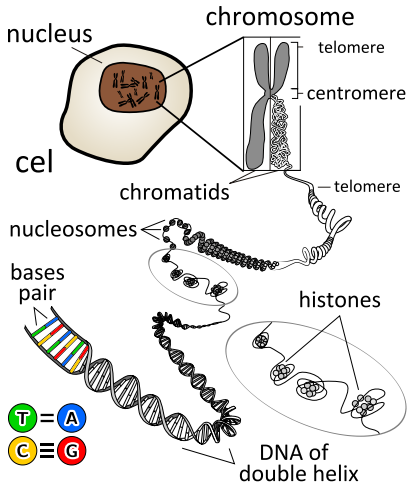


Figure 8.1.35 : Packing of DNA into the chromosome. https://commons.wikimedia.org/wiki/F...omosome_en.svg .

8.1.7: Summary

This chapter provides an in-depth overview of nucleic acids, one of the four fundamental classes of biological macromolecules, emphasizing their central role in storing and transmitting genetic information. It begins by contrasting DNA and RNA in terms of their chemical composition—highlighting that DNA contains deoxyribose and the bases adenine, guanine, cytosine, and thymine, while RNA uses ribose and replaces thymine with uracil. The discussion covers the nucleoside and nucleotide building blocks and explains how the sugar-phosphate backbone is constructed through the formation of phosphodiester bonds, emphasizing the directionality of nucleic acid synthesis (5′ to 3′).

The chapter then explores the structural organization of DNA and RNA. DNA's iconic double helix, stabilized by Watson–Crick base pairing and base-stacking interactions, is examined along with its major and minor grooves, which serve as critical recognition sites for proteins. In contrast, RNA, though generally single-stranded, adopts complex secondary and tertiary structures that enable diverse functions beyond mere information storage.

Additionally, alternative base pairing modes—including reverse Watson–Crick, wobble, Hoogsteen, and reverse Hoogsteen interactions—are discussed, illustrating how nucleic acids can form noncanonical structures that influence replication, transcription, and repair. The synthesis and dynamic behavior of nucleic acids are also addressed, including the mechanism of nucleotide incorporation by polymerases and the energetic contributions from pyrophosphate hydrolysis. Finally, the chapter contextualizes these structural features within the historical framework of nucleic acid research, noting landmark discoveries and the evolution of our understanding of DNA and RNA structures.

This comprehensive review lays the foundation for further exploration into the functional roles of nucleic acids in cellular processes, highlighting both the simplicity of their repeating structures and the complexity arising from their dynamic conformations and interactions.

8.1.8: References

Börner, R., Kowerko, D., Miserachs, H.G., Shaffer, M., and Sigel, R.K.O. (2016) Metal ion induced heterogeneity in RNA folding studied by smFRET. *Coordination Chemistry Reviews* 327 DOI: 10.1016/j.ccr.2016.06.002 Available at: https://www.researchgate.net/publication/303846502_Metal_ion_induced_heterogeneity_in_RNA_folding_studied_by_smFRET

Hardison, R. (2019) B-Form, A-Form, and Z-Form of DNA. Chapter in: R. Hardison's *Working with Molecular Genetics*. Published by LibreTexts. Available at: https://bio.libretexts.org/Bookshelves/Genetics/Book%3A_Working_with_Molecular_Genetics_(Hardison)/Unit_I%3A_Genes%2C_Nucleic_Acids%2C_Genomes_and_Chromosomes/2%3A_Structures_of_Nucleic_Acids/2.5%3A_B-Form%2C_A-Form%2C_and_Z-Form_of_DNA

Lenglet, G., David-Cordonnier, M-H., (2010) DNA-destabilizing agents as an alternative approach for targeting DNA: Mechanisms of action and cellular consequences. *Journal of Nucleic Acids* 2010, Article ID: 290935, DOI: 10.4061/2010/290935 Available at: https://www.hindawi.com/journals/jna/2010/290935/

Mechanobiology Institute (2018) What are chromosomes and chromosome territories? *Produced by the National University of Singapore.* Available at: https://www.mechanobio.info/genome-regulation/what-are-chromosomes-and-chromosome-territories/

National Human Genome Research Institute (2019) The Human Genome Project. *National Institutes of Health*. Available at: https://www.genome.gov/human-genome-project

Wikipedia contributors. (2019, July 8). DNA. In *Wikipedia, The Free Encyclopedia*. Retrieved 02:41, July 22, 2019, from https://en.Wikipedia.org/w/index.php?title=DNA&oldid=905364161

Wikipedia contributors. (2019, July 22). Chromosome. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:18, July 23, 2019, from en.Wikipedia.org/w/index.php?title=Chromosome&oldid=907355235

Wikilectures. Prokaryotic Chromosomes (2017) In MediaWiki, Available at: https://www.wikilectures.eu/w/Prokaryotic_Chromosomes

Wikipedia contributors. (2019, May 15). DNA supercoil. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:40, July 25, 2019, from en.Wikipedia.org/w/index.php?title=DNA_supercoil&oldid=897160342

Wikipedia contributors. (2019, July 23). Histone. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:19, July 26, 2019, from en.Wikipedia.org/w/index.php?title=Histone&oldid=907472227

Wikipedia contributors. (2019, July 17). Nucleosome. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:17, July 26, 2019, from en.Wikipedia.org/w/index.php?title=Nucleosome&oldid=906654745

Wikipedia contributors. (2019, July 26). Human genome. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:12, July 27, 2019, from en.Wikipedia.org/w/index.php?title=Human_genome&oldid=908031878

Wikipedia contributors. (2019, July 19). Gene structure. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:16, July 27, 2019, from en.Wikipedia.org/w/index.php?title=Gene_structure&oldid=906938498

8.2: Nucleic Acids - RNA Structure and Function

Learning Goals (ChatGPT o1, 1/30/25)

1. **Differentiate RNA from DNA:**
   - Describe the key chemical differences (e.g., ribose vs. deoxyribose; uracil vs. thymine) and explain how these differences influence the biological functions of RNA.
2. **Analyze RNA Structural Complexity:**
   - Explain how RNA can form secondary (e.g., stem-loop, hairpin) and tertiary structures, and evaluate the role of noncanonical base pairing and chemical modifications in these processes.
3. **Understand RNA Transcription and Processing:**
   - Illustrate the process of transcription from DNA to RNA, including the formation of heteronuclear RNA and subsequent splicing events that yield mature mRNA.
4. **Classify RNA Types Based on Function and Size:**
   - Distinguish between coding RNAs (e.g., mRNA) and noncoding RNAs, and further classify noncoding RNAs into short (<200 nt) and long (>200 nt) categories with examples such as tRNA, rRNA, miRNA, snoRNA, lncRNA, and circRNA.
5. **Evaluate the Role of RNA in Protein Synthesis:**
   - Describe the functions of messenger RNA in coding for proteins and the roles of transfer RNA and ribosomal RNA in translation, including the molecular interactions that facilitate these processes.
6. **Investigate RNA–Protein Interactions:**
   - Analyze specific examples of RNA–protein complexes (e.g., ribosome components, spliceosomal snRNPs, toxin-antitoxin systems) to understand how RNA structure and binding partners influence cellular functions.
7. **Explore RNA-Mediated Gene Regulation:**
   - Explain how noncoding RNAs such as microRNAs and siRNAs regulate gene expression through mechanisms like mRNA degradation, translational inhibition, or transcriptional modulation.
8. **Apply Thermodynamic Concepts to RNA Folding:**
   - Interpret the concept of RNA's thermodynamic folding landscape and compare it to protein folding, discussing how environmental factors and protein binding can lead to dynamic structural rearrangements.
9. **Examine RNA's Role in Cellular and Disease Contexts:**
   - Discuss the significance of RNA molecules in cellular processes (e.g., splicing, ribosome function, synaptic regulation) and their emerging roles in diseases such as cancer, cardiovascular disorders, and neurological conditions.
10. **Integrate Bioinformatics and Structural Tools:**
    - Utilize available computational programs and databases (e.g., lncipedia, miRBase, CircAtlas) to predict RNA secondary structures and explore experimentally determined RNA–protein complexes via interactive 3D models.

Each goal is designed to promote critical thinking, link structure to function, and encourage practical engagement with current research tools in RNA biology.

8.2.1: RNA: Structure and Function

Ribonucleic acids are very similar in chemical structure to DNA, except they contain ribose instead of deoxyribose. They also have the pyrimidine base uracil instead of thymine, as shown in Figures 1 and 2 above. These two small changes (but mostly the first) confer on it a very different set of biological functions than DNA. This should not surprise us, and the basis of all chemistry and biochemistry is that chemical structure determines chemical and biochemical functions and activities. In the previous section, we discussed how RNA can adopt complex tertiary structures, which requires the presence of more noncanonical base pairs and chemical modification of bases. In this section will explore the plethora of different types of RNA structures and their functions.

The sequence of RNA is made from DNA through a process called transcription (converting the information of DNA, a nucleic acid, into RNA, another nucleic acid). RNA can form double-stranded helices, but typically, these are viral in origin. DsRNA is a pathogen-associated molecular pattern (PAMP) that binds Toll-like receptor 3 (TLR3), as seen in Chapter 5.5. If both strands of DNA are transcribed, the resulting strands can anneal form dsRNA. In addition, a single strand of RNA can fold on itself if the 5' and 3' ends are complementary to form a stem-hairpin loop. Figure 8.2.1 shows a stem-loop from a messenger RNA (4QOZ) when it is bound to a specific RNA binding protein (not shown).

Figure 8.2.1 A stem-loop from a messenger RNA (4QOZ) when it is bound to a specific RNA binding protein (not shown).

Larger ss-RNA can form tertiary structures with many regions of intrastrand hydrogen bonds forming secondary structures, as shown in Figure 8.2.2 for one type of RNA called a transfer RNA

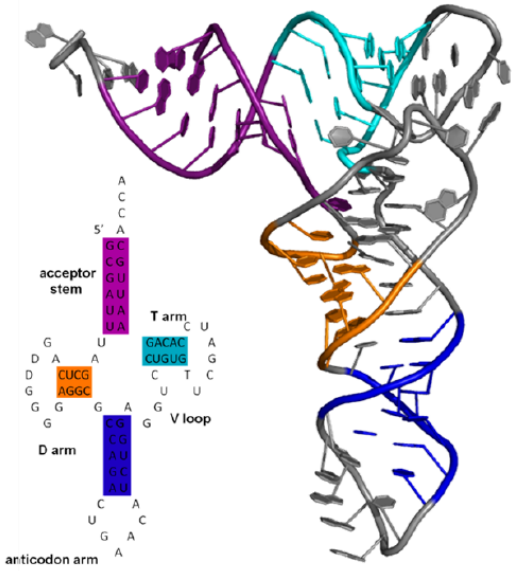


Figure 8.2.2: Secondary and tertiary structure of tRNA showing unpaired (gray) and hydrogen-bonded secondary structure regions (color). Kyle Schneider https://commons.wikimedia.org/w/inde...curid=12309266

Figure 8.2.3 a computed model for secondary structure within a much larger single RNA molecule S11 that is part of the ribosome. The figure shows color-coded differences in accessibility when the S11 RNA is free (blue) and bound to the protein NSP2 (red), which induces structural rearrangements.

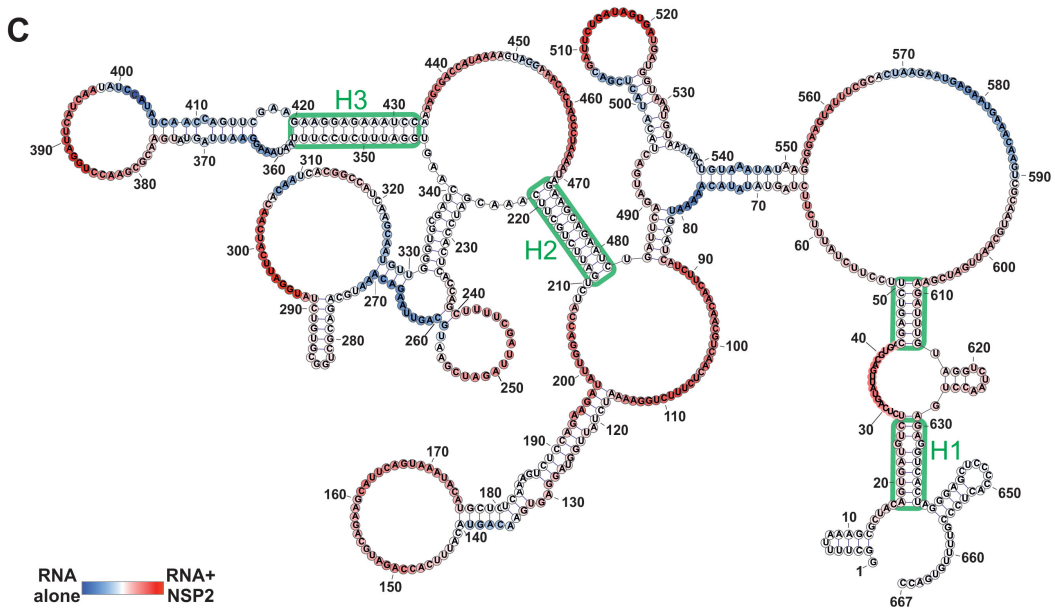


Figure 8.2.3: Model showing secondary structure on ribosomal S11 RNA and accessibility changes on binding the protein NSP2.

You can imagine that a different set of intrachain H-bonded double-stranded regions could easily form, with the most likely determined by sequence, local environment, and protein binding partners. Each RNA molecule would have a thermodynamic folding landscape similar to that of protein. Programs are available to determine secondary structures from RNA sequences. Each RNA molecule would have a thermodynamic folding landscape similar to that of protein. Also, the structures are dynamic, as we can see with proteins.

Another feature that complicates RNA is that many different types of RNA are made from DNA using RNA polymerases. They are loosely divided into two types of RNA. One is **coding RNA**, which contains the sequence information that will be translated into a protein sequence. The other type is called **noncoding RNA**. These RNAs regulate many cellular processes, including transcription which produces coding RNA.

8.2.1.1: Coding RNA

The DNA template from which the coding sequence of a translatable RNA is produced is called a **gene.** The coding RNA, which has the exact sequence that is translated into protein, is called **messenger RNA (mRNA)**. The exact sequence of RNA in mRNA that encodes a protein is derived from a longer contiguous DNA sequence in the nucleus from which sections called **intervening sequences** or **introns** have been removed. The coding sequences of DNA, which are separated by introns, are called **exons**. When coding DNA is transcribed, a long contiguous sequence containing both exons and introns is transcribed into one long primary transcript called **heteronuclear RNA**. The introns in the heteronuclear RNA are removed in a splicing reaction catalyzed by a large complex called the **spliceosome** to form mRNA. The process is illustrated in Figure 8.2.4. The first RNA sequence made is the heteronuclear RNA.

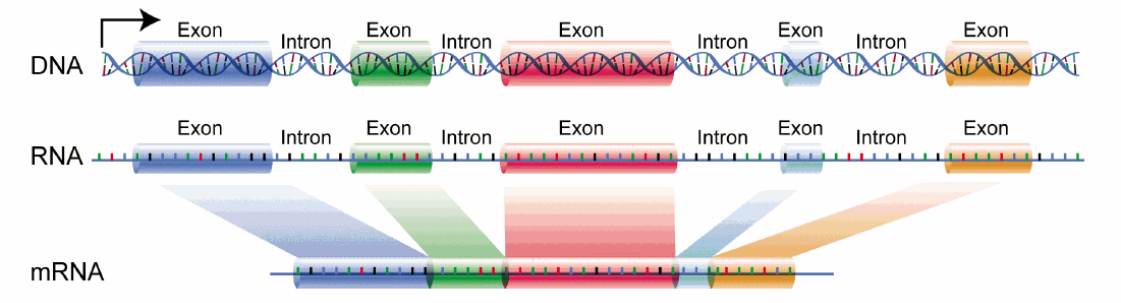


Figure 8.2.4: https://commons.wikimedia.org/wiki/F...ns_introns.gif. This work is in the **public domain** in the United States

The long single-stranded mRNA molecule binds to ribosomes and nanomachines, which orchestrate the translation of the mRNA sequence into a protein sequence. Around 20,000 human genes produce an even larger number of mRNA that arise from

differential splicing of the primary transcript.

8.2.1.2: Noncoding RNA (ncRNA)

Not long ago, few thought about possible RNA transcripts from non protein-coding regions of the genome, except for two types of RNA required to translate mRNA. These two are ribosomal RNAs (rRNA) found in ribosomes and transfer RNAs, to which amino acids are esterified and transferred to a growing protein chain of the ribosome. Many more classes have been discovered and given names that confuse someone more used to protein structures. One way to classify noncoding RNAs (ncRNAs) is based on size.

- short noncoding RNAs (sncRNAs) are <200 nucleotides
- long noncoding RNAs (lncRNAs)are >200 nucleotides

These function to regulate gene expression at both the transcription and post-transcriptional levels. Some have catalytic functions. Some affect chromosome structure and chemical modification.

8.2.1.2.1: Long Noncoding RNAs (lncRNAs)

There may be between 16,000 to over 100,000 human lncRNAs encoded into the genome, which adds much complexity to our understanding of the function of RNA transcripts. An online lncipedia is a database of searchable lncRNA sequences. There are many types of lncRNAs. The first we will consider is ribosomal RNA.

a. Ribosomal RNA (rRNA):

These RNAs fit the simple definition of lnRNAs (>200 nucleotides and are not protein- ing), but most would not think of them as lncRNAs since they have always been in their own category of a nonprotein-coding gene. rRNAs vary in length from between 1500 and 3000 nucleotides long in bacteria and about 1800 and 5000 nucleotides long in humans and are the core structure of ribosomes, the nanomachines that translate bound mRNA into a protein sequence.

Figure 8.2.5 shows an interactive iCn3D model of the structure of 23S rRNA of the large ribosomal subunit from Deinococcus radiodurans (2O44) (long load time).

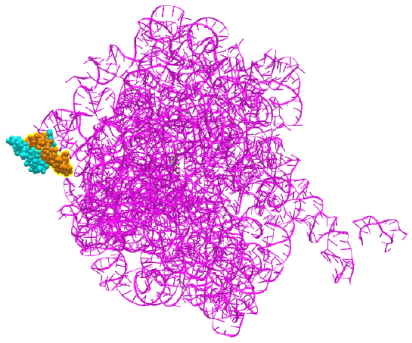


Figure 8.2.5: Structure of 23S rRNA of the large ribosomal subunit from Deinococcus radiodurans (2O44). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/icn3d/share.html?Tb2BVwSTe6LviPou5

The red (highlighted yellow) spacefill is the 5' start of the rRNA. The chain has a complex tertiary structure, like a protein sequence, and ends at the cyan spacefilling 3' end. It has 2880 nucleotides.

b. Other types

Let's focus on more classic examples of long noncoding RNAs (i.e., not rRNA). One way to categorize them is based on the position in the genome that encodes them. The different types include long **intergenic** noncoding RNAs (lincRNA), intronic lncRNAs, antisense RNAs (as lncRNAs), and other variants. These are illustrated in Figure 8.2.6, where the lncRNA is pink.

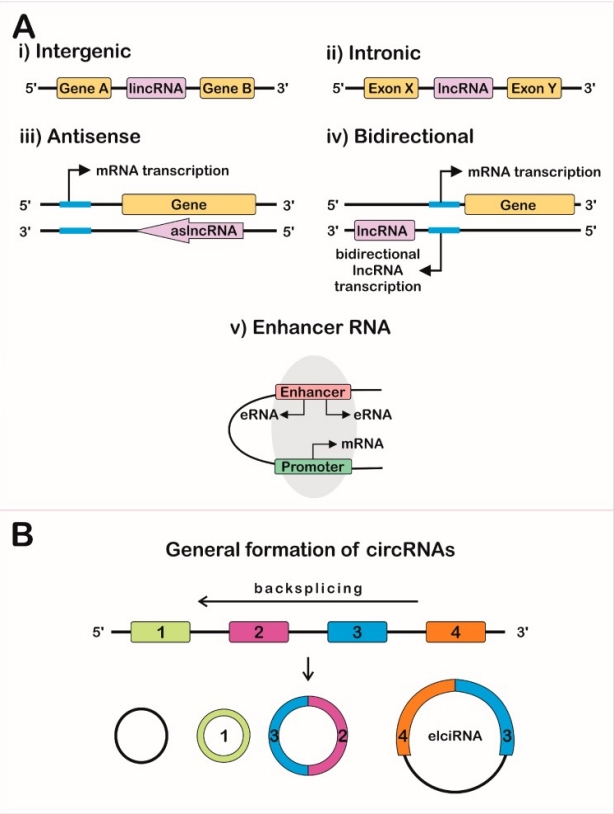


Figure \(\PageIndex{6}\): https://en.Wikipedia.org/wiki/Long_non-coding_RNA. Creative Commons Attribution-Share Alike 3.0 Unported

Another variant is exon and intron-containing circRNAs (EIciRNAs), as illustrated in panel B in Figure 8.2.13. These are presumably produced from pre-mRNA for a given mRNA and appear to regulate gene expression through RNA-RNA interactions with U1 snRNA, which starts the assembly of the spliceosome on pre-mRNA when it binds to the 5′ pre-mRNA splice site.

circRNA are found throughout the biological world and have been associated with diseases such as cancer, cardiovascular disease, and brain disorders like Alzheimer's. Their possible functions include binding to miRNAs and RNA binding proteins, altering their activities, and DNA, regulating its transcription. CircAtlas is a database of circRNAs that have been experimentally validated by two techniques

Now let's consider specific examples of long non-coding RNAs (lncRNAs), which are often bound to target proteins.

- mamRNA (a lnc RNA)

The lncRNA named mamRNA (**M**mi1 **a**nd **M**ei2-associated RNA) binds two proteins, Mmi1 and Mei2 in *Schizosaccharomyces pombe* that control the balane between meisois and mitosis in yeast. (*Schizosaccharomyces pombe* is a "fission" yeast that divides by fission, not budding.) The MamRNA has two variants of length 550 and 700 nucleotides. The binding of mamRNA leads to the ubiquitinylation of the Mei2 in the complex. Mmi1 R is an RNA-binding protein that binds to a modified version of adenosine methylated at N6 and is found internally in mRNA. Mei2 (meiosis protein 2) is necessary of meiosis. The binding of mamRNA leads to the ubiquitinylation of the Mei2 in the complex, targeting it for proteolysis. Mei2 concentrations relatively increase, shifting yeast from mitosis to meiosis. Figure 8.2.7 shows a cartoon depicting these interactions.

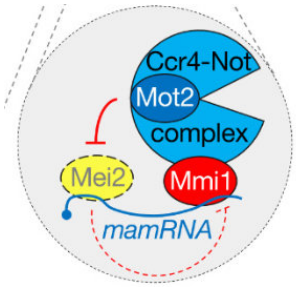


Figure 8.2.7: Binding of the lncRNA mam2 to yeast proteins Mei2 and Mmi1. Andric, V. et al. Yeast. Non-coding RNA 2021, 7, 34. https://doi.org/10.3390/ncrna7020034. Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Figure 8.2.8 shows an interactive iCn3D model of the S. pombe Mei2 RRM3 protein domain bound to the Mei2 binding "domain" of mamRNA (6YYM), which in this structure is only eight nucleotides long (not the full length this lncRNA which is 550 and 700 nucleotides long).

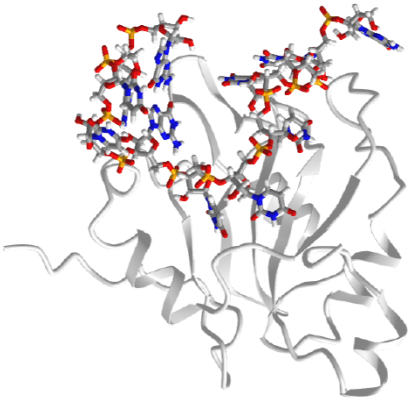


Figure 8.2.8: S. pombe Mei2 RRM3 protein domain bound to an eight nucleotide section of the yeast lncRNA mam (6YYM). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...5xmQjwAmQL9on7

- ToxI - a lnRNA inhibitor of the endonuclease ToxN

Those with a more chemistry-centric background might be surprised that viruses also "infect" bacteria. These viruses are called **bacteriophages**. It is estimated that there are over 10^{30} in nature. Some covalently incorporate into genomes where they reside and are incorporated permanently into the genome. They are a main driver of bacterial genome evolution as they shape the bacteria's immune response and adaptation.

One very interesting example is the type III toxin-antitoxin (TA) system in E. Coli. It consists of a toxin, ToxN, which is a nuclease that cleaves internally after the second A in a AAA sequence. It acts on mRNA, but especially pre-mRNA sequences. It is inhibited by the binding of a lncRNA called ToxI (toxin inhibitor). The RNA sequence of the ToxI inhibitor has 36 "domain" repeats of a pseudoknot, one of which is sufficient to inhibit the ToxN. The ToxN endonuclease cleaves the ToxI lncRNA as it assembles the complex. It also cleaves its mRNA. Figure 8.2.9 shows an interactive iCn3D model of the protein toxin (ToxN):lncRNA (ToxI), which is a shortened version of 29 nucleotide section from *Pectobacterium atrosepticum* (**2xdb**).

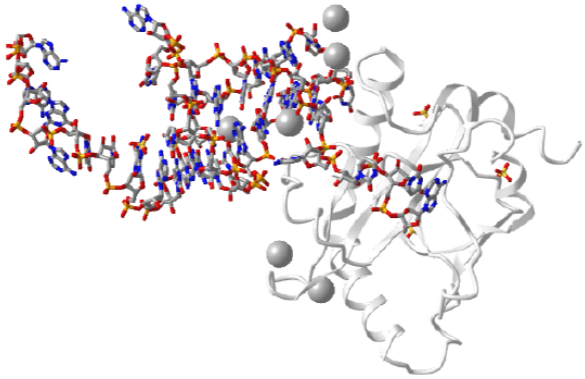


Figure 8.2.9: Protein toxin (ToxN) - 29 NT fragment of lncRNA antitoxin ToxI complex (**2xdb**). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...KjNPGghiUyAXr9

As we discussed in Chapter 3.2: The Structure of Proteins- An Overview, some long noncoding RNAs have noncanonical open reading frames (ncORFs) that are translated into protein, adding an element to the complexity of RNA.

8.2.1.2.2: Short Noncoding RNA

Short noncoding RNAs (sncRNAs) are less than 200 nucleotides in length. By definition, this would include **transfer RNAs (tRNAs),** which bring to the ribosome amino acids covalently attached to their 3' end of the tRNA for incorporation into a growing protein chain during the translation of mRNA. As with rRNA for lncRNAs, these are really in a class of their own. Others include **small nuclear RNAs (snRNAs)** involved in splicing, **small nucleolar RNAs (snoRNAs)** involved in the modification of rRNAs, and **microRNAs (miRNAs)**, involved in the inhibition of translation and transcription, **PIWI-interacting RNAs (piRNAs)**, and endogenous **small interfering RNAs (siRNAs).** It is difficult to remember the subtle differences among these, making them difficult to understand. We will tell their stories with a few targeted examples.

a. Transfer RNA:

Transfer RNAs act as adapter molecules between transcription and translation. They are between 76 and 90 nucleotides long and have a cloverleaf shape. An enzyme, aminoacyl-tRNA synthase, covalently attaches a select amino acid at its 3' end. Another end of the tRNA hydrogen bonds through 3 nucleotides (the anticodon) to a triplet nucleotide (the codon) on the mRNA that encodes a specific amino acid at that triplet position. Figure 8.2.10 shows an interactive iCn3D model of the structure of yeast phenylalanine tRNA (1EHZ)

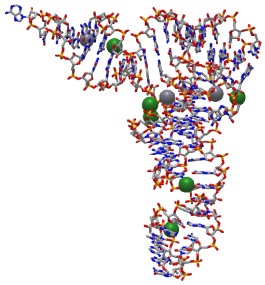


Figure 8.2.10: Yeast phenylalanine tRNA (1EHZ) (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/icn3d/share.html?iAWpKknDWbQm5hHL6

b. Small nuclear RNA (snRNA):

The spliceosome is a nanoparticle that catalyzes the removal of introns from pre-mRNA in eukaryotes (prokaryotes appear devoid of introns). The yeast spliceosome has a molecular weight of 1.3 million and contains five small ribonucleoproteins (RNPs) with many other associated proteins. Each of the 5 RNPs has a small nuclear RNA (U1, U2, U4, U5, and U,6) enriched in uracils. U6 is highly conserved and is directly involved in catalysis. Figure 8.2.11 shows an interactive iCn3D model of the core structure of the U6 small nuclear ribonucleoprotein complex with most of the U6 RNA bound.

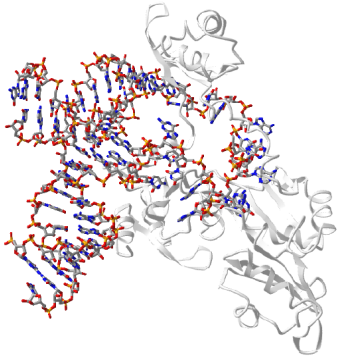


Figure 8.2.11: Core structure of the U6 small nuclear RNA - protein (ribonucleoprotein) complex (4N0T). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...P3SLVRAKhP7W57

c. MicroRNAs (miRNAs) and small inhibitory RNAs (siRNAs)

MicroRNAs (miRNAs) control the expression of thousands of genes in plants and animals. They are single-stranded but fold on themselves to form a stem-hairpin. The miRBase is a microRNA database containing almost 40,000 miRNA sequences. miRNAs are highly conserved and are found in animals, plants, and some unicellular eukaryotes. They interact with the 3′ untranslated regions of mRNAs and inhibit or prevent their translation. Several key proteins, **RNA polymerase II, Drosha and Dicer** are involved in the canonical pathway while the others appear to be independent of **Drosha**, which is a ribonuclease III double-stranded (ds) RNA endoribonuclease.

Dicer is a dsRNA) endoribonuclease, which cleaves long dsRNAs and short hairpin pre-microRNAs (miRNA) into fragments of either 21-23 nucleotides (short interfering RNA) or 19-25 nucleotides (microRNAs). Each has two nucleotides that are unpaired at the 3' end. These bind to the enzyme complex RISC (RNA-induced silencing complex), which then targets them to mRNA complementary to the siRNA/miRNA (RISC), causing cleavage of the mRNA and hence inhibiting translation.

Small inhibitory RNAs (siRNAs) are very similar to miRNA (to the point that differentiating between them is somewhat arbitrary). They both engage in RNA interference (RNAi) of mRNA translation. Here are some reported differences:

- The substrate for dicer cleavage is dsRNA (that could be added exogenously) of length 30-100+ for siRNA, but the actual pre-miRNA of length 7-100 nucleotides that may contain hairpins with some mismatches (i.e., not a perfect stem and hairpin)
- The final RNA after dicer processing is double-stranded for both and 21-23 nucleotides long for siRNA and 19-25 for miRNA
- siRNAs are perfectly complementary to the target mRNA,s while miRNAs, which are not necessarily perfectly complementary, bind typically to the 3' untranslated end of the mRNA
- Because of the perfect complementarity to target mRNA, siRNA interact with only one mRNA while miRNAs, given that they are not perfectly complementary to their target sequences, can bind different mRNAs

- Given their higher affinity binding, the siRNA leads to dicer endonuclease cleavage of the target mRNA. In contrast, inhibition of mRNA translation by miRNAs arises from binding of the miRNA to the mRNA or, if the match between the miRNA and mRNA is high enough, endonuclease cleavage of the mRNA.

Figure 8.2.12 shows canonical and several alternative pathways for their transcription and processing from the noncoding miRNA genes.

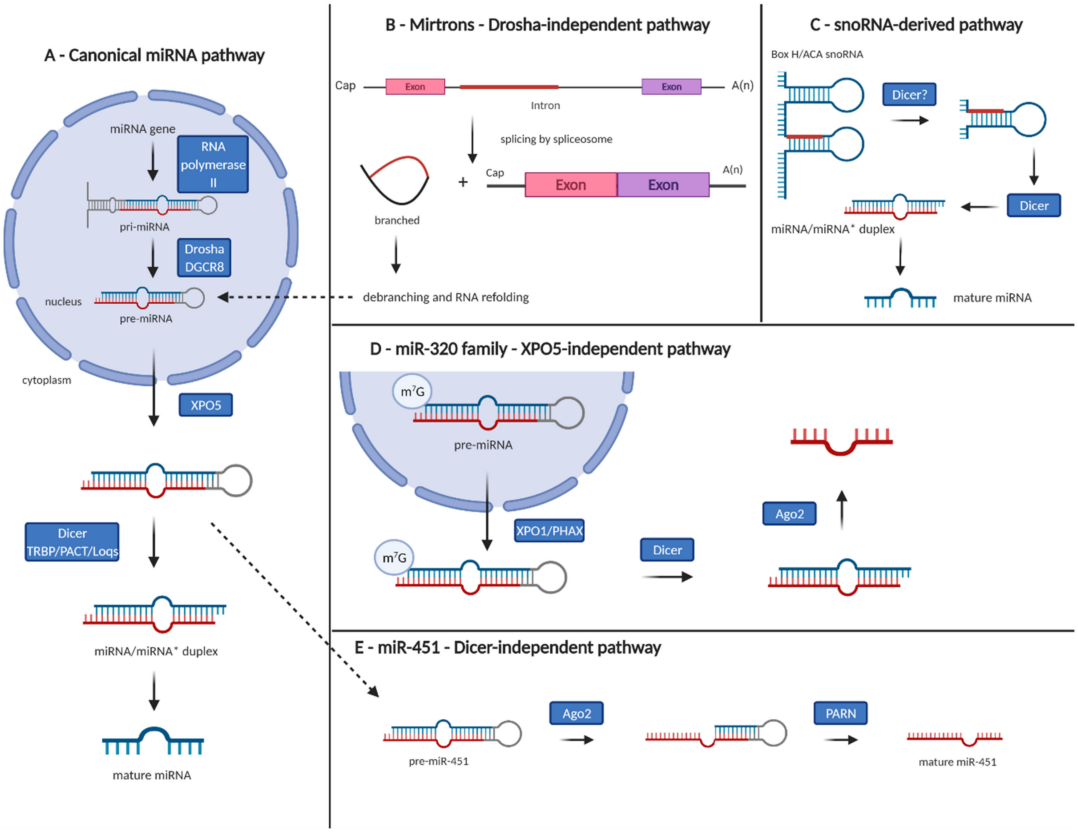


Figure 8.2.12:

Figure 1. Canonical and non-canonical pathways of microRNA biogenesis. (A) Canonical pathway—microRNA gene is transcribed by RNA polymerase II into primary microRNA (pri-miRNA), cleaved by microprocessor complex Drosha/DGCR8, and precursor microRNA (pre-miRNA) is exported from the nucleus to the cytoplasm by Exportin 5 (XPO5) and further processed by Dicer and its partners into 18–25 nucleotide long microRNA duplex with 2-nucleotide 30 overhangs. The guide strand is subsequently bound by the Argonaute proteins 1-4 (AGO1-4) and retained in the microRNA-induced silencing complex to target mRNAs for post-transcriptional silencing. (B) Mirtrons—generated through mRNA splicing independently of the Drosha-mediated processing step. (C) Small nucleolar RNA-derived microRNAs—Drosha-independent pathway. (D) Exportin 5-independent transport of pre-miRNAs from the nucleus to the cytoplasm has been described in the case of miR-320 family. (E) Dicer-independent processing of miR-451—pre-miR-451 is directly loaded into AGO2, cleaved, and trimmed by poly(A)-specific ribonuclease PARN to produce mature miR-451. Gregorova et al. Cancers 2021, 13, 1333. https://doi.org/10.3390/cancers13061333. Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

siRNA that are perfect matches to specific mRNA can be easily designed and purchased for translation inhibition and through gene silencing studies. Both miRNAs and siRNA are potentially therapeutic as it is much simpler to design a drug that targets a mRNA sequence (a 1D sequence target) than a protein active site (a 3D target). Additionally, they can be used to inhibit protein synthesis of target proteins that don't have a "druggable" active site.

The protein Argonaute is involved in miRNA and siRNA silencing of genes through their mRNAs in the RISC (RNA-induced silencing complex). RISC contains the protein argonaute 2 (AGO2) bound to a "**guide**" RNA, which is either microRNA (miRNA) or short interfering RNA (siRNA). The miRNA or siRNA directly interacts with the "target" - the mRNA.

- *Example miRNA:*

Figure 8.2.13 shows an interactive iCn3D model of human Argonaute2 Bound to a Guide (miRNA) and Target RNA (4W5O). The two RNA sequences are 5' UUCACAUUGCCCAAGUCUUU 3' and 5' CAAUGUGAAA 3'.

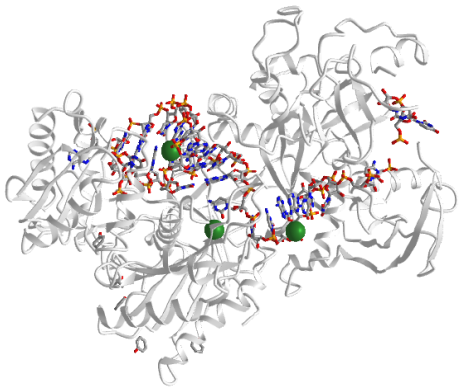


Figure 8.2.13: human Argonaute2 Bound to a Guide and Target RNA (miRNA) (4W5O). (Copyright; author via source). Click the image for a popup or use this external link: **https://structure.ncbi.nlm.nih.gov/icn3d/share.html?9fvyQmPkk4Kctei59**

Recent Updates: October 9, 2024

2024 Nobel Prize in Physiology and Medicine for discovery of miRNAs

Victor Ambros and Gary Ruvkun won the Nobel Prize (October 7, 2024) for discovering miRNA and its function in the roundworm *C. elegans*, which has fewer than 1000 cells. Their basic research on miRNA's role in protein synthesis regulation was extended to other organisms. Humans express over 1000 miRNAs that regulate translation.

Ruvkun found that two small RNAs, lin-4 (22 nucleotides) and let-7 (21 nucleotides),, are required to progress from larval to adult stages in the worm. They were nonhomologous but complementary to parts of the 3'-untranslated region (UTR) of mRNAs for proteins that are downregulated in development. Let-4 is found in many animal species, including humans, and binds to the 3'-UTR of the mRNA for the protein lin-41. Ambrose discovered that a small RNA lin-4 from a non-protein encoding gene was complementary to regions in the 3'-UTR of the mRNA for a protein lin-14.

The seed sequences of miRNAs are a conserved 7-mer at positions 2-7 from the miRNA 5'-end. This part of the miRNA must be exactly complementary to the 3'-UTR of the target mRNA, but the other bases don't have to match exactly.

The sequence of let-7 miRNA from C. elegans is 5'-UGAGGUAGUAGGUUGUAUAGU-3'. Figure 8.2.*i* below shows the aligned sequences of a let-7 miRNA and lin-41 mRNA (A) and analogs made for NMR structure determination (B,C).

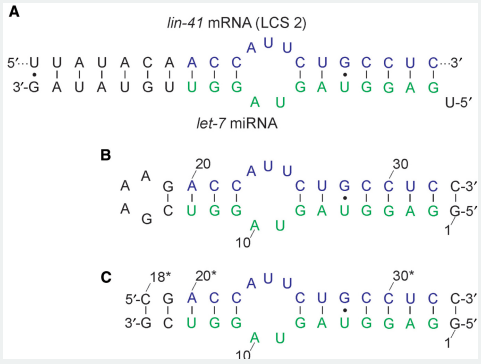


Figure 8.2.*i*: Sequences of a let-7 miRNA and lin-41 mRNA (A) and analogs.

(**A**) Schematic representation of the complex of *let-7* miRNA with 3′-UTR of the *lin-41* mRNA (LCS 2). The residues in green and blue match the sequences used in the monomolecular and dimeric constructs. (**B**) The 33-nt monomolecular RNA construct mimicking the complex. (**C**) The dimeric RNA construct. The numbering used in (B) has been preserved to facilitate comparing the two constructs. The residues originating from the *lin-41* mRNA sequence are labeled by

asterisk. *Cevec, M, Thibaudeau, C and Plavec, J. Nucleic Acids Research*, Volume 36, Issue 7, 1 April 2008, Pages 2330–2337, https://doi.org/10.1093/nar/gkn088. Creative Commons CC-BY-NC license.

Two analog structures preserved most of the matching sequences from the miRNA and mRNA (green for the let-7 miRNA and blue for the lin-41 mRNA. The seed sequence of the miRNA was completely preserved. One was a single strand that formed a stem-loop structure. Figure 8.2.*j* shows an interactive iCn3D model of the solution structure of a let-7 miRNA:lin-41 mRNA complex from C. elegans (2JXV) using the single-stranded analog.

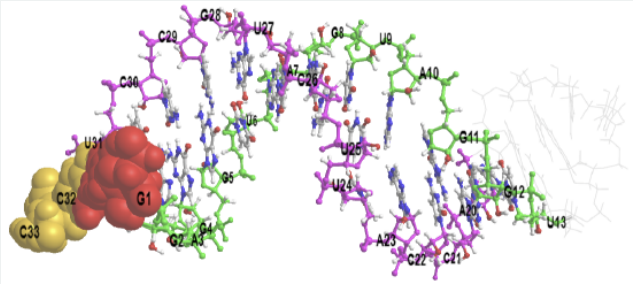


Figure 8.2.*j*: solution structure of a let-7 miRNA:lin-41 mRNA complex from C. elegans (2JXV). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...xJh4paMJRiTQ47

G1 (red spacefill) is the 5'-end and C33 (yellow) is the 3'-end of the single-stranded analog. The green bases match those from the figure above. For clarity, the matching base partners of lin-41 mRNA are shown in cyan instead of blue.

The model shows two stems with an asymmetric loop containing 3 Us and 2 A. A GU wobble base is seen in the first stem between U6 and G28.

- *Example: siRNA (Small interfering RNA)*

Virus genomes are ultimately decoded into new viruses by the host replication, transcription, and translation machinery. Host cells have evolved ways to silence mRNAs from viruses. Unfortunately, viruses, in response, evolve ways to suppress host RNA silencing. Many viral proteins are used by the host to suppress silencing. The viral p19 protein preferentially binds to host short interfering RNAs (siRNAs) than to microRNAs (miRNAs). A single mutation in the viral p19 proteins changes selectivity, allowing it to bind to a specific human miRNA called miR-122. This shows the subtle complexities of protein:RNA interactions. Figure 8.2.14 shows an interactive iCn3D model of the viral suppressor of RNA silencing protein and a 21 residue small interfering RNAs (6BJV)

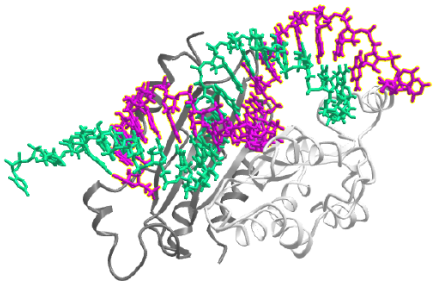


Figure 8.2.14: Viral suppressor of RNA silencing protein p19 mutants and a double-stranded 21 nucleotide small interfering RNAs (6BJV). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...VXzSgezTEPcPt9

- *Example: piRNA (a specific miRNA)*

Piwi proteins are RNA-binding proteins in plants and animals and are similar in structure to argonaute. They bind a guide RNA called piwi-interacting RNAs (piRNAs) and lead to the silencing of sequences called transposable elements that can move around the genome. piWi has endonuclease activity and can cleave mRNA. The piRNAs are just one type of miRNA. Figure 8.2.15 shows an interactive iCn3D model of Ephydatia fluviatilis (a sponge) PiwiA with a guide (piRNA) and-target RNA(7KX9)

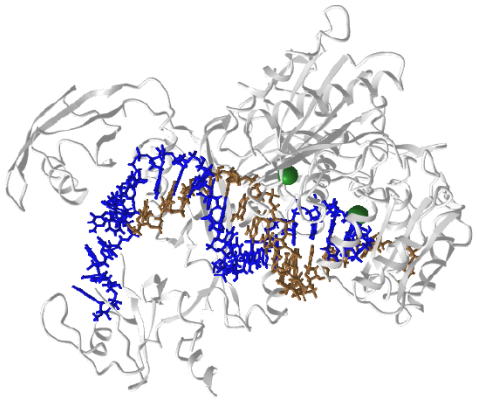


Figure 8.2.15: Ephydatia fluviatilis PiwiA-piRNA-target complex (7KX9). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/icn3d/share.html?j9ikApUgsVvqsiM5A

lncRNAs and miRNAs in the brain

After the decoding of the human genome, many have been struggling to understand how the complexity of the human brain (large size, greater connectivity among neurons) arises, given that we appear to have only around 20,000 protein genes encoded by the genome (not counting small proteins of less than 100 amino acids). Long noncoding RNAs (lncRNAs) and miRNAs appear significant pieces of this puzzle. Their ability to regulate transcription during development may hold the key. Additional roles on these RNAs outside of transcriptional regulation are being discovered. Some are transported away from the nucleus to serve other functions in axons, dendrites, etc. For example, the lncRNA Gm38257 binds to proteins that structure the synapse (components of the spectrin/ankyrin complex) instead of simply regulating gene expression.

The repertoire of miRNA appears to be significantly increased in "intelligent" organisms such as humans and octopi. For example, a large increase (179) in miRNAs occurs in proceeding in the evolutionary scale from mice (which have about 24,000 protein-encoding genes) to humans (around 20,000). miRNAs and lncRNAs may be involved (causative or correlative?) with brain disease. An example is the miR-124 is significantly elevated (3.5X) times higher in hippocampal cells from mouse models of Alzheimer's compared to normal mice. Altered expression of the lncRNA named Gomafu, RNCR2 or MIAT) appears to affect certain psychiatric diseases.

c. small nucleolar RNA

The nucleolus is a small nuclear structure that helps assemble the ribosomal RNAs synthesized in the nucleus. They then are transported through the nuclear membrane into the cytoplasm, where they combine with proteins translated from mRNA in the cytoplasm to form complete ribosomes. As described below, rRNA is chemically modified by enzymes (much like the post-translational modification of proteins). One such modification is 2'-O-methylation in archaea and eukaryotes. A class of small nucleolar RNAs (snoRNAs) that vary from 10-21 base pairs are called C/D RNAs, and they "guide" the modification. Hence, they are also called "guide" RNAs. These snoRNAs bind to 3-4 proteins into ribonucleoproteins. Figure 8.2.16 shows an interactive iCn3D model of the box C/D ribonucleoprotein 40 nt snoRNA "guide" and a 10 nucleotide RNA target substrates. It appears that the maximal duplex RNA formed (from the guide and target) is 10 base pairs long.

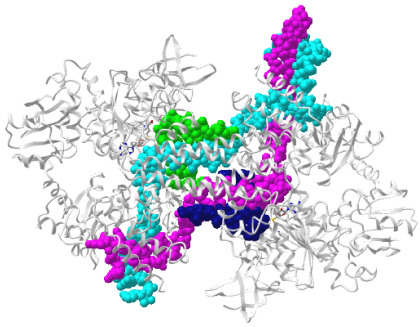


Figure 8.2.16: Box C/D ribonucleoprotein with a small nucleolar RNA (snoRNA) 12 nucleotide "guide" and 13 nucleotide target RNA for 2'-O-methylation (5GIO). (Copyright; author via source). Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/icn3d/share.html?q45tA8LEko87mYTw6

8.2.2: Summary

This chapter provides an in-depth exploration of RNA's unique chemical and biological roles, emphasizing how small differences in structure lead to diverse functions that are essential for cellular life.

Chemical Basis and Structural Diversity:
RNA differs from DNA primarily in that it contains ribose instead of deoxyribose and uracil in place of thymine. These seemingly minor differences not only alter its chemical properties but also empower RNA to adopt a wide range of structures—from simple stem-loop hairpins to complex tertiary conformations. Such structural variability is driven by noncanonical base pairing and chemical modifications, much like the dynamic folding observed in proteins.

Transcription and RNA Processing:
The chapter outlines the transcription process where RNA polymerases synthesize RNA from a DNA template. For coding RNAs, the initial heteronuclear transcript undergoes splicing to remove introns, resulting in mature messenger RNA (mRNA) that serves as the blueprint for protein synthesis. The intricacies of RNA processing are critical for ensuring that the correct genetic information is conveyed and translated.

Classification of RNA Types:
RNA molecules are broadly categorized into coding and noncoding RNAs. While mRNA carries the code for protein synthesis, noncoding RNAs (ncRNAs) serve regulatory and structural functions. The ncRNAs are further divided by size:

- **Short noncoding RNAs (<200 nucleotides):** Include tRNAs (adapters in translation), small nuclear RNAs (snRNAs) that are essential for splicing, microRNAs (miRNAs) and small interfering RNAs (siRNAs) involved in gene regulation, and other specialized molecules like piRNAs and snoRNAs.
- **Long noncoding RNAs (>200 nucleotides):** These include the well-known ribosomal RNAs (rRNAs) that form the core of the ribosome, as well as a diverse array of lncRNAs involved in transcriptional regulation, chromatin organization, and even direct protein interactions.

RNA–Protein Interactions and Functional Implications:
The chapter highlights numerous examples of RNA interacting with proteins to regulate cellular processes. From the assembly of the ribosome to RNA-based toxin-antitoxin systems in bacteria, these complexes underscore the importance of RNA's structure in mediating its interactions. Detailed structural models illustrate how RNA–protein binding can affect processes like mRNA splicing, translation, and gene silencing.

RNA in Gene Regulation and Disease:
A significant portion of the discussion is dedicated to the roles of miRNAs and lncRNAs in regulating gene expression. These molecules control the translation and stability of mRNAs and are implicated in various diseases, including cancer, cardiovascular disorders, and neurodegenerative conditions. Notably, recent advances—highlighted by the awarding of the 2024 Nobel Prize in Physiology and Medicine—underscore the importance of miRNAs in developmental processes and disease pathogenesis.

Integration of Computational and Experimental Tools:
The chapter also emphasizes modern techniques used to predict RNA secondary structures and analyze RNA–protein complexes. Databases such as miRBase and lncipedia, along with interactive 3D structural models, allow biochemists to visualize the dynamic nature of RNA and its myriad interactions within the cell.

In summary, this chapter equips students with a comprehensive understanding of RNA's multifaceted roles—from its chemical foundation and diverse structural forms to its critical involvement in gene regulation and disease. Through the integration of molecular details and modern analytical tools, junior and senior biochemistry majors gain valuable insights into how RNA not only translates genetic information but also orchestrates a complex network of regulatory processes within the cell.

8.2.3: References

Börner, R., Kowerko, D., Miserachs, H.G., Shaffer, M., and Sigel, R.K.O. (2016) Metal ion induced heterogeneity in RNA folding studied by smFRET. *Coordination Chemistry Reviews* 327 DOI: 10.1016/j.ccr.2016.06.002 Available at: https://www.researchgate.net/publication/303846502_Metal_ion_induced_heterogeneity_in_RNA_folding_studied_by_smFRET

Hardison, R. (2019) B-Form, A-Form, and Z-Form of DNA. Chapter in: R. Hardison's *Working with Molecular Genetics*. Published by LibreTexts. Available at: https://bio.libretexts.org/Bookshelves/Genetics/Book%3A_Working_with_Molecular_Genetics_(Hardison)/Unit_I%3A_Genes%2C_Nucleic_Acids%2C_Genomes_and_Chromosomes/2%3A_Structures_of_Nucleic_Acids/2.5%3A_B-Form%2C_A-Form%2C_and_Z-Form_of_DNA

Lenglet, G., David-Cordonnier, M-H., (2010) DNA-destabilizing agents as an alternative approach for targeting DNA: Mechanisms of action and cellular consequences. *Journal of Nucleic Acids* 2010, Article ID: 290935, DOI: 10.4061/2010/290935 Available at: https://www.hindawi.com/journals/jna/2010/290935/

Mechanobiology Institute (2018) What are chromosomes and chromosome territories? *Produced by the National University of Singapore*. Available at: https://www.mechanobio.info/genome-regulation/what-are-chromosomes-and-chromosome-territories/

National Human Genome Research Institute (2019) The Human Genome Project. *National Institutes of Health*. Available at: https://www.genome.gov/human-genome-project

Wikipedia contributors. (2019, July 8). DNA. In *Wikipedia, The Free Encyclopedia*. Retrieved 02:41, July 22, 2019, from https://en.Wikipedia.org/w/index.php?title=DNA&oldid=905364161

Wikipedia contributors. (2019, July 22). Chromosome. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:18, July 23, 2019, from en.Wikipedia.org/w/index.php?title=Chromosome&oldid=907355235

Wikilectures. Prokaryotic Chromosomes (2017) In MediaWiki, Available at: https://www.wikilectures.eu/w/Prokaryotic_Chromosomes

Wikipedia contributors. (2019, May 15). DNA supercoil. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:40, July 25, 2019, from en.Wikipedia.org/w/index.php?title=DNA_supercoil&oldid=897160342

Wikipedia contributors. (2019, July 23). Histone. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:19, July 26, 2019, from en.Wikipedia.org/w/index.php?title=Histone&oldid=907472227

Wikipedia contributors. (2019, July 17). Nucleosome. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:17, July 26, 2019, from en.Wikipedia.org/w/index.php?title=Nucleosome&oldid=906654745

Wikipedia contributors. (2019, July 26). Human genome. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:12, July 27, 2019, from en.Wikipedia.org/w/index.php?title=Human_genome&oldid=908031878

Wikipedia contributors. (2019, July 19). Gene structure. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:16, July 27, 2019, from en.Wikipedia.org/w/index.php?title=Gene_structure&oldid=906938498

8.3: Nucleic Acids - Comparison of DNA and RNA

Learning Goals (ChatGPT o3-mini, 2/1/25)

1. **Explain Chemical Modifications:**
   - Describe how intentional chemical modifications—such as methylation and hydroxymethylation—alter the structure and function of DNA and RNA, including their roles in regulating transcription and translation.
2. **Distinguish Epigenetic Versus Epitranscriptomic Changes:**
   - Compare how DNA methylation (and histone modifications) contribute to epigenetic regulation, while analogous modifications in RNA contribute to the emerging field of epitranscriptomics.
3. **Analyze Mutation Mechanisms:**
   - Explain how spontaneous chemical reactions, such as the hydrolytic deamination of cytosine, lead to point mutations and discuss the role of repair enzymes (e.g., uracil-DNA glycosylases) in maintaining genomic integrity.
4. **Evaluate the Impact of Chemical Agents:**
   - Assess how external mutagens (like nitrous acid and alkylating agents) cause point mutations and structural rearrangements in DNA, and predict their potential consequences for genomic stability.
5. **Compare Uracil and Thymine Roles:**
   - Discuss the chemical rationale behind using uracil in RNA and thymine in DNA, emphasizing the stabilizing effect of the methyl group in thymine to prevent erroneous repair of deaminated cytosine.
6. **Examine Backbone Linkage Chemistry:**
   - Evaluate why DNA and RNA employ phosphodiester bonds instead of alternative linkages (e.g., carboxylic acid esters or amides) and how the chemical properties of these bonds contribute to the stability and overall structure of nucleic acids.
7. **Understand Sugar Structure and Flexibility:**
   - Describe how the presence or absence of a 2'-OH group in ribose versus deoxyribose influences molecular stability, backbone flexibility, and the ability to form extended double-helical structures.
8. **Interpret Sugar Puckering and Helical Conformations:**
   - Analyze the effects of ribose puckering (C3'-endo vs. C2'-endo) on the formation of A-RNA versus B-DNA helices, and explain how these conformational differences affect nucleic acid-protein interactions.
9. **Differentiate Base Pairing Modes:**
   - Contrast Watson-Crick and Hoogsteen base pairing in DNA, discuss how their dynamic equilibrium is influenced by structural modifications, and understand the implications for protein recognition and damaged DNA repair.
10. **Integrate Structure-Function Relationships:**
    - Synthesize the comparative structural features of DNA, RNA, and proteins to explain how these differences have been evolutionarily selected to optimize genetic information storage, replication fidelity, and catalytic activity.

These goals are intended to guide your study of how subtle chemical differences in nucleic acids impact their structure, stability, and biological functions, thereby deepening your understanding of molecular biology at a fundamental level.

Now that we understand the structures of DNA and the structures and various functions of RNA, we can more fully explore how their chemical similarities and differences contribute to different functions.

8.3.1: Chemical modifications of DNA and RNA

Post-translation modifications of proteins alter their structural/functional properties. Likewise, intentional chemical modifications of nucleic acid bases alter both their structures and potentially their transcriptional and translational status. Figure 8.3.1 shows

common modifications of bases in DNA.

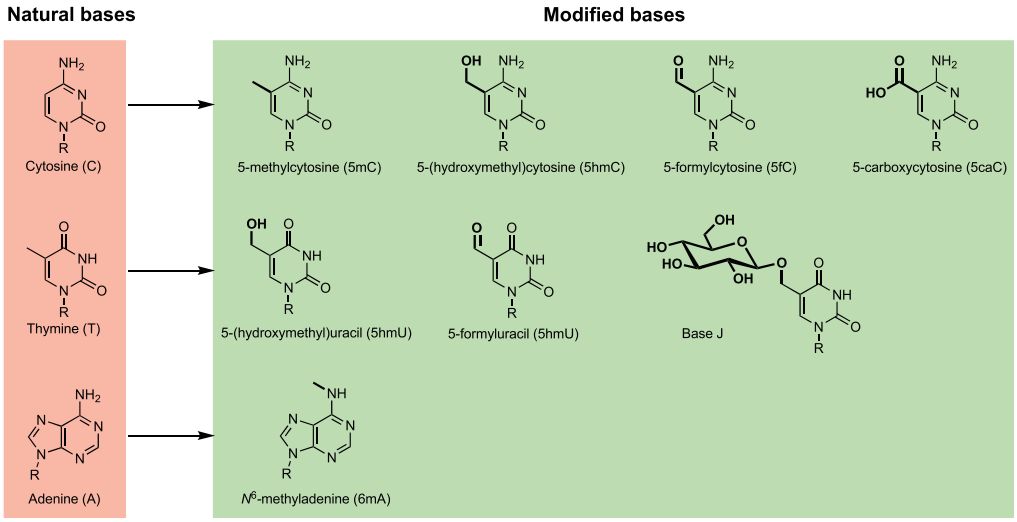


Figure 8.3.1: Common modifications of bases in DNA. Matthew K.Bilyard et al. Current Opinion in Chemical Biology. Volume 57, August 2020, Pages 1-7. https://doi.org/10.1016/j.cbpa.2020.01.014. Under a Creative Commons license

Likewise, RNA is chemically modified. Figure 8.3.2 shows common modifications of bases in RNA. Methylation and subsequent hydroxylation to hydroxymethyl are common to both DNA and RNA. Methylation of DNA often represses the transcription of the DNA into RNA. Hence, it has huge potential to alter gene transcription. Such changes to the DNA are called **epigenetic** modifications. These changes can also be passed down to future generations and affect a cell's phenotype. Histone proteins involved in DNA packing into nucleosomes can also be methylated and acetylated, altering the interaction of the DNA with the nucleosome core and further packing, again affecting transcription.

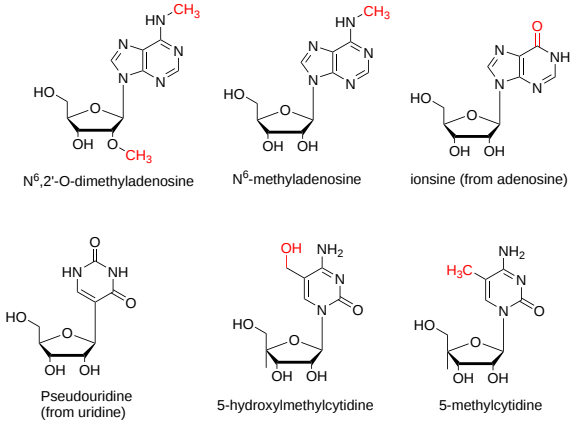


Figure 8.3.2: Common modifications of bases in RNA

Chemical modification to RNA can also change the reading of the genome. The **epitranscriptome** refers to the collective chemical modifications to RNA, and its understanding is part of a new field, **epitranscriptomics**.

8.3.2: Mutations

Mutation can arise from the chemical modification of bases. Uracil in RNA is a demethylated form of thymine in DNA. In RNA, AU base pairs replace AT base pairs. Why the need for uracil in RNA? The question could be rephrased as to why there is a need for thymine, with its extra methyl group, in DNA. It's useful to think about the consequence of replacing a single H in a molecule with a $-CH_3$. Take HOH (i.e., water) as an example. Our bodies are over 60% water. We drink liters of water with a concentration of 55 M each day. Yet if we drink 0.07 L of methanol (CH_3OH), half of us would die! Let's probe some consequences of the U (no $-CH_3$) and T (with $-CH_3$) changes in DNA. It can get confusing, but remember that the normal base pairs in DNA are AT, but AU base pairs also form (they are the norm in RNA). The $-CH_3$ substituent on thymine does not affect its base pairing.

a. Spontaneous deamination of cytosine in DNA

Why are we now discussing cytosine in DNA? One reason is that the most common mutation in DNA is a C to T replacement. One way that happens is through the spontaneous hydrolytic deamination of cytosine in DNA to uracil, which we have presumed to be found only in RNA. The mechanism for this deamination and subsequence conversion of a GC to an AT base pair is shown in Figure 8.3.3. The inset box shows a simplified mechanism for spontaneous deamination.

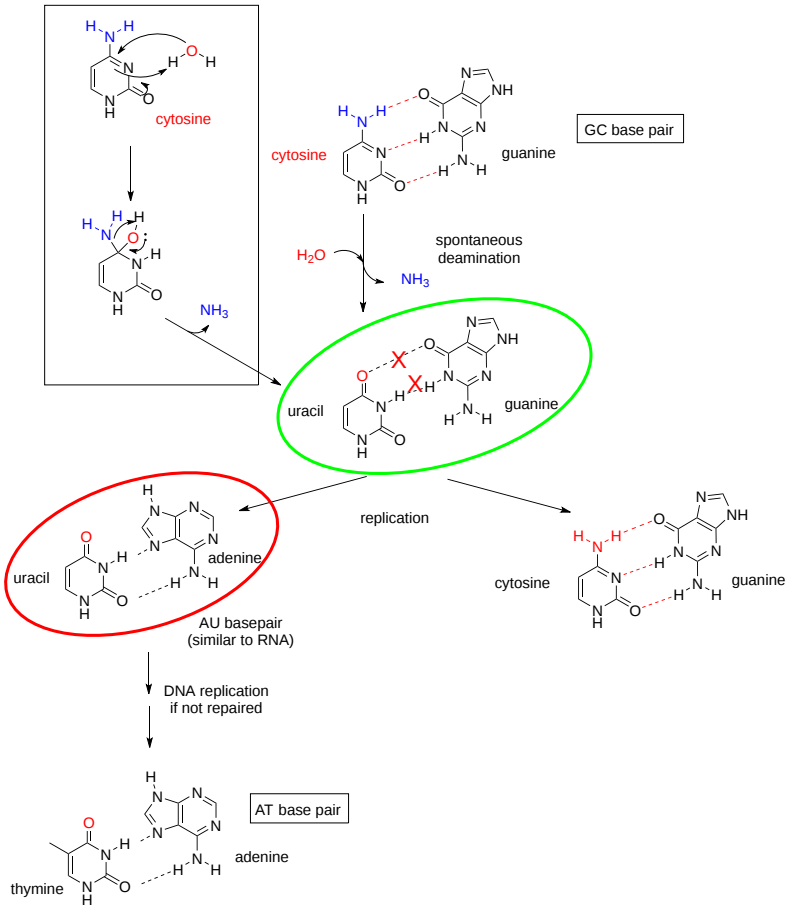


Figure 8.3.3: GC to AT base pair mutation on spontaneous hydrolytic deamination of cytosine in DNA.

Hence, a possible consequence of the deamination reaction is a GC to AT base pair mutation if the uracil in DNA is not removed before DNA replication. Fortunately, the enzyme uracil-DNA glycosylases can remove any uracils found in DNA, leaving an abasic site, which can be fixed with DNA repair enzymes.

We can now ask why T and not U in DNA. Pretend you are a DNA repair enzyme and see a UA base pair in DNA. How can you tell if the UA base pair is correct and intended to be there or if it should be a CG base pair that underwent deamination? The most common uracil-DNA glycosylases remove the uracil whether it is across from guanine, the correct base but which can not hydrogen bond with uracil (in the **green oval** in Figure 8.3.3), or if it is across from adenine, the wrong base (in **red oval**), which is present after a round of replication. Evolution has addressed this problem by adding a methyl group to uracil to form thymine and using that base, which forms a base pair with adenine. Now, no decision on which base across from a uracil (guanine if the uracil arose from deamination) or across from a "uracil-like" thymine (adenine) is correct.

b. Other mutations

Since we are considering chemical modifications to DNA and mutations, giving a more expanded background on them is appropriate. In addition to mutations caused by spontaneous hydrolytic cytosine deamination, mutations can also arise by adding a wrong base during DNA replication, by chemical damages caused by radiation or chemical modifying agents. How many mistakes in replication are made? You would be ecstatic if you received a 99% on an examination. That's not good enough for DNA replication. In Cell Biology by the Numbers, they calculate it this way. Assume the replication /repair is so good that it takes 10^8 replications to make a mistake (an error rate of 10^{-8}/BP). Assume also there are 3 x 10^9 base pairs in the human genome. This leads to a mutation rate 10-100 mutations/genome/generation or about 0.1-1 mutations/genome/replication. Not bad!

Figure 8.3.4 shows how common point mutations might randomly arise.

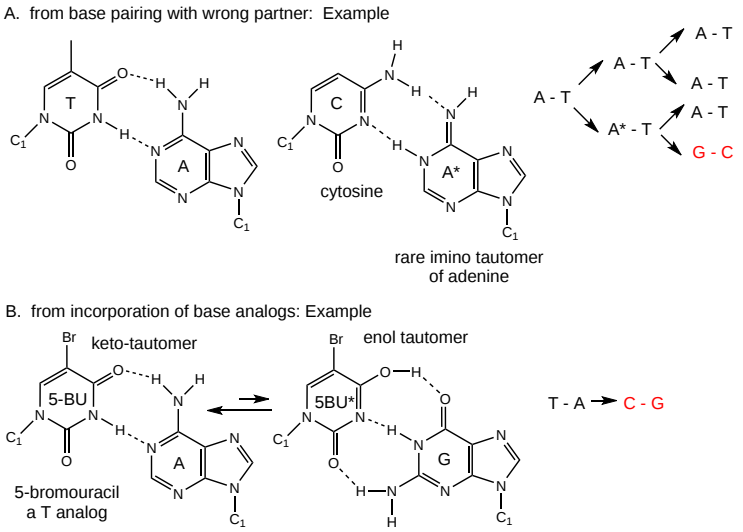


Figure 8.3.4: How common point mutations might arise randomly

Chemical agents also can cause point mutations. Figure 8.3.5 shows point mutations arising from oxidative deaminations (not hydrolytic) by nitrous acid/nitrosamines and from alkylating agents.

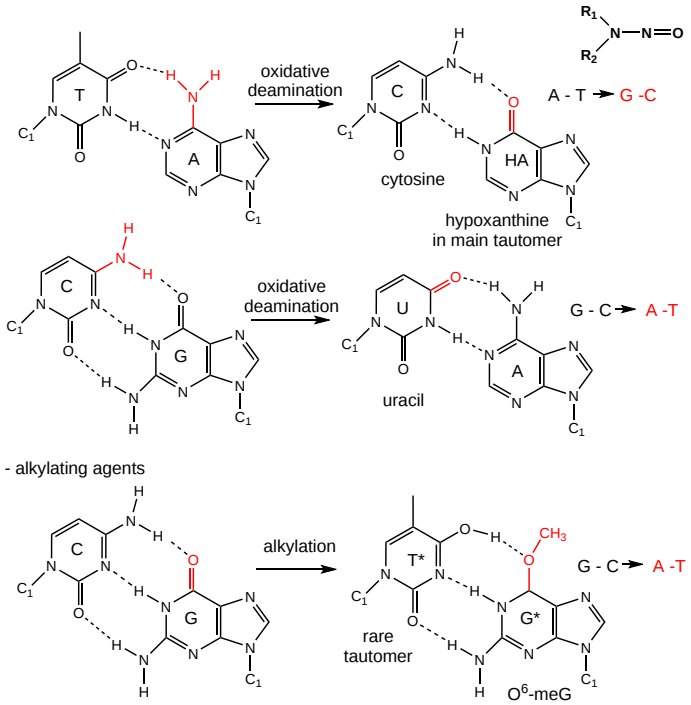


Figure 8.3.5: Nitrous acid/nitrosamines and alkylating agent point mutations

Figure 8.3.6 shows a variety of alkylating agents with mutagenic potential.

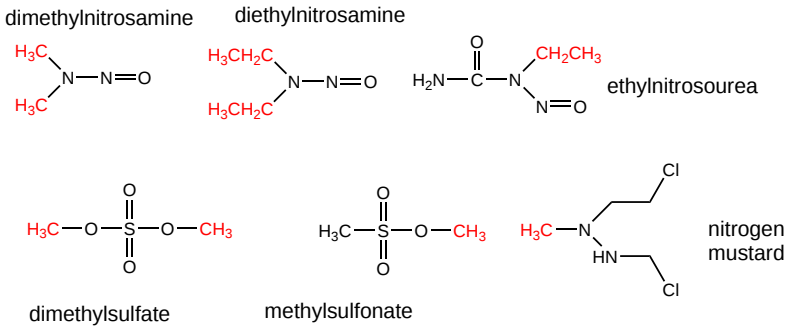


Figure 8.3.6: Alkylating agents with mutagenic potential.

Finally, large-scale changes in chromosome structure can also occur, as shown in Figure 8.3.7, usually with profound consequences.

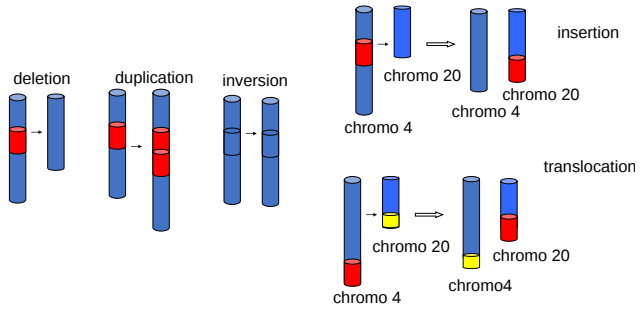


Figure 8.3.7: Large scale structural rearrangements in DNA

8.3.3: Why DNA and RNA - A Chemical Perspective

Asking a "why" question (like above) in the sciences is inappropriate as teleological questions are more philosophical or religious. Yet we will in this section, in part, to be in the company of Alexander Rich, who wrote a very cool article entitled "Why RNA and DNA have different Structures".

Given that RNA expresses catalytic activities and can carry genetic information (some viruses have ds and ss RNA as their genome), it has been suggested that early life might have been based on RNA. DNA would evolve later as a more secure carrier of genetic information. Inspecting the chemical properties of DNA, RNA, and proteins shows them to have attributes needed for their expressed function. Let's examine each for structural features that might be important for function.

a. Why does DNA lack a 2' OH group (found in RNA), which has been replaced with hydrogen? This required the evolutionary creation of a new enzyme, ribonucleotide reductase, to catalyze the replacement of the OH in a ribonucleotide monomer to form the deoxyribonucleotide form. One possible explanation is offered in the figure below. DNA, the main carrier of genetic information, must be an extremely stable molecule. An OH present on C'2 could act as a nucleophile and attack the proximal P in the phosphodiester bond, leading to a nucleophilic substitution reaction and potential cleavage of the bond. RNA, an intermediary molecule whose concentration (at least as mRNA) should rise and fall based on the need for a potential transcript, should be more labile to such hydrolysis. Figure 8.3.8 shows a possible reaction diagram for the internal cleavage of RNA. (The reaction would probably proceed with no actual intermediate but just a transition state.

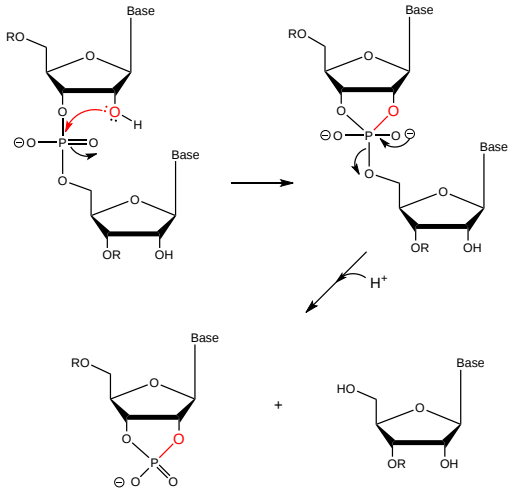


Figure 8.3.8: Internal cleave of RNA using the C'2-OH as an intramolecular nucleophile

b. Why do both DNA and RNA contain a phosphodiester link between adjacent monomers instead of more "traditional" links such as carboxylic acid esters, amides, or anhydrides? One possible explanation is given below. Nucleophilic attack on the sp^3 hybridized P in a phosphodiester is much more difficult than for a more open sp^2 hybridized carboxylic acid derivative. In addition, the negative charge on the O in the phosphodiester link would decrease the likelihood of a nucleophilic attack. The negative charges on both strands in ds-DNA probably help keep the strands separated, allowing the traditional base pairing and double-stranded helical structure to be observed. The cleavage of the phosphodiester link in DNA and a hypothetic ester link is shown in Figure 8.3.9. Again, the reaction of the phosphodiester shows a pentavalent intermediate, but most likely, the reaction proceeds directly from the transition state.

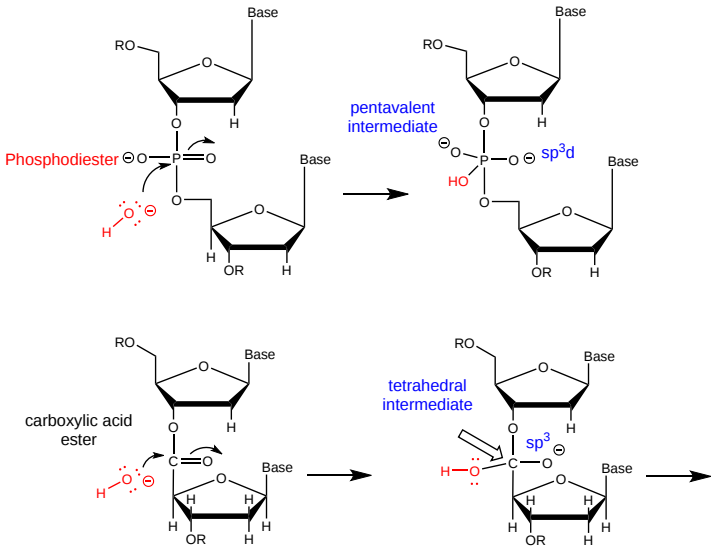


Figure 8.3.9: cleavage of the phosphodiester link in DNA and a hypothetic ester link

c. Why is DNA found as a repetitive double-stranded helix, but RNA is usually found as a single-stranded molecule that can form complicated tertiary structures with some ds-RNA motifs?

Another reason for the absence of the 2' OH in DNA is that it allows the deoxyribose ring in DNA to pucker just the right way to allow extended ds-DNA helices (B type). The pucker in deoxyribose and ribose can be visualized by visualizing a single plane in the sugar ring defined by the ring atoms C1', O, and C4'. If a ring atom points in the same direction as the C4'-C5' bond, the ring atom is defined as **endo**. If it points in the opposite direction, it is defined as **exo**. In the most common form of double-stranded DNA, B-DNA, the iconic extended double helix you know, C2' is in the endo form. It can also adopt the C3' endo form, forming another less common helix, a more open ds-A helix. In contrast, steric interference prevents ribose in RNA from adopting the 2'endo conformation. It allows only the 3'endo form, precluding the occurrences of extended ds-B-RNA helices but allowing more open, A-type helices.

Figure 8.3.10 shows another comparison between the A-RNA and B-DNA double helices and the C'3 and C'2 endo forms of the ribose

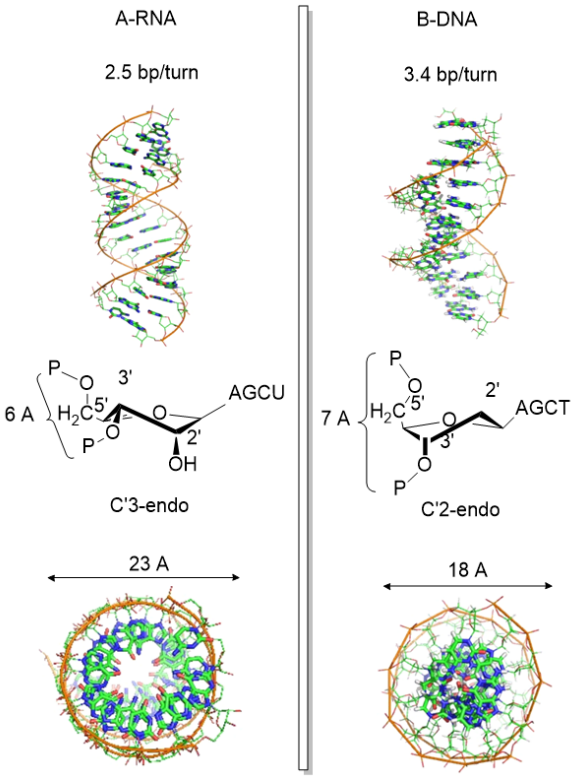
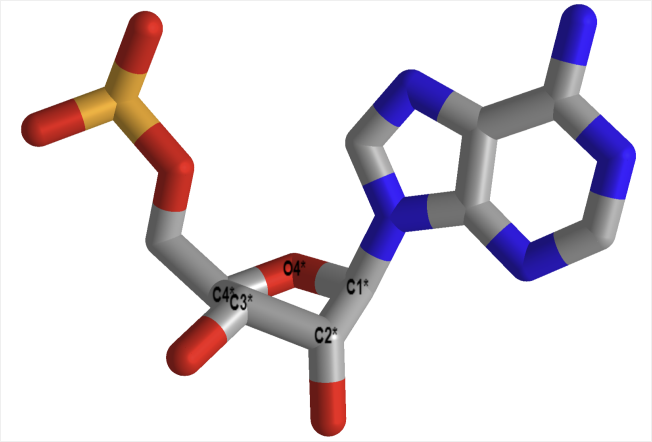
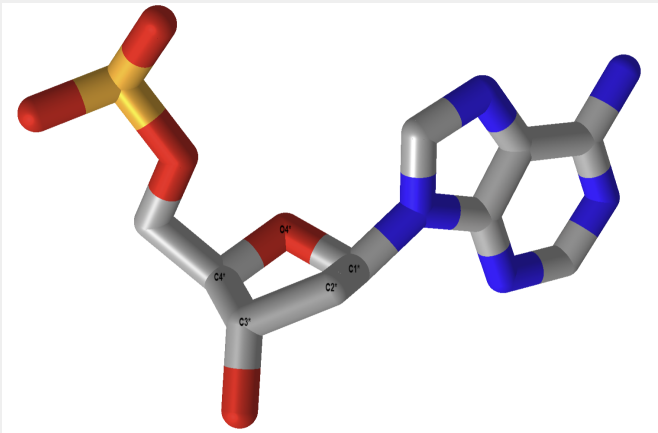


Figure 8.3.10: after Zhou et al. Nature Structural and Molecular Biology. doi:10.1038/nsmb.3270

Figure 8.3.5 shows interactive iCn3D models of the pentoses in a strand of A-RNA (413D), double-stranded, left, and B-DNA (1BNA), double-stranded, right.

C'3-endo ribose, A-RNA (413D, double stranded)	C'2 endo ribose, B-DNA (1BNA, double stranded)
	
Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...KPueqrBADczh26	Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...BEn5nqsCQG2JH6

d. What about the molecular dynamics of A-RNA and B-DNA?

The information above suggests that the sugar ring of DNA is conformationally more flexible than the ribose ring of RNA. This can be inferred from the observation that dsDNA can adopt B and A forms, which requires a switch from the 2' endo in the B form to the 3'endo form in the A form. The smaller H on the 2'C would offer less steric interference with such flexibility. The rigidity in

ribose is associated with a smaller 5'O to 3'O distance in RNA, leading to a compression of the nucleotides into a helix with a smaller number of base pairs/turn.

The increased flexibility in DNA allows rotation around the C1'-N glycosidic bond connecting the deoxyribose and base in DNA, allowing different orientations of AT and GC base pairs with each other. The normal "anti" orientation allows "Watson-Crick" (WC) base pairing between AT and GC base pairs, while the altered rotation allows "Hoogsteen" (Hoog) base pairs. Figure 8.3.11 shows the different orientations for an AT base pair.

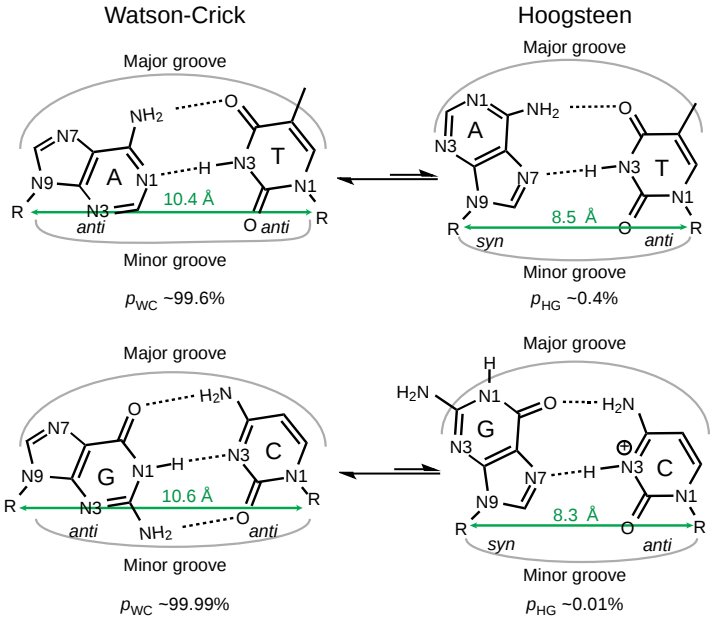


Figure 8.3.11: Xu, Y., McSally, J., Andricioaei, I. *et al.* Modulation of Hoogsteen dynamics on DNA recognition. *Nat Commun* **9,** 1473 (2018). https://doi.org/10.1038/s41467-018-03516-1Creative Commons Attribution 4.0 International License. http://creativecommons.org/licenses/by/4.0/.

The Watson-Crick (WC) and Hoogsteen (HG) base pairs in B-DNA are in a dynamic equilibrium, with the equilibrium greatly favoring the WC form, as indicated by the arrows in the figure above. In a DNA:protein complex, the WC ↔ HG equilibrium can favor the WG form for AT and GC^+ forms (in the latter, the C is protonated) when those base pairs are also involved in protein recognition. They can also occur more frequently in damaged DNA. In contrast, molecular dynamic studies show that the HG base pairs A-U and GC^+ are strongly disfavored in ds A-RNA.

One type of DNA damage is methylation on N1-adenosine and N1-guanosine. This modification prevents normal Watson-Crick base pairing, but for DNA, these modified bases can still engage in Hoogsteen base pairing, preserving the overall structure of dsDNA and its ability to carry genetic information stably. This same methylation occurs normally in post-transcriptional modified RNA. Hence, N1 adenosine and N1 guanosine methylation prevent any base pairing in the modified RNA. These properties make DNA a better carrier of molecular information and offer another way to regulate RNA's structural and functional properties.

Hoogsteen base pairs can be found in distorted dsDNA structures (caused by protein:DNA interactions) and normal B-DNA. Figure 8.3.12 shows a Hoogsteen base pair between dA7 and dT37 in the MAT α 2 homeodomain:DNA complex (pdb 1K61). Note that the dA base in the Hoogsteen base pair is rotated syn (with respect to the deoxyribose ring) instead of the usual anti, allowing the Hoogsteen base pair.

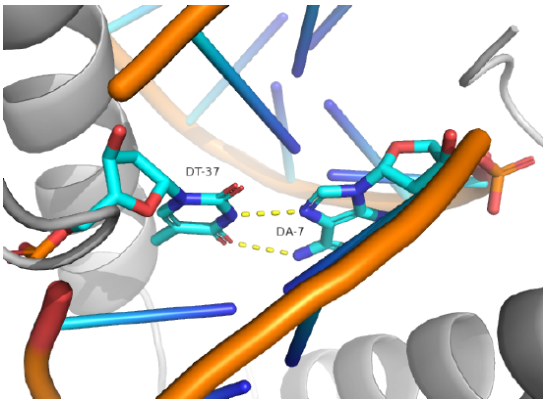


Figure 8.3.12: Hoogsteen base pair between dA7 and dT37 in the MAT α 2 homeodomain:DNA complex (pdb 1K61)

8.3.4: A Structural Comparison

Now, let's review the structures adopted by the three major macromolecules: DNA, RNA, and proteins. DNA predominately adopts the classic ds-BDNA structure, although it is wound around nucleosomes and "supercoiled" in eukaryotic cells since it must be packed into the nucleus. Prokaroytic DNA is typically packed into a more amorphous nuclear region, the nucleoid, through interactions with other proteins that also facilitated supercoiling. It is, in effect, a dynamic molecular condensate.

The extended ds-BDNA helical form arises partly from the significant electrostatic repulsions of two strands of this polyanion (even in counter-ions). Given its high charge density, it is unsurprising that it forms complexes with positive proteins and does not adopt complex tertiary structures. RNA, conversely, can not form long B-type double-stranded helices (due to steric constraints of the 2'OH and the resulting 3'endo ribose pucker). Rather, it can adopt complex tertiary conformations (albeit with significant counter-ion binding to stabilize the structure) and, in doing so, can form regions of secondary structure (ds-A RNA) in the form of stem/hairpin forms. Proteins, with their combination of polar charged, polar uncharged, and nonpolar side chains, offer little electrostatic hindrance in adopting secondary and tertiary structures. RNA and proteins can adopt tertiary structures with potential binding and catalytic sites, making them ideal catalysts for chemical reactions. Given its four nucleotide alphabet, RNA can also carry genetic information, making it an ideal candidate for the first evolved macromolecules enabling the development of life. Proteins with abundant organic functionalities eventually supplanted RNA as a better choice for life's catalyst. DNA, with its greater stability, would supplant RNA as the choice for the primary carrier of genetic information (Figure 8.3.13):

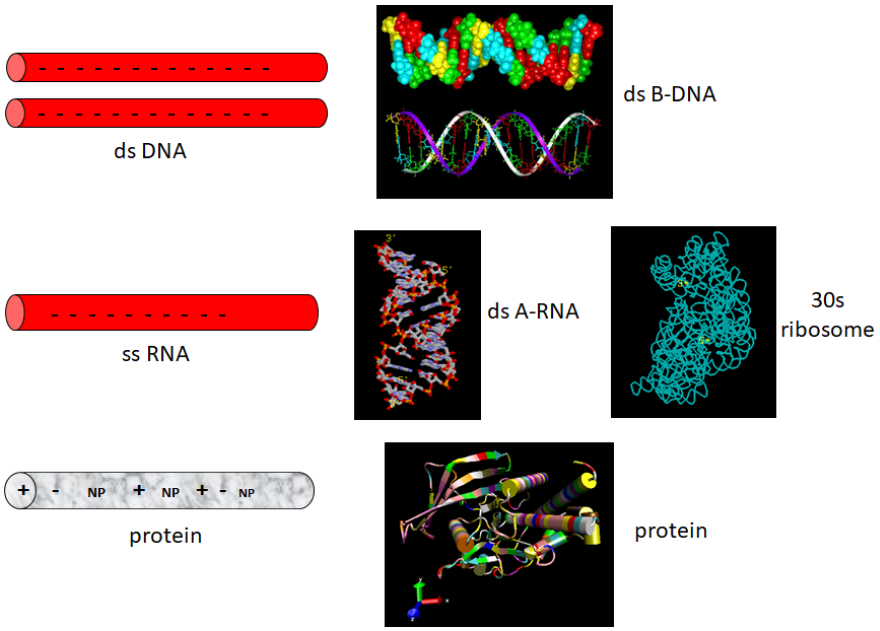


Figure 8.3.13: Summary comparison of DNA, RNA, and Protein structures

A final note on the simplicity of the dsDNA structure. A mutation causing a single base pair change in DNA does **not** change the iconic ds-stranded DNA structure. If it did, DNA would not be a reliable molecule to store and read out the genetic blueprint. In

contrast, a single mutation in the DNA leading to a single amino acid substitution may lead to a protein with altered structure and function. This could be deleterious or even fatal to the organism. On the other hand, the new protein structure might have new functionalities that allow adaptation to new environments or allow new types of reactions. Evolution would favor the latter.

8.3.5: Summary

This chapter examines how subtle chemical variations and modifications in nucleic acids—DNA and RNA—lead to distinct structural and functional outcomes that are essential for life. Designed for junior and senior biochemistry majors, the chapter integrates concepts from chemical modifications and mutation mechanisms to explain why DNA and RNA serve different roles in the cell.

Chemical Modifications and Regulation:
Both DNA and RNA undergo intentional chemical modifications that can alter their physical structure and influence gene expression. In DNA, common modifications such as methylation (and subsequent hydroxymethylation) are central to epigenetic regulation. These modifications can repress transcription and are heritable, impacting cell phenotype over generations. RNA modifications, collectively referred to as the epitranscriptome, similarly affect RNA stability, processing, and translation, highlighting the dynamic regulatory potential of these molecules.

Mutation Mechanisms and DNA Repair:
The chapter discusses how chemical modifications can also lead to mutations. For example, the spontaneous hydrolytic deamination of cytosine to uracil in DNA represents a major source of point mutations, potentially converting GC to AT base pairs if not corrected. DNA repair enzymes, such as uracil-DNA glycosylases, play a critical role in identifying and excising these aberrant bases. Additionally, errors during DNA replication and damage induced by chemical agents (e.g., nitrous acid, alkylating agents) contribute to mutations ranging from single base changes to large-scale chromosomal rearrangements.

Structural Basis for Functional Differences:
A significant portion of the chapter is devoted to understanding why DNA and RNA, despite their chemical similarities, adopt very different structures and fulfill distinct roles:

- **Sugar Chemistry and Backbone Stability:**
  DNA's deoxyribose lacks a 2'-OH group, making it more chemically stable and less prone to self-cleavage compared to RNA, which contains ribose with a reactive 2'-OH. This difference is crucial for DNA's role as a long-term, reliable repository of genetic information, whereas RNA's inherent lability suits its function in transient information transfer and catalysis.
- **Phosphodiester Linkages:**
  Both nucleic acids use phosphodiester bonds to link nucleotides. The inherent chemical resistance of these bonds to nucleophilic attack—enhanced by their negative charge—contributes to the overall stability of the DNA double helix and the more dynamic secondary and tertiary structures found in RNA.
- **Sugar Puckering and Helical Forms:**
  The conformational flexibility of the sugar ring is a key determinant of nucleic acid structure. In DNA, the deoxyribose can adopt different puckers (C2'-endo for B-DNA, C3'-endo for A-DNA), which influences the overall helix geometry. In contrast, the ribose in RNA is more restricted, favoring the C3'-endo conformation, which precludes extended double-stranded helices and promotes the formation of complex tertiary structures.
- **Base Pairing Dynamics:**
  The chapter also explores the equilibrium between Watson-Crick and Hoogsteen base pairing in DNA. Although Watson-Crick pairing predominates in stable B-DNA, dynamic shifts to Hoogsteen pairings can occur—particularly in protein-DNA complexes or damaged regions—affecting recognition and repair mechanisms. In RNA, such alternative pairing is less common due to structural constraints imposed by the ribose.

Comparative Structural Overview:
In a broader context, the chapter compares the structural attributes of DNA, RNA, and proteins. DNA's robust double-helical structure, which is maintained despite mutations, underscores its reliability as a genetic archive. RNA's versatility in folding into complex structures enables it to perform catalytic and regulatory roles, while proteins, with their diverse side chains, are well-suited to function as dynamic catalysts and structural components. This comparison reinforces the evolutionary specialization of these macromolecules for their respective biological roles.

In summary, the chapter provides a comprehensive exploration of how chemical modifications and intrinsic structural features govern the stability, function, and evolutionary utility of DNA and RNA. Understanding these principles is fundamental for

appreciating the molecular basis of gene regulation, mutation, and the overall maintenance of genetic integrity in biological systems.

8.3.6: References

Börner, R., Kowerko, D., Miserachs, H.G., Shaffer, M., and Sigel, R.K.O. (2016) Metal ion induced heterogeneity in RNA folding studied by smFRET. *Coordination Chemistry Reviews* 327 DOI: 10.1016/j.ccr.2016.06.002 Available at: https://www.researchgate.net/publication/303846502_Metal_ion_induced_heterogeneity_in_RNA_folding_studied_by_smFRET

Hardison, R. (2019) B-Form, A-Form, and Z-Form of DNA. Chapter in: R. Hardison's *Working with Molecular Genetics*. Published by LibreTexts. Available at: https://bio.libretexts.org/Bookshelves/Genetics/Book%3A_Working_with_Molecular_Genetics_(Hardison)/Unit_I%3A_Genes%2C_Nucleic_Acids%2C_Genomes_and_Chromosomes/2%3A_Structures_of_Nucleic_Acids/2.5%3A_B-Form%2C_A-Form%2C_and_Z-Form_of_DNA

Lenglet, G., David-Cordonnier, M-H., (2010) DNA-destabilizing agents as an alternative approach for targeting DNA: Mechanisms of action and cellular consequences. *Journal of Nucleic Acids* 2010, Article ID: 290935, DOI: 10.4061/2010/290935 Available at: https://www.hindawi.com/journals/jna/2010/290935/

Mechanobiology Institute (2018) What are chromosomes and chromosome territories? *Produced by the National University of Singapore*. Available at: https://www.mechanobio.info/genome-regulation/what-are-chromosomes-and-chromosome-territories/

National Human Genome Research Institute (2019) The Human Genome Project. *National Institutes of Health*. Available at: https://www.genome.gov/human-genome-project

Wikipedia contributors. (2019, July 8). DNA. In *Wikipedia, The Free Encyclopedia*. Retrieved 02:41, July 22, 2019, from https://en.Wikipedia.org/w/index.php?title=DNA&oldid=905364161

Wikipedia contributors. (2019, July 22). Chromosome. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:18, July 23, 2019, from en.Wikipedia.org/w/index.php?title=Chromosome&oldid=907355235

Wikilectures. Prokaryotic Chromosomes (2017) In MediaWiki, Available at: https://www.wikilectures.eu/w/Prokaryotic_Chromosomes

Wikipedia contributors. (2019, May 15). DNA supercoil. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:40, July 25, 2019, from en.Wikipedia.org/w/index.php?title=DNA_supercoil&oldid=897160342

Wikipedia contributors. (2019, July 23). Histone. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:19, July 26, 2019, from en.Wikipedia.org/w/index.php?title=Histone&oldid=907472227

Wikipedia contributors. (2019, July 17). Nucleosome. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:17, July 26, 2019, from en.Wikipedia.org/w/index.php?title=Nucleosome&oldid=906654745

Wikipedia contributors. (2019, July 26). Human genome. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:12, July 27, 2019, from en.Wikipedia.org/w/index.php?title=Human_genome&oldid=908031878

Wikipedia contributors. (2019, July 19). Gene structure. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:16, July 27, 2019, from en.Wikipedia.org/w/index.php?title=Gene_structure&oldid=906938498

8.4: Chromosomes and Chromatin

Learning Goals (ChatGPT o3-mini, 2/1/25)

1. **Differentiate Chromatin Types:**
   - Compare and contrast heterochromatin and euchromatin in terms of staining properties, spatial distribution within the nucleus, gene activity, and associated proteins.
2. **Understand the Cell Cycle Context:**
   - Describe how chromatin structure changes during the cell cycle—from interphase (with dispersed, transcriptionally active chromatin) to mitosis (when discrete, highly condensed chromosomes are formed).
3. **Describe Chromosome Structure and Organization:**
   - Explain how eukaryotic DNA is organized into chromosomes, including the formation of sister chromatids, centromeres, and the significance of karyotypes in identifying chromosomal abnormalities (e.g., trisomy 21).
4. **Explain DNA Packaging Mechanisms:**
   - Analyze the role of supercoiling, nucleosome formation, and histone proteins in compacting the long DNA molecule into the confined space of the nucleus.
5. **Interpret Higher-Order Chromatin Structure:**
   - Illustrate the hierarchical organization of chromatin—from the nucleosome "beads on a string" to the formation of 30 nm fibers, loop domains, and ultimately chromosome territories within the nucleus.
6. **Examine the Function of Telomeres:**
   - Discuss the structure and function of telomeres, including how telomerase and the end replication problem contribute to chromosome integrity and cellular aging.
7. **Understand Chromatin Dynamics and Gene Regulation:**
   - Evaluate how chromatin looping, topologically associating domains (TADs), and the spatial organization of enhancers, silencers, and promoters influence gene transcription.
8. **Apply Concepts of Phase Separation:**
   - Explore how phase separation and microemulsion-like behavior help segregate transcriptionally active euchromatin from inactive heterochromatin, and the role of RNA and protein interactions in this process.
9. **Contrast Eukaryotic and Prokaryotic DNA Organization:**
   - Compare the linear, histone-packaged chromatin of eukaryotes with the circular, supercoiled DNA of prokaryotes, highlighting the implications for gene regulation and genomic organization.

These goals are intended to guide your understanding of the multiple layers of chromatin organization—from the molecular interactions within nucleosomes to the three-dimensional architecture of the nucleus—and how these structural features relate to essential cellular functions such as gene expression and cell division.

8.4.1: Chromatin

When stained and viewed in a microscope, eukaryotic nuclear DNA in nondividing cells is observed in two different states, **heterochromatin** (dark areas) and euchromatin (light areas), as shown in Figure 8.4.1.

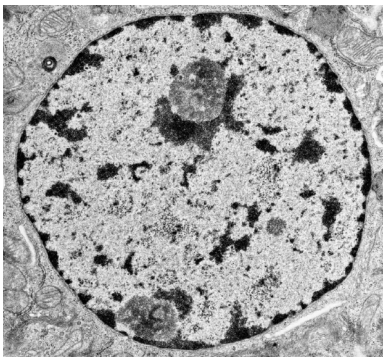


Figure 8.4.1: Nucleus showing heterochromatin (dark) and euchromatin (light). http://medcell.med.yale.edu/histolog...ochromatin.php

The heterochromatin is darkly stained and found along the inner side of the nuclear envelope and inside the nucleus. Euchromatin doesn't stain well since it is more dispersed and is found throughout the nucleus. The image above is of a cell in interphase, a part of the cell cycle in which the cell is in between cell divisions. Here are some differences between heterochromatin and euchromatin:

Heterochromatin:

- contains genes that are transcriptionally inactive and not expressed;
- is not found in prokaryotic cells;
- in humans, one of the X chromosomes is silenced in heterochromatin while the other is transcribable and in euchromatin;
- protects the tightly packed in it from nucleases.

Euchromatin:

- contains about 90% of genome DNA;
- contains chromosomes that are unfolded into nucleosomal DNA resembling beads on a string and is characterized by looser association with packaging histone proteins;
- is the only form of DNA in bacteria;
- is transcriptionally active, so it is open to the binding of RNA polymerase and transcription factors;
- has some genes that can be transcriptionally silenced by moving them into heterochromatin.

Look as long as you wish, but you won't see the iconic pictures of chromosomes in Figure 8.4.1. They are there but dispersed. They are only discretely visible at certain times in the cell cycle when cells are preparing to divide.

Cell Cycle

A short cell cycle review would be helpful for those with a more chemistry-centric background. Somatic (non-germ cells) spend only part of their lives preparing for and dividing into two cells in **mitosis**. Figure 8.4.2 shows a simple model for the cell cycle.

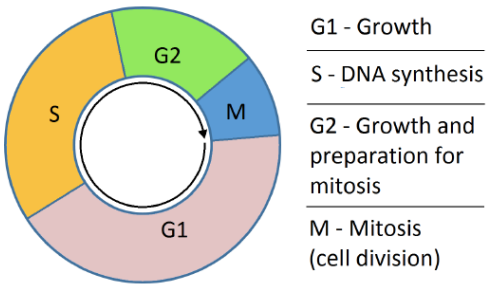
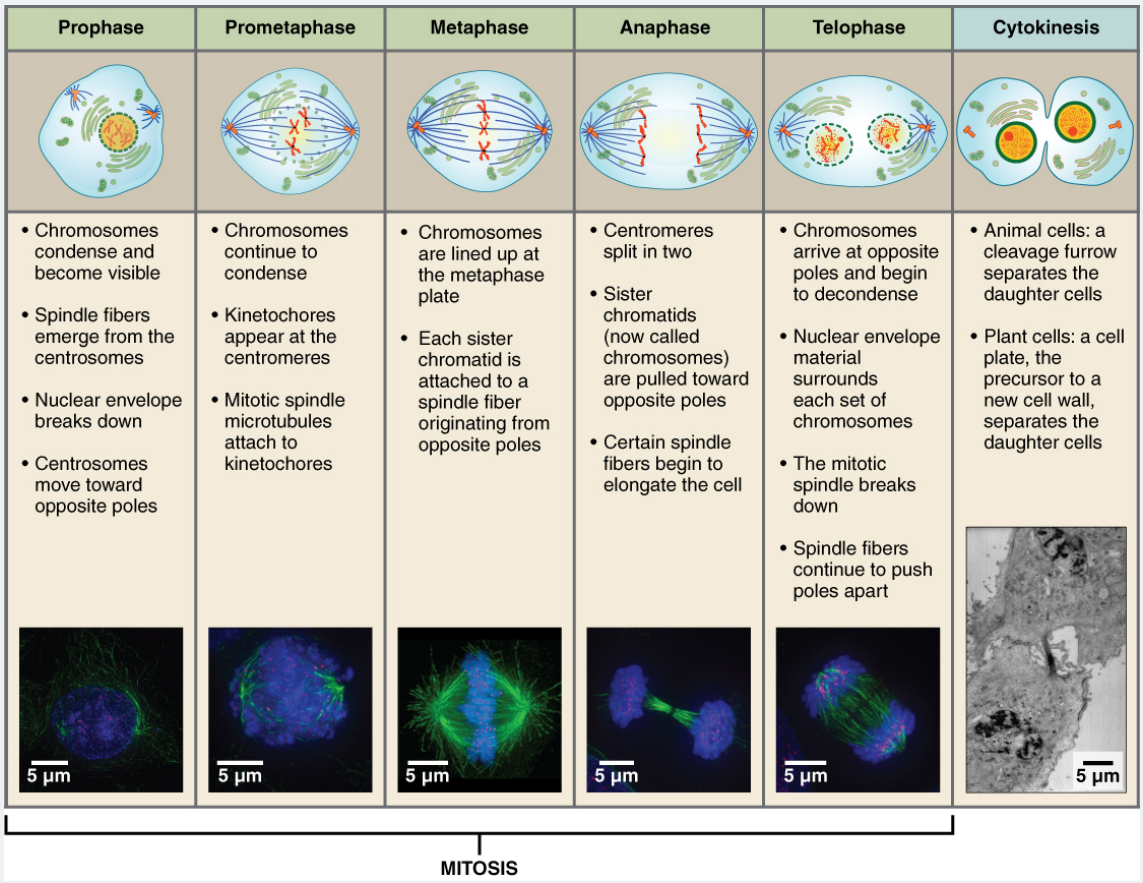


Figure 8.4.2: The Cell Cycle. https://commons.wikimedia.org/wiki/F...cle_simple.png. Creative Commons Attribution-Share Alike 3.0 Unported

Before a cell gets ready to divide, it is in **interphase** (between mitotic phases), encompassing the cell cycle's **G1, S, and G2** phases. In interphase, cells grow and replicate their DNA. **Mitosis** or cell division occurs after G2 and consists of a series

of new discrete phases, including **prophase, metaphase, anaphase, and telophase**, after which the cell divides in a process called cytokinesis. Table 8.4.1 below shows the stages of mitosis and cytokinesis, which occur after interphase.



Let's start with a discussion of the structure of chromosomes as classically observed in mitotic cells. Then, we will look more closely at chromatin's seemingly amorphous and more complicated structures.

8.4.2: Chromosomes

Within eukaryotic cells, DNA is organized into long linear structures called **chromosomes.** A **chromosome** is a "thread-like" structure in the nucleus of animal and plant cells. It consists of a single but long molecule of double-stranded DNA (two ss-DNAs) with a myriad of bound proteins. The proteins bind to and condense the DNA molecule to prevent it from becoming an unmanageable tangle. Before typical cell division (mitosis), these chromosomes are replicated by DNA polymerase to make two identical chromosomes, one for each future daughter cell. The two identical chromosomes, called **sister chromatids**, bind to each other at a common structure called the **centromere**. A replicated chromosome (sister chromatids) bound to each other at the centromere is shown in Figure 8.4.3. Each chromosome in the sister chromatid structure represents one chromatid.

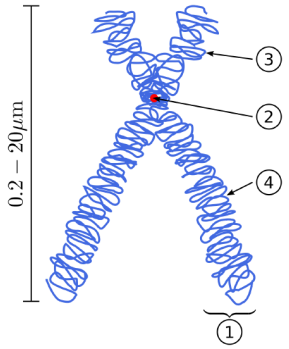


Figure 8.4.3: **Diagram of Replicated and Condensed Eukaryotic Chromosome (sister chromatids).** (1) Chromatid – one of the two identical parts of the chromosome after the S phase. (2) Centromere – the point where the two chromatids are joined together. (3) The short arm is termed *p*; the Long arm is termed *q*. *Image* by: Magnus Manske, Dietzel65, and Tryphon

Each cell normally contains 23 pairs of chromosomes in humans, for a total of 46. Each pair member is similar in twenty-two of these pairs, called *autosomes*. The 23rd pair, the sex chromosomes, differ between males and females. Females have two copies of the X chromosome, while males have one X and one Y chromosome. Figure 8.4.4 shows a DNA **karyotype** with 22 pairs of autosomes and one pair of sex chromosomes. Karyotypes are prepared using standardized staining procedures that reveal characteristic structural features for each chromosome, usually from white blood cells. The karyotype of human cells shown in Figure 4 below shows an extra copy of chromosome 21 (trisomy 21), which causes Down's Syndrome.

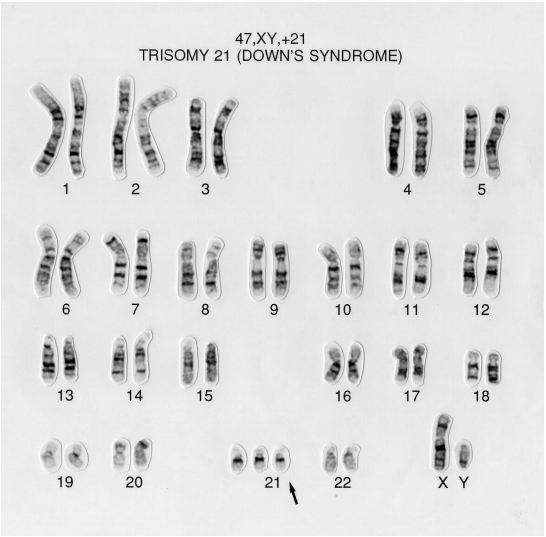


Figure 8.4.4: Karyotype of the human chromosome from a human male with Down's Syndrome. https://wellcomecollection.org/works/wmcdanw6. **Creative Commons Attribution 4.0 International License**

Each species has a unique chromosomal complement. For example, chickens have 39 pairs of chromosomes (78 total) with 38 autosomal pairs and one pair of sex chromosomes (Z and W). ZW chickens are female, and ZZ chickens are male. The total length of the human genome is over 3 billion base pairs. The total length of the human genome is over 3 billion base pairs. The genome also includes mitochondrial DNA.

Eukaryotic organisms (animals, plants, fungi, and protists) store most of their DNA inside the cell nucleus as linear nuclear DNA. However, DNA in the mitochondria and chloroplast are circular (with no ends). Bacteria and Archaea cells do not have organelle structures and, thus, store their DNA only in a region of the cytoplasm known as the *nucleoid region*. Prokaryotic chromosomes consist of *double-stranded circular DNA*.

8.4.3: Packing of DNA: Supercoiling

The genome of a cell is often significantly larger than the cell itself. For example, if the DNA from a human cell containing 46 chromosomes were stretched out in a line, it would extend more than 6 feet (2 meters)! How is it possible that the genetic information not only fits into the cell but fits into the cell nucleus? Eukaryotes solve this problem by combining supercoiling and packaging DNA around the histone family of proteins (described below). Prokaryotes do not contain histones (with a few

exceptions). Prokaryotes tend to compress their DNA using **nucleoid-associated proteins (NAPs)** and **supercoiling,** as shown in Figure 8.4.5.

DNA supercoiling refers to the over- or under-winding of a DNA strand and is an expression of the strain on that strand. Supercoiling is important in some biological processes, such as compacting DNA and regulating access to the genetic code. DNA supercoiling strongly affects DNA metabolism and possibly gene expression. Additionally, certain enzymes, such as **topoisomerases,** can change DNA topology to facilitate DNA replication or transcription functions.

In a "relaxed" double-helical segment of B-DNA, the two strands twist around the helical axis once every 10.4–10.5 base pairs of sequence. Adding or subtracting twists, as some enzymes can do, imposes strain. If a DNA segment under twist strain were closed into a circle by joining its two ends and then allowed to move freely, the circular DNA would contort into a new shape, such as a simple figure-eight, as shown in Figure 8.4.5. Such a contortion is a *supercoil*. The noun form "supercoil" is often used in the context of DNA topology.

Figure 8.4.5: DNA Supercoiling. The supercoiled structure of linear DNA molecules with constrained ends. The helical nature of the DNA duplex is omitted for clarity. I*mage by: Richard Wheeler*

Positively supercoiled (overwound) DNA is transiently generated during DNA replication and transcription and, if not promptly relaxed, inhibits (regulates) these processes. The simple figure eight is the simplest supercoil and is the shape a circular DNA assumes to accommodate one too many or one too few helical twists. The two lobes of the figure eight will appear rotated either clockwise or counterclockwise with respect to one another, depending on whether the helix is over- or underwound. For each additional helical twist being accommodated, the lobes will show one more rotation about their axis. As a general rule, the DNA of most organisms is negatively supercoiled.

Lobal contortions of circular DNA, such as the rotation of the figure-eight lobes above, are called *writhe*. The above example illustrates that twisting and writhing are interconvertible. Supercoiling can be represented mathematically by the sum of *twist* and *writhe*. The *twist* is the number of helical turns in the DNA, and the *writhe* is the number of times the double helix crosses over on itself (these are the supercoils). Extra helical twists are positive and lead to positive supercoiling, while subtractive twisting causes negative supercoiling. Many topoisomerase enzymes sense supercoiling and either generate or dissipate it as they change DNA topology.

Because chromosomes may be very large, segments in the middle may act as if their ends are anchored. As a result, they may be unable to distribute excess twist to the rest of the chromosome or to absorb twist to recover from underwinding—the segments may become *supercoiled*, in other words. In response to supercoiling, they will assume an amount of writhe, just as if their ends were joined.

Supercoiled circular DNA forms two major structures: a *plectoneme* or a *toroid*, or a combination of both. A negatively supercoiled DNA molecule will produce either a one-start left-handed helix, *the toroid*, or a two-start right-handed helix with terminal loops, the *plectoneme*. *Plectonemes* are typically more common, and this is the shape most bacterial plasmids will take. For larger molecules, it is common for hybrid structures to form – a loop on a toroid can extend into a plectoneme, as shown in Figure 8.4.6. DNA supercoiling is important for DNA packaging within all cells and also plays a role in gene expression.

Figure 8.4.6: Bacterial DNA Supercoiling. Atomic force microscopy (AFM) visualization of torsionally relaxed (A), and negatively supercoiled (B) bacterial plasmids pBR322. (C) Electron microscopy image of the E. coli chromosomal DNA displaying a hybrid toroidal-plectoneme structure. Image A and B from Witz, G. and Stasiak, A. (2009) Nucleic Acids Research 38(7):2119-2133. Image C from Prokaryotic Chromosomes

In addition to forming supercoiled structures, circular chromosomes from bacteria have been shown to undergo the processes of *catenation* and *knotting* upon the inhibition of topoisomerase enzymes. *Catenation* is the process by which two circular DNA strands are linked together like chain links, whereas *DNA knotting* is the interlooping structures occurring within a single circular DNA structure. These are illustrated in Figure 8.4.7. *In vivo*, the action of topoisomerase enzymes is critical to keep knots and catenoids from tangling the DNA structure.

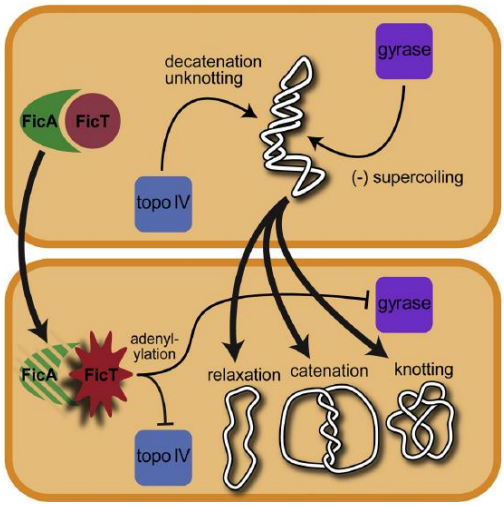


Figure 8.4.7: DNA Catenation and Knotting. The upper structure shows the negatively supercoiled form of bacterial DNA. The inhibition of topoisomerase enzyme activity leads to the relaxation, catenation, and knotting of the chromosomal structure. *Harms, A. et al. (2015) Cell Reports 12(9):1497-1507*.

Mitochondrial and Chloroplast DNA

Mitochondrial and Chloroplast DNA are circular, suggesting a bacterial origin for both organelle structures. Sequence alignments further support the *endosymbiotic theory*, which proposes that early "eukaryotic" organisms engulfed bacteria and subsequently became symbiotic to their eukaryotic counterpart rather than being digested.

In the cells of eukaryotic organisms, the vast majority of the proteins present in the mitochondria (numbering approximately 1500 different types in mammals) are coded for by nuclear DNA. However, sequencing of the human mitochondrial genome has revealed 16,569 base pairs encoding 13 proteins, as shown in Figure 8.4.8. Many of the mitochondrially produced proteins are required for electron transport during the production of ATP.

Figure 8.4.8: Mitochondrial Genome. Mitochondria are organelle structures containing a double membrane, thought to have originated as an independent prokaryotic organism that was originally engulfed by a eukaryotic organism, where it became a symbiotic counterpart. Mitochondria contain circular chromosomal DNA that shares high sequence similarity with alphaprotobacteria. The human mitochondrial genome contains 16,569 base pairs encoding 13 proteins and ribosomal RNA (rRNA) components. *Images adapted from The National Human Genome Research Institute and Shanel, Knopfkind, and JHC*

8.4.4: Histones and Nucleosomes

Within eukaryotic chromosomes, chromatin proteins, known as **histones**, compact and organize DNA. These compacting structures guide the interactions between DNA and other proteins, helping control which parts of the DNA are transcribed.

Histones are highly basic proteins in eukaryotic cell nuclei that package and order the DNA into structural units called **nucleosomes**. They are the chief protein components of chromatin, acting as spools around which DNA winds and playing a role in gene regulation. Without histones, the unwound DNA in chromosomes would be very long (a length-to-width ratio of more than 10 million to 1 in human DNA). For example, each human diploid cell (containing 23 pairs of chromosomes) has about 1.8 meters of DNA wound on the histones, and the diploid cell has about 90 micrometers (0.09 mm) of chromatin.

There are five major families of histones: H1/H5, H2A, H2B, H3, and H4. Histones H2A, H2B, H3, and H4 are the core histones, while H1/H5 are the linker histones.

The core histones all exist as dimers, which are similar in that they all possess the histone fold domain: three alpha helices linked by two loops (Figure 4.13). This helical structure allows for interaction between distinct dimers, particularly in a head-tail fashion (also called the handshake motif). The resulting four distinct dimers then come together to form one octameric nucleosome core, approximately 63 Angstroms in diameter. Around 146 base pairs (bp) of DNA wrap around this core particle 1.65 times in a left-handed super-helical turn to give a particle of around 100 Angstroms across, called a nucleosome as illustrated in Figure 8.4.9.

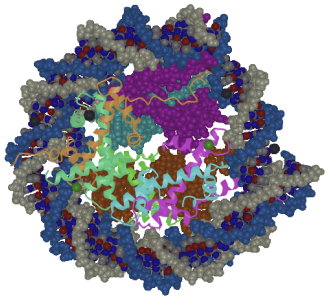
Figure 8.4.9: Nucleosome Core Structure. Histones H2A and H2B dimerize, and Histones H3 and H4 dimerize. Two dimers of each join to form a histone core octamer. The DNA double helix winds 1.65 times around the octamer core, forming the nucleosome structure. *image adapted from Nucleosome Structure*

The linker histone H1 binds the *nucleosome* at the entry and exit sites of the DNA, thus locking the DNA into place and allowing the formation of higher-order structures (Figure 4.14). The 10 nm fiber or beads on a string conformation is the most basic formation. This involves wrapping DNA around nucleosomes with approximately 50 base pairs of DNA separating each pair of nucleosomes (also referred to as linker DNA).

The nucleosome contains over 120 direct protein-DNA interactions and several hundred water-mediated ones. Direct protein–DNA interactions are not spread evenly about the octamer surface but rather located at discrete sites. These are due to the formation of two DNA binding sites within the octamer: the α1α1 site, which uses the α1 helix from two adjacent histones, and the L1L2 site formed by the L1 and L2 loops. Salt links and hydrogen bonding between both side-chain basic and hydroxyl groups and main-chain amides with the DNA backbone phosphates form the bulk of interactions with the DNA. This is important, given that the ubiquitous distribution of nucleosomes along genomes requires it to be a non-sequence-specific DNA-binding factor. Although nucleosomes tend to prefer some DNA sequences over others, they can bind practically to any sequence, which is thought to be due to the flexibility in the formation of these water-mediated interactions. In addition, non-polar interactions are made between protein side-chains and the deoxyribose groups, and an arginine side-chain intercalates into the DNA minor groove at all 14 sites where it faces the octamer surface. The spatial distribution and strength of DNA-binding sites about the octamer surface distort the DNA within the nucleosome core. The DNA is non-uniformly bent and also contains twist defects. The twist of free B-form DNA in solution is 10.5 bp per turn. However, the overall twist of nucleosomal DNA is only 10.2 bp per turn, varying from 9.4 to 10.9 bp per turn.

The histone tail extensions constitute up to 30% by mass of histones but are not visible in the crystal structures of nucleosomes due to their high intrinsic flexibility and have been thought to be largely unstructured (Figure 4.14). The N-terminal tails of histones H3 and H2B pass through a channel formed by the minor grooves of the two DNA strands, protruding from the DNA every 20 bp. The N-terminal tail of histone H4, on the other hand, has a region of highly basic amino acids (16-25), which, in the crystal structure, forms an interaction with the highly acidic surface region of an H2A-H2B dimer of another nucleosome, being potentially relevant for the higher-order structure of nucleosomes. This interaction is thought to occur under physiological conditions and suggests that acetylation of the H4 tail distorts the higher-order chromatin structure.

Figure 8.4.10 shows an interactive iCn3D model of the human nucleosome (3afa). The histones are shown as cartoons. H2A is red, H2B orange, H3 is magenta, and H4 is cyan. There are two copies of each histone. In each pair, one is shown in spacefill and one in cartoon. The DNA backbone are shown in blue and gray, and the bases in cpk colors.



NCBI iCn3D Figure 8.4.10: Human nucleosome (3afa). (Copyright; author via source).
Click the image for a popup or use this external link: https://www.ncbi.nlm.nih.gov/Structu...3e95a06119eb93

The packing of DNA from dsDNA to the metaphase chromosomes is schematically shown in Figure 8.4.11. The DNA double helix's formation represents the chromosome structure's first-order packaging. The formation of nucleosomes represents the second level of packaging for eukaryotic chromosomes. *In vitro* data suggests that nucleosomes are then arranged into either a solenoid structure, which consists of 6 nucleosomes linked together by the Histone H1 linker proteins, or a zigzag structure similar to the solenoid construct. Both the solenoid and zigzag structures are approximately 30 nm in diameter. The solenoid and zigzag structures reported from *in vitro* data have not yet been confirmed to occur *in vivo.*

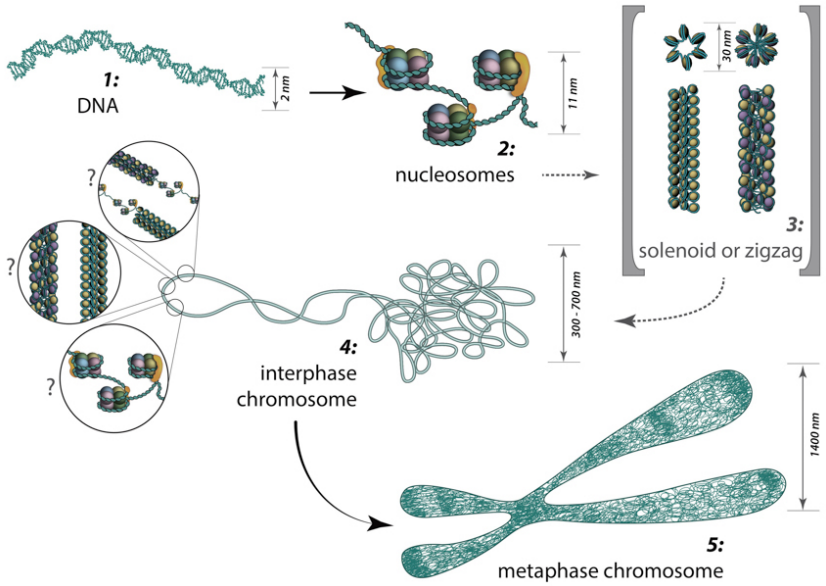


Figure 8.4.11: Chromosome Structure. (1) DNA double helix is approximately 2 nm in diameter. (2) The nucleosome core structure is approximately 11 nm in diameter. (3) The solenoid/zigzag structure is approximately 30 nm in diameter and is proposed to form chromosome loops (4) during cellular interphase and more condensed chromosome territories (5) during mitosis. *Image by MBInfo*

8.4.5: Telomeres

At the ends of the linear eukaryotic chromosomes are specialized regions of DNA called telomeres. The main function of these regions is to allow the cell to replicate chromosome ends using the enzyme telomerase, as the enzymes that normally replicate DNA cannot copy the extreme 3′ ends of chromosomes. Figure 8.4.12 shows a cartoon showing telomeres and their extension.

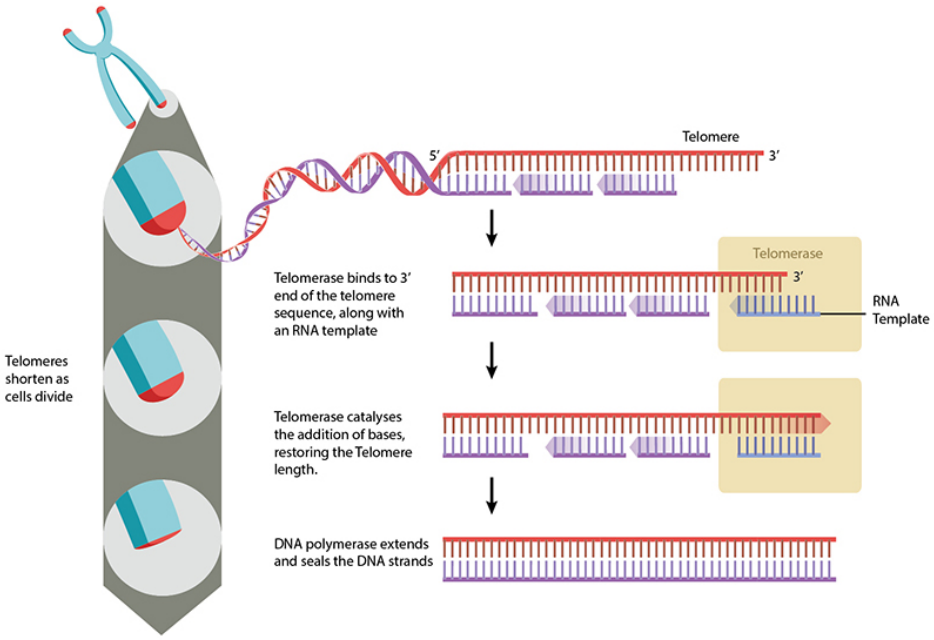


Figure 8.4.12: Telomer structure and extension. https://www.mechanobio.info/genome-r...are-telomeres/. Creative Commons Attribution-NonCommercial 4.0 International License.

These specialized chromosome caps also help protect the DNA ends and stop the DNA repair systems in the cell from treating them as damage to be corrected.

In human cells, telomeres contain 300-8000 repeats of a simple **TTAGGG** sequence. The repetitive **TTAGGG** sequences in telomeric DNA can form unique higher-order structures called **quadruplexes**. Figure 8.4.13 shows an interactive iCn3D model of

parallel quadruplexes from human telomeric DNA (1KF1). The structure contains a single DNA strand (5'-AGGGTTAGGGTTAGGGTTAGGG-3') which contains four TTAGGG repeats.

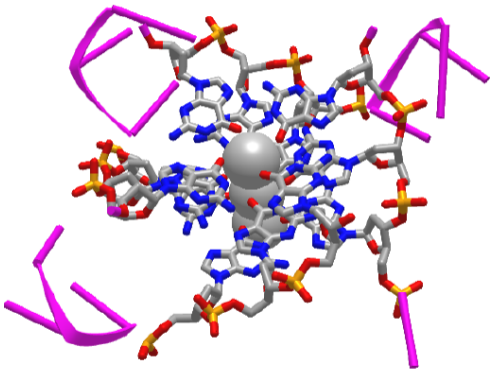


Figure 8.4.13: A buried phenylalanine in low molecular weight protein tyrosyl phosphatase (1xww) (Copyright; author via source).

Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...y5joFHDgWJQsQ6

Rotate the model to see three parallel layers of quadruplexes. In each layer, four noncontiguous guanine bases interact with a K^+ ion. Hover over the guanine bases in one layer, and you will find that one layer consists of guanines 4, 10, 16, and 22, which derive from the last G in each of the repeats in the sequence of the oligomer used (5'-AGGGTTAGGGTTAGGGTTAGGG-3'). These quadruplexes certainly serve as recognition and binding sites for telomerase proteins. The guanine-rich telomere sequences, which can form quadruplexes, may also function to stabilize chromosome ends

The double-stranded DNA is unwound during DNA replication, and DNA polymerase synthesizes new strands. However, as DNA polymerase moves in a unidirectional manner (from 5' to 3'), only the **leading strand** can be replicated continuously. For the complementary **lagging strand**, DNA replication is discontinuous. In humans, small RNA primers attach to the lagging strand DNA, and the DNA is synthesized in small 5'-3' stretches of about 100-200 nucleotides, which are termed **Okazaki fragments.** The RNA primers are removed and replaced with DNA, and the Okazaki DNA fragments are ligated together. At the end of the lagging strand, it is impossible to attach an RNA primer, meaning that a small amount of DNA will be lost each time the cell divides. This 'end replication problem' has serious consequences for the cell as the DNA sequence cannot be replicated correctly, with the loss of genetic information. Hence, most telomeres have 3´ overhangs. Bacteria DNA, which is circular, does not have the problem.

To prevent this, telomeres are repeated hundreds to thousands of times at the end of the chromosomes. Each time cell division occurs, a small section of telomeric sequences is lost to the end replication problem, thereby protecting the genetic information. At some point, the telomeres become critically short. This decrease leads to cell senescence, where the cell cannot divide, or apoptotic cell death. Telomeres are the basis for the **Hayflick** limit, the number of times a cell can divide before reaching senescence.

Telomeres can be restored by the enzyme **telomerase**, which extends the telomere's length. Telomerase activity is found in cells that undergo regular division, such as stem cells and lymphocyte cells of the immune system. The enzyme has two major subunits. One is the **telomerase reverse transcriptase (TERT)** catalytic enzyme, an RNA-dependent DNA polymerase. Almost all DNA polymerases use DNA as a template for replication. Some RNA viruses that use RNA as their genetic information, like HIV, encode their own **reverse transcriptase**, which directs the polymerization of a DNA copy of the viral RNA genome as part of its life cycle.)The virus that causes the Covid-19 pandemic is called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It is also an RNA virus, but unlike HIV, it uses and encodes an RNA-dependent RNA polymerase.)

The second subunit of telomerase is the **telomerase RNA (TR)**, which contains a template from which the new telomer is made. The enzyme makes many DNA copies from this to create a multitude of DNA repeats in the telomeres. Figure 8.4.14 shows the structure of the telomerase RNA used to build new telomers. Ovals show proteins that form part of the complex. 'For this discussion, the most important is TERT, telomerase reverse transcriptase, the RNA-dependent DNA polymerase which synthesizes new telomeric DNA from the template sequence of the RNA.

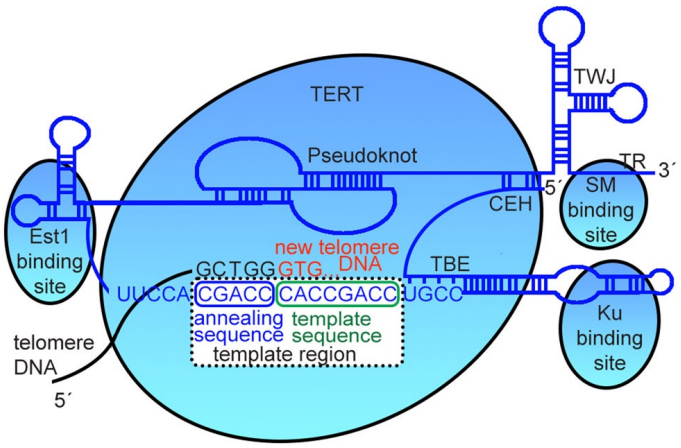


Figure 8.4.14: Telomerase RNA organization for the yeast Lachancea sp. The most relevant binding sites for protein-RNA interactions are indicated by blue ellipses (Est1, SM, and KU), and the bigger one is the contact region with the telomerase reverse transcriptase (TERT). TBE template boundary element, TWJ three-way junction, CEH core enclosed helix. Adapted from Waldl et al. The template region is located between the EST1 binding site and Ku binding hairpins. Peska, V *et al. Sci Rep* **11,** 12784 (2021).Creative Commons Attribution 4.0 International License. http://creativecommons.org/licenses/by/4.0/.

Trace the black single-stranded telomeric DNA strand from the 5' end (bottom left) to its 3' end containing the terminal sequence **5'GCTGG3'**. Now start tracing the telomerase RNA starting with its 5' end, between the CEH and SM binding site in Figure 8.4.14 It bends sharply near TBE and continues to the right through the Ku binding site, forming stem-loops. It then continues into the template region containing a template sequence (3'CACCGACC5'), from which new telomeric DNA is made (5'GTGGCTGG3' - a slightly different repeat than the human repeat), and an annealing sequence (3'CGACC5'), which is complementary to the 3' end of the existing telomeric DNA (5'GCTGG3'). The RNA continues in the 3' direction through the Est1 bind site region, through the pseudoknot, and ends at the 3' end past the SM binding site.

Telomeres can also be extended through the Alternative Lengthening of Telomeres (ALT) pathway. In this case, telomeres are switched between chromosomes by homologous recombination rather than being extended. As a result of the telomere swap, one set of daughter cells will have shorter telomeres, and the other set will have longer telomeres.

A downside to telomere extension is the potential for uncontrolled cell division and cancer. Abnormally high telomerase activity has been found in most cancer cells, and non-telomerase tumors often exhibit ALT pathway activation. In addition to the potential for losing genetic information, cells with short telomeres are at a higher risk for improper chromosome recombination, leading to genetic instability and aneuploidy (an abnormal number of chromosomes).

8.4.6: Chromatin Structure

During interphase, distinct chromosomes as shown in Figure 8.4.4 are not observed. Rather, each chromosome occupies a spatially limited, roughly elliptical domain known as a **chromosome territory (CT)**. Each chromosome territory is comprised of higher-order chromatin units of ~1 Mb each. These units are likely built from smaller loop domains containing the solenoid/zigzag structural motifs. On the other hand, 1Mb domains can themselves serve as smaller units in higher-order chromatin structures. With the development of high-throughput biochemical techniques, such as 3C (chromosome conformation capture) and 4C (chromosome conformation capture-on-chip and circular chromosome conformation capture), numerous spatial interactions between neighboring chromatin territories have been described as shown in Figure 8.4.15.

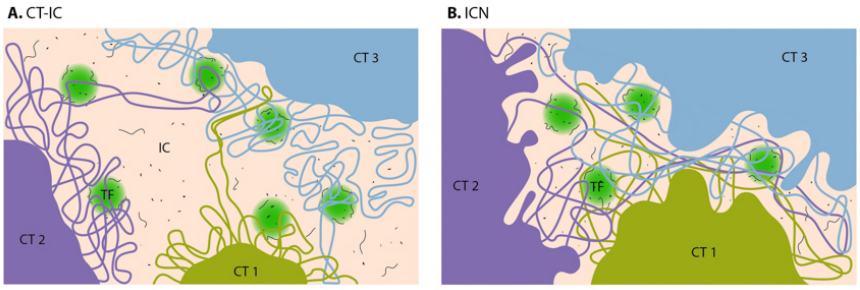


Figure 8.4.15: Computer Models of Chromosome Territory (CT) Structure. In the CT-IC model, the space between discrete CTs can be visualized in light and electron microscope and is called an interchromatin compartment (IC). Transcription factories (TF, green color) are localized predominantly in the perichromatin region. In the ICN model, the interchromatin compartment is not apparent. Instead, intermingling decondensed chromatin loops occupy the space between CTs, which often share the same transcription factories. *Image by MBInfo*

Chromosome territories are known to be arranged radially around the nucleus. This arrangement is both cell and tissue-type specific and is also evolutionarily conserved. The radial organization of chromosome territories correlated with their gene density and size. In this case, the gene-rich chromosomes occupy interior positions, whereas larger, gene-poor chromosomes tend to be around the periphery. Chromosome territories are also dynamic structures, with genes able to relocate from the periphery towards the interior once they have been 'switched on'. In other cases, genes may move in the opposite direction or simply maintain their position.

Figure 8.4.16 shows nano- to more micro-folding structures of chromosomes in the nucleus. The top left shows the most zoomed-in view, where DNA is wound around histone complexes (nucleosomes), which condense to form 10 nm fibers. These condense further into **Topologically Associated Domains (TAD), which** separate into **Compartments A and B**. These then pack into discrete **territories**. Individual chromosomes occupy their own chromosome territories in the nucleus.

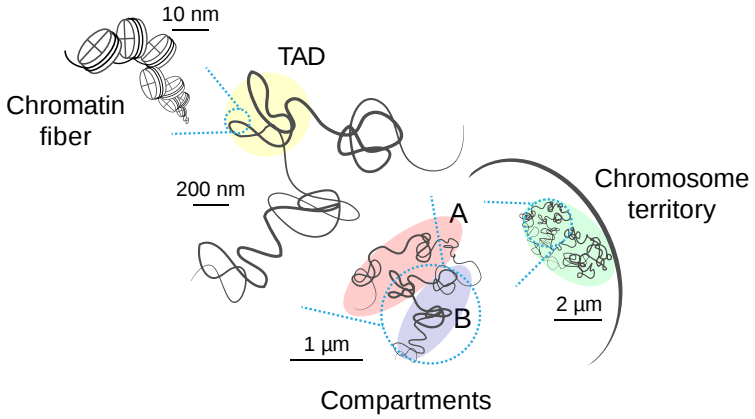


Figure 8.4.16: Schematic view of chromosome folding inside the nucleus Szabo Q et al. Sci Adv. 2019 Apr 10;5(4):eaaw1668. doi: 10.1126/sciadv.aaw1668. PMID: 30989119; PMCID: PMC6457944.Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC)

This may seem very complex, and it is. Still, it is somewhat analogous to protein folding, which starts with a linear primary sequence and moves into more complicated secondary structures, secondary structure motifs, domains, tertiary structures, and quaternary structures, which display varying degrees of symmetry.

Because these different types of folding and compaction of chromosomes are difficult to visualize, we will present several different representations of these nano- to microstructures to help your understanding. Figure 8.4.17 shows scaling factors and differential structures of chromosomes and chromatin.

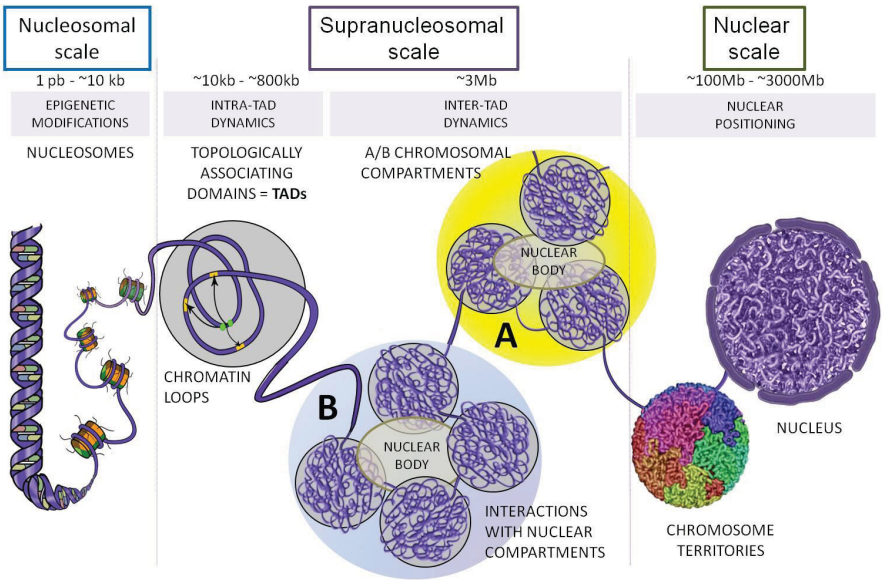


Figure 8.4.17: Schematic representation of genome organization in mammals. a, V.; Baudement, M.-O.; Lesne, A.; Forné, T. Contribution of Topological Domains and Loop Formation to 3D Chromatin Organization. *Genes* **2015**, *6*, 734-750. https://doi.org/10.3390/genes6030734. Creative Commons Attribution license (**http://creativecommons.org/licenses/by/4.0/**).

The large-scale A compartment is gene-rich and actively transcribed, best described as euchromatin. In contrast, B compartments are gene-poor and best represent heterochromatin. At the subscale level, boundaries between TADs are transcriptionally rich and can be separated from other TADS by heterochromatin "islands."

Another representation of chromatin organization that emphasizes TADs is shown in Figure 8.4.18

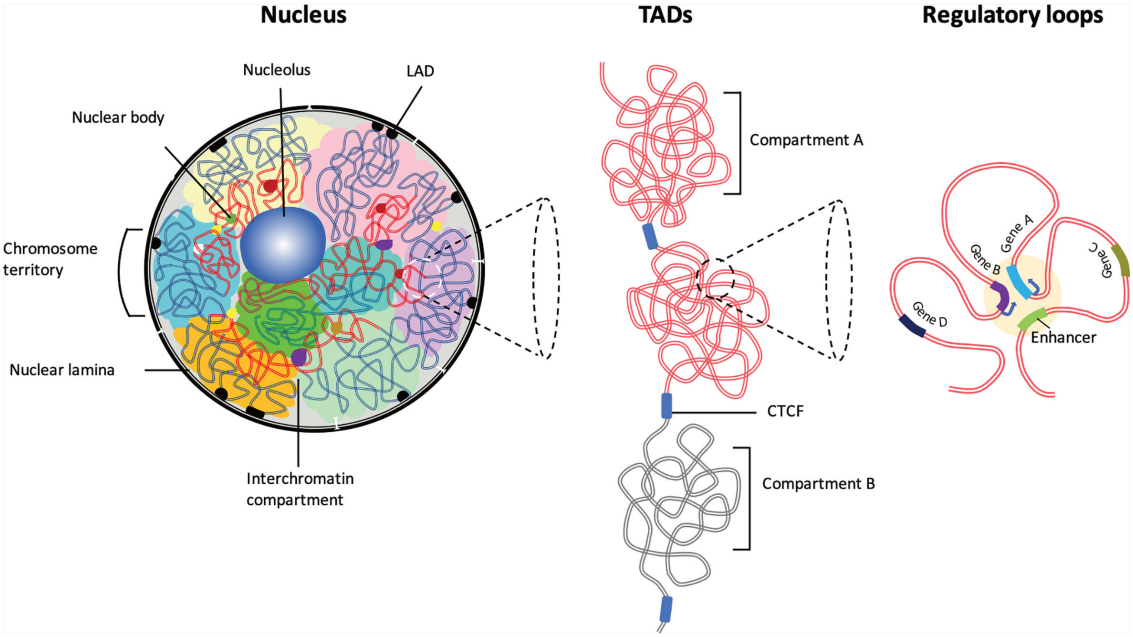


Figure 8.4.18: The hierarchical organization of the 3D chromatin inside the interphase nucleus. Chromatin interactions occur predominantly between compartments with similar biochemical or functional properties. The majority of the chromatin interactions are intra-chromosomal. Preferential self-interactions within the heterochromatin and euchromatin (A and B compartments) regions result in the formation of topologically associating domains (TADs), demarcated by boundary elements enriched with CTCF/Cohesin proteins. Jagan M. R. Pongubala and Cornelis Murre. Front. Immunol., 29 March 2021 | https://doi.org/10.3389/fimmu.2021.633825. Creative Commons Attribution License (CC BY).

The TAD boundaries in Figure 8.4.18 show a blue CTCF between compartments, with each TAD consisting of many interacting loops. Figure 8.4.19 shows a more detailed view.

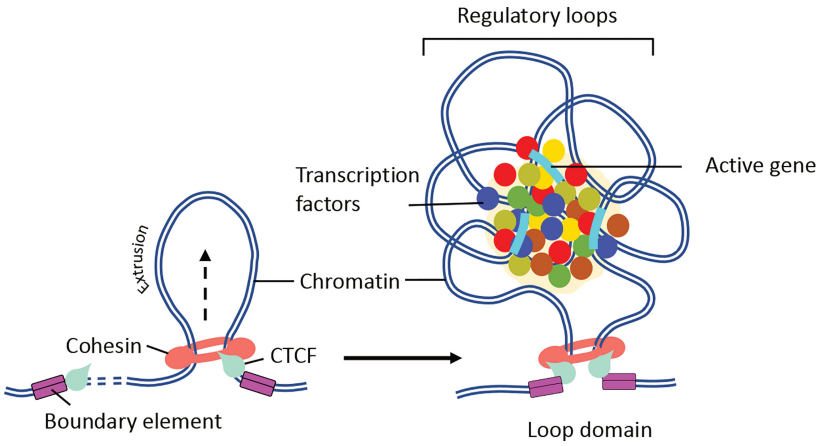


Figure 8.4.19: Jagan M. R. Pongubala and Cornelis Murre. Front. Immunol., 29 March 2021 | https://doi.org/10.3389/fimmu.2021.633825. Creative Commons Attribution License (CC BY).

A protein complex containing cohesin (which forms a ring) and CTCF is found at sites where loops of relationally transcribable DNA are extruded through the ring. The extruded loops interact with others to form a cluster of regulatory loops containing genes with similar potential for transcription (activated if they end up in Compartment A or repressed if in Compartment B). Cohesin is a functional complex that forms a ring that traps sister chromatids. During anaphase, the complex is cleaved, and the sister chromatids separate. CTCF is a chromatin-binding factor with many associated activities.

Mechanism of chromatin loop formation: TADs contain varying numbers of chromatin loops generated through loop extrusion by CTCF/cohesin
complexes. (Right panel): In the presence of NIPBL and MAU2, the cohesin complex is loaded onto the DNA. Then, cohesin extrudes chromatin until a pair of convergent CTCF binding sites is reached. (Right panel) The N-terminus of CTCF and convergent positioning of the CTCF-DNA complex stabilizes cohesin binding and stalls chromatin extrusion, leading to the establishment of higher-order chromatin organization. The intervening DNA between two convergent CTCF sites leads to the formation of a loop domain, which adopts various complex shapes comprised of multiple regulatory loops. The internal structure of the loop domain is likely determined by polymer chromatin-chromatin self-interactions, which may be further stabilized by phase separation. The contacts within the loop domains facilitate the targeting of enhancers to specific genes (104). The black arrow depicts the direction of loop extrusion."

Special elements in the DNA called **enhancers** and **silencers** of gene transcription have long been known to influence gene transcription. These are cis sequences (i.e., on the same molecule of DNA, not trans factors like separate proteins that bind to promoters) and can be quite distant from the proximal promoter sequence, which controls the transcription of a target gene. How might they work from such a large distance from the promoter? One obvious answer is the DNA folds in 3D space, so the enhancers and silencers are close to each other in 3D space. Chromatin loop formation facilitates such promoter and enhancer/silencer interactions. Evidence suggests that specific enhancers and silencers are housed in loops in specific TADs, limiting their effects to a subset of genes. Enhancers and promoters seem to interact only within TADs, which suggests that TADS are a fundamental folding "domain" based on gene regulation. If boundaries between different TADS were removed, promoters and enhancers in one TAD might affect transcription in other TADS, leading to aberrant gene expression. Chromatin loop formation facilitates interactions between promoter and enhancer/silencer elements. Figure 8.4.20 shows how enhancers and silencers can be brought close in space to promoters and their genes through TAD formation.

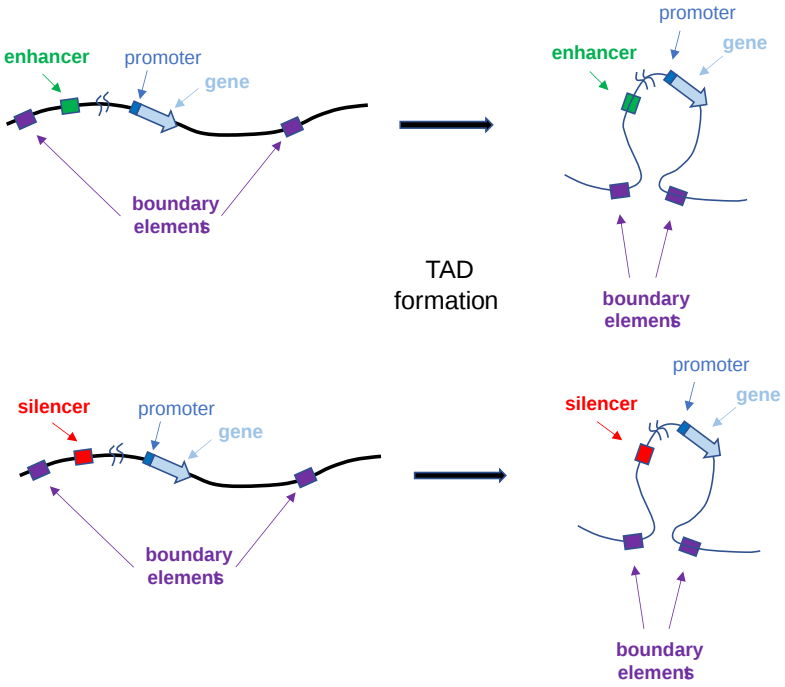


Figure 8.4.20: TADs bring enhancers and silencers close to promoters and their genes

TADs appear highly conserved in mammals and comprise most (90%) of the genome. The median size is about 880 Kb. The Boundaries between TADs have CCCTC-binding factor (CTCF) and the structural maintenance of chromosomes (SMC) cohesin complex. In Drosophila, TADs are organized by epigenetic state (methylation, condensation). For example, some TADS are transcriptionally active with epigenetic histone modifications (for example, trimethylation of histone H3 lysines 4 and 36 or H3Kme3 and H3K36me3) that activate transcription). Others are transcriptionally repressed (enriched in H3K27me3 and containing Polycomb group (PcG) proteins), and some are more classically representative of heterochromatin.

8.4.7: Mechanisms of Separation of Euchromatin and Heterochromatin

What interactions stabilize TADs, compartments, heterochromatin, and euchromatin from a chemistry focus? The easiest way to conceptualize their separation in the nucleus is to use the idea of phase separations. Figure 8.4.21 shows a cartoon representation of the separation of heterochromatin and euchromatin in the zebrafish embryo. The euchromatin is shown as more centrally located and dispersed, with red dots indicating transcriptionally active sites. In the late blastula stage, gene expression dramatically increase,s and the nucleolus and heterochromatin are not seen.

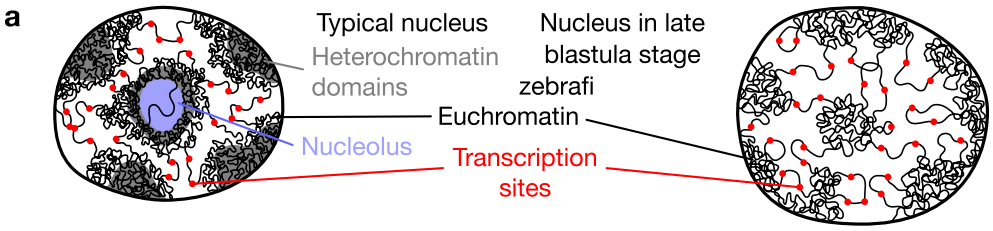


Figure 8.4.21: Sketch of nuclear compartmentalization in a typical nucleus and the nucleus of a late blastula (sphere) stage zebrafish embryo. Hilbert, L., Sato, Y., Kuznetsova, K. *et al.* Transcription organizes euchromatin via microphase separation. *Nat Commun* **12,** 1360 (2021). https://doi.org/10.1038/s41467-021-21589-3. Creative Commons Attribution 4.0 International License. http://creativecommons.org/licenses/by/4.0/.

Analyses show that transcription forms regions enriched in RNA, RNA binding proteins, and, accordingly, transcriptionally active chromatin separated from transcriptionally inactive heterochromatin. Given the dispersion of the DNA required for transcription, the regions enriched in RNA are also depleted in "stainable" DNA. Micrographs showing the depletion of DNA in the actively

transcribed regions are shown in Figure 8.4.22 Note the clear lack of DNA in the white rectangles, which contain high levels of RNA and RNA polymerase II (as a proxy for protein).

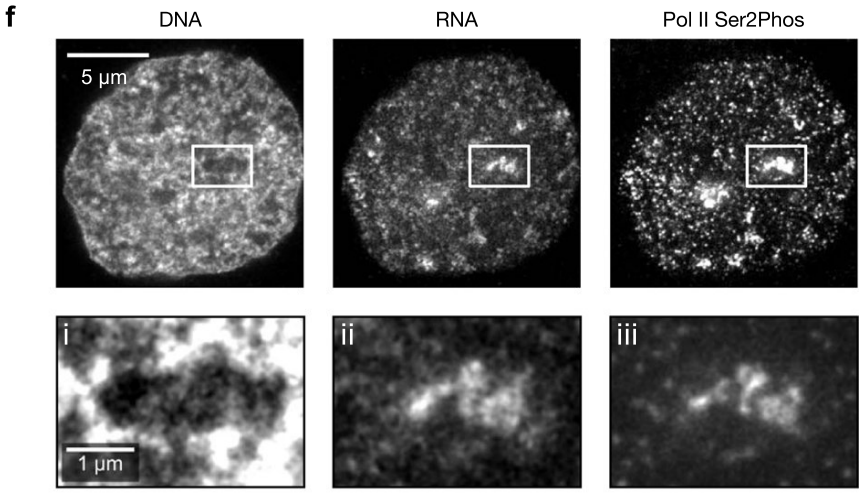


Figure 8.4.22: (F) Representative three-color micrographs showing DNA, RNA, and transcriptional activity (Pol II Ser2Phos) in a nuclear mid-section after transcription onset

It appears that a high concentration of RNA drives the "phase" separation of heterochromatin and euchromatin. Hence, euchromatin might act as an "oil in water" type microemulsion, with the euchromatin core stabilized by "tethered" RNA acting as an amphiphile. Ribonucleoproteins (RNPs) can be modeled as phase-separated droplets or condensates. Figure 8.4.23 shows how increasing RNA leads to the formation of euchromatin "domains," which stay dispersed in the presence of continuing RNA formation.

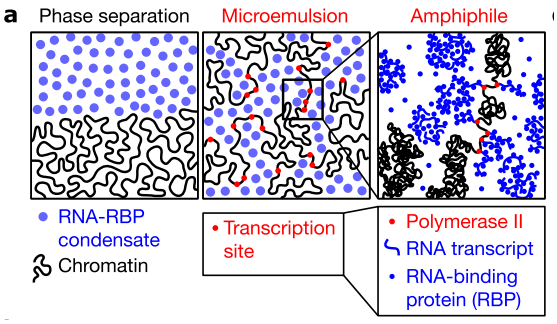


Figure 8.4.23: (a) Cartoon representation of conventional phase separation and a microemulsion. The right panel focuses on the amphiphile in the microemulsion. Hilbert, L., Sato, Y., Kuznetsova, K. *et al.* Transcription organizes euchromatin via microphase separation. *Nat Commun* **12,** 1360 (2021). https://doi.org/10.1038/s41467-021-21589-3. Creative Commons Attribution 4.0 International License. http://creativecommons.org/licenses/by/4.0/.

A final summary cartoon that describes the formation and separation of chromosomal DNA into euchromatin and heterochromatin in Zebra fish is shown in Figure 8.4.24

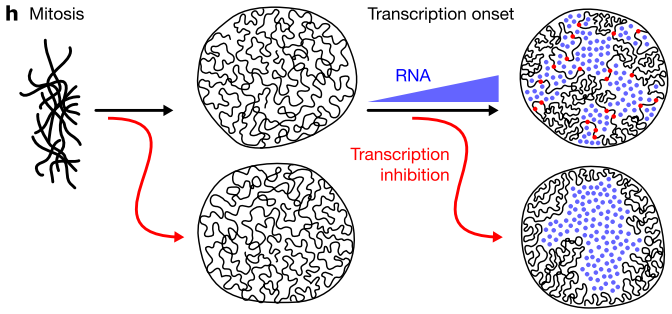


Figure 8.4.24: (h) Sketch summarizing the experimental observations to this point. Hilbert, L., Sato, Y., Kuznetsova, K. *et al.* Transcription organizes euchromatin via microphase separation. *Nat Commun* **12,** 1360 (2021). https://doi.org/10.1038/s41467-021-21589-3. Creative Commons Attribution 4.0 International License. http://creativecommons.org/licenses/by/4.0/.

8.4.8: Summary

This chapter delves into the complex organization of chromatin in eukaryotic cells, integrating structural, functional, and regulatory aspects relevant to gene expression and cellular division. It begins by distinguishing between the two observable states of chromatin in the nucleus—heterochromatin and euchromatin—highlighting that heterochromatin, which appears darkly stained, is densely packed and transcriptionally inactive, whereas euchromatin is loosely organized, occupies most of the nuclear space, and is transcriptionally active.

A review of the cell cycle establishes the context in which chromatin organization changes dramatically. During interphase, DNA is decondensed, allowing active transcription, while in mitosis, chromosomes condense into the iconic structures used for proper segregation during cell division. The chapter explains that eukaryotic chromosomes are linear structures composed of long DNA molecules complexed with proteins. This organization is critical for managing the vast amount of genetic material—if uncondensed, the human genome would span over two meters in length.

At the molecular level, DNA is compacted by supercoiling and packaged around histone proteins to form nucleosomes. The nucleosome, the fundamental unit of chromatin, comprises an octamer of core histones around which DNA winds approximately 1.65 times, forming the "beads on a string" structure. Linker histones such as H1 further stabilize higher-order structures, leading to the formation of 30 nm fibers and, eventually, organized chromosome territories within the nucleus.

The chapter also explores the organization of specific chromosomal regions, such as telomeres—specialized repetitive sequences that protect chromosome ends and mitigate the end replication problem. Telomeres are maintained by telomerase or, alternatively, by recombination-based mechanisms in some cells, and play a key role in cellular aging and genome stability.

Advanced topics include the higher-order folding of chromatin into topologically associating domains (TADs) and compartments, which spatially segregate regions of active (euchromatin) and inactive (heterochromatin) transcription. This three-dimensional organization, facilitated by protein complexes like cohesin and CTCF, not only promotes efficient regulation of gene expression through long-range interactions between enhancers, silencers, and promoters, but also reflects the dynamic nature of chromatin remodeling in response to transcriptional activity.

Finally, the concept of phase separation is introduced as a mechanism for the segregation of chromatin into distinct nuclear domains. The chapter illustrates how RNA and protein interactions contribute to the formation of transcriptionally active microenvironments, further emphasizing the intricate interplay between chromatin structure and cellular function.

In summary, this chapter provides a comprehensive understanding of chromatin—from its fundamental nucleosome structure and dynamic folding into higher-order domains to its crucial role in regulating gene expression, maintaining genome integrity, and orchestrating cellular processes during the cell cycle.

8.5: References

This page is a draft and is under active development.

4.4 References

Börner, R., Kowerko, D., Miserachs, H.G., Shaffer, M., and Sigel, R.K.O. (2016) Metal ion induced heterogeneity in RNA folding studied by smFRET. *Coordination Chemistry Reviews* 327 DOI: 10.1016/j.ccr.2016.06.002 Available at: https://www.researchgate.net/publication/303846502_Metal_ion_induced_heterogeneity_in_RNA_folding_studied_by_smFRET

Hardison, R. (2019) B-Form, A-Form, and Z-Form of DNA. Chapter in: R. Hardison's *Working with Molecular Genetics.* Published by LibreTexts. Available at: https://bio.libretexts.org/Bookshelves/Genetics/Book%3A_Working_with_Molecular_Genetics_(Hardison)/Unit_I%3A_Genes%2C_Nucleic_Acids%2C_Genomes_and_Chromosomes/2%3A_Structures_of_Nucleic_Acids/2.5%3A_B-Form%2C_A-Form%2C_and_Z-Form_of_DNA

Lenglet, G., David-Cordonnier, M-H., (2010) DNA-destabilizing agents as an alternative approach for targeting DNA: Mechanisms of action and cellular consequences. *Journal of Nucleic Acids* 2010, Article ID: 290935, DOI: 10.4061/2010/290935 Available at: https://www.hindawi.com/journals/jna/2010/290935/

Mechanobiology Institute (2018) What are chromosomes and chromosome territories? *Produced by the National University of Singapore.* Available at: https://www.mechanobio.info/genome-regulation/what-are-chromosomes-and-chromosome-territories/

National Human Genome Research Institute (2019) The Human Genome Project. *National Institutes of Health*. Available at: https://www.genome.gov/human-genome-project

Wikipedia contributors. (2019, July 8). DNA. In *Wikipedia, The Free Encyclopedia*. Retrieved 02:41, July 22, 2019, from https://en.Wikipedia.org/w/index.php?title=DNA&oldid=905364161

Wikipedia contributors. (2019, July 22). Chromosome. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:18, July 23, 2019, from https://en.Wikipedia.org/w/index.php?title=Chromosome&oldid=907355235

Wikilectures. Prokaryotic Chromosomes (2017) In MediaWiki, Available at: https://www.wikilectures.eu/w/Prokaryotic_Chromosomes

Wikipedia contributors. (2019, May 15). DNA supercoil. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:40, July 25, 2019, from https://en.Wikipedia.org/w/index.php?title=DNA_supercoil&oldid=897160342

Wikipedia contributors. (2019, July 23). Histone. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:19, July 26, 2019, from https://en.Wikipedia.org/w/index.php?title=Histone&oldid=907472227

Wikipedia contributors. (2019, July 17). Nucleosome. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:17, July 26, 2019, from https://en.Wikipedia.org/w/index.php?title=Nucleosome&oldid=906654745

Wikipedia contributors. (2019, July 26). Human genome. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:12, July 27, 2019, from https://en.Wikipedia.org/w/index.php?title=Human_genome&oldid=908031878

Wikipedia contributors. (2019, July 19). Gene structure. In *Wikipedia, The Free Encyclopedia*. Retrieved 06:16, July 27, 2019, from https://en.Wikipedia.org/w/index.php?title=Gene_structure&oldid=906938498

8.6: Enzymes for Genetic modifications

Learning Goals (ChatGPT o3-mini, 2/1/25)

1. **Describe the Mechanism of Restriction Endonucleases:**
   - Explain how Type II restriction enzymes recognize specific palindromic sequences in double-stranded DNA and catalyze double-strand breaks, generating predictable fragment sizes.
2. **Understand the Restriction-Modification System:**
   - Illustrate how bacteria use restriction enzymes together with DNA-methyltransferases to differentiate between self and non-self DNA, thereby protecting their genomes from viral infection.
3. **Analyze the Catalytic Mechanism of Type II Restriction Enzymes:**
   - Describe the role of divalent metal ions (e.g., Mg^{2+}) in activating water for nucleophilic attack on the phosphodiester bond and discuss the stereochemical evidence (inversion of configuration) that supports a direct hydrolytic mechanism.
4. **Interpret Enzyme Specificity and Nomenclature:**
   - Discuss how restriction enzymes are named based on their organism of origin and how isoschizomers and neoschizomers differ in their cleavage patterns and evolutionary origins.
5. **Compare and Contrast Gene Editing Technologies:**
   - Compare traditional restriction enzyme-based cloning methods with modern CRISPR-Cas9 gene editing, emphasizing their mechanisms, applications, and limitations.
6. **Explain the Structure and Function of CRISPR-Cas9:**
   - Describe the architecture of Cas9, including its two catalytic nuclease domains (HNH and RuvC) and the roles of the guide RNA (crRNA and tracrRNA, or the synthetic sgRNA) in conferring target specificity.
7. **Understand the Allosteric Regulation of Cas9:**
   - Analyze how the binding of guide RNA triggers conformational changes in Cas9 that prime the enzyme for DNA binding and cleavage, and explain the role of the protospacer adjacent motif (PAM) in target recognition and cleavage specificity.
8. **Evaluate the Applications and Challenges of CRISPR Gene Editing:**
   - Assess how CRISPR-Cas9 is used in disease diagnosis, therapy, and genetic modifications, and discuss strategies for minimizing off-target effects and ensuring efficient delivery into target cells.
9. **Integrate Concepts in Protein Structure and Enzyme Mechanism:**
   - Relate the structural features of both restriction enzymes and Cas9 to their catalytic mechanisms, emphasizing how protein conformation, cofactor binding, and specific amino acid residues contribute to their function.

These goals are intended to guide your study and critical analysis of gene editing systems, helping you connect fundamental enzyme mechanisms with their revolutionary applications in biotechnology and medicine.

It is difficult to read newspapers and newsmagazines without encountering the CRISPR-Cas9 gene editing system that has the potential to make gene editing routine in disease diagnosis, treatment, and cure, as well as in genetic modification of organisms to improve their quality and quantity for food and natural product production. In this chapter section, we will explore the mechanism of restriction enzymes that made gene cloning possible as well as the CRISPR-Cas gene editing system.

8.6.1: Restriction Endonucleases

A ***restriction enzyme***, ***restriction endonuclease***, or ***restrictase*** is an enzyme that cleaves DNA into fragments at or near specific recognition sites within molecules known as restriction sites. Restriction enzymes are one class of the broader endonuclease group of enzymes. Restriction enzymes are commonly classified into five types, which differ in their structure and whether they cut their DNA substrate at their recognition site or if the recognition and cleavage sites are separate. To cut DNA, all restriction enzymes

make two incisions, once through each sugar-phosphate backbone (i.e., each strand) of the DNA double helix. Here, we will focus on the Type II restriction enzymes routinely used in molecular biology and biotechnology applications.

As with other restriction enzymes, Type II Restriction Enzymes occur exclusively in unicellular microbial life forms—mainly bacteria and archaea (prokaryotes)—and are thought to protect these cells from viruses and other infectious DNA molecules. Inside a prokaryote, the restriction enzymes selectively cut up *foreign* DNA in a process called ***restriction digestion;*** meanwhile, host DNA is protected by a modification enzyme (a methyltransferase) that modifies the prokaryotic DNA and blocks cleavage. Together, these two processes form the ***restriction-modification system***.

The first Type II Restriction Enzyme discovered was HindII from the bacterium *Haemophilus influenzae* Rd. The event was described by Hamilton Smith (Figure 7.23) in his Nobel lecture, delivered on 8 December 1978:

‘"In one such experiment we happened to use labeled DNA from phage P22, a bacterial virus I had worked with for several years before coming to Hopkins. To our surprise, we could not recover the foreign DNA from the cells. With Meselson’s recent report in our minds, we immediately suspected that it might be undergoing restriction, and our experience with viscometry told us that this would be a good assay for such an activity. The following day, two viscometers were set up, one containing P22 DNA and the other Haemophilus *DNA. Cell extract was added to each and we began quickly taking measurements. As the experiment progressed, we became increasingly excited as the viscosity of the* Haemophilus *DNA held steady while the P22 DNA viscosity fell. We were confident that we had discovered a new and highly active restriction enzyme. Furthermore, it appeared to require only* Mg^{2+} *as a cofactor, suggesting that it would prove to be a simpler enzyme than that from* E. coli *K or B.*

After several false starts and many tedious hours with our laborious, but sensitive viscometer assay, Wilcox and I succeeded in obtaining a purified preparation of the restriction enzyme. We next used sucrose gradient centrifugation to show that the purified enzyme selectively degraded duplex, but not single-stranded, P22 DNA to fragments averaging around 100 bp in length, while Haemophilus *DNA present in the same reaction mixture was untouched. No free nucleotides were released during the reaction, nor could we detect any nicks in the DNA products. Thus, the enzyme was clearly an endonuclease that produced double-strand breaks and was specific for foreign DNA. Since the final (limit) digestion products of foreign DNA remained large, it seemed to us that cleavage must be site-specific. This proved to be case and we were able to demonstrate it directly by sequencing the termini of the cleavage fragments.’"*

Figure (\PageIndex{35}\): Hamilton Smith and Daniel Nathans at the Nobel Prize press conference, 12 October 1978 (reproduced with permission from Susie Fitzhugh). Original Repository: Alan Mason Chesney Medical Archives, Daniel Nathans Collection. Image from: Pingoud, A., Wilson, G.G., and Wende, W. (2014) Nuc Acids Res 42(12):7489-7527.

Restriction enzymes are named according to the taxonomy of the organism in which they were discovered. The first letter of the enzyme refers to the genus of the organism and the second and third to the species. This is followed by letters and/or numbers identifying the isolate. Roman numerals are used to specify different enzymes from the same organism. For example, the enzyme ‘HindIII’ was discovered in *Haemophilus influenzae*, serotype d, and is distinct from the HindI and HindII endonucleases also present within this bacterium. The DNA-methyltransferases (MTases) accompanying restriction enzymes are named similarly, and given the prefix ‘M.’. When there is more than one MTase, they are prefixed ‘M1.’, ‘M2.’, etc, if they are separate proteins or ‘M1∼M2.’ when joined.

Restriction Enzymes that recognize the same DNA sequence, regardless of where they cut, are termed ***‘isoschizomers’*** (iso = equal; skhizo = split). ***Isoschizomers*** that cut the same sequence at different positions are called ***‘neoschizomers’*** (neo = new).

Isoschizomers that cut at the same position are frequently, but not always, evolutionarily drifted versions of the same enzyme (e.g. BamHI and OkrAI). ***Neoschizomers,*** on the other hand, are often evolutionarily unrelated enzymes (e.g.EcoRII and MvaI).

Type II Restriction Enzymes are a conglomeration of many different proteins that, by definition, have the common ability to cleave duplex DNA at a fixed position within, or close to, their recognition sequence. This cleavage generates reproducible DNA fragments and predictable gel electrophoresis patterns, properties that have made these enzymes invaluable reagents for laboratory DNA manipulation and investigation. Almost all Type II Restriction Enzymes require divalent cations, usually Mg^{2+}, as essential components of their catalytic sites. Ca^{2+}, on the other hand, often acts as an inhibitor of Type II Restriction Enzymes.

The recognition sequences of Type II Restriction Enzymes are ***palindromic***, with two possible types of palindromic sequences. The ***mirror-like palindrome*** is similar to those found in ordinary text, in which a sequence reads the same forward and backward on a single strand of DNA, as in GTAATG. The ***inverted repeat palindrome*** is also a sequence that reads the same forward and backward, but the forward and backward sequences are found in complementary DNA strands (i.e., of double-stranded DNA), as in GTATAC (GTATAC being complementary to CATATG). ***Inverted repeat palindromes*** are more common and have greater biological importance than ***mirror-like palindromes.*** The position of cleavage within the palindromic sequence can vary depending on the enzyme and can produce either single-stranded overhanging sequences (sticky ends) or blunt-ended DNA products. Table 8.6.8 below shows examples of staggered and blunt end cuts by restriction enzymes.

EcoR1	.
Sma1	.

Table 8.6.8: Staggered and blunt end cut sequences by EcoR1 and Sma1

The host can use methylation to protect its own genome from cleavage. For example, the methylation of the EcoRI recognition sequence by the M.EcoRI methyltransferase (MTase), changes the sequence from GAATTC to GAm6ATTC (m6A = N6-methyladenine). This modification completely protects the sequence from EcoRI cleavage.

Type II Restriction Enzymes initially bind non-specifically with the DNA and proceed to slide down the DNA scanning for recognition sequences as shown in Figure (\PageIndex{36}\):. Upon binding to the correct palindromic sequence, the enzyme associates with the metal cofactor and mediates catalytic cleavage of the DNA using the mechanism of ***strain distortion*** and ***catalysis by approximation***.

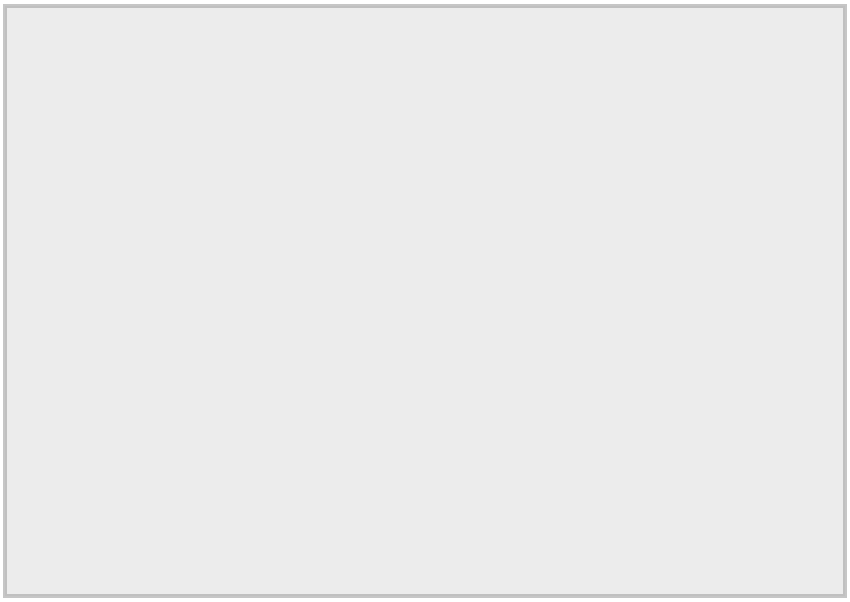


Figure (\PageIndex{36}\): DNA Recognition and Cleavage by Type II Restriction Endonucleases. (A) Pictorial view of an EcoRV dimer scanning nonspecifically along the DNA until a specific binding site is recognized. This causes coupling with the metal cofactor and strain distortion of the DNA. Hydrolysis of the phosphodiester bond is mediated and the DNA cleavage products are released from the enzyme. (B) shows a space-filling model of EcoRV DNA recognition and cleavage. Figure (A) from Pingoud, A., Wilson, G.G., and Wende, W. (2014) Nuc Acids Res 42(12):7489-7527. and Figure (B) from Thomas Splettstoesser

One of the most important questions regarding the catalytic mechanism of a hydrolase is whether hydrolysis involves a covalent intermediate, as is typical for the proteases described previously. This can be decided by analyzing the stereochemical course of the reaction. This was done first for EcoRI and later for EcoRV. Both enzymes were found to cleave the phosphodiester bond with inversion of the chiral center at the phosphorus, which argues against the formation of a covalent enzyme–DNA intermediate. Thus, it is proposed that cleavage involves the direct nucleophilic attack of the substrate by a water molecule, as shown in Figure (\PageIndex{37}\) below.

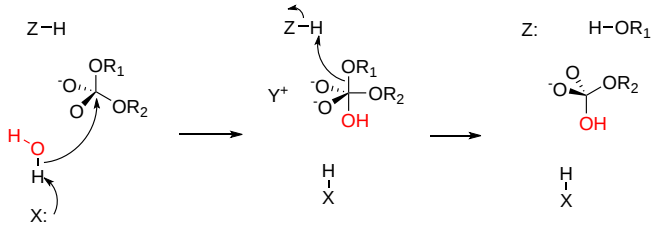


Figure (\PageIndex{37}\): A General Mechanism for DNA Cleavage by EcoRI and EcoRV. An activated water molecule attacks the phosphorous in-line with the phosphodiester bond to be cleaved, which proceeds with an inversion of configuration. X, Y, and Z are a general base, a Lewis acid, and a general acid, respectively. Figure adapted from:

Pingoud, A., Wilson, G.G., and Wende, W. (2014) Nuc Acids Res 42(12):7489-7527.

Type II restriction enzymes typically form a homodimer when binding with DNA, as shown in the crystal structure of ***Bgl*II** in Figure 7.26B. ***Bgl*II** catalyzes phosphodiester bond cleavage at the DNA backbone through a phosphoryl transfer to water. Studies on the mechanism of restriction enzymes have revealed several general features that seem true in almost all cases. However, the actual mechanism for each enzyme is most likely some variation of this general mechanism (Figure 7.25). This mechanism requires a base to generate the hydroxide ion from water, acting as the nucleophile and attacking the phosphorus in the phosphodiester bond. Also required is a Lewis acid to stabilize the extra negative charge of the pentacoordinate transition state phosphorus, as well as a general acid or metal ion that stabilizes the leaving group (3'-O^-). Two divalent metal cofactors are required in some Type II Restriction Enzymes (such as in EcoRV and BamHI). In contrast, other enzymes only require one divalent metal cofactor (such as in EcoRI and BglII).

Structural studies of endonucleases have revealed a similar architecture for the active site with the residues following the weak consensus sequence Glu/Asp-$(X)_{9\text{-}20}$-Glu/Asp/Ser-X-Lys/Glu. ***Bgl*II's** active site is similar to other endonucleases', following the Asp-$(X)_9$-Glu-X-Gln sequence. In its active site, a divalent metal cation, most likely Mg^{2+}, interacts with Asp-84, Val-94, a phosphoryl oxygen, and three water molecules. One of these water molecules can act as a nucleophile because of its proximity to the scissile phosphoryl group (Figure 7.26A). The nucleophilic water molecule is positioned for attack onto the phosphoryl group by a hydrogen bond with the side chain amide oxygen of Gln-95 and its contact with the metal cation. Interaction with the metal cation effectively lowers its pK_a, promoting the water's nucleophilicity as shown in Panel A of Figure (\PageIndex{38}\) below (from Pingoud, A., Wilson, G.G., and Wende, W. (2014) Nuc Acids Res 42(12):7489-7527). During hydrolysis, the divalent cation can stabilize the 3'-O^- leaving group and coordinate proton abstraction from one of the coordinated water molecules

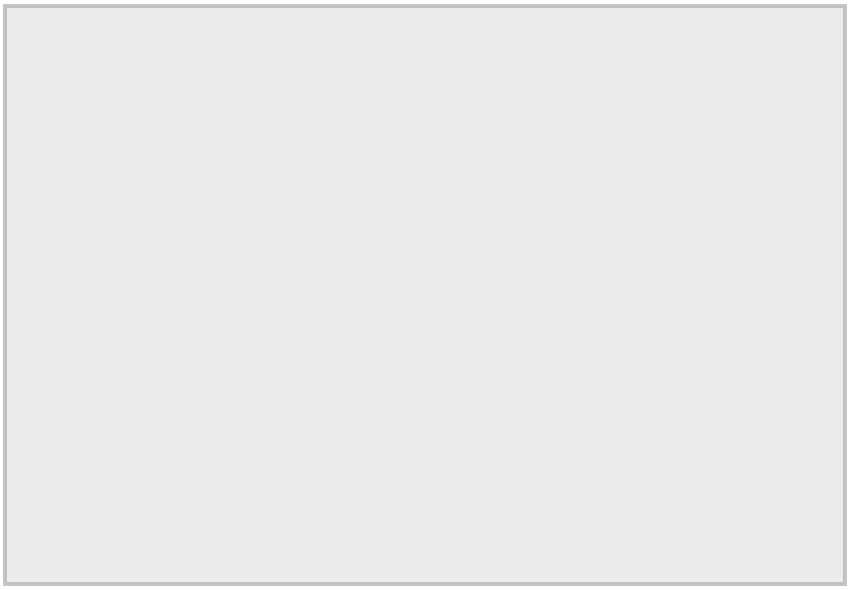


Figure (\PageIndex{38}\): Proposed Reaction Mechanism for the Type II Restriction Endonuclease, *Bgl*II. (A) Schematic diagram of the catalytic mechanism demonstrating the utility of Mg2+ ions and polar amino acid residues within the active site to activate and position a water molecule for nucleophilic attack on the phosphodiester bond of the DNA substrate. (B) Crystal structure of the BglII dimer with double-stranded DNA and (C) Coordination of the Mg2+ cofactor within the active site of the BglII enzyme.

Figures from G Williams

8.6.2: CRISPR-Cas 9

The **CRISPR** (clustered regularly interspaced short palindromic repeats) operon was initially discovered as part of the adaptive immune system of bacteria and archaea, which must defend themselves against viruses (bacteriophages) and unwanted plasmids transferred from both bacteria. It would be ideal for bacteria to recognize previous exposure to viruses and their nucleic acids as the basis of their immunological memory system. Given the tendency of viral DNA to integrate into the host genome (which allows later transcription and translations of the viral genes in the process of new virus production), immunological memory could be based on that viral integrated DNA. Without going into detail, viral DNA can be integrated between two direct repeats in the bacterial genome. DNA from different viruses from previous exposures is also incorporated in the same fashion. One site of integration is the CRISPR operon. The DNA of the CRISPR operon contains both protein-coding and noncoding regions which are transcribed and processed to form at least three RNA molecules, as shown in Figure (\PageIndex{24}\) below.

- a coding Cas 9 mRNA this is translated to produce the **Cas** 9 (**C**RISPR-**as**sociated protein);
- a noncoding cr-RNA (CRISPR RNA)
- a noncoding tracr-RNA (trans-activating CRISPR RNA)

Figure (\PageIndex{24}\): DNA of the CRISPR operon

The two mature noncoding RNAs eventually associate to form a binary complex. When using CRISPR-Cas 9 in eukaryotic gene editing applications, the two noncoding RNAs are covalently combined into one large **s**ynthetic **g**uide RNA (**sg-RNA**), described later in this section. The Cas 9 protein is an endonuclease that cleaves both strands of bound target dsDNA in a blunt-end fashion at specific sequences. This occurs after the DNA binds to two arginines (1333 and 1335) in Cas9 through a short (3-5+ bases) recognition **protospacer adjacent motif** (**PAM**) located three base pairs from the cleavage site. The DNA must also bind in a complementary and specific fashion to the **protein-bound** noncoding cr+tracr-RNAs (or a single sg-RNA molecule for gene editing applications). Binding and cleavage of target DNA would render DNA from an invading bacteriophage inactive.

Basic research into the bacterial CRISPR system has led to revolutionary and explosive eukaryotic applications of this gene editing system. The hope is that CRISPR technology will give us a precise and incredibly cheap way to do gene therapy in diseased cells and organisms. Given its role in transforming our ability to edit the genome and potentially cure genetically based diseases, we will explain its mechanism.

We have discussed the structure and function of many proteins. Protein enzymes are key to life as they catalyze almost all biological reactions. Most key enzymes are regulated. The activity of Cas 9 must be carefully controlled. Think of the

consequences if the enzyme cleaves promiscuously at off-site targets! This section will help you understand several critical features of this enzyme:

1. How does the enzyme find its correct target site, a 20 nucleotide DNA sequence, and a proximal PAM site, among all the possible alternative sites? Think of how many PAM sequences must be in the host DNA genome!
2. How can the enzyme be "turned" on when it finds its target site and remains off when free, but more importantly when bound off-site?

First, we will discuss the apo- form of the enzyme without bound substrate and RNA.

Apo- and Holo-Cas 9

This section will focus on the Type II-A Cas9 from *Streptococcus pyogenes* (SpyCas9 or SpCas9). Cas 9 is an endonuclease that cleaves both strands of DNA 3 base pairs from a DNA motif, NCC/NGG, called PAM. It has two distinct lobes. The nuclease lobe (NUC), amino acids 1-56 and 718-1368, has two different nuclease domains for the two cleavages. The recognition or receptor lobe (REC), amino acids 94-717, interacts with the RNA molecules. There is also an arginine-rich bridge helix (57-93).

The enzyme has two catalytic nuclease domains:

- HNH-like nuclease domain cleaves the "target" DNA strand, which is complementary to the RNA that confers specificity to the enzyme. The key catalytic residues are His 840 and Asn 854. It also contains a Mg ion;
- Ruv-like domain cleaves the complementary "non-target" strand with key active site residues Asp 10, Glu 762, Asp 986, and His 983. It also contains a bound Mn ion. The two lobes are separated by two linkers, amino acids 712-717, and an arginine-rich bridge (basic helix - BH), amino acids 628-658.

The overall structure of the apoenzyme (without bound RNA and DNA,pdb id 4cmp) is shown in Figure (\PageIndex{25}\) below, which shows the NUC domain (light blue) with the two catalytic domains (HNH and Ruv), the REC domain (orange) and the BH helix (red).

Figure (\PageIndex{25}\): Apoenyzme Cas9 (without bound RNA and DNA (4cmp)

A close up view showing the two catalytic sites is shown in Figure (\PageIndex{26}\) below.

Figure (\PageIndex{26}\): Two catalytic sites in Cas9

Figure 8.6.27 shows an interactive iCn3D model of Streptococcus pyogenes Cas9 in complex with guide RNA and target DNA (4OO8) (long load time). The Cas9 enzyme is shown as a gray transparent surface with an underlying cartoon rendering. The DNA is shown as colored sticks. The RNA is shown as a cyan cartoon.

NCBI iCn3D Figure 8.6.27: **Streptococcus pyogenes Cas9 in complex with guide RNA and target DNA (4OO8)**. (Copyright; author via source).

Click the image for a popup or use this external link: https://structure.ncbi.nlm.nih.gov/i...RjzJBFVt5qRjS7 (long load time)

A comparison of the crystal structure of the apo-Cas 9 and the ternary Cas 9: sgRNA:DNA target strand complex shows a significant conformational change on binding nucleic acids. The structure of the holoenzyme (ternary complex) is shown in Figure (\PageIndex{28}\) below.

Figure (\PageIndex{28}\): Structure of the holoenzyme Cas9 (ternary complex)with bound guide RNA and DNA

The extent of the conformation change between apo- and holo-Cas 9 enzymes can be seen by examining the distance between D435 and E 944/945 in Figure (\PageIndex{29}\) below. The importance of this change will be described later.

Figure (\PageIndex{29}\): Distance between D435 and E 944/945 going from the apo-Cas9 (left) to the holo-Cas 9 (right) enzyme

Figure (\PageIndex{30}\) below shows the pathway from the transcription of the relevant CRISPR genes (coding and noncoding) to the assembly of the ternary complex and the blunt end cut of the target DNA strand three nucleotides from the PAM sequence.

Figure (\PageIndex{30}\): Pathway from the transcription of the relevant CRISPR genes (coding and noncoding) to the assembly of the ternary complex and the blunt end cut of the target DNA strand three nucleotides from the PAM sequence

Figure (\PageIndex{31}\) below shows an expanded view of the ternary complex.

Figure (\PageIndex{31}\): Expanded view of the ternary complex of Cas9 with guide RNA and DNA

Mechanism of DNA binding and cleavage

The above figures do not speak to the mechanism of the binding processes that form the ternary complex. Kinetic and structural studies have been conducted to elucidate the mechanism of binding and cleavage and address the following questions:

- which binds first, the RNA or DNA?
- What are the consequences of the profound conformational changes on the formation of the ternary complex?

The **specificity** of target DNA binding depends both on enzyme:PAM DNA and enzyme:sgRNA (or tracr- and crRNA) interactions. It should seem improbable that the trinucleotide PAM DNA sequence (NGG in *S. pyogenes*), which interacts with a pair of arginines (R 1333, R 1335) through H-bonding, as shown in the images above, and other local sites in Cas 9 would provide the sole or even the majority of the binding interactions. Figure (\PageIndex{32}\) below shows the Args:PAM interaction (pdb code 4un3)

Figure (\PageIndex{32}\): Args:PAM interaction in holo-Cas9 (4un3)

Hence it is most likely that RNA binds first. Indeed, it does with the tracRNA implicated in the recruitment of Cas and the crRNA providing specificity for target DNA binding. The resulting Cas9:RNA binary complex could then search the relevant DNA genome. That would include the DNA of the bacteriophage in viral infection or eukaryotic DNA if the CRISPR DNA operon with the genes for Cas 9 and a sg-RNA was transfected into the eukaryotic cell. After RNA binding, the enzyme changes conformation and allows loose DNA binding through Cas 9: PAM interactions.

Studies have shown that the apo form can also bind DNA, but it does so loosely and indiscriminately. It dissociates quickly, and binding is affected by generic polyanions such as the glycosaminoglycan heparin, which indicates its nonspecific nature. Once bound, both off-target and target DNAs would then be surveyed. If a target DNA contained a PAM sequence, the complex would undergo another conformational change to position the HNH and Ruv nuclease catalytic residues and locally unwind the duplex DNA to make the blunt-end cuts.

Cas 9 binding to the PAM site would promote better interaction of the unwound DNA and the bound RNA. If no PAM was present, no catalytically effective Cas 9:target DNA would form. This prevents off-site cleavage. These allosteric changes and controls are vital to the function of the endonuclease. Here are some findings that support this proposed mechanism:

- the conformation of apo Cas 9 is catalytically inactive;
- on binding RNA to form a binary complex, Cas 9 undergoes a dramatic conformational change, mostly in the REC lobe. However, on binding DNA in a nonspecific fashion, the conformational changes are much smaller. This suggests that most changes in conformation occur before DNA binding. In a way, RNA acts as an allosteric activator of the enzyme (as well as the major source of binding specificity to target DNA). Conformational changes can be determined directly by comparison of crystal structures or spectral techniques such as fluorescence resonance energy transfer (FRET) between two different attached fluorophores.
- Cas 9: RNA interactions lead to ordering of the region of the RNA that interacts with the DNA PAM sequence and adjacent deoxynucleotides (a "seed sequence"), allowing the Cas 9:RNA complex to scan and interact with potential DNA targets with PAM sequences;
- Once a PAM site is found, conformational changes lead to unwinding of the dsDNA, which allows heteroduplex formation between the crRNA and the target DNA strand;
- since Cas 9 recognizes a variety of DNA target sequences (but of course, only a specific PAM sequence), the binding of the target sequence depends on the geometry, not the sequence, of the target DNA;

- since binding of off-target DNA to the Cas 9:RNA complex occurs but with very infrequent cleavage, binding and cleavage are very distinct steps;
- on specific DNA binding, the HNH catalytic site moves near the sessile DNA bond site. Crystal structures show that the active site His is not sufficiently close to facilitate cleavage, suggesting that binding a second metal ion (see below) may be necessary. Molecular dynamics studies show that the HNH domain is "remarkably plastic."

Figure (\PageIndex{33}\) below shows an animation that illustrates the relative conformational changes going from the apo Cas 9 to the binary Cas 9:sgRNA complex to the ternary Cas 9: sgRNA: target DNA complex. The NUC catalytic domain is shown in light blue, the REC (receptor or RNA binding domain) in orange, sgRNA in red, and the target DNA in green. Note again that on binding RNA to form a binary complex, Cas 9 undergoes a dramatic conformational change, mostly in the REC lobe. The pdb protein sequences shown were aligned using pdbEfold.

Figure (\PageIndex{33}\): Animation illustrating conformational changes going from the apo Cas 9 to the binary Cas 9:sgRNA complex to the ternary Cas 9: sgRNA: target DNA complex

A potential abbreviated catalytic mechanism for the Ruv nuclease domain is shown in Figure (\PageIndex{34}\) below. The red arrows indicate the second set of electron movements. His 983 acts as a general base to abstract a proton from the water, making it a more potent nucleophile. An intermediate trigonal bipyramidal phospho-intermediate is formed, and the preceding transition state, is stabilized by the proximal Mg^{2+} ion (an example of electrostatic or metal ion catalysis). The magnesium is positioned through its interaction with negatively charged carboxyl groups of Asp 10, Glu 762, and Asp 986.

Figure (\PageIndex{33}\): Abbreviated catalytic mechanism for the Ruv nuclease domain of Cas9

A second metal ion might be recruited to the Ruv site to further facilitate the cleavage of the DNA. The HNH catalytic site has a structure (beta-beta-alpha) and conserved His in common with a class of nucleases that require one metal ion. In contrast, the Ruv catalytic site does not have this common secondary structural motif. It has a critical histidine, which are both common features found in endonucleases that use two metal ions.

CRISPR and Eukaryotic Gene Editing

How could blunt-end cutting of both DNA strands by Cas 9 lead to the holy grail of specific eukaryotic gene editing with no off-site effects? Cutting the DNA genome seems like a bad idea. It is potentially so bad that many DNA repair mechanisms have evolved to fix the cut. These include homologous recombination. If corrective DNA is supplied, as well as the components of the CRISPR system, a cell could effectively add the corrective DNA after the double-stranded cut and repair a deleterious mutation. Consult a molecular biology textbook for more insight into homologous recombination.

Mutations in the PAM sequence prevent Cas9 nuclease activity. Hence, the NGG PAM sequence is vital for the above interactions and activities. This would seem to limit the utility of CRISPR-Cas 9 in eukaryotic gene editing until one realizes that the GG dinucleotide has a 5.2% frequency of occurrence in the human genome, corresponding to over 160 million occurrences. Even then, it might not occur in a desired gene target. Cas 9 nuclease from other bacteria extends the range of activity of the CRISPR/Cas system as they interact with other PAM sequences (NNAGAA and NGGNG for *S. thermophilus* and NGGNG for *N. meningidtis*). Likewise, mutations in the *S. pyogenes* PAM (NGG) have been made as well. A D1135E mutation retains but increases the specificity for the normal NGG PAM site. D1135V, R1335Q, and T1337R mutations alter the optimal PAM recognition site to NGAN or NGNG.

CRISPR editing can be easily used to knock out specific genes. In addition, if cells are transfected with a plasmid with many target sequences, the system can edit multiple genes in one experiment. This would be very useful in studies of diseases linked to multiple genes. Since the cost of CRISPR reagents (plasmids, RNAs) is so inexpensive, and the specificity of editing is so high, the great excitement about CRISPR use for gene editing in human disease and for modification of plant and fungal genomes is warranted.

Other systems have been developed to bind to and cleave a target DNA sequence. They typically contain a protein that binds to a specific DNA target and an associated endonuclease that cleaves within the target DNA site. Typical prokaryotic restriction enzymes bind to and cut at a specific nucleotide sequence (for example, Eco R1 cleaves at G/AATTC palindromic sequences) to form sticky ends. The protein itself binds to this DNA recognition site. Other examples are based on the structure of known transcription factors. Libraries of genetically engineered proteins with Zn finger DNA binding domains (designed for specific DNA target sequences) fused to endonucleases have been created. Other examples are proteins called TALENs (transcription activator-like effector nucleases). These are fusion proteins containing a TAL effector DNA-binding domain and a nuclease. In each case, a 3D-folded protein is the specific target DNA recognition molecule. Think how much easier it is to make a 1D-DNA recognition element, a simple linear RNA sequence, which would adopt the correct 3D structure on binding its complementary target.

One major problem in using CRISPR for gene editing must be solved: how to get the CRISPR components in the correct cells in an organism. In effect, it's the same problem faced by small drug designers, only the components are much larger. *Ex vivo* applications, when diseased cells are removed from the body, repaired by CRISPR, and then reinjected, are likely to have more success. In these cases, electroporation would allow the uptake of Cas 9 and the sg-RNA. I*n vivo* therapy has included adeno-associated viruses in which Cas 9 and sg RNA genes could be encapsulated. This technique, used for other gene delivery systems, can be tolerated immunologically. However, this system allows for continual gene expression, which is undesirable for gene editing. After an initial "fix" of a mutant gene, continued expression of the CRISPR-Cas 9 genes would increase the chances for off-target cutting. A more recent approach is to deliver the mRNA in artificial lipid nanoparticles that can be taken into cells. Once free and translated into protein and sg RNA inside the cell, gene editing can occur before the RNA and protein are degraded.

8.6.3: Summary

Chapter Summary: Restriction Enzymes and CRISPR-Cas9 Gene Editing

This chapter explores two revolutionary tools that have transformed molecular biology and biotechnology: the classical restriction endonucleases and the modern CRISPR-Cas9 gene editing system.

It begins by examining the fundamental principles of restriction enzymes, particularly the widely used Type II enzymes. These enzymes recognize specific palindromic DNA sequences and cleave both strands at precise sites, generating reproducible fragment patterns essential for gene cloning and analysis. The chapter details the restriction-modification system in prokaryotes, where bacteria and archaea protect their own genomes via DNA methylation while using these enzymes to degrade foreign DNA. Emphasis is placed on the catalytic mechanism—where divalent metal ions, usually Mg^{2+}, activate a water molecule to directly hydrolyze the phosphodiester bond with inversion of configuration—and on how enzyme nomenclature and classifications (including isoschizomers and neoschizomers) reflect evolutionary relationships and cleavage patterns.

The discussion then transitions to the CRISPR-Cas9 system, which has rapidly gained attention due to its unprecedented precision and versatility in gene editing. Originating as part of the adaptive immune system in bacteria, CRISPR-Cas9 utilizes a guide RNA (either as separate crRNA and tracrRNA or as a fused single guide RNA) to direct the Cas9 endonuclease to a target DNA sequence adjacent to a protospacer adjacent motif (PAM). Detailed structural insights reveal that Cas9 consists of distinct lobes—the nuclease (NUC) lobe housing the HNH and RuvC catalytic domains, and the recognition (REC) lobe responsible for RNA binding. Binding of the guide RNA triggers significant conformational changes that “activate” Cas9, enabling it to specifically bind and cleave target DNA with blunt-end cuts.

Furthermore, the chapter outlines the critical factors that govern Cas9 specificity and activity, including the allosteric changes induced upon PAM binding and the kinetic distinction between nonspecific DNA association and target recognition. It also addresses practical considerations in applying CRISPR-Cas9 for eukaryotic gene editing, such as delivery strategies (e.g., viral vectors, lipid nanoparticles) and the importance of minimizing off-target effects.

In summary, this chapter integrates the biochemical principles underlying restriction enzymes and CRISPR-Cas9, illustrating how understanding enzyme mechanisms, structure-function relationships, and regulatory features has not only advanced basic molecular biology but also paved the way for transformative applications in disease diagnosis, treatment, and genetic engineering.

8.6.3.0.1: 7.4 References:

1. Wikipedia contributors. (2020, April 21). Nucleophile. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:39, April 26, 2020, from en.Wikipedia.org/w/index.php?title=Nucleophile&oldid=952368939
2. Oregon Institute of Technology (2019) Organic Chemistry II (Lund). In Libretexts. Retrieved 10:58 am, April 27, 2020 from: https://chem.libretexts.org/Courses/Oregon_Institute_of_Technology/OIT%3A_CHE_332_--_Organic_Chemistry_II_(Lund)
3. Wikipedia contributors. (2020, April 12). Bond cleavage. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:15, April 27, 2020, from en.Wikipedia.org/w/index.php?title=Bond_cleavage&oldid=950494652
4. Wikipedia contributors. (2020, February 24). Arrow pushing. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:25, April 27, 2020, from en.Wikipedia.org/w/index.php?title=Arrow_pushing&oldid=942438883
5. Wikipedia contributors. (2020, April 16). Acid dissociation constant. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:48, April 27, 2020, from en.Wikipedia.org/w/index.phptitle=Acid_dissociation_constant&oldid=951313744
6. Farmer, S., Reusch, W., Alexander, E., and Rahim, A. (2016) Organic Chemistry. Libretexts. Available at: https://chem.libretexts.org/Core/Organic_Chemistry
7. Ball, et al. (2016) MAP: The Basics of GOB Chemistry. Libretexts. Available at:https://chem.libretexts.org/Textbook_Maps/Introductory_Chemistry_Textbook_Maps/Map%3A_The_Basics_of_GOB_Chem

istry_(Ball_et_al.)/14%3A_Organic_Compounds_of_Oxygen/14.10%3A_Properties_of_Aldehydes_and_Ketones
8. McMurray (2017) MAP: Organic Chemistry. Libretexts. Available at:https://chem.libretexts.org/Textbook_Maps/Organic_Chemistry_Textbook_Maps/Map%3A_Organic_Chemistry_(McMurry)
9. Soderburg (2015) Map: Organic Chemistry with a Biological Emphasis. Libretexts. Available at:https://chem.libretexts.org/Textbook_Maps/Organic_Chemistry_Textbook_Maps/Map%3A_Organic_Chemistry_With_a_Biological_Emphasis_(Soderberg)
10. Ophardt, C. (2013) *Biological Chemistry.* Libretexts. Available at:https://chem.libretexts.org/Core/Biological_Chemistry/Proteins/Case_Studies%3A_Proteins/Permanent_Hair_Wave
11. Soderberg, T. (2016) *Organic Chemistry with a Biological Emphasis*. Libretexts. Available at:https://chem.libretexts.org/Textbook_Maps/Organic_Chemistry_Textbook_Maps/Map%3A_Organic_Chemistry_with_a_Biological_Emphasis_(Soderberg)
12. Ball, et al. (2016) MAP: The Basics of General, Organic, and Biological Chemistry. Libretexts. Available at:https://chem.libretexts.org/Textbook_Maps/Introductory_Chemistry_Textbook_Maps/Map%3A_The_Basics_of_GOB_Chemistry_(Ball_et_al.)
13. Clark, J. (2017) Organic Chemistry. Libretexts. Available at:https://chem.libretexts.org/Core/Organic_Chemistry/Amides/Reactivity_of_Amides/Polyamides
14. Wikipedia contributors. (2018, December 28). Metabolism. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:28, December 29, 2018, fromen.Wikipedia.org/w/index.php?title=Metabolism&oldid=875751739
15. Ball, Hill, and Scott. (2012) Enzyme Activity, section 18.7 from the book Introduction to Chemistry: General, Organic and Biological (v1.0) retrieved on Dec 31, 2018 fromhttps://2012books.lardbucket.org/books/introduction-to-chemistry-general-organic-and-biological/s21-07-enzyme-activity.html
16. Wikipedia contributors. (2018, November 29). Mechanism of action. In *Wikipedia, The Free Encyclopedia*. Retrieved 05:00, January 1, 2019, fromen.Wikipedia.org/w/index.php?title=Mechanism_of_action&oldid=871201209
17. Mótyán, J.A., Tóth, F., and Tőzsér, J. (2013) Research Applications of Proteolytic Enzymes in Molecular Biology. *Biomolecules 3*(4), 923-942; https://doi.org/10.3390/biom3040923
18. Wikipedia contributors. (2020, April 11). Adenylate kinase. In *Wikipedia, The Free Encyclopedia*. Retrieved 19:28, May 4, 2020, from en.Wikipedia.org/w/index.php?title=Adenylate_kinase&oldid=950311736
19. Ahern, K., Rajagopal, I., and Tan, T. (2019) Biochemistry Free and Easy. Available at Oregon State University (http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy) and Libretexts (https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Free_For_All_(Ahern%2C_Rajagopal%2C_and_Tan)/04%3A_Catalysis/4.03%3A_Mechanisms_of_Catalysis)
20. Wikipedia contributors. (2020, April 16). Serine protease. In *Wikipedia, The Free Encyclopedia*. Retrieved 14:32, May 6, 2020, from en.Wikipedia.org/w/index.php?title=Serine_protease&oldid=951309456
21. Wikipedia contributors. (2020, April 16). Restriction enzyme. In *Wikipedia, The Free Encyclopedia*. Retrieved 15:12, May 16, 2020, from en.Wikipedia.org/w/index.php?title=Restriction_enzyme&oldid=951351229
22. Pingoud, A., Wilson, G.G., and Wende, W. (2014) Type II restriction endonucleases - a historical perspective and more. *Nuc Acids Res* 42(12)7489-7527. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4081073/pdf/gku447.pdf
23. Wikipedia contributors. (2019, July 25). BglII. In *Wikipedia, The Free Encyclopedia*. Retrieved 20:48, May 16, 2020, from en.Wikipedia.org/w/index.php?title=BglII&oldid=907885716
24. De la Peña, M, GarcÍa-Robles, I., and Cervera, A. (2017) The Hammerhead Ribozyme: A Long History for a Short RNA. *Molecules* 22(1):78. Retrieved from: https://www.mdpi.com/1420-3049/22/1/78/htm
25. Jakubowski, H. (2019) Biochemistry Online. *Libretexts.* Available at: https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Online_(Jakubowski)